Leiden University

ICT in Business and the Public Sector

Assessment of Generative AI Systems:

Defining Evaluation Criteria and Framework Design

Name:	Alicia Tasselli
Student ID:	s4228340
Date:	06/02/2026
1st supervisor:	Prof. dr. ir. Joost Visser
2nd supervisor:	Peter van der Putten
Company supervisor:	Alexander Trap and Silke van der Burg

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Einsteinweg 55
2333 CC Leiden
The Netherlands

Abstract

Background: As Generative AI development and adoption accelerates across industries, the technology becomes an important factor in mergers and acquisitions. Currently, no publicly documented standard framework is available for the evaluation of Generative AI systems during due diligence, which impacts the work of consultants, who face challenges in conducting GenAI-related assessments due to lack of standardised criteria, fragmented literature, and knowledge gaps in understanding GenAI technical dimensions.

Aim: This research proposes the design and development of an assessment framework for Generative AI featuring assessment dimensions, scoring metrics, maturity levels, technical reference information and best practices, to support consultants in delivering an informed and structured Generative AI evaluation during technology due diligence assessment.

Method: The study uses the *research through design* methodology, as the research is the artefact creation, combined with design thinking to structure the project phases into immersion, ideation and prototyping, and testing phases. The Generative AI framework is created in partnership and support from EY-Parthenon (EY-P); its intended application for technology due diligence assessment was internally validated through mentor feedback, and usability testing with consultants. Due to confidentiality, the second and third deliverables of this research were developed only for internal use and reference of EY-P.

Results: Generative AI assessment framework, with three key deliverables: (1) a scoring framework for assessing targets' Generative AI system during technology due diligence engagements; (2) a supportive slide deck with instructions and template slides that can be used in project deliverables, supporting the framework; (3) an informative guide containing supplementary technical information about GenAI. The first and second deliverables were validated through usability tests with 8 end-users and resulted in a TAM (Technology Acceptance Model) score of 3.06/4, indicating user acceptance.

Conclusion: The resulting Generative AI assessment framework consists of 9 dimensions and 39 subdimensions, divided into 3 categories, providing a holistic assessment of Generative AI. Additionally, it demonstrates the relationship of all of those components in order to have the most complete evaluation of the technology, to the best of the author's knowledge, is not yet described in existing academic work or publicly available industry materials.

Contents

Chapter 1

Introduction

Since the launch of ChatGPT by Open AI in November 2022, Generative AI (GenAI) has been considered a new technology evolution phase, redefining public perception and usage of AI [Kung et al., 2023]. GenAI usage has increased significantly since its release, ranging from personal use to full native GenAI companies and services. The variety of GenAI use cases is expanding, impacting industries and economic sectors with AI products being increasingly embedded into company processes, consumer-facing applications, health research, finance investments and more [Farhat et al., 2024].

Since then, companies have engaged in a competitive race to adopt the technology in their operations. According to a McKinsey study, larger companies (e.g. with revenue of 5 billion dollars or more) are found to be more successful in adopting AI in their workflow, with 39% scaling AI and 10% having already fully scaled AI [Singla et al., 2025], compared to smaller companies (e.g. with revenue of less than 100 million dollars) where 25% are scaling AI and 5% have fully scaled AI. As GenAI adoption grows, new challenges arise, including the change in technical roles that must be redefined to keep pace with the quick evolution of the technology. As an example, the role of CTO has been reshaped according to an EY-Parthenon survey [EY-Parthenon, 2025], as they need to keep up with the market pressure to incorporate GenAI capabilities into products, manage the increased pace of product innovation, maintain continuous product modernization, consider increasing operational risks, cybersecurity concerns, and governance frameworks while doing so.

Beyond adoption challenges, organisations report barriers in GenAI utilisation and deployment, including difficulties in managing risks (32%), implementation challenges (27%), identifying use cases (22%), lack of adoption strategies (21%) and trouble in choosing the right technologies (17%) [Deloitte, 2025].

Companies exploring GenAI adoption and transformation face a dual challenge. First, they must keep pace with the rapid evolution of AI technology, including frequent releases of new models and versions. Second, they need to understand how to define a fit-for-purpose Generative AI implementation. This includes identifying the metrics, parameters, and criteria that define an effective GenAI model, determining relevant use cases, establishing best practices, and developing a targeted business strategy that enhances their operations.

The current scenario for organizations that are using Generative AI in at least one function is also complicated by AI specific risks, with ongoing work to mitigate them [Singla et al., 2025]. These include inaccuracy risks (54%), cybersecurity (51%), regulatory compliance (43%), intellectual-property infringement (38%), unauthorised or unintended action technologies (28%) and, explainability (28%).

In order to address challenges of adoption, implementation and risk, there is a need for GenAI evaluation that currently lacks standard industry criteria because the paradigm evolves so rapidly. In due diligence assessment, this situation becomes more complex, as evaluation methodologies need to address both the foundational and service-models level and the use-case level, which is context-dependent and application-specific. The motivation for this research stems from a gap in academic literature. While various isolated approaches exist to assess different aspects of the technology at the foundation level, the field lacks a holistic assessment framework that can be adapted to diverse use-cases scenarios of the technology usage in organizational contexts.

A systematic literature review reveals this gap, with existing research proposing methodologies for evaluating Generative AI based on different aspects, including: model architecture classification and evaluation [Bandi et al., 2023], input-output modalities [Kim et al., 2024, Gozalo-Brizuela and Garrido-Merchan, 2023, Feuerriegel et al., 2024], precision-based metrics [Singh et al., 2025], trustworthiness

concepts [Huang et al., 2025], risk-based evaluation [Raut and Singh, 2024], and user experience design [Kim et al., 2024]. Although these works demonstrate in-depth approaches in evaluating specific aspects of Generative AI dimensions, they are narrow in scope. Existing research lacks a unified framework that integrates multiple concepts and assessment perspectives that is applicable to diverse Generative AI implementations and use cases.

This scenario is particularly challenging for due diligence (DD) consultants who need to evaluate companies technically. More specifically, in technology due diligence (TDD), which is the process of thoroughly examining a company before a potential investment [Berkman, 2013], the absence of an industry-wide evaluation framework means GenAI assessments lack rigour and defined baseline criteria. Without a framework, there is no standardised and objective methodology to answer clients' questions in due diligence projects. This is critical due to the nature of due diligence transactions in which each engagement depends on a unique project scope, with varied clients' investment theses and strategic fit considerations. Ultimately, the answers to those investment questions impact transaction outcomes and without standardised criteria, the consistency and quality of due diligence assessments may be compromised.

This leads consultants to raise the critical question during technology due diligence assessments: "Is the company really leveraging Generative AI or is it engaging in Generative AI washing?" Additional questions consultants have include: How can they assess companies creating service offerings with GenAI or utilizing capabilities powered by Generative AI? How can they define that the GenAI product has proper technology capabilities? How can they define its quality? What are the standards for each use case of GenAI? These questions illustrate the core problem.

The challenges are exemplified in strategy and transaction context, where this research was conducted in collaboration with EY-Parthenon (EY-P) to develop and validate a Generative AI assessment framework with the support of technology due diligence consultants.

Based on the literature review conducted and research of publicly available materials of industry practice, no comprehensive GenAI assessment tool or scoring methodology to support a standard and objective evaluation was identified. Particularly for due diligence transactions, this situation compromises assessment results by creating unstructured approaches for technical assessments and potential misalignment of GenAI capabilities with business or strategic objectives. Further implications arise due to the economic impact of consultants needing to understand Generative AI criteria at a deeper level to perform their assessments but lack standardised methodologies.

This thesis seeks to address these concerns by creating a standardised assessment framework that bridges academic technical knowledge and the practical needs of consultants in technology due diligence transactions. More specifically, it aims to provide an assessment framework for Generative AI technology that measures maturity across multiple dimensions. This assessment framework would ultimately enhance the work of consultants during technology due diligence, helping them perform more precise evaluations, improve GenAI alignment with business strategy, and potentially increase value creation. Bridging theory and practice, the framework and guidelines are designed to be applied in a due diligence transaction, enhancing the potential practical contribution for future research in GenAI.

1.1 Research objective

The research objectives were defined by aligning academic and business needs. More specifically, this work aims to define the relevant assessment dimensions, create a scoring-based assessment framework, establish best practices for GenAI adoption, and account for different use cases.

The research aims to develop a comprehensive, standardized assessment framework for evaluating Generative AI systems in technology due diligence projects.

Supporting the research objective, the assessment framework is designed to be generic for any GenAI system encountered in due diligence evaluations, integrating broad academic assessment components with DD-specific structure, process, and evaluation topics focused on GenAI concepts and concerns.

1.2 Research questions

The research questions were defined to answer all the problem questions posed in the problem introduction surrounding the creation of the Generative AI assessment framework and guide the project towards answering those questions along its development.

RQ How can the application of GenAI usage in organizations be assessed and evaluated in its development and deployment within due diligence projects?

During the development of the project the described sub questions are expected to be answered:

SQ.1 How can GenAI technology be defined and categorised to support practical understanding and assessment of its technological capabilities?

SQ.2 How can we design a comprehensive framework to categorise and define the assessment parameters for GenAI technologies based on technical aspects?

SQ.3 What are the key parameters and criteria that determine GenAI models quality?

SQ.4 How can we quantitatively assess valid unique value proposition characteristics of GenAI models while understanding their technical moats and replication risks?

SQ.5 How can organizations strategically implement GenAI solutions across different use cases to maximize business alignment and ensure measurable ROI?

1.3 Company project

The research project is in partnership with EY-Parthenon, as a research internship, allowing collaboration with the Due Diligence, Transaction and Strategy sector. The goal of this partnership is gain field experience of due diligence by being immersed in the environment and create an assessment framework specifically for generative AI for due diligence transactions, to better support the team of consultants and mitigate current problems and the lack of an evaluation tool. Two consultants from EY-P, who are experienced in AI Strategy work and Due Diligence, will act as mentors providing shared supervision and support during the internship for the completion of the project. As of the completion of this project, there have been no cases of due diligence projects of native technology companies with Generative AI systems.

1.4 Thesis structure

The thesis document is organised according to the research methodology, as follows: Chapter 1 accounts for the research problem, objectives, and questions, Chapter 2 describes the methodology, Chapter 3 reviews related literature, Chapter 4 describes the immersion in the research problem and data collection, Chapter 5 details the ideation process , Chapter 6 narrates the prototyping process of the framework, and Chapter 7 presents the validation of the framework content and methodology through usability tests. Finally, Chapter 8 presents the discussions of the research and Chapter 9 the conclusion.

Chapter 2

Methodology

This chapter describes the research methodology, explaining the research through design approach, how design thinking structured the project phases, and the data analysis activities defined for each phase.

2.1 Research methodology applied

The methodology chosen is research through design (RtD) [Zimmerman et al., 2007], a constructive research approach in which the process of designing artefacts generates knowledge and helps solve real-world problems. Stappers and Giaccardi [2017] state that creating a prototype is in itself a means of generating knowledge.

The RtD methodology fits our goals, as the artefacts creation is the research, and the outcome is the GenAI assessment framework. Zimmerman et al. [2010] defined RtD as an approach that uses design practice methods and processes as a means of examination. For our research, this means translating technical literature knowledge into an assessment tool that bridges theoretical concepts with practical application.

Additionally, we combined RtD (including academic research and documenting the process) with design thinking as our operational process, providing structure to the project phases. Design thinking is a user-centred methodology that focuses on the problems faced by the user and on implementing a solution for the root cause, consisting of four phases: clarify, ideate, develop, and implement [Han, 2021].

For this project, we structured the work in two parts: thesis proposal and thesis development. For the first part, we added two preliminary activities: empathise and define. For the second part, we adapted the design thinking terminology of the four project phases to our context, reframing clarify as immersion, ideate as ideation (noun form), develop as prototyping, and implement as testing.

The process and its phases are illustrated in Figure 2.1. The empathise and define activities were performed during the thesis proposal, before the start of the internship at EY-P, but already in collaboration with the company. The subsequent phases immersion, ideation, prototyping and testing were performed during the thesis internship at the company.

2.1.1 Proposal phase: empathising and defining

The proposal phase had the objective of empathising with the user, understand the research and define the research objectives.

- Initial understanding of the problem context and alignment with the partner company EY-Parthenon;
- Define objectives and research questions.

2.1.2 First phase: immersion

The first phase had the objective of immersing in the context of the problem including literature, data collection and gather quantitative and qualitative data.

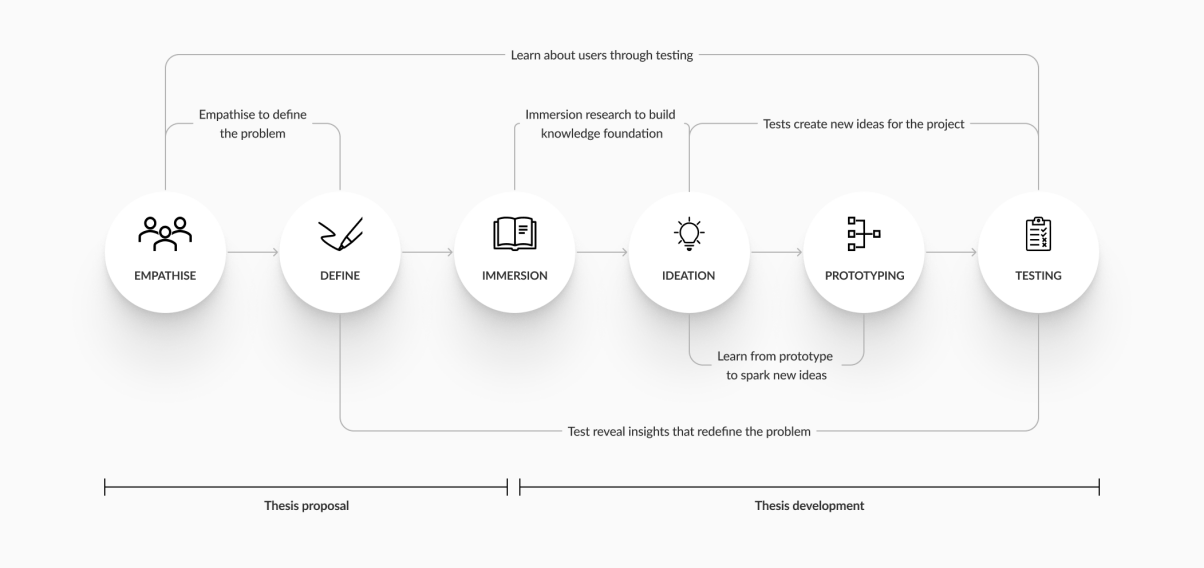

Figure 2.1: Research methodology and its phases

Activities

- Perform literature review;
- Define data collection methodology, interview scripts, analysis methodology, in alignment with the research questions and objectives;
- Perform review of EY-Parthenon documentation regarding GenAI/AI initiatives and materials;
- Gather information about consultants experience in TDD and GenAI knowledge;
- Select the most relevant work that can be used as a base for the artefact;
- Analyse all the gathered data and validate the problem scope with the focal-point person.

Literature review

- **Qualitative:** Data collection of related work on 1) metrics, criteria, parameters and categorisations for models, data and training of GenAI; 2) best practices in developing GenAI; 3) use cases for GenAI in different business contexts and necessities;
- **Data collection tools:** EBSCO, Google Scholar and Leiden University catalogue.

Internal data collection

- **Quantitative:** Questionnaire for company consultants: to understand the current situation of consultants regarding GenAI knowledge, usage, problems, difficulties, needs and wants;
- **Qualitative and exploratory:** Participate (shadowing) a TDD project in order to understand in depth the technology due diligence evaluation, additionally gather the main problems and gaps in their current process;
- **Qualitative:** Research documentation from EY-Parthenon: reports, past projects with GenAI, current framework templates for assessing AI and related documents with the current framework.

2.1.3 Second phase: ideation

The second phase had the objective of applying the findings from the first phase to ideating and prototyping the solution for the research problem.

Activities

- Gather the data collected and identify the most relevant models, metrics, criteria, parameters, quality, and use cases for the GenAI framework;
- Create indeces and scales (scores) for the most relevant concepts of GenAI in the due diligence context.

Ideation data

- **Qualitative analysis:** Select most relevant data collected from immersion phase, most relevant work to base the framework and guidelines, indexes and score types

2.1.4 Third phase: prototyping

The third phase had the objective of applying the ideation into practice and actively prototype the assessment framework.

Activities

- Brainstorm the framework and the deliverable of the guidelines, including sessions with the mentors;
- Design the framework and guidelines, including co-design session with mentors;
- Design iterations of the framework, adhering to requirements.

2.1.5 Fourth phase: testing

The fourth phase had the objective of testing the prototype with consultants, the end-users, with a business case to validate its usability, methodology, and acceptance within them.

- Perform usability test with consultants to validate the framework methodology and gather feedback;
- Validation method Technology Acceptance Model (TAM);
- Iteration loops between feedback from user tests and adjustments to the deliverables;

The testing protocol and requirements are described in detail in Chapter 7.

Chapter 3

Literature review of related work

This chapter has the goal of discussing and explaining the studies that are part of the literature review of this research. All of those topics contributed to building the knowledge foundation and immersion relating the central theme of this research, Generative AI, resulting in a diverse collection of works. Topics covered in the related work body are Generative AI: definition; classification; evaluation methods and metrics; explainability; transparency and openness; data types, sources and implications; security, risks and legal; business impact and innovation; and finally design principles and user experience.

The related work will be described throughout this chapter following the order mentioned above.

3.1 Generative AI definition

Establishing a definition of Generative AI is important as it is the fundamental topic of this project. Feuerriegel et al. [2024] define Generative AI as:

> "The term generative AI refers to computational techniques that are capable of generating seemingly new, meaningful content such as text, images, or audio from training data."

The author further elaborates that Generative AI can have different purposes, even acting as professionals such as writers and illustrators, but the system ultimately acts as an assistant for humans on a question-answering basis, generating multi modal outputs.

However, this traditional definition is content-focused, and can be further expanded due to include the diverse capabilities that current Generative AI models have, such as generating source code, digital artefacts, realistic video creation all through prompting, as well as the capabilities of Agentic AI implementations.

Besides its capabilities, the technology can also be defined by its technical specifications, including model type, training method, output type, data requirements, and computational needs.

Contrary to this definition, Ronge et al. [2025] propose a perspective on Generative AI beyond technical characteristics, which differentiates it from simple generating models. The author defines four factors that define what are unique to GenAI systems: multimodality, interaction, flexibility, and productivity, given the argumentation that Generative AI is capable of processing different types of inputs and most importantly, receiving natural human language.

Furthermore, Ronge et al. [2025] observes how the different types of GenAI users interact with the technology, highlighting diverse discourse terminology used by different publics. The general public often uses the term "GenAI"; semi-technical users refer to system-level and model-level terms; and technical participants instead refer to models name directly with "Generative model". This terminology diversity reveals that the user-friendly characteristics of Generative AI systems are bringing the general public closer to the technology, since they are using terminology in a looser, broader sense. This suggests that compared to other technologies, such as the personal computer, GenAI has achieved faster adoption by the general public.

3.2 Generative AI classification

The classification of GenAI systems varies across the literature. The most common approaches found in research to classify Generative AI systems are based on model, architecture types, and input-output modalities.

Among different studies, Bandi et al. [2023] have the most extensive classification of Generative AI models based on seven architecture types. This classification system groups models into Variational Autoencoders (VAE), Generative Adversarial Networks (GANs), Diffusion Models, Language Models, Transformers, Normalizing Flow models and Hybrid models. The research also provides architecture components and training methods used for each different type of model.

In addition to the architecture classification, Bandi et al. [2023] created a diagram based on the output modality of models. This visualization organizes the diverse categories and sub-categories of outputs that generative AI systems are able to produce. The main categories include: image; text; source code; speech; video; 3D; graph; tabular; and molecular structure. Each category is complemented by sub-categories, for example, image model capabilities include synthesis, style transfer, image segmentation, while language models can perform translation and caption generation.

Similarly to the architecture and output modality classification approach, Kim et al. [2024] propose a taxonomy in which systems are divided into modalities and further sub-divided into architecture types. Unlike other papers, they organised the models by year of announcement, showing their evolution timeline from 2012 to 2024. The four modalities are text-based systems, image-based systems, audio-based systems, and multi-modal-based systems, while the three architectures present for each modality are encoder, encoder-decoder, and decoder.

A different approach to modality classification is provided by Gozalo-Brizuela and Garrido-Merchan [2023] that created a taxonomy of Generative AI models based on the combination of input and generated output formats. The taxonomy is divided into text-to-image; text-to-3D; image-to-text; text-to-video; text-to-audio; text-to-text; text-to-code; text-to-science and other models. This taxonomy shows the state of the art of generative AI in 2023 and can be directly compared to Bandi et al. [2023] in the same year. The difference between the approaches is that Gozalo-Brizuela and Garrido-Merchan [2023] taxonomy is less detailed in combinations of input and output types, but provides model names, whereas Bandi et al. [2023] provides more combinations of output types and granularity, but fewer model examples.

As an alternative to technical taxonomies, Feuerriegel et al. [2024] classify generative AI systems based on outputs related to the applicability of these modalities. They provide a functional view of output models, directly connecting modalities to business problems. The classification is based on output modalities, model level, system level, and application level, interconnecting these information points. The output modalities are text-generation, image/video generation, speech/music generation, and code generation. The first model level, brings examples of models for each output modality. The middle system level describes what functionally the interface interaction provides, such as agents, search engines, bots, and system types. And the last application level shows which business problems it actually solves, interconnecting output to real-world problems application.

These classification approaches have similarities and distinctions between them. The modality-based classification of Kim et al. [2024] and Feuerriegel et al. [2024] is similar by grouping Generative AI models based on input-output modalities, while Kim et al. [2024] complements Bandi et al. [2023] classification taxonomy grouping models by architecture types. Both Bandi et al. [2023] and Gozalo-Brizuela and Garrido-Merchan [2023] classify systems by combination of modalities, though the former offers more examples of output types and the latter provides specific model examples.

In contrast to the other classification methodologies, the distinction of Feuerriegel et al. [2024] is by providing a practical application level to the output modalities, associating applications to system output, to connect technical information to real-world problem solving contexts.

3.3 Generative AI evaluation metrics and methods

This section examines related work that provides diverse Generative AI evaluation methodologies and perspectives. These range from foundational metric taxonomies, quantitative and qualitative metrics to specialised approaches featuring trustworthiness, risk-based, and human-centred usability.

The foundational evaluation approaches for Generative AI systems are presented by Singh et al. [2025], that describes diverse general model metrics. Including precision-based metrics, output evaluation, human evaluation techniques, red teaming / blue teaming, and LLM as a judge. The author further specifies that a combination of evaluation techniques is optimal and that they need to be tailored to specific applications.

Building on these general model metrics, Modake and Patil [2024] presents quantitative and qualitative metrics, providing more granularity to evaluate Generative AI systems. Precise quantitative or general model-level metrics are divided into text, image, and computational. Subjective qualitative metrics include human evaluation and subjectivity considerations, providing a holistic view on the approaches.

However, according to the author, challenges in evaluation are present namely data bias, scalability, and ethical concerns, and should be addressed during evaluation. There are emerging trends to mitigate current evaluation challenges, such as automated evaluation frameworks, dynamic and context-aware evaluations, integration of Human-AI collaboration, and multimodal evaluation techniques.

Moving from general metrics to specialised holistic system evaluation, Huang et al. [2025] created TrustGen, a dynamic benchmarking platform to evaluate the trustworthiness of Generative AI applications. This approach differs from previous ones as the authors created a novel assessment tool with conceptually in depth evaluation of generative AI tailored to specific concerns and context that affect the trustworthiness value of models.

The main assessment dimensions for Generative AI LLM models of TrustGen are: truthfulness, safety, fairness, robustness, privacy with subdimensions of hallucination, sycophancy, honesty, jailbreak, toxicity, exaggerated safety, stereotype, disagreement, and preference. The platform also covers image-to-text and text-to-video models. Additionally, Huang et al. [2025] defines eight guidelines for trustworthy generative AI covering design, intended use, human oversight, responsibility for model behaviour, robustness, harmlessness, reliability, accuracy, privacy and data protection. The article is for the creation of the platform, but it also provides the technical details and results of its creation, serving as a comprehensive source material on how to evaluate generative AI.

From a holistic risk perspective, Rauh et al. [2024] defines a risk-based evaluation, and also comments on gaps in current safety evaluations, with a focus on audio and video modalities, risk coverage, and context. The study performed an analysis on the number of existing risk evaluations per model modality type. They identified a discrepancy on the number methods per type, with text-based models being well covered, containing 159 different types of risk in evaluations, while audio, image and multimodal-based models had less than 36 risk types each. As a result, Rauh et al. [2024] proposed an approach to improve risk evaluation by closing the modality coverage through repurposed evaluations. They achieve this by increasing the risk coverage through automated evaluations and leveraging language models systems. The result was the increase in risk content coverage, evaluating not only the model but the context where is being used. In addition, they suggest an interdisciplinary approach to safety evaluation by utilizing data from other areas and qualitative analysis of how AI are used.

Another downstream task specific perspective is proposed by Kim et al. [2024], that created an evaluation of UX/UI of GenAI on the application-level based on usability heuristics. Their method of evaluation consists of comparing the GenAI system against their ten defined heuristics and identifying problem points that may create usability issues to the user. An interesting output of this research is the result of their UI/UX evaluation of Generative AI systems. In total, 22 different systems (i.e., including text, image, audio and multimodal based GenAI) were evaluated against their heuristics, in which 100% did not scored on the 5th heuristic capability, and 72% did not score on the 9th and 10th heuristic. Those percentages show a significant absence of user support provided by GenAI and demonstrates a possible significant industry gap in those aspects.

Each work provides a different perspective on GenAI evaluation ranging from qualitative and quantitative metrics to human collaboration approaches, risk-based, and usability heuristics. This diversity reveals that there is no straightforward set of metrics to define Generative AI evaluation; rather, using complementary approaches can provide a comprehensive evaluation across multiple dimensions.

3.4 Generative AI explainability

With the increased creation of Generative AI models, a new concept emerged: explainable artificial intelligence (XAI) for Generative AI. Kalota [2024] explains the importance and need for explainability for GenAI particularly understanding its decision-making process and reasoning process. This transparency ultimately enables users to trust the model answers and utilize model outputs in their own decision making process, even if they are not an expert in the area. The author defines XAI as:

> "XAI is a set of processes and techniques that allow humans to understand and trust the output of machine learning algorithms."

Explainability can be reached through the multidisciplinary interaction of fields and use of techniques that allows the model's inner workings to be transparent and understandable to the user, such as black-box, white-box, model-specific, and model-agnostic techniques originally developed for machine learning but applicable to GenAI.

Alongside techniques that are being used and adapted to XAI, there are also metrics to measure it, being cited by Modake and Patil [2024] such as SHAP (SHapley Additive exPlanations) and LIME

(Local Interpretable Model-Agnostic Explanations), that while they were not created specifically for GenAI, they could provide numerical interpretation of GenAI outputs and improve its transparency.

Similarly, Schneider [2024] provides reasons why explainability is important for GenAI alongside risks, concerns, and emerging desiderata required to reach this paradigm. The author argues that the importance of XAI is based on the concept that users should be able to understand the quality of models outputs, how models reached their answer, and their thinking process. Models should also provide support for users to control their outputs, not the other way around. Additionally, users should be informed by models and its creators that the technology has its limitations and associated risks to ensure a safe use.

Schneider [2024] further defines key reasons why transparency of GenAI systems is necessary. Reasons for XAI include user-related motives as the need to adjust, in order to meet with user preferences and requirements; and the need to verify, in order to ensure correct outputs and maintain accuracy, systems should identify errors and hallucinations. Other reasons, relates to the nature of GenAI applications, as they can be used in diverse tasks, systems should be able to cover diverse possible outputs and anticipate use cases, while maintaining special care with high impact applications, especially in sensitive or high-risk sectors (e.g., healthcare).

To support transparent systems, Schneider [2024] defines requirements to achieve XAI through desiderata specifications. They include: verifiability; lineage; interactivity and personalization; dynamic explanations; costs; alignment criteria; security; and uncertainty. All of those desiderata specifications have the goal of improving the quality of how GenAI systems are structured and reach the ultimate XAI paradigm.

Overall, explainability principles are important to ensure that we can understand how models reach their answers, their thinking and decision making process. This understanding increases trust and encourages users to apply critical thinking when reviewing models outputs, which is important given the global increase of GenAI models and applications available.

3.5 Generative AI transparency and openness

With the advancements of foundation models, Bommasani et al. [2024] proposed a 'Foundation Model Transparency Reports' to evaluate and standardize their content for developers, based on the previously created 'Foundation Model Transparency Index', containing 100 requirements for model transparency. After creating this index, they compared their requirements against 6 governmental policies (i.e., Canada code of conduct, EU AI Act, G7 Code of Conduct, US Executive Order, US FM Transparency Act, and White House Commitments) and found that only 37 of their 100 transparency topics were covered by those policies. One of the contributions of this research is to propose a standardised report of transparency for foundation models developers, which will enhance the overall adoption and compliance of these requirements, improving not only the accountability of models but also later possible penalties from non-compliance.

To increase transparency of models, Liesenfeld and Dingemanse [2024] created a framework to score the openness of Generative AI models (i.e., text, audio and video modalities) using 14 criteria on availability, documentation and access. This openness judgement was created due to the new EU AI Act requirements about transparency and accountability, and the authors observed that while many providers defined their models as open, the term was being broadly used, as they judge that they are mainly open-weight. Further, the authors believe that the reason behind this "open-washing" of GenAI is to avoid regulatory and legal inspection about their training and data used. They benchmarked 40 text generators models that were defined as open, using Chat-GPT model as their close reference model, with an openness score of 0.5/14 in their benchmark. The benchmark results revealed a median score of 5/14 points with the top scoring model 'OLMo 7B instruction' achieving a 12.5/14 score openness, and the lowest-scoring model, 'Xwin-LM' with a score of 1/14.

The initiatives of these researchers emphasize the importance of developer accountability and regulatory compliance. Both the Foundation Model Transparency Report Bommasani et al. [2024] and the Openness benchmark Liesenfeld and Dingemanse [2024] share similar goals of increasing the transparency of GenAI development, through different lenses and criteria. The former proposes a standardised report with comprehensive criteria that increases transparency coverage across companies. The latter illustrates the extent to which GenAI companies actually meet openness criteria, which has a direct impact on disclosure requirements introduced by the new EU AI Act. The EU Act requires certain criteria and applies requirements depending on the type of company and product, and their benchmark research demonstrated that the researched companies do not meet them at the time of research. Transparency overall

is a baseline requirement for GenAI companies, considering the multi-modal generation capabilities of models and to prevent risks and security vulnerabilities, regulations must hold developers accountable for their products.

3.6 Generative AI data types, sources, and implications

Data serves as the foundation for understanding GenAI models. Data type and source allow us to understand how, and based on what, models are created, trained, and the quality of their sources of information especially because GenAI models can be adapted for specific domain areas depending on the training data and method.

Foundation models are flexible in that regard, as they can undergo pre-training on broad data and then fine-tuning with specific domain data, including proprietary data, for specific tasks. The quality of GenAI model training can be enhanced through proper data collection, preprocessing, protocols, techniques and initiatives that help improve this process and tackle industry problems with training data. This section discusses training data formats, data sources, synthetic data, data transparency, and regulatory and ethical considerations.

Training data for Generative AI can present different formats. The author Stober [2025], defines different possible sources of data to train generative AI models including: publicly accessible data collections; commercially licensed data collections; self-collected data from publicly accessible sources; self-collected data from sources accessible through payment; self-digitalised media; self-made recordings; in-house recorded data; data extracted from primary data; and synthetic data. Since training data can bed originated from diverse sources it is important that developers document those sources [Stober, 2025] and also consider implications of its use [Ungureanu and Amironesei, 2023].

Among the data sources identified by Stober [2025], synthetic data is a particular one due to its characteristics and role in helping solve data scarcity. Goyal and Mahmoud [2024] provides a definition of synthetic data, approaches to generate it, and use case of models for different types of data generation.

Synthetic data generation has the following creation process: an original database goes through data analysis and preprocessing, then a specific synthetic approach and model create the synthetic dataset. The main advantage of this process is that the synthetic data preserves underlying patterns and behaviours of the original data, enabling businesses to produce cost-effective realistic datasets without revealing sensitive information. The main advantages of this technique related to addressing data privacy concerns and algorithm biases while increasing accuracy of models, facilitating collaboration of stakeholders and maintaining confidentiality of information. Goyal and Mahmoud [2024] also provides a review on synthetic data generation techniques approaches, including machine learning, rule-bases systems, mathematical and probabilistic methodologies, and differential privacy.

Regardless of their type, all formats of training data are subject to legal and regulatory considerations. Ungureanu and Amironesei [2023] explores the legal consequences and implications of the data used in GenAI models. The authors defines a classification of legal data types: data that is exploited under different licenses; data that is not subject to intellectual property rights; and data protected by intellectual property rights but not authorised for use. Across those categories there different types of data: open source; voluntary provided; protected by intellectual mechanisms; and synthetic data. The classification of data types helps us understand its source and also the legal implication of its use.

To address these legal considerations and ensure transparency, Stober [2025] comments on the importance of keeping the source documentation of the data used in models, distinguishing between training and input data. Additionally, the author proposes a list of recommendations surrounding documentation that highlights the importance of this practice also due to copyright issues and to improve the quality of models through transparency, which can be achieved by publishing the source list, a search portal or API.

Beyond legal compliance and documentation requirements, Xiang [2024] examines training data from an ethical perspective, addressing its biases and implications. A critical point is that any data used to train models inherently could demonstrate biases from its source, since it is a by-product of society itself. The author argues that developers have the responsibility of identifying possible biases and implement mechanisms to diminish their impact when training models. Some counteract measures are suggested such as reinforcement learning, guidelines, and prompting strategies to reduced biased data. Extending Stober [2025] taxonomy of data sources, Xiang [2024] suggest that developers prioritize collecting data from sources in which individuals willingly uploaded data, including for financial compensation, and utilize public source of information when appropriate. This also connects with the EU AI Act transparency

obligations and Bommasani et al. [2024] report to increase transparency among models, encouraging developers to improve their data collection practices and reportage of them.

To comply with the transparency requirements discussed above, mechanisms to recognize data sources and AI generated content are necessary. Various authors mention possible techniques to recognize data sources, with fingerprinting appearing as an emerging method. Stober [2025] identifies various techniques including: data-sheets; URLs and download stamps; unique identifiers; metadata; and fingerprinting. However, fingerprint can be used in different ways as a recognition of data sources either in the input or in the output as recognition of AI generated content. For training data documentation, Stober [2025] proposed a broad case recognition of data to document it, through fingerprinting, to be used directly in AI systems in their training process suitable for different data types. For training data provenance, Yu et al. [2020] proposed a novel fingerprint mechanism specific for AI-generated images for source recognition, adding a level of security, helping to avoid fake data generation and overall easier recognition of data.

Collectively, these studies provide a broad vision on the different aspects of Generative AI and data. The technical aspect of data sources taxonomy and synthetic data, that are subjective to regulations and must comply with legal elements of licensing and intellectual property. Documentation and source recognition mechanisms, combined with ethical concerns of data bias coordinate to improve transparency of data sources. Together, these perspectives establish that transparent, ethical, and compliant data in Generative AI systems are fundamental to ensure responsible GenAI development.

3.7 Generative AI risks, security and regulations

The quantity of Generative AI-related risks is extensive and complex. Zeng et al. [2024] performed a structured review in eight governmental, and sixteen private regulations and policies finding 314 unique risks types. Because of the extensive risk possibilities, the author proposes a unified language to identify types of risks divided into four broad categories: system operational risks; content safety risks; societal risk; and legal rights-related risks. The risk taxonomy created by the author, aside from showing the granularity of potential risks, also compares the risks covered in USA, EU and China regulations, highlighting the intersections of risks covered between those countries, and serving as a guidance material for understanding the subject.

Complementary research on risks by Cheng and Liu [2024] provides another organization of risks, dividing them into five categories: ethical risk; intellectual property infringement; privacy and data protection; market monopoly; and cybercrime and data security. Similarly to Zeng et al. [2024], this research compares EU and US regulatory approaches, where the EU takes a heavier regulatory approach compared to the US, that expects self regulation and assessment from companies. Those contrasting regulation approaches can influence Generative AI development and risks insurance, and ultimately individuals and companies that make use of GenAI solutions will have to manage diverse regulations. In order to mitigate differences, the author suggests that regulations should be embedded during development of Generative AI while also keeping ethical AI central. Also, to maintain the balance between technological development and governance regulation. This suggestion could help mitigate risks when users interact with GenAI systems and support organizations in complying and avoiding fines.

Examining harms from a user-centric perspective, Rauh et al. [2024] defines an overview of harms from Generative AI system with risks areas, definitions and examples. The study defines six main risks areas: representation & toxicity harms; misinformation harms; information & safety harms; malicious use; human autonomy & integrity harms; and socioeconomic & environmental harms. This definition of harms complements, from another perspective the works of Zeng et al. [2024], that proposes a extensive risk taxonomy, and Cheng and Liu [2024] that provides a organization of risks in 5 areas. Rauh et al. [2024] has a lens focused on the human-AI harms and its implications, including economical, for users. The author illustrates the harms with examples such as GenAI leaking private images from training data or its capability of generating deep fake images, without regulation, that can lead to economical or personal harms or even engage the user into unsafe behaviours.

However, there is also a way to mitigate risks by increasing security during models development and training. Feretzakis et al. [2024] presents privacy risks associated with LLM models, especially related with training data for models, and provides an overview of data privacy risks and privacy- preserving techniques to mitigate them. The study also illustrated a workflow for a privacy-preserving AI in which diverse techniques can be utilised based on the application domain, associated risk levels, and privacy requirements. The workflow consists of four stages: data collection; followed by federated learning and model training; later there is the model deployment; and the last steps include homomorphic encryption and secure multi-party computation.

Beyond privacy-preserving workflow, Feretzakis et al. [2024] proposes other techniques, describing their key strengths and limitations, for different use cases. For example, utilizing privacy-enhancing technologies and blockchain. In addition, mentions techniques specific for preventing data memorization including: selective forgetting and scrubbing, privacy filters, privacy fine-tuning, real-time privacy audits, and using synthetic data.

Similarly, Xiang [2024] cautions about bias in GenAI generated content and about who is responsible for it. Considering the notion that as models are made by humans they are going to have inherently some degree of bias. The author propose techniques to improve GenAI models, specifically for biases in generate content reinforcement learning is suggested, including also mitigating skewed data and extreme cases. However, not only utilizing techniques alone are sufficient, acknowledging biases and providing transparency to the user about model limitations can help further mitigate them.

Multiple authors, Kim et al. [2024], Ayemowa et al. [2024] and Xiang [2024], mentioned the prompt filter layer technique to improve security. Those authors discuss the implementation of a filter later between the user input and the GenAI model, with the goal of recognizing personal sensitive information, filtering the content of inputs before sending out to the model to be processed. This technique prevents users sensitive information from being accessed by models or used as training data.

A regulatory approach to privacy and security concerns is taken by Golda et al. [2024], who defines 5 perspectives of GenAI implications and cites data protection laws from multiple countries. The five perspectives of concerns are: user; ethical, regulatory and law; technological; and institutional. This approach highlights how different entities are involved in ensuring privacy and security of GenAI systems and its users, each with a responsibility. For each perspective, actions are suggested to improve security, for example users should provide consent, control and insurance. Whereas the ethical part should be concerned with bias, developers responsibility and policy makers. For technological perspective, privacy preserving techniques to mitigate implications are mentioned, similarly to Feretzakis et al. [2024] and Xiang [2024]. For institutions, it is recommended to have data handling procedures, governance structure, risk mitigation and reversal specific to address possible harms, which is particularly important for the security aspect of models.

Within the regulatory and law perspective, there are classic data protection laws, and intellectual property rights considerations. The author Golda et al. [2024] cites global data protection laws that provide users rights about their personal information and how they can be used by companies. Those laws have a few elements in common but each country provides different data subject rights. Common elements across those regulations include principles of transparency, security, confidentiality, and limitation of data used among others. The reviewed laws are GDPR from European Union, CCPA from California USA, LGPD from Brazil, PDPA from Singapore, and POPIA from South Africa. In line with this, Xiang [2024] also recommends verifying if generated data (i.e., output) is compliant with regulatory laws, reinforcing the importance of privacy in GenAI models.

To translate regulatory compliance requirements into actions developers can act on, Hacker et al. [2023] establish three policies for regulating GenAI models. First, the policy of minimum standards, defined in line with the art. 14 of the AI Act, for all Large Generative AI Models (LGAIMs) are: they need to have transparency rules, risks must be allocated to developers, data governance needs to be present along rules on cybersecurity, sustainability and content moderation. Second, the rules for high-risk systems, applied only to LGAIMs with high-risk use cases, in line with the AI Act (Art. 11, Annex IV AI Act) are: data provenance, data curation, performance metrics, incidents and mitigation strategies should be reported by LGAIM developers and employers. Alongside this, they should also report if parts of its content were generated using LGAIM. Third, rules for collaboration are defined to bridge and enable compliance of the first two policies. The authors suggest that data curation should be representative and balanced to avoid discrimination and to tackle this risks as soon as possible in the deployment pipeline. In addition to this, centralised monitoring is suggested through allowing users to flag models output and provide feedback to developers, and that those user flags should be prioritised by developers to create a content moderation process flow.

Together these works present specific GenAI risks, propose techniques to mitigate privacy and security problems, identify applicable laws and regulations. The risk taxonomies developed by Zeng et al. [2024], Cheng and Liu [2024], and Rauh et al. [2024] demonstrate the range of concerns and harms that GenAI models can inflict on users and companies including operational, ethical, and socioeconomic consequences. These identified risks can be managed through the presented solutions for responsible development and deployment cycle, proposed by Feretzakis et al. [2024] and Xiang [2024] which utilize privacy-preserving techniques, bias mitigation methods and prompt filtering, reinforced by Kim et al. [2024], to reduce harms. The technical measures are embedded within regulatory frameworks that establish laws, policies

and rules, defined by Golda et al. [2024] and Hacker et al. [2023], where proper regulatory oversight can help shape how GenAI are created and maintained, while ensuring compliance, privacy, and security for users.

3.8 Generative AI business impact and innovation

Kanbach et al. [2024] created six propositions of GenAI impact on innovation across three categories, using as inspiration the business model innovation (BMI) framework proposed by Kraus et al. [2022]. This BMI framework has three categories: value creation innovation (e.g., new capabilities and technologies), new proposition innovation (e.g., new offerings, customers, and market) and value capture innovation (e.g., new revenue models and cost structures). From a review of 513 sources and industry cases, the authors defined three impact categories: innovation activities, work environment and information infrastructure, than applied the BMI framework to organize them. The propositions state: (I) GenAI will level the playing field across expertise areas, technologies and resources; (II) the combination of GenAI, critical thinking and factual knowledge will increase the degree of innovation of business models; (III) GenAI will affect business models via innovations on value creations; (IV) GenAI will have the most impact on white-collar knowledge workers; (V) GenAI will redefine the necessary skill-set of job roles, from creators to editors; and (VI) GenAI will be used by companies, its not a matter of if but how it will be used.

The propositions on GenAI impact on innovation are valid and reflect how the industry already has been adapting to the new GenAI technologies and possibilities, however to reach the levelled playing field across all expertises areas, stated in proposition I, will require much industry investment and quantifiable outcomes to effectively equally impact all fields. Similarly, the required skill-set for new adapted job roles, stated in proposition V, could also create discrepancies between workers that will have and not have it, since not easily all will have proper access to training and resources in the same level.

3.9 Generative AI design principles and user experience

The author Weisz et al. [2024] proposes six design principles for developing Generative AI applications, with four associated strategies each on how to implement them. These principles aim to counteract current issues of the technology from the user's perspective. The principles include: design responsibly; design for mental models design for appropriate trust and reliance; design for generative variability; design for co-creation; and design for imperfections. The authors principles have an emphasis on helping and enabling the user in diverse circumstances, showing the importance of a good design that is user-centric, helping resolve issues, reduce harms, estabilish an effective communication, consider users goals, educate users to be skeptical of outputs, proposes variability, enable productive collaborations, and sets expectations.

Similarly, Kim et al. [2024] proposed usability heuristics for UX/UI of GenAI, including: (1) Visibility of the System Status; (2) Match Between the System and the Real World; (3) User Control and Freedom; (4) Consistency and Standards; (5) error prevention; (6) Recognition Rather than Recall; (7) Flexibility and Efficiency of Use; (8) Aesthetic and Minimalist Design; (9) Help Users Recognize, Diagnose, and Recover from Errors; and (10) Help and Documentation. While, the research had the goal of proposing a UI/UX heuristic evaluation, its content serves a dual purpose: to perform the heuristics evaluation and as a usability guideline for Generative AI development, that could help GenAI applications improve their user experience levels. Also considering the heuristics gap identified in their research, described earlier in Section 3.3.

All of those design and heuristics principles are important to lay a foundation of what a good user-centred generative AI application should have, but most importantly have at its core the foundation of instructions on how to use and interact the technology, alarming users of its risks and setting realistic expectations of their outputs. We consider those things to be important aspects for GenAI models that can help distinguish themselves across the multitude of models available for users in the market.

Chapter 4

Immersion phase

This chapter has the goal of describing and explaining the activities performed during the immersion phase of this research, which ultimately lays the foundation for the prototyping phase. We immersed in the research problem to gather information from academic and company sources and based on that defined the necessary requirements for our prototype.

The chapter follows the immersion process chronologically, presenting first the literature review methodology, second the internal data collection, third the findings from participating in a technology due diligence project, and fourth the analysis of the results of an internal questionnaire about GenAI knowledge. Finally, we discuss the findings of this phase and how they impact the next project phase.

4.1 Literature review methodology

The objective of the literature review was to understand the current academic landscape regarding GenAI, the work being produced by fellow authors, and to find materials that helped answer our research question. In order to achieve that, we selected 140 items of related work on the topic, across various publication types, including journals, surveys, articles, papers, reports, and books. From these, 126 were accessible, and 121 were reviewed in depth, with 35 papers ultimately being chosen to compose the related work reference, providing an essential basis for developing the assessment framework. Chapter 3 presents the complete literature review of related work.

None of the 121 reviewed publications presented a comprehensive assessment framework specific for GenAI that integrated multiple assessment dimensions, or that was due diligence oriented. However, within the 35 selected papers, several provided taxonomies, evaluation metrics, or evaluation tools for GenAI and they were used as the primary reference for the content of the proposed framework.

Database sources and search strategy

The systematic literature search was conducted using Web of Science, EBSCO, Google Scholar and the Leiden University Library Catalogue between February and May 2025. The search was limited to publications of the last five years (2020-2025) to focus on recent content about Generative AI and to limit publications about traditional AI and Machine Learning. Publicly available publications were prioritised, and some were accessed through the Leiden University student account; no paywalled publications were purchased.

The inclusion criteria were to select papers that were peer-reviewed and provided technical explanations of relevant topics of Generative AI, with a focus on AI evaluation, technical concepts, frameworks, best practices and business applications. The keywords search resulted in a high number of health-related or medicine-focused papers with GenAI. These were excluded from the selection because our scope focused on technical and business related publications.

During the review process, we tagged papers with comments about their themes. This resulted in a high-level categorization of the overall themes of the 140 papers initially selected. Some of the tags included: foundational concepts; evaluation methodologies; evaluation gaps; categorization framework; terminology; taxonomy; general purpose; ethical concerns; best practices; use cases; metrics; GenAI integration policies; building GenAI; hallucination; GenAI application; AI principles; deep generative models; XAI; safety evaluation; data transparency; foundational models; data practices; risk and bias; privacy-preserving techniques; AI generated images; dynamic benchmarking; synthetic data generation;

and risks use cases. Additionally, in this tagging system the papers that were connected to each other were tagged with "connected to paper number x" for easier cross-referencing.

Below is an explanation of the precise steps taken to perform the literature review:

Definition of search keywords

Based on research question and objectives, keywords were defined to support the systematic literature review search. During the literature review, additional keywords were added to complement the initial list, based on experimentation with word combinations that provided relevant search results.

List of search keywords

The complete list of research keywords used in our literature review (terms marked with * were the most productive):

- GenAI definition*
- GenAI categorization*
- GenAI practical understanding
- GenAI technological capabilities
- GenAI basic concepts
- Defining GenAI basic concepts
- Evaluation framework for GenAI
- Evaluate GenAI quality
- Parameters for GenAI quality
- Types of GenAI
- GenAI metrics*
- GenAI criteria
- Parameters for models generation, training data collection and usage policies
- Retrieval augmented generation
- Generative models technical disclosure*
- GenAI bias and risks
- Assessment of Generative AI models
- Generative AI data
- Generative AI training data types
- Generative AI embedding levels
- Generative AI best practices development
- Guide on how to develop a good Generative AI
- Generative AI toolkit
- Best practices when use third party Generative AI applications
- Best practices for generative AI model development

It is important to note that some papers appeared under multiple research terms.

Initial review of papers

The initial scan of papers was performed by analysing the abstract, results and conclusions sections, and defining if the paper produced relevant content to our research. 140 papers were considered relevant and within the theme, these were added into a private database that contained all the literature references.

Construction of the reference database

In the reference database, properties were used to structure the organization of papers, and for easier maintainability, backtracking and access of reference materials. The properties used were: number ID; type; comments (i.e., keywords to recognize the paper); source database; research term used; status (i.e., to be read, read, not useful, added to future work, and added to the mind map); URL link; and citation in APA and BibTeX style.

In-depth read of papers

From the initially identified papers, 14 were inaccessible and 5 were excluded as not relevant, and the remaining 121 papers were part of the in-depth analysis.

We proceeded to read each paper in detail in search for content related to technical evaluations, best practices, and value creation in Generative AI or that could be relevant to constructing the deliverables. In this step, 35 papers were selected to be included in the related work.

Selected papers were prioritised if they had produced frameworks, assessment metrics, best practices, guidelines, visual components, artefacts or tables that could be directly used as building blocks for our assessment framework development. Additionally, technical papers that proposed taxonomy of risks, legal laws, emerging evaluations or trends. Our goal was to construct the framework based on existing peer-reviewed literature, composing different works into one singular artefact.

Within the search keywords, the most productive terms were:

1. "Generative models technical disclosure" (9 selected papers) - This term was particularly useful, as it resulted in different papers that did not appear under 'generative AI technical evaluation' keywords search alone;
2. "GenAI categorization" (5 selected papers);
3. "GenAI metrics" (4 selected papers);
4. "GenAI definition" (3 selected papers).

Categorization of most useful papers

Following the in-depth read of papers, we organised the reference works into five categories. These included:

1. Works that provided definitions of technical concepts about Generative AI such as data, architecture, metrics;
2. Works that created assessment frameworks, or provided visual materials, to be used as reference for the second phase;
3. Works that mitigated current problems providing solutions to concerns regarding Generative AI, such as transparency and generated data recognition;
4. Works that described new signals or innovations regarding GenAI, so we would have a view on the current advancements of the existing literature
5. Works that established explicit guidelines and best practices regarding diverse topics of GenAI

Notes were added to each paper including visual materials (e.g., diagrams/frameworks) when available. From this selection, the most useful were saved for future reference to support the development of the framework.

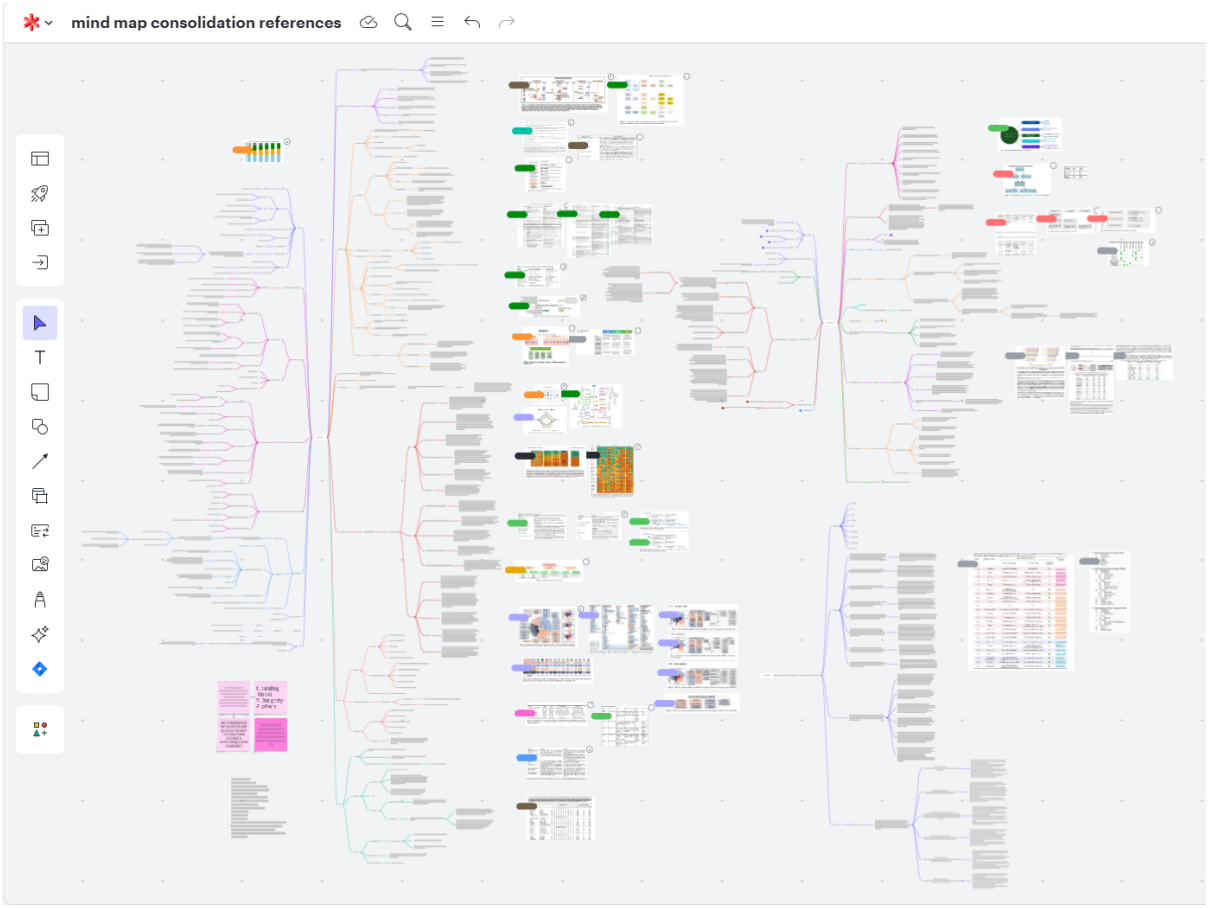

Figure 4.1: Literature references displayed in a mind map to support the ideation and prototyping phases

Connection between papers

After completing the initial categorization of most useful papers, the next step taken was to visually connect all findings in a mind map, illustrated in Figure 4.1. This was particularly helpful in visualizing the branches of topics, subtopics, and how papers correlated to each other. Out of the 35 reference papers included in the literature review, 28 were incorporated into the mind map that was used as a working tool throughout all of the phases of the thesis. With this tool we were able to input images alongside branches, that supported the construction of the related work big picture and GenAI concepts.

This was fundamental for managing the connections between references and topics. In this mind map visualization, three main categorizations emerged organically. The first category included concepts to be used for the assessment framework, Figure 4.2, including definitions of Generative AI, architecture, data, metrics, risks and legal considerations. The second category contained papers that defined GenAI best practices, illustrated by Figure 4.3. The third and last category included two papers that contained specific guidelines on GenAI trustworthiness and UX practices, illustrated by Figure 4.4.

4.2 Internal data collection

We conducted internal research within the company in parallel to the literature review during the immersion phase. Relevant sources included internal and external documents, informal conversations with colleagues, and an internal questionnaire about generative AI knowledge. Additionally, to learn first-hand about the TDD process, we shadowed a technology due diligence project in the company and observed how a consultancy team executes a project.

In the following sections, we explain how we conducted the activities and the results obtained.

Figure 4.2: Mind map with base concepts of Generative AI

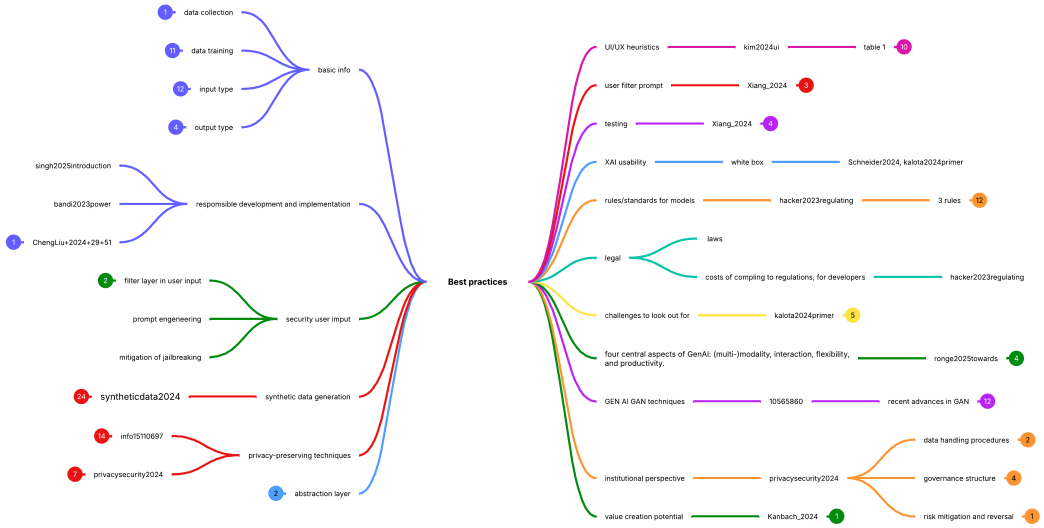

Figure 4.3: Mind map with best practices for Generative AI

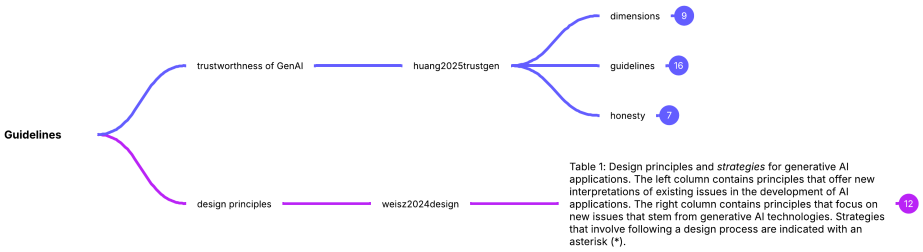

Figure 4.4: Mind map with guidelines for Generative AI

4.2.1 Documentation review

We performed an internal documentation review, looking for information related to AI and Generative AI and specifically for existing technical assessment frameworks. The search was done in two ways: inside-out by researching internal documentation, and outside-in by searching for EY-P GenAI initiatives in publicly available content.

We found that the company has relevant frameworks for assessing technology, including software, cybersecurity, data quality, and AI/Machine Learning, but none for Generative AI.

A note to be made is that within the company, consultants gain experience and knowledge with time and frequently reuse project materials for new projects (e.g., past project decks) or by consulting colleagues who worked on similar projects. It was found that existing assessment frameworks are utilised as a guide in evaluations and are tweaked when necessary to fit specific project scope.

This research branch, which required spending substantial time researching the company's internal documentation, talking to people of interest and consulting colleagues for materials, illustrates the complexity of managing knowledge accessibility in big companies, since more often than not, finding materials was not a straightforward search.

Beyond documentation, we also examined how consultants interact with AI tools. Internally, the use of AI-related tools is allowed and promoted, especially their own GPT tool that the company developed as its own client-zero case, which can handle sensitive client and internal data.

The company also provides regular technical training to consultants on how to use GenAI tools, including use cases, best practices and innovations to improve productivity in projects and in daily work. They also maintain an AI Hub page, data risk guidelines for AI systems, a technology repository that includes diverse solutions, and a monthly newsletter on AI. Furthermore, AI-related events are organised, such as AI Week.

Considering the number of AI-related initiatives and high level of interaction of consultants with the technology, we observed that more data were needed to quantify the specific technical knowledge and experience of consultants about GenAI, with the objective of developing the framework aligned with their technical familiarity. For that, the questionnaire, in Section 4.3, was designed.

This internal research confirmed the initial request for a Generative AI assessment, as the company lacks a structured scope and assessment tools to support technology due diligence of Generative AI, and and we were able to collect existing framework materials to be used as reference for the next phase.

4.2.2 Internal conversations

As part of the exploration activities, we conducted five informal conversations with two managers and three senior consultants. All of these colleagues participated in AI-related initiatives internally or worked in AI/GenAI projects. The content of those chats was documented through field notes after each interaction.

The conversations revealed that diverse parallel initiatives were happening related to AI and Gen AI within different teams. A few of them include developing a use case library, value creation framework for technologies, understanding the internal use of Generative AI in people's work, and encouraging the internal use of AI tooling for various use cases, with attention to data security and confidentiality.

Those conversations were an important source of information, both for referring us to other people to talk to and to documents that helped the exploration, and for revealing that, as a general basis, people do

not know what all the ongoing initiatives related to Generative AI are, and where to find them. Overall, the information is somewhat dispersed and not always directly accessible.

The conversations added a new requirement for the project, regarding communication and conciseness of the materials created in this project, in order to deliver an accessible framework.

4.2.3 Shadowing technology due diligence

During the first month of the research internship, a short two-week technology due diligence project started, and we had the opportunity to join as an observer to gain insight on how a consulting team works. While shadowing the project, the goal was to learn the step-by-step process taken by a team of consultants when approaching a technology due diligence evaluation. This knowledge would help tailor the assessment deliverable, taking into account the process of a technical DD, and translating the findings into a process map to support the research implementation later on.

Technology due diligence process mapping

One of the goals was to understand the process of a technology due diligence. Below is a high-level description of the activities that occur in such a project.

The activities of the shadowed technology due diligence included: proposal document; sending the information request list; introductory session with the client; data analysis; interviews with management, experts, or others; structuring the report; hypothesis drafting; software code-scan; interim status updates and other team organization-related activities.

In this process, there are three main points in which experience, frameworks, and tools are particularly relevant. During (i) the creation of information-request list which, in order to add more value to the project, needs to be specific to the project scope and the consultants need to know what data is relevant for their analysis; (ii) data analysis, where technical knowledge and expertise combined with assessment frameworks, standards, and metrics are important; and (iii) management interviews, in which expertise and experience asking the right questions is crucial to get the necessary information and analyse the quality of the responses provided by the target.

Observations and insights from shadowing

The dynamics of the due diligence project were structured, with specific tasks to be completed and assessments to write. Consultants worked mainly based on their experience, using assessment frameworks, reusing past projects, and consulting with peers that may have worked in similar cases.

A consideration about project deliverables is that all projects use a standard template structure and layout, in which the content varies based on the scope of each project. This included specific standard templates for the technical assessments results as well and those are commonly reused across projects that perform those assessments.

It was observed that experience is really an important factor when executing projects. The main activities of a TDD including defining the project's hypotheses and writing interview agendas, which support the validation of the hypothesis, are essential to the successful completion of a project. These activities require consultants to know what data is needed, what to look for, what are the standard metrics, this information is necessary to perform a quality assessment and produce a quality client deliverable.

In view of that, even though project scopes can be very diverse and unique, making it difficult to have a full standardised approach to their structure, the technical assessment are mainly standard across projects with their defined dimensions, content and results template.

The experience in shadowing the TDD provided insights about the process taken to complete the engagement and we were able to map this process to use as input information for the construction of our assessment framework.

4.3 Questionnaire about GenAI knowledge and the technology due diligence process

The objective of creating this internal questionnaire was to gain insight on the level of technical knowledge of GenAI that consultants have. In addition, inquire about their methodology and tooling used during

Role	Count	%	Age Group	Count	%
Sr. Consultant	12	57.14	25 - 34	16	76.19
Partner	3	14.29	35 - 44	2	9.52
Manager	2	9.52	45 +	2	9.52
Consultant	2	9.52	18 - 24	1	4.76
Sr. Manager	1	4.76			
Intern	1	4.76			
Total	**21**	**100.00**	**Total**	**21**	**100.00**

Table 4.1: Respondent Demographics

a due diligence technology evaluation and gather insights on problems and pain points in their current assessment processes.

In consulting, everyone has their own methods, so the goal was also to find common needs and pains to take into account surrounding tooling, challenges, and technical understanding of GenAI, and use those inputs in the second phase during the ideation and prototyping of our assessment framework.

4.3.1 Questionnaire design

The questionnaire was designed in two parts, first initial questions for all respondents and second with different questions for two types of user case. The first use case comprehended questions for consultants that have only ever done TDD and the second use case was for consultants that have done an AI/GenAI DD. In Appendix B the questionnaire is described in full.

The questionnaire contained 5 sections, each with the goal of finding specific information, described below:

1. **Demographics:** gather background about the seniority of roles and the possible impact in their answers;
2. **Generative AI knowledge:** followed the related works topics to evaluate precisely the level of knowledge consultants had and find gaps;
3. **Technical evaluation process:** initial questions to distribute the questionnaire into the two use cases possibilities;
4. **Technical Evaluation process - first use case:** only TDD work , with a hypothetical scenario question was added on how to approach a technical DD on GenAI, to get information on how consultants who had never done an AI/GenAI DD would approach it;
5. **Technical Evaluation process - second use case:** AI/GenAI TDD, inquiring about specifics about the evaluation of GenAI/AI technology.

The questionnaire was designed with a question structure to ensure that the findings were aligned with the methodology and data collection, so that the analysis is coherent and consistent.

The questionnaire was shared with consultants in Software Strategy Group of the company via internal communication. The number of responses was 21, with 12 respondents for the first use case (i), who only performed a technical evaluation, and 4 respondents for the second use case (ii), who performed a GenAI/AI evaluation; the rest did not have experience with either.

4.3.2 Questionnaire results

Bellow is the recollection of the results from the questionnaire. It includes the direct numbers gathered and numerical interpretations.

Section 1 - Demographics

The respondents demographics are presented at the table 4.1, where most respondents were senior consultants followed by partners, with the bigger age group being in between 24 to 34 years.

Section 2 - General generative AI knowledge

The graph 4.5 illustrates the results of the Generative AI knowledge section of the questionnaire. Comparing the difference in scores between strategy concepts and technical concepts. The highest five average score concepts were Ethics& bias (4), Data quality impact (3.86), Business value identification (3.86), LLM basics (3.81) and Privacy implications (3.71).

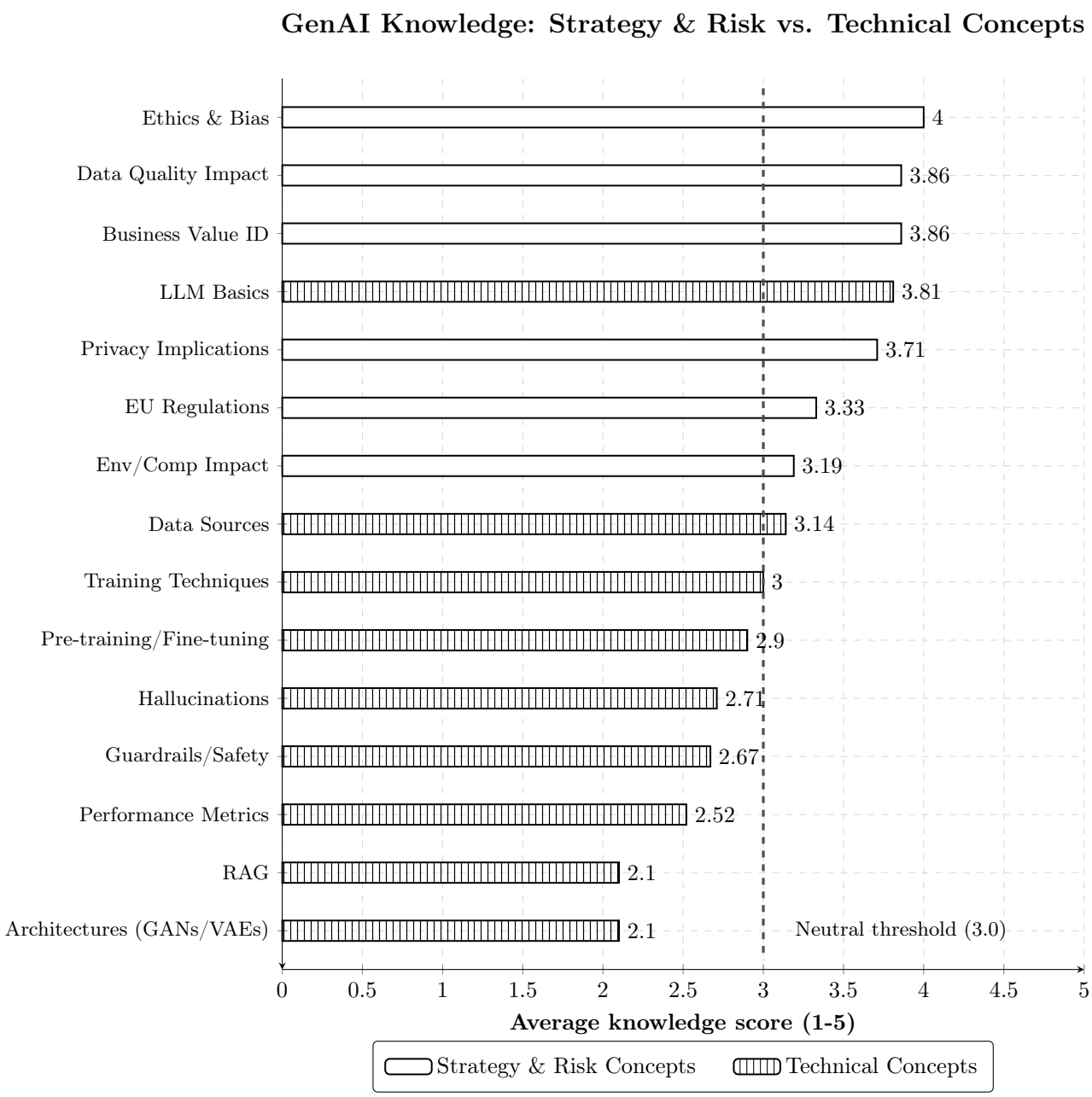

Figure 4.5: Questionnaire results: strategy & risk vs. technical concepts knowledge score

In particular, four technical elements about GenAI had low scores across responses:

- **Performance metrics:** I understand how to evaluate generative AI performance using appropriate metrics - 47.6% disagree;
- **RAG:** I can explain how Retrieval-Augmented Generation (RAG) works and its benefits over stand-alone generative AI - 42.9% strongly disagree;
- **Architectures:** I can distinguish between different generative AI architectures (e.g. transformer-based models, diffusion models, GANs, VAEs) - 38.1% strongly disagree and disagree;

DD Range	Consultant	Intern	Manager	Partner	Sr. Consultant	Sr. Manager	Total
0	–	–	–	–	2	–	**2**
1-5	2	1	1	–	3	–	**7**
6-10	–	–	–	–	1	–	**1**
11-15	–	–	–	1	3	–	**4**
16-20	–	–	–	–	2	–	**2**
20+	–	–	1	2	1	1	**5**
Total	**2**	**1**	**2**	**3**	**12**	**1**	**21**

Table 4.2: Technology DD Contributions by Role

- **Hallucinations:** I understand the concept of hallucinations in generative AI and strategies to mitigate them - 33% disagree.

Section 3 - Technical evaluation process

How many Technical DD's have you contributed on?

The number of technology due diligence distribution among roles is showed in table 4.2, the role category that most participated in TDD were senior consultants. Five respondents participated in more than 20 TDD, 7 respondents ranged between 1 to 5 and four respondents between 11-15.

What aspects do you focus on a technical evaluation?

12 responses included: 33% Roadmap; 25% Architecture, technical and AI; 17% Product strategy and R&D organization; and 8% Miscellaneous topics: roadmap, strategy, AI, tech organization, model performance, hosting& security, technical maturity, technology, technical architecture, software solution, GenAI solution and technical aspects.

Have you ever performed a technical (maturity) assessment of (Gen)AI- native companies?

19 responses: 79% answered no and 21% answered yes.

Section 4 - Technical evaluation process (continuation)

What were the tools used in the evaluation process?

12 responses included: 67% Past projects; 50% Benchmarks; 42% Internal GPT; 33% Excel template, google search and experience; 17% ChatGPT, SSG; and 8% AI assessments, frameworks, projects and own experience, internal GPT and copilot.

What are the most challenging steps and main pain points you encountered during the technology due diligence and evaluation process?

11 responses included:

- 27.3% Data quality/availability, unstructured, polluted, incomplete data or lack data;
- 9.1% Data integration, business alignment, technical/ hosting;
- 9.1% Finding a benchmark/framework that is appropriate to different situations/ contexts;
- 9.1% Defining performance of the AI model; how effective is the actual output. Additionally, defining how it can be deployed on additional use cases;
- 9.1% Sometimes lack of specific knowledge/information.

What did you miss during the technical assessment, or what additional information or resources would have benefited your evaluation process?

12 responses include:

- 33.3% Benchmarks/Frameworks - larger/better/clear frameworks, up to date information, more comprehensive benchmarks;
- 33.3% Data/information - amount, insufficient and more accurate data, access to more information;
- 8.3% More training on tech topics;
- 8.3% Past projects experience base.

Would you feel confident conducting a GenAI maturity assessment of a target organization with your own experience and knowledge?
14 responses: 50% no and 50% maybe.

Can you briefly outline your hypothetical approach to this assessment?
11 responses included framework development and adaptation, asking questions, understanding the AI logic, and other activities.

- "I would ask EYQ for a framework to do a GenAI maturity assessment. Look at it, tweak it to the specific case and go from there";
- "Yes i think i would begin to assess the Data governance model, AI infrastructure assessment (e.g. data pipline, data ingestion etc.), AI R&D org etc";
- "Try to come up with a framework to assess the maturity, and related data requests. Ask for data or interview with the organization. Assess the framework";
- "Current state, competitive best practices, comparison, recommendations";
- "Leverage the AI playbooks and templates developed by AI as a starting point to identify the key areas to evaluate, and develop a questionaire evalutating them on these domains";
- "Use an outlined framework and assess based on that using information from management interviews";
- "First define the type of AI product we're assessing, what the actual use cases are, how this compares to other relevant solutions, what tech stack it is built on, how it is trained, understanding limitations";
- "I would assess: the model itself (proprietary vs off-the-shelf), determine what AI techniques and technologies are used (e.g., diffusion or other type of model, as well as tech stack), i'd assess data inputs (how is the model trained/ pipelines, what are the data sources and how fit-for-purpose are those), assess how well the model is maintained, assess how well data governance processes are set up and adhered too".

Section 5 - GenAI /AI evaluation process

What model of GenAI or AI did you evaluate? (E.g. Transformers, Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Classification or Regression Models, Random Forrest, etc.)
3 responses included: ML model with curation process, traditional AI models, ML, MN, classification models, random forests, KNN.

In the evaluation, what were the main priorities when assessing the GenAI/AI model?
3 responses included: competitiveness, replicability, AI Act readiness, AI maturity, quality of the model, FI, adequacy of responses and accuracy of data.

Were there any differences between a GenAI/AI technical DD and other technologies evaluations? Considering process and content.
3 responses included: DD context, frameworks used, management interviews, experience of the team, tech nuances. In addition: don't treat (Gen)AI as a tool but as a colleague; help it learn

What were the tools used in the evaluation process? (Eg. framework/excel template, benchmarks, google search, EYQ, past projects, own experience)
4 responses included:

- 40% frameworks/benchmarks/tools;
- 40% experience/ expert input;
- 20% AI maturity assessment;
- 20% scoring model.

What are the most challenging steps and main pain points you encountered during the GenAI/AI due diligence and evaluation process?
3 responses included: understanding the quality of the output, translating the technical product to

business value, creating new frameworks, getting and finding information and lack of expertise in the team.

What did you miss during the GenAI assessment, or what additional information or resources would have benefited your evaluation process?
2 responses included: more examples to sparkle ideas, experience, on-boarding training, and adequacy of evaluation forms.

4.3.3 Analysis of questionnaire results and findings

The findings collected from the questionnaire demonstrated correlation between experience, GenAI knowledge and confidence, and also provided valuable insights regarding current methodology and tooling for performing technical due diligence projects, and common pain points that consultants experience in those engagements.

Considering the overall findings, three main correlations between variables were found. First, a correlation (r) of 0.54 between Generative AI knowledge average score and number of contributions in technology due diligence (experience). Indicating that the higher the number of TDDs performed, the greater the GenAI knowledge. Second, a correlation of 0.4 between role seniority and Generative AI knowledge average score, suggesting that seniority is associated with higher knowledge, however, practical TDD experience (r = 0.54) builds more knowledge. Third, a correlation of 0.59 between role seniority and number of TDD performed (experience), as it is expected that with higher seniority there is more experience. These correlations were calculated from the 21 respondents answers, while the sample size is small the insights provided align with the findings from the previous Section 4.2.

In the second section, a strategic versus technical knowledge gap was found in responses scores. As illustrated in Figure 4.5, respondents demonstrated higher knowledge levels about strategy & risks concepts, with average score of 3.71, rather than technical elements about GenAI, with average score of 2.52, with particularly low scores on RAG (2.10), architectures (2.1) and performance metrics (2.52). Indicating that respondents know more about why and how GenAI is used rather about how it works. The breakdown of average score per role is: Interns/Consultants averaged 3.26; Senior Consultants averaged 3.04; and Partners averaged 4, the highest averaged among all roles.

In the third section, we found that consultants focus on specific topics when performing technical evaluations: 33% focus on roadmap, 25% focus on architecture, technical aspects and AI when performing technical evaluation and 17% focus on product strategy and organization. The emphasis focus on roadmap shows that technology implementation is the most recurrent topic in TDD, followed by architecture and technical aspects, and later by product strategy and organization. This shows the strategy nature of current TDD scopes mixed with technical concepts.

The question "Would you feel confident conducting a GenAI maturity assessment of a target organization with your own experience and knowledge?" 50% of respondents stated they would not feel confident and the other 50% stated that maybe, indicating a capability gap important to our research. Analysing individual answers and roles: 2 partners responded maybe; 1 manager responded maybe; 4 senior consultants responded no and another 4 responded maybe; 1 consultant no and another 1 consultant maybe; and 1 intern no. However, in this question a correlation is observed, "no" confidence averaged 3.05 in the knowledge score and "maybe" confidence averaged 3.32 in knowledge score, demonstrating that higher knowledge impacts their confidence and ultimate ability to conduct GenAI assessments with their experience. No explicit correlation was found between level of seniority and confidence, considering that even the most senior roles, partners and managers, indicated maybe in their confidence.

Additionally, the following question asked "can you briefly outline your hypothetical approach to this assessment?". There were four main approaches themes within the answers. First, 4 respondents explicitly answered about developing or using a framework to access the maturity of GenAI. Second, 4 respondents focused on technical evaluation components such as data governance, AI infrastructure, R&D organization, assess current state, best practices, define type of AI being assessed and its underlying logic, use cases, tech stack, training methods and limitations. Third, one respondent gave a comprehensive answer of technical evaluation components, including assessing the model, AI techniques and technologies, data inputs, maintainability and data governance processes. Fourth, 3 respondents mentioned information gathering activities of asking data and interview the organization/ management.

This question provided valuable insights that consultants recognize the need for a specific GenAI framework, that the assessment cannot be done just relying on one framework but rather is a combination of multiple data input across diverse technical evaluation components and data gathering sources. This shows alignment with the finding from Subsection 4.2.1, and explicitly reinforces the need of creating a

centralised singular assessment framework, embedded in other evaluation processes, that would provide a comprehensive means to close the current capability gap in GenAI knowledge.

Even tough the 14 'maybe' and 'no' confidence answers, consultants demonstrated a elevated level of answers in their hypothetical of the assessment of GenAI, showing that they know what its needed to perform it, they just lack the supporting tool itself and additional knowledge.

The question "What were the tools used in the evaluation process?" had two sets of answers for each TDD type. For technology due diligence, the most used tools in the evaluation process were: 67% past projects, 50% benchmarks and 33% excel templates. The highest percentage indicates that consultants rely on previous engagements, and reuse materials across projects as common practice, as also found during the shadowing experience in Subsection 4.2.3. And the second highest number, also indicates that evaluation benchmarks are highly consulted during engagements to support their evaluations, as indeed having market benchmarks with precise numbers helps compare and analysing technical findings against others.

Compared to GenAI/ AI due diligence, among the 4 responded (21% of participants) the most used tools in the evaluation process were: 40% frameworks, 40% experience / expert input, 20% scoring models and, 20% AI maturity assessment. Consultants working in GenAI/AI similarly rely on frameworks and scoring model tools, but also heavily on experience and expert input. This difference in approach between the two due diligence types, indicates the importance in closing the knowledge gap and improve consultants' capabilities regarding GenAI knowledge. It also shows, that the creation of an assessment framework would fit within consultants current tooling, and we could assume that integrating this new material would have no hard adoption barriers.

The question "What are the most challenging steps and main pain points you encountered during due diligence and evaluation process?" had two set of answers for each TDD type. For technology due diligence, it resulted in relevant responses including: Finding a benchmark/framework that is appropriate to different situations/ contexts; Defining performance of the AI model; how effective is the actual output; defining how it can be deployed on additional use cases; and Sometimes lack of specific knowledge/information.

For the GenAI/AI due diligence and evaluation process, answers to the previous question, from the 4 participants included: understanding the quality of the output, translating the technical product to business value, creating new frameworks, getting and finding information and lack of expertise in the team.

These set of answers reveal similar pain points, mostly related with finding benchmarks and creating frameworks, how effective is the output and understanding its quality, lack of specific knowledge, information and expertise in the team. Additionally, TDD respondents said defining performance of the AI model, while AI/GenAI TDD pain is in translating the technical product to business value, indicating that those who engaged in the latter type of due diligence has not a capability gap that extends beyond the technical evaluation to include but how to communicate and correlate those results back into strategy.

Additionally, in the question "What did you miss during the assessment, or what additional information or resources would have benefited your evaluation process?". For technical assessment, respondents indicated: 33.3% larger/better/clear frameworks, with up to date information, and more comprehensive benchmarks. For GenAI assessment, 2 responses included: more examples to sparkle ideas, experience, on-boarding training, and adequacy of evaluation forms.

Those two sets of answers quantify users need for specific resources, including updated information, better clear frameworks, adequacy of evaluation forms. While TDD respondents requested factual responses, GenAI/AI requested more examples to sparkle ideas, experience and training. Indicating that there is a need for greater support regarding GenAI-related engagements

All of those points were taken for requirements of what the GenAI framework should have in order to mitigate the specific pain points regarding GenAI evaluation. The pain point "translating the technical product to business value" implies that even consultants that technical results and are able to complete the assessment, have difficulties in translating this findings. The pain point "getting and finding information" aligns with the conversation findings, in Subsection 4.2.2, that knowledge is dispersed and hard to access.

The central implication is that consultants generally know about the strategy of GenAI but may lack the technical capabilities to evaluate the technology during due diligence engagements, further validating the research problem and objectives.

4.4 Outcome of immersion phase

Across the immersion phase, we validated the research problem in multiple occurrences. Our first finding is that there is no GenAI assessment framework. This is validated by the literature review performed in Section 4.1 and the internal documentation review done in Subsection 4.2.1. Our second finding was about the knowledge gap between GenAI usage, discussed in Section 4.2, experience and GenAI knowledge, discussed in Subsection 4.3.3, that showed that there is a high adoption and AI-related initiatives in the company, however consultants score low (average 2.52) on technical concepts of Generative AI, and that those who have more experience in TDD have a lower technical knowledge gap. The third finding is the validation of an explicit need for a Generative AI assessment framework, from the questionnaire 4.3.3, that consultants in the hypothetical scenario said they would use a framework to perform the GenAI assessment but currently lack a centralised one.

4.4.1 Application of findings into next project phase

An important finding is that there is effectively no standard approach or scope for technology evaluation, every project has a base structure of the most important things to focus on, but it depends on the particularities of the target and the objectives of the TDD. This further validates the research question and reinforces the research problems faced by users (consultants) that Generative AI does not have a set industry standard for evaluation. This makes it necessary to create a deliverable that continues to be relevant with new advancements, providing a solid assessment of the basis of the technology and that can be maintained within the company.

Shadowing the technology due diligence contributed to better illustrate how it could be possible to work alongside the normal "process map" taken by consultants and to integrate the proposed deliverables of this project into it. One of the main priorities in designing the assessment guide is that it will be a widely used tool in the company; thus, the goal is to make its use seamless for them, utilising a delivery format that is most familiar to the user. Chapter 5 provides further information on the design process of the deliverable.

This incorporated another requirement for the implementation of this project, the framework materials should be kept to the minimum and interlinked so that they are easily accessible and available for consultants to reference and use in technology due diligence projects, improving their experience in future interactions.

This helped clarify that the framework should be consistent, easily accessible for consultants, and have components that can be easily reused across different project engagements.

4.4.2 Framework requirements

Upon completing the immersion phase, we defined the requirements for the framework to be used into the next project phase, including the following:

- **R1** Accessibility and conciseness across framework materials (from Subsections 4.2.1, 4.2.2);
- **R2** Alignment with existing due diligence processes (from Subsections 4.2.3);
- **R3** Modular and reusable components across projects (from Subsection 4.2.3, Section 4.3);
- **R4** Technical depth balanced with consultant knowledge level (from Section 4.3);
- **R5** Standard templates for assessment, following company layouts (from Subsection 4.2.3, Section 4.3);
- **R6** Integration of quantifiable metrics and benchmarks (from Section 4.3);
- **R7** Additional support materials beyond the framework (e.g., information request list, management interview questions) (from Subsection 4.2.3);
- **R8** Address specific pain points identified by consultants (from Section 4.3);
- **R9** Framework should support maintenance and updates (from Sections 4.1, 4.3);
- **R10** Framework should support diverse experience and knowledge levels (from Section 4.3);
- **R11** Support the translation of technical concepts into business implications (from Section 4.3).

Chapter 5

Ideation phase

Following the immersion phase, described in Chapter 4, the ideation phase used the research through design methodology and applied design thinking methods to the ideation and prototyping of the framework. The steps taken in this phase included analysing requirements, organizing the supporting material, and defining the structure of the assessment framework.

5.1 Selecting reference material for prototyping

During the ideation phase, we revised the material collected in the literature review, Section 4.1, to use the references as building blocks to define and visualize the high level topics that would compose the framework.

As explained during the methodology review, the selection criteria for references prioritised choosing papers that had created guidelines, frameworks, visuals, indices, or scoring systems that could be directly applied or adapted to our assessment. The objective was to reuse peer-reviewed work as the basis for our framework to ensure that its content was accurate and to provide credibility by building upon academic research rather than developing the content of the assessment from scratch.

The framework underwent multiple iterations, for accountability the exact reference material used in the framework is in Appendix E.

5.2 Preparation and organization of materials

The ideation process primarily involved preparation work, including re-organizing the selected data for the ideation, reviewing the questionnaire results from Subsection 4.3.3 and creating a visual process map from Subsection 4.2.3. In addition to the requirements gathered from the questionnaire results, there were other requirements provided by the mentors that supported this project internally at the company, the combined requirements enabled us to determine the structure and format of the framework.

The process of this organisation is described in three main activities. First, in the ideation process, the references from the mind map 4.1 were retrieved and inputted into a spreadsheet, organised using the topic structure from the mind map, with the goal of finding correlations between the sources and of structuring the content from the mind map by theme. For that we used another visualization method, to refine the source material and correlate how sources within each topic complemented each other, in a systematic format. During this activity, we reorganised and consolidated content and were able to observe how topics could be cross-referenced.

The items marked with * were the topics that provided the most cross-correlation potential. In the first marked topic (*), we bundled together architecture, input and output, and metrics, aligning all those data points in each other. This provided a more comprehensive landscape of how metrics are fit for different output and input variations. The second marked topic (*) represented a substantial recollection of best practices from those different sources, across diverse topics.

The tabs of the spreadsheets and the sources for each topics organised in the spreadsheet were:

- GenAI criteria [Ronge et al., 2025];
- Model documentation [Xiang, 2024];
- Emerging trends in evaluation [Modake and Patil, 2024];

- Data [Stober, 2025];
- Synthetic data [Goyal and Mahmoud, 2024];
- Architecture types correlated with models input and output types [Bandi et al., 2023, Ayemowa et al., 2024, Kim et al., 2024, Gozalo-Brizuela and Garrido-Merchan, 2023, Feuerriegel et al., 2024]*;
- Benchmark of input, output and generative AI models with tasks [Bandi et al., 2023];
- Specific evaluation metrics [Bandi et al., 2023];
- Explainability [Kalota, 2024, Schneider, 2024];
- Trustworthiness [Huang et al., 2025];
- Openness [Liesenfeld and Dingemanse, 2024];
- Risks [Zeng et al., 2024, Rauh et al., 2024];
- Responsible generative AI implementation [Singh et al., 2025, Bandi et al., 2023, Cheng and Liu, 2024];
- Guidelines on trustworthiness, design principles and UX heuristics Huang et al. [2025], Weisz et al. [2024], Kim et al. [2024];
- Best practices across data, security, testing, privacy preserving techniques, rules and standards for models. institutional considerations [Feretzakis et al., 2024, Golda et al., 2024, Xiang, 2024, Stober, 2025, Kim et al., 2024, Hacker et al., 2023]*;
- Value creation [Kanbach et al., 2024].

Second step in the ideation process, included reviewing the questionnaire results, on Generative AI Knowledge, to identify the biggest knowledge-gaps topics that indicated what technical areas consultants lacked knowledge the most in order to guide the prototyping. Detailed findings are in Section 4.3.

Last step, we collected extra requirements for usability and tooling, that were added to the ones defined in Section 4.4.2. For example, the company utilises the Microsoft ecosystem so it was ideal that the framework created used existing Microsoft tools. Additionally, because technical knowledge levels vary among consultants, the framework needed to be as user-friendly as possible so users could understand the assessments without requiring much background knowledge of GenAI. Furthermore, visualizing the integration of the framework within the typical due diligence process was important to guarantee that the assessment would be easily used by consultants during TDD, and eventually become standard practice by default.

5.3 Process map to support ideation

To address the requirement of integrating the framework into existing workflows, we created a visual process map to integrate the framework process and the typical due diligence process, as shown in Figure 5.1. This map was based on the technology due diligence map created previously, see Subsection 4.2.3.

In this map, we created use cases and they were ideated by predicting what would be the most frequent project use cases for the TDDs, and to brainstorm the possible different usages of the framework. The idea was to predict the project scopes and prepare the material of the framework to match possible scenarios, from the perspective of a consultant as user. Four use cases were defined: (i) evaluating a model technically; (ii) evaluating an integration; (iii) developing GenAI and (iv) value creation. Alongside these use cases, a rough definition of the assessment dimensions for the technical GenAI assessment was established, based on the analysis of the reference material made previously, Subsection 5.2.

The framework process map focused on providing visualization of the steps that a user would need to take to be able to go through the evaluation of GenAI, in the four use cases, from start to finish. The first version of the process map provided clarity to understand the relationship points between a normal technology due diligence and the framework process. This helped identify which reference materials are useful for each step, further refining the framework and the use of the supporting reference materials.

The initial use cases identified in the first process map, 5.1, were reviewed during a co-design session with mentors and the professor advisor, who both analysed the process map, identified improvements, and provided validation.

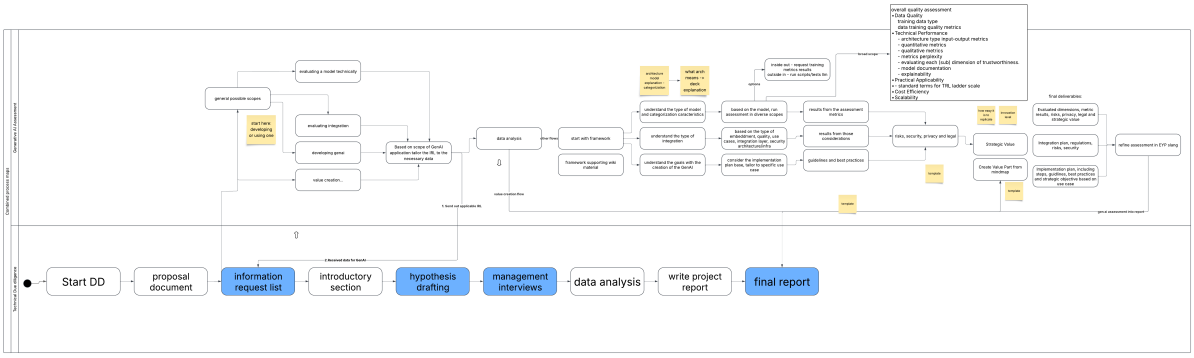

Figure 5.1: Framework process map integrated with TDD flow, first version

5.4 Outcome of ideation and transition into prototyping phase

The ideation phase connected all the gathered information from the immersion, establishing a solid base for the prototyping activities. This included the two sets of requirements: from the questionnaire outcome, Subsection 4.4.2, and from those identified during the preparation activities, Subsection 5.2. Additionally, we correlated related work sources, defined a process map for four framework use cases, hypothesized the consultants-framework interaction, and refined the map of the overall assessment process **??**.

With those elements, the next chapter 6, accounts for the five prototype iterations of the framework, each refined co-design sessions, aligning the selected related work and business requirements.

Chapter 6

Prototyping phase

This Chapter will explain the prototyping phase, including co-design sessions, requirements, design decisions for the framework format, the prototype evolution with the iterations of its versions, the process of creating the framework metrics, the refinements made based on the feedback from company mentors and finally the core challenges faced while prototyping.

6.1 Beginning of prototyping: decision on format and assessment dimensions

Following the preparation work, the prototyping ideation began, closely supported by the mentors, with co-design sessions, weekly meeting and feedback iterations.

The first design choice to be made was the actual framework format, with the requirement of being accessible in a familiar format for consultants, we decided to present the content in a slide deck, since it is the most used format company-wide. In addition, client's project reports also follow slide deck templates and layout, and in order to easily integrate the assessment within the standard due diligence process, it would facilitate the use of the framework in the slide configuration.

During immersion in selection of the ideation data, recurrent topics about Generative AI were identified that ultimately became the initial assessment dimensions for the framework. The topics included: *model documentation, model data, output quality, explainability, trustworthiness, risks, related laws, and human-AI*. Based on the research, we defined those dimensions as the most important aspects to consider when evaluating Generative AI for a technology due diligence assessment, as it provided the fundamental basis of generative AI technology, insights on its underlying quality, and compliance considerations. All of the assessment dimensions, were chosen to align with the normal TDD project scope, but considering the required GenAI perspective on them.

The beginning of the prototyping process established a solid foundation for developing the assessment framework, ensuring it had a structured technical knowledge base and met the practical requirements needed.

6.2 Framework prototype iterations

The chosen format for the framework, as mentioned earlier, was a slide deck. The material of the framework went through multiple iterations across this phase, each in order to improve its format, layout, structure, align content and the requirements. Due to these iterations, we modified the format of how to approach the input and output of the assessment iteratively, in this section we will describe the iterations and the prototype versions changes.

The contents of the slide deck, the chapters and the assessment dimensions topics changed with the evolution of the framework, multiple iterations on the structure and content of the slide deck and of the framework were made during the prototyping process, until we reached the prototyping version 5 that was used for the usability tests in Section 7.2.

Dimension	Subdimension
Profiling	GenAI use case; Architecture type; Model, input and output type; Prescribed tasks
Model documentation	Model training/testing criteria; Model data; Guidelines and regulations; Guidelines biases and ethical issues; Data profile & organization; Data source& provenance; Data practice & training
Model data assessment	Data quality metrics; Data volume assessment; Ownership and rights; Ethical considerations
Evaluating Generative AI model Technically	Output quality; Diversity; Trustworthiness; Explainability; Openness level; Human-centric; Ethical
Evaluating UI/UX usability	
Evaluating Generative AI integration	Use and goal of integration; Embedment level (Architecture); Abstraction layer; Security; Data management; Regulations and legal; TDR ladder scale;
Risks, security and legal	

Table 6.1: Prototype version 1

6.2.1 Prototype version 1

This version of the prototype was created based on the four use cases structured on the process map illustrated in 5.1, the use cases were taken from the process map along its process:

- **Generative AI evaluation:** When evaluating a model technically the step by step of the process is 1) tailor the IRL to the necessary data; 2) data analysis; 3) model type and categorization; 4) perform technical assessments based on the scope; 4.1) inside out requesting training metrics results or 4.2) inside out running scripts and tests on LLM; 5) gather results from the technical assessment; 6) assess risks, security, and legal; and 7) strategic value creation;
- **Integration evaluation:** When evaluating the integration of Generative AI with a system the step by step process is 1) tailor the IRL to the necessary data; 2) data analysis; 3) define the integration type; 4) evaluate the embedment quality and use cases; 5) gather results from integration assessments; 6) assess risks, security and legal; and 7) strategic value creation;
- **Development of Generative AI technology:** When assessing a Generative AI development or developing one the step by step process is 1) tailor the IRL to the necessary data; 2) data analysis; 3) understand the goals with the Generative AI dev; 4) define implementation plan and use cases; 5) refer to development guidelines and best practices; 6) assess risks, security and legal; and 7) strategic value creation;
- **Value creation strategy:** When creating or assessing a value creation strategy the step by step process is 1) tailor the IRL to the necessary data; 2) data analysis; 3) define strategic value; 4) apply value creation frameworks.

The four use case processes were combined with the technical referenced content in 5.2 were integrated to form the first draft of the structure of the assessment, illustrated in 6.1.

That first version of prototype reflects the slide deck sections organization, format of the material. The first assessment structure was as follows:

1. Generative AI technical information;
2. Profiling of Generative AI project case;
3. Model documentation: Model Data Assessment;
4. Evaluating Generative AI models technically: model architecture, type of mode, data quality, technical performance, practical applicability, cost efficiency, scalability;
5. Evaluating Integration: map integration, quality, use caes, integration layer, security, architecture and infrastructure;
6. Developing Generative AI: implementation plan, use case, guidelines and best practices;

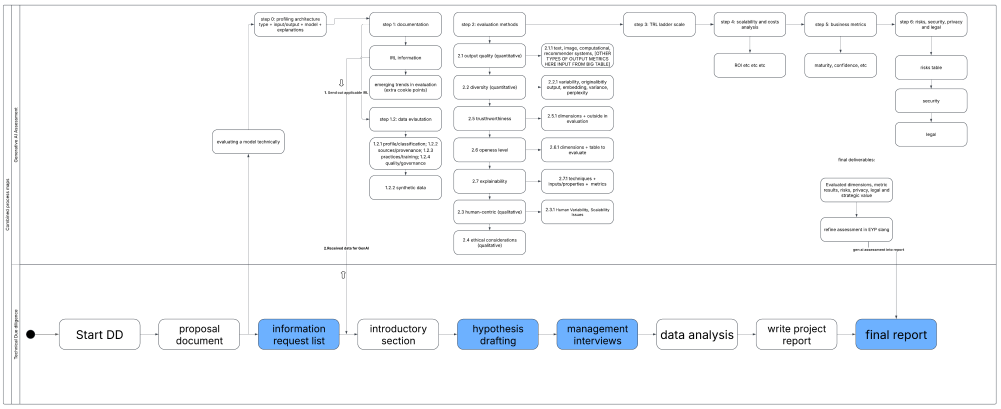

Figure 6.1: Framework process map integrated with TDD flow, second version

7. Value creation Generative AI: strategic value;
8. Risks, security and legal;

This version was revised in sessions with the mentors and was iterated in version 2.

6.2.2 Prototype version 2

The first prototype version provided structure for the storyline, but upon examination revealed usability issues. The mentors from the company, reviewed the framework acting as users, and had trouble to differentiate the goal of the diverse evaluations since they had similar goals. And ultimately, in practice, providing people with four assessment options could create confusion and impact their usability of the framework, increasing the framework adoption time.

The initially identified use cases were revisited and modified considering the evolution of the prototyping process and the feedback received from the mentors, that the use case (i) evaluating a model technically and (ii) evaluating an integration, contained many duplications and would be better to merge them under one use case for conciseness purposes. And instead of having four separate use cases entrance points, we thought it best to create one integrated flow for the whole assessment, with all assessment items streamlined into one process streamlined. The second process map, illustrated in Figure 6.1, reflects the changes made in the assessment flow.

This new version prioritised the technical assessment of Generative AI use case but sacrificed the granularity that the other parallel use cases provided to the framework. Even though the 4 use cases provided a granular coverage of possible scenarios, they reduced the usability of the framework, adding more complexity than was necessary to its users, creating confusion on how and when to choose each use case.

Additionally, an UX/UI usability assessment was added following the usability heuristics assessment template created by [Kim et al., 2024]. We also added separate sections for development guidelines, best practices, and value creation, marked in * in the list below. They were not included in the second version of the process map as they are additional information provided to the users, rather than part of the "mandatory" process of the assessment. The second prototype version is illustrated in Table 6.2.

With the new process definition for the assessment, the first slide deck was created, its content index is described below:

1. Generative AI technical information;
2. Profiling of Generative AI project case;
3. Model documentation: Model Data Assessment;
4. Evaluating Generative AI models Technically: Output quality, Diversity, Trustworthiness, Explainability, Openness, Human-centric, Ethical;
5. Evaluating UI/UX Usability*;
6. Evaluating Generative AI ladder scale;

Dimension	Subdimension
Profiling	GenAI use case; Architecture type; Model, input and output type; Prescribed tasks
Model documentation	Model training/testing criteria; Model data; Guidelines and regulations; Guidelines biases and ethical issues; Data profile & organization; Data source& provenance; Data practice & training
Model data assessment	Data quality metrics; Data volume assessment; Ownership and rights; Ethical considerations
Evaluating Generative AI model Technically	Output quality; Diversity; Trustworthiness; Explainability; Openness level; Human-centric; Ethical
Evaluating UI/UX usability	
Evaluating Generative AI integration	Use and goal of integration; Embedment level (Architecture); Abstraction layer; Security; Data management; Regulations, legal; TDR ladder scale;
Risks, security and legal	

Table 6.2: Prototype version 2

7. Scalability and costs;
8. Business evaluation metrics;
9. Developing Generative AI: Developing Guidelines*;
10. Best practices Generative AI*;
11. Risks, security and legal;
12. Value creation Generative AI*.

6.2.3 Prototype version 3

In next refinement iteration, the mentors pointed out the importance of having a clear storyline that tells a concise story about the assessment and does not leave the users wondering what to do next. In order to achieve that, we polished the content of the slides. First we needed to improve the storyline, version two had many chapters, all independent from each other, deeming difficult to follow the flow of the assessment, how to execute it, and understand how topics related to each other. For that, it was important to revise the storyline with the goal that the slides explained themselves, reading as a guide, and could be used independently without relying on outside support.

To solve this requirement, we revised the content of the previous "Evaluating Generative AI models Technically", and created three overarching categories, adding a higher layer to the assessment dimensions. The three categories are: Operational, grouping dimensions about the technical specs of GenAI systems; Quality, grouping dimensions that provided specific insights and criteria about the model beyond technical specs; and Compliance, grouping dimensions to assure security and compliance of models.

Additionally, the "Generative AI technical information" was moved to the end of the slides, so users could visualize immediately the assessment itself, rather than supporting GenAI information. Besides that, we added a usage guide section, that explained the assessment framework methodology, its steps and the expected output. The third prototype version, reflecting those changes, is illustrated in Table 6.3.

Those new modifications let to a new slide structure, described bellow:

1. Usage guide;
2. Operational assessment dimensions: Model documentation, Data, Output quality metrics;
3. Quality assessment dimensions: Model Trustworthiness, Model Explainability;
4. Compliance assessment dimensions: Security, Governance, Legal, Human-AI and ethics;
5. Generative AI usability assessment: UI/UX evaluation;

Category	Dimension	Subdimension
Operational	Model documentation	Model data; Data lineage and privacy compliance; Model training/testing criteria; Technical architecture and specifications; Use case and operational boundaries; Solution ownership and replicability; Guidelines legal and regulation; Guidelines biases and ethical
Operational	Data	Data quality metrics; Data volume assessment; Ownership and rights; Ethical considerations
Operational	Output quality	Latency; Throughput; Accuracy; F1 score; Precision
Quality	Model trustworthiness	Truthfulness; Safety; Fairness; Robustness; Privacy
Quality	Model explainability	Foundational XAI sources & methods; Explanation scope & integration coverage; Human-centered explanation quality; Interactivity & personalization; Model access & transparency; Security& ethical safeguards; Sample difficulty & resource adaptation; Method-specific quality metrics
Compliance	Security	Foundation model; Privacy& data protection; Prompt security & input manipulation; Output security & content control
Compliance	Governance	Data handling procedures; Governance structure; Risk mitigation and reversal
Compliance	Legal	Automated decision making; Discriminatory activities; Unauthorised privacy violations; Deception; Advice in heavily regulated industries; Types of sensitive data; Autonomous operation; Manipulation; Hate/toxicity
Compliance	Human AI and ethics	Human control & agency; Human-AI collaboration; User experience & satisfaction; Ethical alignment & fairness; Explainability & trust

Table 6.3: Prototype version 3

6. Generative AI development guide: Developing generative AI roadmap, TRL scale;
7. Generative AI reference guide:
8. Technical information: technical terminology, architecture types, training types, synthetic data generation, environmental and ethical concerns;
9. Risks, legal and security: GenAI risks, GenAI security considerations, legal regulations (EY, global, USA);
10. Generative AI guidelines, principles and best practices: trustworthiness guidelines, design principles, and overall best practices;
11. Value creation assessment: profiling of company current state vs desired state, resources allocations.

Regarding the construction of the slide deck, we leveraged visual elements and slide templates from diverse proprietary materials from the company, re-utilizing them to lower the amount of work necessary to create the slide deck and to comply with formatting requirements of the company. We used templates as the base for the slides layout, to display the information necessary for each assessment dimension content.

In this iteration, we also started developing the scoring for each assessment dimension, inputting the reference material gathered into the slides and also creating the maturity levels for some assessment dimensions. However, it is import to note, that in this version each dimension had very simple maturity levels, and they were diverse from each other, varying in content and form. For this reason, we will not describe how and what were those initial maturity levels, as they were changed and standardised in the prototype version 5 6.2.5.

Additionally, in this iteration we started the creation of management questions for each assessment dimension. We created them as support material for the execution of the assessment, complying with the requirement of additional materials to support the framework assessment process. They followed the structure of the assessment domains, for each subdimensions supporting questions were created, that consultants could use those questions directly when performing management interviews. The writing style and format, complied with the provided rules from the company.

6.2.4 Prototype version 4

The prototype version 4, was a improvement iteration of the version 3. First there were some issues, that were pointed out during feedback sessions with the mentors: the layouts for the scoring sheets of assessment dimensions were unstandardised, the management questions were scattered across the slide deck in the beginning of each dimension, and the slides containing the technical reference scores for some assessment domains (i.e., output metrics, LLM trustworthiness, and explainability) were disrupting the storyline flow, due to the technical nature of its content.

To solve those issues, first we standardised all the scoring sheets templates that contained the description of the assessment subdimensions for each dimension. Adding a table with the name of the subdimension, its description, the input of the maturity level score to be filled in after the execution of the assessment. Second, we centralised in one section all the supporting management questions, and third we moved to the appendix all the technical reference slides to improve the readability of the slides. Additionally, general layout improvements were made to comply with the company rules and guidelines on how to format slides, including action titles and subtitles.

There were some additional modifications regarding the topics of the assessment dimensions: we switched diversity to architecture in the operational assessment; we moved the output metrics to the quality assessment; and we simplified the human-AI, since the ethics content was already present in the governance and legal assessment, creating repetition.

The architecture dimension was added because it was an important topic, that was found repetitively throughout the literature review. Initially it was not included to avoid duplication with the normal architecture TDD flow, as architecture is part of the usual tech due diligence scope, but upon review and clarification we decided to included it with AI-specific architecture evaluation topics. The fourth prototype version, reflecting those changes, is illustrated in Table 6.4.

The new modifications resulted in a new assessment slide structure:

1. Usage guide - explaining how to use the framework;

Category	Dimension	Subdimension
Operational	Model documentation	Model data; Data lineage and privacy compliance; Model training/testing criteria; Use case and operational boundaries; Solution ownership and replicability; Guidelines legal and regulation; Guidelines biases and ethical
Operational	Data	Data quality metrics; Data volume assessment; Ownership and rights; Ethical considerations
Operational	Architecture	Technical architecture and specifications
Quality	Output metrics	Performance foundation; Operational foundational; Quality health; Task-specific
Quality	Model trustworthiness	Truthfulness; Safety; Fairness; Robustness; Privacy
Quality	Model explainability	XAI sources & methods; Explanation scope & integration coverage; Human-centered explanation quality; Interactivity & personalization; Model access & transparency; Security& ethical safeguards; Sample difficulty & resoure adaptation; Method-specific quality metrics
Compliance	Security	Foundation model; Privacy& data protection; Prompt security & input manipulation; Output security & content control
Compliance	Governance	Data handling procedures; Governance structure; Risk mitigation and reversal
Compliance	Legal	System & operational risks; Content safety risks; Societal risks; Legal & rights related risks
Compliance	Human ai	Human control & agency; Human-AI collaboration; User experience & satisfaction; Ethical alignement & fairness; Explainability & trust

Table 6.4: Prototype version 4

2. Detailed assessment dimensions;
3. IRL questions;
4. Operational assessment: model documentation, data quality, architecture;
5. Quality assessment: output metrics, model trustworthiness, model explainability;
6. Compliance assessment: security, governance and legal, human-AI;
7. Usability assessment;
8. Development guide;
9. Reference material;
10. Appendix.

6.2.5 Prototype version 5

Even with the improvements made in version 4, the company mentors still had difficulty reviewing the prototype slides and fully comprehending its content and how to execute the GenAI assessment following the material of the slide deck.

With this finding, we decided to separate the framework content, creating three supporting materials, each with clear specific, content to support the assessment process.

Initially, in the ideation of the framework, our strategy consisted of using the references works, selected as the building blocks, to develop the framework and its diverse assessment components. However, every reference material had different approaches and assessment structures that created confusion when a user, in this case the mentors, went through the assessment flow in the slide deck and had to actually understand the methodology for scoring each subdimension, that both had and lacked standardization between them.

Category	Dimension	Subdimension
Operational	Model documentation	Model data; Data lineage and privacy compliance; Model training/testing criteria; Use case and operational boundaries; Solution ownership and replicability; Guidelines legal and regulation; Guidelines biases and ethical
Operational	Data	Data quality metrics; Data volume assessment; Ownership and rights; Ethical considerations
Operational	Architecture	Technical architecture and specifications
Quality	Output metrics	Performance foundation; Operational foundational; Quality health; Task-specific
Quality	Model trustworthiness	Truthfulness; Safety; Fairness; Robustness; Privacy
Quality	Model explainability	XAI sources; Explanation scope; Human-centered explanation quality; Personalization; Security safeguards; Adaptation; Method-specific quality metrics
Compliance	Security	Foundation model; Privacy& data protection; Prompt security & input manipulation; Output security & content control
Compliance	Governance	Data handling procedures; Governance structure; Risk mitigation & reversal
Compliance	Legal	System & safety risks; Societal risks; Legal & rights related risks
Compliance	Human AI and ethics	Human control & agency; Human-AI collaboration; User experience & satisfaction; Ethical alignement & fairness; Explainability & trust

Table 6.5: Prototype version 5

This created a usability issue and also a scoring normalization problem. The mentors also raised the question of how consultants should normalize the score of each individual subdimension and interpret the results if each assessment consisted of different approaches.

To solve this, during a co-design meeting with the mentors, we decided to create a spreadsheet scoring model, similar to the ones existent in the company. To achieve this, we had to normalise and standardise all the assessment content into one format that supported a clear assessment flow and scoring methodology. This work is described in Section 6.3.

To build this, we kept the reference material content as our source of information, and created maturiy level based on reference, maturity levels with progression between 1, 3, and 5 levels that allowed a levelled scoring throughout the assessment dimensions. The choice for 1,3,5 maturity levels instead of 1-5 was because this is the standard used at EY-P scoring frameworks and this pattern allows for more flexibility to consultants when scoring the framework, which is needed since the assessment framework needs to fit diverse use cases, industries and solutions. The fifth prototype version, the last version of this phase, is illustrated in Table 6.5.

In this prototype iteration we took out all the extra information within the assessment dimensions of the slide deck from version 4, and left just the scoring sheet templates, presented in table format, with the subdimensions descriptions, space for filling in maturity level scoring and commentary. The previous version contained the methodology for scoring each subdimension across the dimensions; this information was transferred to the scoring model spreadsheet and normalised to a standard scoring methodology that contained specific descriptions for each maturity level.

Additionally, several components along all the technical information of GenAI previously from earlier versions was transferred to a new slide deck acting as an informative guide for consultants' reference. Reallocated material includes the best practices, development guide, UI/UX heuristics evaluation, and TRL ladder scale. The value creation assessment was removed entirely as other teams at the company were working in similar initiatives.

In this version, the previous slide deck from version 4 was reduced from 72 slides to 32, keeping only the most essential information: the detailed assessment dimensions with only 3 slides per category. The supporting management questions were kept the same from the previous structure. In addition, the appendix, which previously contained the technical reference material for assessment completion, changed to contain additional template slides to be used in project reports.

The framework materials became three: (1) slide deck with assessment templates to support the framework, (2) scoring spreadsheet with scoring input, and (3) slide deck with GenAI reference material,

development guide, best practices and guidelines.

The new slide deck with the assessment template structure:

1. Usage guide: how to apply this framework;
2. Detailed assessment dimensions;
3. Operational: model documentation, data, architecture;
4. Quality: output metrics, model trustworthiness, model explainability;
5. Compliance: security, governance, legal, human AI and ethics;
6. Supporting management questions;
7. Appendix: template slides for project reports.

The new scoring spreadsheet structure:

- Scoring input sheet;
- Scoring output sheet;

The new slide deck with GenAI reference material:

- Guidelines;
- Development guide of generative AI;
- Technical reference information.

This marked the fifth prototype iteration in this phase, this version was subsequently used for the testing phase in Chapter7.

6.3 Development of metrics and maturity levels

This subsection describes the maturity metrics developed for the assessment framework. This process started during the third prototype iteration, Subsection 6.2.3, when we started developing scoring for the framework subdimensions. After defining the process map and the framework structure, we started developing scoring metrics for some subdimensions of the framework. Until the fourth prototype iteration, the scoring methodologies for the subdimensions were misaligned and did not follow a specific pattern. For this reason, in prototype version 5, Subsection 6.2.5, a two new requirements were identified due to previously unstandardised scoring; one the need for numerical maturity levels and specific metrics in the assessment dimensions to ensure that assessment was consistent across projects and two that the technical evaluation scoring was within a specific range, from 1 to 5.

To address this requirement for standardised metrics and maturity levels across all assessment dimensions, we used Claude AI and Perplexity to develop the framework metrics. This involved a prompting methodology that instructed the tools to use only the information provided as references (i.e., literature source files) and to create specific metrics for each assessment dimension with a set of requirements. See Appendix F for the prompts used in prototype version 5. Earlier prompts are not included.

This prompting approach involved three steps: (1) retrieving technical information about the dimension topic from the reference source material; (2) inputting the research objective and questions for contextualization; and (3) requesting the generative AI tool to create five-level metrics to assess the technical information considering the source material, objectives, and research questions.

The objective of using AI tools was to create metrics grounded in the source material, to maintain reference content as the primary source of information for the assessment framework (and other deliverables), and to address project timeline constraints while maintaining methodological rigour.

To ensure output quality, we established a human-AI collaborative approach that leveraged the analytical and information synthesis capabilities of the tools whilst maintaining human oversight for revision and validation. All outputs were reviewed by the researcher to ensure the quality of the generated metrics, and iterative modifications were made as necessary.

6.4 Transition between prototyping and testing project phases

The last weeks of the prototyping phase were dedicated to reviewing and receiving feedback from the mentors and advisor. During those feedback iterations, the size and content of the framework were not easily digestible by the EY-P mentors, who approached the framework from a consultancy perspective rather than an academic one. As a result, it became clear that was necessary to simplify and condense the framework in order to make it comprehensible to all EY-P consultants.

This finding redefined how to proceed with the framework, demonstrating that it needed to be less technical, more concise, and easier for consultants to understand. The team decided to standardize all the assessment dimension methodologies, which had previously varied, and utilize the maturity level scoring approach across the whole framework. All of those changes were reflected in the prototype version 5 6.2.5.

This marked the transition from the prototyping to the testing phase with more direction and context. This new perspective shifted the original testing approach: from testing the content and different methods of performing each singular assessment to testing the usability, in the next project phase, and ultimately the success of the maturity-level assessment method for the framework.

Chapter 7

Testing phase

Following the prototyping, and the new testing approach with the standardized scoring approach, Section 6.4, the testing focuses on validating the usability of the framework prototype using a design thinking approach.

This chapter describes mock case creation, individual usability testing sessions, test results, and subsequent framework refinement.

The testing method provides users with a task to complete during the testing session, and the protocol asks users to share their thoughts while interacting with the framework. In addition, the facilitator observes users with minimal interference, documents their reasoning and interaction with the framework while completing the task, and reinforces the think-aloud protocol if needed.

The use of the protocol for human-computer interaction was adapted by Lewis [1982] from the original workin cognitive psychology by [?]. This adaptation result showed that when users were asked to verbally express their thoughts while interacting with systems, it provided information about their mental model and approach to its interface. For this reason, although not formally integrated, this combination of method and protocol is often used in testing scenarios, as it enables researchers to gather valuable insights that would not be accessible otherwise.

Applying this protocol in our context enables a deeper understanding of the usability of the framework from the user's perspective, helps to discover diverse approaches to complete the same task, and provide insights into whether the level of Generative AI technical knowledge may or may not affect the task completion success.

7.1 Mock case creation for test

The creation of a mock case to support the test dynamic was suggested by the advisor as a means to illustrate the use and applicability of the assessment framework for the participating consultants. Using mock cases for assessments is a common practice in EY-P, as they are applied in interviews and internal trainings.

The criteria for creating the mock case were that it needed to mimic a TDD scenario and contain sufficient information to serve as a robust case example. The mock case was designed as a buy-side transaction, sufficiently realistic to engage the test subjects during the test dynamic. The scenario was designed to closely simulate the real-world practice, featuring a fictional company using generative AI technology in the application layer.

Creating a mock case for a technology due diligence from scratch is time-consuming, and the we considered it a adequate use Generative AI to support this process. The tool used was Perplexity, with the Claude 4.0 Sonet model, incorporating both web and academic sources. See Appendix F.2.1.

7.1.1 Acceptance of the created mock case

The acceptance criteria (AC) for the mock case were defined as:

- **AC1:** The scenario is fully realistic and does not requires a suspension of disbelief;
- **AC2:** Complexity and size are adequate given the time constraints for the test dynamic;
- **AC3:** The technology description is plausible and has no mention of non-existing technology;

- **AC4:** The case does not include window dressing scenarios of Generative AI use in the mock company.

The provided case options generated by Perplexity were analysed against the defined acceptance criteria, and the quality of the selected case was double-checked and confirmed by the advisor.

It is important to note that the prompting followed a trial-and-error approach, including varying levels of research context and background information to observe the effect it would have on the mock case output. A total of sixteen different prompts were tested, using both Perplexity pro search and the Perplexity deep search for comparison, before choosing the one whose output provided the most alignment with the acceptance criteria. Upon selecting the optimal prompt, the mock case output was iteratively refined to arrive at the final mock case scenario that is described in full in Appendix C.

The rationale for the refining process of the Perplexity generated output that led to the final choice of mock case used during the testing dynamic sessions, followed a evaluation of each iteration output until the mock case was deemed acceptable. Between the mock company options a logistics company was chosen as it is a business that is broadly know and realistic that they could be improving with Generative AI.

Four versions of the mock case were necessary to arrive at the final mock case. First version was a good generic use case in the logistics transportation that was realistic but the text style was too factual, lacking a narrative in the beginning to provide better contextualization of the case, although the mock company financials were interesting. Two sections of the case were considered irrelevant for the test scope (Technical Challenges & risks and Competitive landscape). Second version, showed an improved narrative, but too picture perfect, lacking the most challenging part of non-AI native companies turning AI which is the migration phase. Third version was more aligned with reality of non-AI companies migrating to utilizing a AI platform/application.

Upon review, the usability test dynamic changed scope, from a group session to a individual session, where users would assess only one assessment dimension out of the nine in the framework. The chosen assessment dimension was explainability. Due to this change, the case needed to be further tailored and fine-tuned to align with the testing. In the next section, 7.2 we have the complete description of the testing dynamic.

The change in scope, meant another iteration on the third version. The content of the version was adequate against most of the acceptance criteria, but was overly long, containing more than 6 pages of content, unfit for an individual testing session.

For this reason to comply with the AC2, a new smaller version was created based on the content of third version. The fourth tailored version contained information to provide the user enough context to complete the explainability assessment, half of the assessment subdimensions in the smaller mock case were present, providing a realistic situation with more focused information on the explainability but without giving away the whole assessment making it too easy on the user.

In Appendix F.2.1, only the optimal prompt and its subsequent iterations in Perplexity are documented. For replication purposes, all the other prompts whose output did not sufficiently comply with the acceptance criteria were deliberately not included. In addition, the prompts used to create the fourth version and the full text of the mock case are described.

7.2 Individual usability testing sessions

Initially, the testing approach was to be a group setting with 4-6 participants. They would pair up to apply the framework to specific assessment dimensions, and then conclude with a group discussion. This approach was intended to provide an overview of the whole framework in one session, optimizing time, stimulating discussion, and allowing opinions to be exchanged.

Due to scheduling constraints, the format was shifted to individual sessions, where we could still test the framework methodology and receive feedback even with a different format. Furthermore, the results of those testing sessions could be extrapolated to all nine assessment dimensions, as they have the same format.

To evaluate both the usability and the acceptance of the framework, a mixed-methods approach was used that combined qualitative observations from facilitator notes and session recording with quantitative evaluation through a standard TAM questionnaire.

Specifically, to quantify the results of the test, the constructs from TAM1, TAM2 and TAM3 created by [Davis et al., 1989] were used as input to create the questionnaire. The TAM constructs applied were: Perceived Usefulness, Perceived Ease of Use, Output Quality, Result Demonstrability, and Behavioural

Intention. To support the creation of the questionnaire, Perplexity AI was used, the complete prompt descriptions are available on Appendix F.2.3.

The individual test involved providing a task to the participants for them to execute. The chosen task was to perform the framework's explainability dimension assessment, based on the information from the mock case, using the scoring sheet and the slide deck materials. Explainability refers to the ability to understand, interpret, and provide human-comprehensible explanations for how models create their outputs. This dimension was chosen because while it is an unknown concept to most non-technical people, it can is intuitively understandable. Also, all participants regularly interact with Generative AI interfaces, providing them with experience in understanding how GenAI provides outputs.

Additionally, by focusing on only one assessment dimension, it allowed us to perform the test working with the limited availability of participants schedules and understand the usability of the framework, iterate on the received feedback, and improve the framework based on the findings.

The total materials used for the testing session were: testing agenda in Subsection 7.2.1, version 5 of the framework in table 6.2.5 with the scoring card, the mock case in AppendixC, and the TAM questionnaire in Appendix D.

To validate the dynamic of the testing, a dry run test was performed with the thesis advisor to validate the testing script, content, and timing using version 4 of the prototype described in **??**. The dry run demonstrated that the agenda and task for the test were appropriate and could be completed in the estimated time, under 60 minutes. After this dry run, the individual usability tests were performed with the goal of testing the subjects attitude towards the framework methodology in a structured manner.

7.2.1 Individual usability test agenda

The testing agenda was created to define the structure and timing of the testing session. Perplexity was used to support the revision and improvement of the user session agenda, the complete prompt descriptions can be found in the Appendix F.2.4

Test Session Agenda

Context and Instructions: 5 minutes

- Welcome participant, obtain recording consent and start recording;
- Brief explanation: "You'll be helping us understand how consultants might use a Gen AI assessment framework. Use these materials to conduct your assessment as you naturally would";
- Introduce the think-aloud protocol: "Please share your thoughts out loud as you work through this".

Task execution: 45 minutes

- Present the small mock case and materials simultaneously;
- Request to participant to complete the explainability assessment evaluation based on the the mock case material;
- Facilitator stays as a silent observer, answering questions when requested and, making notes on how the user is tackling the task.

TAM questionnaire: 5 minutes

- Administer 6-question TAM survey immediately after framework experience, asking user to base the answers on the task experience.

Debrief and wrap-up: 5 minutes

- Facilitator asks any final thoughts to the user about the assessment material, approach and other considerations.

7.2.2 Participant profiles

Individual usability tests were performed with seven consultants from EY-P Strategy & Transactions, their profiles are described in Table 7.1. They were recruited based on their profile but also on their availability to participate in the testing sessions. They were contacted by direct message or in person at the office. The number of participants was based on their availability to be in the office in person, with

Table 7.1: Participant Profiles for Individual Usability Testing

ID	Role	Years (at EY-P)	TDD Experience	AI Strategy Projects
P1	Senior Consultant	3.0	Yes	No
P2	Consultant	1.2	No	No
P3	Consultant	0.1	No	Yes
P4	Senior Consultant	3.3	Yes	No
P5	Senior Consultant	3.0	Yes	Yes
P6	Consultant	1.7	No	No
P7	Consultant	1.2	No	Yes

the aim of performing as many testing sessions as possible while managing scheduling and availability constraints.

The level of experience of participants in technology due diligence projects varied: at the time of the testing, the senior consultants had previously participated in TDD projects, and the consultants had not. Three participants had done AI strategy creation projects. Three of the participants had technical backgrounds and expertise including software, coding, cybersecurity, and AI, while the others had a broader experience and skills by having participated in diverse project types in the company.

The sample was intentionally diverse to capture different approaches and perspectives on how they performed the explainability assessment. This diversity also allowed observation of how different participant profiles interacted with the framework's content and the scoring methodology, and how this might impact the TAM constructs observed during the sessions, as defined in Section 7.2.

7.2.3 Test execution

The testing sessions were conducted by the same facilitator in person at the company office during the month of October 2025 to ensure consistency of results. Each testing session followed the predetermined agenda described in 7.2.1 and all seven participants were able to complete successfully the task provided of performing the explainability assessment of the framework.

All participants provided consent to audio-record the session and capture their think-aloud thought process facilitating the results analysis after the sessions. The facilitator kept the silent observer role as established previously in the agenda.

The total testing time allocated for the whole session was 60 minutes, with 45 minutes just for the task execution. Based on recorded audio files, total test sessions ranged from 33 to 70 minutes, with median duration of 45 minutes. The longest test was due to the participant suggesting improvements in the material during the wrap-up part of the agenda, extending the allocated wrap-up time from 5 minutes to 20 minutes.

7.3 Test results

This section is organised as follows: first, remarks about test results collection; followed by the users methodology for completing the task; then, a collection of qualitative considerations derived from the audio recording transcriptions and notes taken during the session; the results of the TAM questionnaire; a list of the necessary improvements for the framework from the test; and finally, an analysis of all the collected information and analysis cross-referencing the qualitative and quantitative findings.

The results of the usability test were collected from two sources: (1) qualitative results from the notes and audio recordings of the testing sessions, and (2) quantitative results from the TAM questionnaire.

We analysed the sources by transcribing the audio files, identifying and collecting relevant findings and quotes, and cross-analysing them to support the next iteration on the framework based on the test results, feedback and suggestions. The transcript analysis was done shortly after each testing session, with data collected and cross-referenced it notes across session, to identify recurring and new findings. After the completion of the tests, the framework went through two further iterations of improvement and final refinements, resulting in the final framework, described in Section 7.5.

7.3.1 Users methodology

During testing sessions, a pattern was observed in how users approached the usability assessment task and how they solved it. Users either read first the case or started by going through the slides. Of the eight users, 3 read the case first and then the slides (Users 3, 5 and 8), in particular User 3 demonstrated a rushed approach when reading the slides and later had to revisit them when looking for information in the provided material. The remaining 5 users (Users 1, 2, 4, 6 and 7) read the slides first and then the case. User 2 was the only participant who started with the slides, identified which slides were applicable to the task, wrote then down, and then referenced the case information as needed, demonstrating a methodical approach that supported smooth task resolution.

General improvements on the slides were suggested, in particular user 5 provided sound feedback in how to improve the sequence and storyline of the slide deck that support the framework.

This indicated that the assessment needed to support different approaches when users interact with the material, considering its realistic use, consultants would either use the slides or the scoring framework spreadsheet first, both entry points need to provide equal information to allow consultants to reference and interact with it without constraints.

Concerning the materials, the scoring spreadsheet and the slide appendix contained the same information but with different level granularity, we purposely provided users with two different scoring formats to validated which was the optimal. The spreadsheet file contained maturity level scores with descriptions for level 1, 3, and 5 and the table in the slide appendix had descriptions for levels 1 to 5, offering more granularity. Taking this into account, it was found that most users expressed a preference for using the spreadsheet, even though some referenced both. Six of the eight users referenced both the spreadsheet and the slide appendix (Users 1, 2, 3, 4, 5 and 6) and two out of eight users referenced only the spreadsheet (Users 7 and 8). It was also found that having two parallel formats confused users in which should be the primary source of reference, the table in the slide appendix or the spreadsheet. Given that most users are more familiar with spreadsheets and engaged with it more, the spreadsheet format was identified as to the best format for the framework scoring.

Regarding the granularity of the descriptions, users with less technical knowledge referenced it when in need of extra information and in case of doubt about how to score, the granular level descriptions in the slide appendix for support during the task execution. In contrast, users with more technical knowledge preferred the less granular levels in the spreadsheet and were more autonomous in scoring the case company in the assessment dimensions. Sometimes their scoring was based on their own knowledge with less dependency on what exactly was described in the maturity levels. However, since users mainly used the spreadsheet as their preferred scoring table format, this suggests that improving the spreadsheet with additional information would be the optimal mix of both approaches, supporting both types of users needs.

Another observation concerned how the users interacted with the material. Users 2, 3, 4, 7 and 8 took an iterative approach, referencing the materials as they went through the execution of the task, while Users 1, 5 and 6 read the whole case, making notes on the case, before proceeding to fill in the score in the assessment framework. This difference in approach showed that users who made notes while reading the case demonstrated greater familiarity with the material, being able to retrieve information quicker than those who did not annotate and had to circle through the case to find relevant information to support the task completion and the assessment content.

In addition, Users 3 and 5 demonstrated enhanced interactivity with the maturity levels, by analysing what the level 5 scenario would look like and then comparing the case content to that to derive their score. A direct quote from the transcript illustrates this: "If this capability is a five, my first thought process would be this looks like a five, but is there any evidence that I would want to move this to a four or three?". This approach helped them visualize the highest maturity scenario, connect the framework with real-world cases and their own knowledge, and decide on the score based on their knowledge mixed with what the maturity levels descriptions.

Two users required more assistance from the facilitator, Users 4 and 7. User 4 relied on the facilitator explanations due to inexperience and non-familiarity with the company slides format. User 7 initially experienced confusion regarding the instructions of the task, which was clarified, and then proceeded to successfully complete the task, both cases required some familiarization with the material and extra support to be able to execute independently the scoring task. This showed that users are able to use the assessment framework with minimal support once the initial learning curve is overcome.

7.3.2 Qualitative considerations collected

The general consensus of the participants was positive regarding the acceptance of the framework and its future utility and the testing sessions were productive to identify issues that needed improvement. Some of the improvement aspects included the framework structure, slide storyline, lack of examples in the descriptions, technical terminology in the descriptions, repetition of content between dimensions and complex wording constructions.

The framework methodology was not immediately clear to some participants. Specifically Users 1, 5, and 7 required additional time to understand the structure and navigation of the framework before starting the task, demonstrating that the storyline of the slides that did not successfully explained the framework methodology. User 8 suggested making the storyline more explicit: "the storyline could even be a little bit more clearer, because the first slide has to explain everything about the framework, everything I'm gonna do..And also the endpoint. And then I have this in mind when I'm doing the assessment.".

This indicated that the framework and the storyline needed to be clear and self-explanatory. The material required improvement to its structure and methodology explanation, to allow any consultant to utilise it independently in their projects without requiring guidance from the facilitator or colleagues in the future.

Most users (Users 1-6) required examples to understand the maturity level descriptions and requested additional context to clarify the technical language used in the framework. User 3 elaborated on this need: "One thing that would make it clear especially for people that are planning to use this without a technical background is if you would put like an example for level one, three and five descriptions". This suggests that technical knowledge needs to be illustrated and explained in a user friendly way, correlating it to concepts more broadly known or with direct examples.

The sentences of the descriptions were difficult for the users comprehend at first and often required multiple reads. Users 5 and 7 reported difficulty due to the technical language, further indicating that they were not sure about what the level descriptions meant. Those issues led to uncertainty when executing the testing task, being highlighted by quote from User 7: "I guess in this case it would not be very helpful (the level descriptions) because I'm not sure what they mean".

However, regarding the descriptions, we had two viewpoints. User 3 had a different opinion from Users 5 and 7, and found the framework quite clear, at the end of the testing session: "I think it iss generally quite clear especially if you would take the time to read over the documentation that you have. Especially now I see here, it's actually quite clear, right? What we have here". Another point made by User 3: "It is all about the wording...about the person reading it, that they actually read it and don't just click through (the slides)". Relating to that, User 6 found the descriptions understandable for an investor who may not be GenAI-native. Considering the dual viewpoint on wording, even though some users did not have issues, it still indicated that the language and wording of the sentences throughout the framework needed revision and to adopt a more simple and user friendly language. This improvement was explicitly suggested by users during the sessions, in order to allow them to read naturally and understand the meaning of sentences without re-reading required.

Related to this, Users 3, 4, 6, and 7 demonstrated difficulty in understanding technical terminology, remaining stuck in certain terms such as XAI sources, desiderata and black box. User 4 required extra support and explanation throughout the task due to lack of knowledge regarding GenAI technology in general. Terminology issues regarding the term XAI (explainable AI) are illustrated by User 4's quote: "XAI sources. I never heard of it. So that I don't know what I should look for", and additionally by User 6 quote: "XAI sources. When I read this it took me a bit of time to understand it was only about the sources. But for non-technical people that is a stretch, how can you explain this to them?". In addition to improving the descriptions, it was found that technical terms should be easily explained in the descriptions to allow a full understanding of their meaning.

With regard to the actual assessment dimensions, three dimensions caused particular confusion during the testing sessions: Method-Specific Metrics, Security Concerns and Personalization. Regarding the first two dimensions, Users 1, 2, and 3 did not understand the relation between Security Concerns and explainability. User 6 remarked: "This does not belong here. This belongs in a security bucket, privacy or compliance.". Users 2 and 7 struggled to understand the Method-Specific Metrics description, with User 2 choosing not to answer it.

Concerning the Personalization dimension, User 2's personal understanding of this dimension that did not match the level descriptions, created conflict. User 3 found the Personalization description similar to the User Understanding dimension, not being able to differentiate two dimensions from each other. Additionally, User 6 interpreted the concept differently and User 8 did not see the personalization at all

TAM questionnaire results								
Question	P1	P2	P3	P4	P5	P6	P7	P8
Perceived Usefulness - To what extent would this generative AI assessment framework improve your ability to evaluate AI implementations effectively in consulting projects?	3	4	4	4	3	3	2	4
Perceived Usefulness - How much would using this assessment framework enhance your confidence when making AI investment recommendations to clients?	3	2	4	4	3	4	3	3
Perceived Ease of Use - How easy did you find it to navigate and apply the assessment framework during the evaluation process?	2	3	3	3	3	3	2	3
Output Quality - How would you rate the quality and reliability of the assessment insights generated using this framework?	3	2	4	3	3	3	3	3
Result Demonstrability - How clearly could you communicate and justify your assessment findings to stakeholders using this framework?	3	3	4	4	2	1	2	3
Behavioural Intention - How likely are you to use this generative AI assessment framework in your future consulting assignments?	4	2	3	4	3	3	3	3

Table 7.2: Qualitative TAM questionnaire test results

in the case, both resulting in low scores in the assessment. Further reiterating this point, User 6 explicitly suggested changes in the Personalization dimension: "You're tackling two things here. You cannot do two things in one dimension". These overlaps in the content demonstrated that the dimensions were not yet MECE compliant, an important requirement of the project. This indicated the need for revision to ensure that dimensions were clearly defined and mutually exclusive.

At the end of the usability test, most users found the framework useful for future use. This is supported by two feedback quotes from the users. A quote from User 3: "I think this is really useful if we would actually start doing this explicitly, especially when there is a big dependence on GenAI in a software company. I think it only makes sense that we have a structured framework to assess this". Another quote from User 8: "I think that's already a really good thing that it stands on its own...it could become part of the story (of a technology due diligence)". Even with the user difficulties and challenges with the wording, descriptions, and technical language of the framework, users still found the assessment important and considered that with improvements it would be a beneficial framework.

7.3.3 Quantitative TAM questionnaire results

The table 7.2 presents the responses to the TAM questionnaire given to the users at the end of each testing session. The purposes was to quantify the results of the test and collect users acceptance regarding the assessment framework artefact they had interacted with. The scoring of the TAM questionnaire ranges from 1 to 4.

The questionnaire results indicated that users thought that the framework would improve their ability in evaluating GenAI in projects, with the highest mean score of 3.38 in perceived usefulness; and that it would enhance their confidence in making AI investment recommendations to clients with the mean score of 3.25. However, participants did not find it easy to apply and navigate the framework during the testing task execution, with the lowest mean score of 2.75 in perceived ease of use. They further rated the output quality with a mean score of 3 and reliability of the insights generated with the framework,

although lacking result demonstrability in clearly being able to communicate and justify findings to stakeholders with a mean score of 2.88. Behavioural intention of users was the second highest mean score of 3.12, indicating that they were highly likely to use the assessment framework in future projects involving GenAI.

The overall mean score of the questionnaire for all six TAM criteria is 3.06 out of 4 indicating overall agreement with the framework, with improvements required in the ease of use and result demonstrability. For the technology acceptance scoring, even with all the qualitative comments throughout the testing sessions, users still demonstrated a high acceptance, with improvements to be made, but the logic and the methodology of the framework was sound and clear once the users passed through the learning curve at the beginning to understand the instructions.

7.3.4 Recollection of improvements for the framework

Following the completion of all eight usability tests, we present a recollection of all the necessary adjustments identified for the framework, to improve its usability and applicability for consultant's future use.

The necessary adjustments:

1. Adjust the scoring method to adopt only the 3-level scale present in the spreadsheet and remove the appendix in the slides;
2. Ensure MECE compliance throughout the entire framework content, particularly between the subdimensions content;
3. Ensure that all subdimensions weight the same and are presented in order of importance in the framework;
4. Move the Security Concerns subdimension to the Security dimension;
5. Improve the explanation of the Method-Specific Metrics, to counter confusion about its meaning;
6. Establish a clear distinction of the Personalization subdimension, removing overlap with other subdimensions.
7. Simplify wording by replacing technical and academic terms to make the content more user-friendly for non-technical users;
8. Standardize the maturity descriptions and clarify their wording;
9. Expand the descriptions of the maturity levels to increase the content provided;
10. Add concrete, short examples for each subdimension;
11. Add an executive summary to the two materials of the assessment, to support consultant's independent interaction with either material;

7.3.5 Analysis of test results and findings

The findings collected from the testing sessions and the TAM questionnaire demonstrated that the framework needed improvement and polishing, and informed the refinement of the framework described in Section 7.4. Despite these issues, users were able to see through the pain points and validate the essence of the artefact and the usefulness for consultants' future projects. This was supported by the TAM results, which showed an overall score of 3.06 out of 4.

Regarding accessibility of the assessment framework, testing illustrated how users with different technical levels and knowledge interacted with the material. Users with less technical knowledge relied more on the framework descriptions, which provided explanations of what was being assessed and the maturity levels, to complete their scoring task. This demonstrated that the framework scoring allowed participants to complete a technical assessment even without prior knowledge of the technical concepts required.

However, the framework also provided guidance to users with more technical knowledge and TDD experience. Even though these participants drew more on their technical background and judgment to justify their scoring decisions, they still used the material as north-star guidance, relying less on the exact descriptions but using it as a reference.

This dual scenario proved that the framework supports different levels of participants knowledge, with interaction being more structured ("black and white") or more interpretative ("gray") depending on each consultant's approach to using it and how much they relied on the framework information.

This is an important finding, as it reiterates the initial problem of this research project, which was to support consultants who lacked GenAI knowledge to assess GenAI technology. An important requirement of the assessment framework was to support consultants with less technical knowledge of Generative AI, providing a standardised and objective scoring methodology for all GenAI assessments in TDD projects.

7.4 Framework refinement

Based on testing results, we revised the framework requirements, Subsection 4.4.2, and the list of necessary changes to begin the refinement iterations on the artefact. Our strategy was to implement all the findings from the test of one dimension across the nine framework dimensions. Testing revealed three main issues in the framework: content repetition between dimensions, technically dense descriptions and complicated writing style, and lack of standardization.

To address this, in the first refinement iteration, we performed a MECE review across the entirety of the framework, incorporating the feedback received in the testing itself (e.g., security concerns from explainability to be grouped under security dimension). After this, we ensured that all subdimensions in each dimension had equal weight.

We then reviewed the structural organization, determining the optimal placement and grouping of the subdimensions, described in Subsection 7.4.1. We achieved this by merging subdimensions with similar content, moving those whose content fit best under another dimension and removing those that had less weight or were not MECE compliant. Those modifications were made to improve the flow of the assessment and to enable users to follow the order of the assessment, analyse one thing at a time, and complete it covering all topics without repetitions or confusion.

Complementing the structural changes, we improved all the framework descriptions, adopting a clearer writing style with simple non-technical terminology, and also following the company business writing guidelines manual.

After completing the first iteration, we performed a second iteration through two co-design sessions with the mentors to ensure alignment with business requirements of technology due diligences projects, to align the assessment storyline, and to polish the framework nomenclature, explained in Subsection 7.4.2. To ensure business alignment, we created investment questions to use as guidance to improving the storyline and the connection between the business "so what" and the technical aspect of the evaluation.

With these questions established, we defined the importance of each dimension and subdimension. From this perspective, we revised the framework structure by visualizing it as modular Lego-bricks, enabling us to reorder the dimensions within categories. With the final structure in place, we revised the nomenclature to ensure a simpler, more cohesive language. This marked the completion of the Generative AI assessment framework refinements, illustrated in Section 7.5.

7.4.1 First iteration

This subsection describes the improvements and changes made from the test version to the second test version. It is structured to provide accountability for the restructuring that was made on the framework, followed by other writing improvements made in this iteration.

The structural changes to the framework's subdimensions are illustrated in the Table 7.3. Through MECE analysis, we identified repetitive or similar content and reallocated items in the most appropriate place. The table shows the operations made in the framework consisting of subdimensions moved, removed, merged & renamed (i.e., within the same dimension), moved & renamed, moved, merged & renamed.

In total, 15 subdimensions were moved, 13 were merged and renamed, 11 were moved, merged and renamed, 1 was only renamed, and 7 were removed. These operations formed part of the MECE revision with the objective of reducing repetition and overlap between subdimensions content, to group related topics together and to improve the logical flow of the assessment. The resulting version 6 of the framework is illustrated in Table 7.4, that is a refinement from the last version 5 described in Chapter 6.

With this MECE-revised structure, the assessment flow became more concise, clear, and usable, ensuring that each subdimension was allocated alongside similar themes rather than unevenly distributed across the framework.

Dimension	Moved	Removed	Merged & Renamed	Moved & Renamed	Moved, Merged & Renamed
Model documentation	3	1	2	0	3
Data	1	0	2	0	1
Output metrics	0	1	3	0	0
Model trustworthiness	0	3	0	0	3
Model explainability	0	0	2	0	0
Security	4	0	2	0	3
Governance	4	0	2	1	0
Legal	2	0	0	0	1
Human–AI	1	2	0	0	0
Total	**15**	**7**	**13**	**1**	**11**

Table 7.3: Subdimension Operations by Dimension

Category	Dimension	Subdimension
Operational	Model documentation	Training data disclosure; Model cards; Intended use & limitations; Privacy & consent; Bias
Operational	Data	Data lineage; Model data; Data quality; Data volume
Operational	Architecture	Technical architecture and specifications; Solution ownership and replicability
Quality	Output metrics	Correcteness; Reliability; Speed; Scalability; Variety
Quality	Model trustworthiness	Hallucination; Sycophancy; Dispargement prevention; Stereotype avoidance; Robustness
Quality	Model explainability	Explanation coverage & adaptively; Explainable AI (XAI) sources; Human-centered explanation; Transparency, access & security; Method-specific
Compliance	Security	Model integrity; Input security; Output safety; Privacy protection
Compliance	Governance	Governance & accountability; Operational risk
Compliance	Legal	Control safety risks; Legal rights & intellectual property
Compliance	Human ai	Human control & agency; Human-AI collaboration; User experience & satisfaction; Personalization

Table 7.4: Prototype version 6: framework refinement, first iteration

Following the re-organization of the framework structure, all the subdimension descriptions were rewritten to integrate the feedback received from the testing sessions. These refinements also included improving all maturity level descriptions with additional content, as requested by users, while maintaining the technical content of the original framework. All the content was derived from the same source materials referenced in Appendix E.

Additionally, this rewrite activity also needed to comply with the requirements established by the company for business writing in TDD, which included guidelines for the grammar, voice, conciseness, clarity and MECE principles. Another requirement was to ensure that the framework's scoring model evaluated consistent items across maturity levels, with clear progression between levels 1, 3, and 5.

Three main directions guided this work:

- Description sentences should be simple and active, with the subdimensions described in one simple sentence, starting with "This subdimension refers to XYZ...", and elaborating only if needed;
- Maturity level sentences should start with: "Target has or does not have XYZ", be written in list style, making a clear split in what is important and what is additional information;
- Sentences that are long, technical and complex should be simplified.

We rewrote all the dimensions, subdimensions, and maturity level descriptions across the framework, which were subsequently reviewed by the mentors from EY-P. This marked the end of the first iteration, then we proceeded to the final refinements of the framework.

7.4.2 Second iteration

This subsection describes the improvements made that led to the final version of the framework and the co-design sessions performed to polish the framework and align business requirements with its technical nature. This second iteration built on the structural changes and writing improvements made in the first one, by focusing on the business alignment and storyline, as described in the beginning of this Section 7.4.

In a co-design session, we went through the entirety of the framework in an exercise of thinking as investors would, writing down the relevance of assessing each dimension. This practice of going through the framework content through the lens of investors helped establish how technical aspects impact the business rationale for performing this assessment and also the possible economic impact of it.

In typical due diligence projects, such questions are created to anticipate questions that investors may ask before a due diligence project and help explain the importance and impact of performing those assessments to non-technical people, and additionally to help refine the framework ensuring alignment with strategy concepts.

To comply with business requirements and polish the storyline, we aligned the framework with the improved investment questions and a clear storyline. It also needs to have a "so what" factor that allows consultants to explain to investors and non-technical clients why assessing Generative AI technology matters, its importance, and the financial or quality consequences of these assessments.

Lastly, we decided to reorganize the categories and dimensions, utilizing the same Lego-brick visualization approach as in the first iteration, treating the dimensions as modular components. With the objective of providing a more logical assessment flow to consultants and to ensure the assessment storyline could also be justified to investors and clients. We also incorporated the feedback received from the testing that the storyline needed to be clearer. The final prototype structure and materials are described in their entirety in Section 7.5.

This re-grouping was made in another co-design session, and changes made were to re-organize the assessment dimensions; for balance, the framework was reduced from 10 to 9 dimensions. To achieve this balance, we dissolved the model documentation subdimensions, this dimension was chosen since its content could easily fit within other dimensions, without creating conflict.

Model documentation subdimensions were redistributed: one to Data, one removed, one to Architecture, and two to Compliance. Additionally, from the AI Governance dimension one subdimension moved to AI security, from the Data dimension one subdimension moved to AI Governance, and finally from AI Security one subdimension moved to AI compliance.

In parallel, several categories and dimensions were renamed to improve readability and define a nomenclature that consultants could easily identify and use, regardless of technical background. This simpler approach was chosen to allow for clearer communication about the framework and to make the names very precise about what each part of the framework is about. Additionally, we refined

Category	Dimension	Subdimension
Technology	Data	Data quality; Data volume; Training data; Lineage
Operational	Architecture	Model card; Intended use & limitations; Infrastructure; Integration; Model moat
Operational	Performance	Accuracy; Reliability; Latency; Scalability; Diversity
Quality	Trustworthiness	Hallucination; Flattery; Discrimination; Stereotype; Fragility
Quality	Explainability	Reasoning; Sources; User understanding; Explanation validation; Model access
Quality	Human-AI	Human-AI collaboration; Human control & oversight; User satisfaction; Personalization
Oversight	AI Security	Model integrity; Input security; Output safety; Content safety risks
Oversight	AI Governance	Governance structure; Data management; Operational risk; AI organization maturity
Oversight	AI Compliance	Legal rights& intellectual property; Bias mitigation; Consent management; Privacy protection

Table 7.5: Prototype version 7: framework refinement, second iteration

the management questions to ensure alignment with new framework structure, and renamed them to "question repository".

Two out of three categories were renamed: Operational to Technology; Quality stayed Quality;and Compliance to Oversight. Alongside, six out of nine dimensions were renamed: Output metrics to Performance; Model Trustworthiness to Trustworthiness; Model Explainability to Explainability; Security to AI Security; Governance to AI Governance; and Legal to AI Compliance. The 7 version of the prototype reflects those changes, illustrated in Table 7.5

These adjustments, the investment questions, the structural changes and the simplified nomenclature ensured that the framework was MECE, descriptions were clear and user friendly, maturity descriptions were levelled ensuring consistent evaluation criteria throughout the framework scoring, resulting in a polished and clear assessment framework tool that complied with company requirements for frameworks and due diligence materials.

7.5 Final version of the Generative AI assessment framework

The final delivery of the framework included three materials: (1) Generative AI Assessment scoring framework, in spreadsheet format; (2) Generative AI Assessment, in slide deck format; and (3) Generative AI Assessment guide, also in slide deck format. Below are the descriptions and contents for each material.

First, the Generative AI Assessment – Scoring framework: containing the framework for assessing targets' Generative AI systems during TDD, illustrated in Table 7.6. The GenAI assessment framework includes three categories, nine assessment dimensions, and 39 subdimensions. The spreadsheet contained, not only the framework itself, which is the main asset of this research, but additional supporting sheets: Instructions, question repository (containing the management questions), scoring input (Table 7.6), scoring output, data request list and guidelines.

Second, the Generative AI Assessment consisted of a slide deck with instructions and template slides that could be used in project deliverables, supporting the framework. Its agenda content included instructions, detailed assessment dimensions, and an appendix.

Third, the Generative AI Informative Guide consisted of supplementary technical information about GenAI, for internal use and reference. Its contents included guidelines, a GenAI readiness and development guide, and technical information about the topic.

Generative AI assessment framework: created for assessing GenAI systems during technology due diligence engagements

Category	Dimension	Subdimension
1. Technology	Data	Data quality
		Data volume
		Training data
		Data lineage
	Architecture	Architectural setup
		Model end-to-end pipeline
		Model moat
		Intended use & limitations
	Performance	Accuracy
		Reliability
		Latency
		Scalability
		Diversity
2. Quality	Trustworthiness	Hallucination
		Flattery
		Discrimination
		Stereotype
		Fragility
	Explainability	Reasoning
		Sources
		User understanding
		Explanation validation
		Model access
	Human-AI	Human control
		Human-AI augmentation
		Personalization
		User experience
3. Oversight	AI Security	Model integrity
		Input security
		Output safety
		Content safety risks
	AI Governance	Governance structure
		AI organization maturity
		Data management
		Operational risk
	AI Compliance	Legal rights & intellectual property
		Consent management
		Privacy protection
		Bias mitigation

Table 7.6: Generative AI assessment framework

Chapter 8

Discussion

8.1 Research questions

This section provides answers to the research question and to the five sub-questions that were complementary to the main research question. We were able to answer the main research question, and sub-questions 1-3, through the literature review, immersion on the problem environment and iterative prototyping of the GenAI assessment framework. While, sub-questions 4-5 were partially answered, they require real TDD application to provide a full answer and complete validation.

RQ

This research aimed to answer: how can the application of GenAI usage in organizations be assessed and evaluated in its development and deployment within due diligence projects? The answer is through an assessment framework created and designed specifically for GenAI, which encompasses the main dimensions identified as essential for evaluating its characteristics: data, architecture, performance, trustworthiness, explainability, human-AI, security, governance and compliance. The resulting framework directly answers the main research question.

SQ.1

Regarding how can GenAI technology be defined and categorised to support practical understanding and assessment of its technological capabilities? This was answered through the selected related work that provided definitions [Feuerriegel et al., 2024, Ronge et al., 2025] and categorizations of Generative AI [Bandi et al., 2023], which formed our theoretical knowledge to understand GenAI and its capabilities, Chapter 4 describes all the literature that helped answer this question.

SQ.2

To address how can we design a comprehensive framework to categorise and define the assessment parameters for GenAI technologies based on technical aspects? We achieved this through the use of design thinking methodology applied specifically to the research problem, progressing through the immersion (Chapter 4), ideation and prototyping (Chapter 5), and testing of the framework (Chapter 7) phases, utilizing tools to support this process. Applying this methodology allowed us to immerse ourselves in the problem context and work iteratively with the users, who contributed during the immersion and testing phase, and the mentors who provided essential feedback to improve the prototype in its iterations.

SQ.3

On the question of what the key parameters are and criteria that determine GenAI models quality? In our framework we chose three dimensions as the key quality indicators of GenAI: trustworthiness [Huang et al., 2025], explainability [Schneider, 2024, Kalota, 2024] and Human-AI [Weisz et al., 2024, Cheng and Liu, 2024]. These dimensions are part of the Quality category of our framework, and are distinct from the Technical and Oversight categories, as they address GenAI's particular capabilities, specifically the trustworthiness and explainability of the generated content, and the human-AI enhancement and

collaboration aspects. In contrast, the Technology and Oversight dimensions could also be applied to other AI implementations even though they contain specific GenAI content in their assessments.

SQ.4

Examining how can we quantitatively assess valid unique value proposition characteristics of GenAI models while understanding their technical moats and replication risks? This question was partially answered through the framework's subdimension content and scoring maturity levels, especially in the Data and Architecture subdimensions. However, to fully answer this sub-question we would need practical implementation of the framework in real TDD projects to observe unique value propositions, technical moats and replication risks across different GenAI systems use cases. Throughout the prototyping and testing phase, we hypothesised that proprietary data with real replication costs were one of the main moats of current GenAI systems, but this notion requires validation. One way to answer this question would be through practical observation, and based on the 39 subdimensions of the framework, create a matrix showing which subdimensions indicate competitive advantage.

SQ.5

Concerning how can organizations strategically implement GenAI solutions across different use cases to maximize business alignment and ensure measurable ROI? Like the previous question, this was partially answered. As we were not able to test this framework in a TDD project, we could not observe how and which subdimensions could enable better GenAI solutions implementations, across diverse use cases to provide better business alignment and ROI impact. Yet, through collaboration and close contact with consultants who were working in value creation engagements during the internship time, we observed how a well defined business problem with alignment of why and how the technology is being used could have major impact in business strategic use of GenAI. Specifically for GenAI, researching use cases and understanding how GenAI could be applied is how companies can achieve competitiveness and differentiation, that ultimately reflect on their ROI.

The deliverables from this research to the company provided some tools to the consultants that can help align technical capabilities to business alignment, such as the "Question repository"and the "GenAI readiness guide" sections. With the rise of foundational models and Agentic AI solutions, we observed that having highly specialised systems, such as foundational models fine-tuned on proprietary data, could provide enhanced business alignment while leveraging the technology in a strategically relevant way, however this notion also requires further research to be validated.

8.2 Main findings

The main findings of this research were that we were able to create an assessment framework specially for GenAI due diligence projects, featuring a comprehensive dimensional structure with three categories, nine assessment dimensions and 39 subdimensions.

This research was originally motivated by consultants question during technology due diligence assessments: "Is the company really leveraging Generative AI or is it engaging in Generative AI washing?". While the research was not able to test the framework in real transactions, its content, structure and dimensions were designed specifically to support consultants in understanding what lies underneath GenAI systems. The scoring model system 1-3-5 defined the maturity levels as non existent, intermediate and expert, provides a structured methodology that in principle consultants can apply to distinguish between GenAI implementations and its underlying real capabilities.

The application of design thinking methodology was effective and well-suited to our research. This research involved a direct collaboration with EY-P and its objective involved creating a solution to a specific group of consultants in the particular context of a technology due diligence assessment. With the application of design thinking, we were able to empathise with the user from the beginning, define the problem, and immerse ourselves in the context through literature and internal data collection. Then, ideate and prototype the solution in co-design sessions and refine the five versions of prototype with direct and constant feedback from the mentors. Subsequently, test the prototype with consultants with diverse profiles and finally finalize the deliverable of the research.

This approach worked, especially because of the challenge of translating technical concepts from different sources, integrating them into a singular framework that needed to have a standard scoring methodology, throughout all of its dimensions. Using design thinking methodology with constant mentor

feedback, co-design sessions and iterative refinement on the feedback, allowed us to change and evolve the framework, with the objective of balancing the specific technical content depth required with a format that was familiar and usable to the end user, all aligned with the due diligence project process and its specific tooling.

During the immersion phase, we gathered requirements for the framework that were collected through the GenAI questionnaire, Section 4.3. Eleven requirements were identified, described in Subsection 4.4.2, that served as our reference throughout the ideation and prototyping phase, and we consider that all of those eleven requirements have been met in the final prototype. Specifically, requirement R10 "Framework should support diverse experience and knowledge levels", demanded that the language used in the framework to be user-friendly to all levels of GenAI knowledge. In the questionnaire results, we observed that the average technical knowledge score of consultants was 2.52, this implied that to meet R10, we we not only needed to create a framework that bridged technical concepts with business implications, but also one that could express said concepts in a language that consultants who were using the tool could understand and additionally use to explain said concepts to clients in projects. This requirement directly aligned with one of the research objectives, "Allow better evaluation and understanding of GenAI technology for consultants".

During testing, we found that the impact of the assessment framework in the specific context in which it was created was satisfactory to the users and helped them complete a more confident and standard evaluation of GenAI in due diligence case. In the test results, through the TAM questionnaire (Subsection 7.3.3), we discovered that its format and content were adequate, and users found that the framework could increase their ability to evaluate GenAI implementations. We also found that participants showed a high score in their intention to use the assessment framework in their future work related to GenAI due diligence projects. Another important finding was that even consultants with little knowledge of GenAI were able to complete the usability test tasks with the help of the framework. This finding will require further testing and real TDD applicability to be further verified, but showed that its format was adequate, and met the research objectives and framework requirements.

Additionally, we found that the materials to support the users throughout the whole process of completing the due diligence evaluation were essential. These materials included the management questions, data request list, scoring calculation formulas, and slides templates ready to be used in project reports. Ultimately, not only the assessment framework with its scoring was important, but also all the materials that complemented it, together enhancing the tools usability and integration with the due diligence process.

To further support the translation of technical concepts to business into "so what" implications, investment questions were defined to illustrate why technical evaluation concepts (e.g., performance, architecture setup, trustworthiness and explainability) are important and effectively determine the quality of a Generative AI system. However, not necessarily all dimensions have the same weight for all GenAI applications. Based on the premise that every GenAI application has a different use case and risk factor, some dimensions of the framework would be more relevant than others, and this categorization could not be achieved, as it would create too much granularity on the framework. Defining the most relevant dimensions for each use case also depends on where the company will be applying the system, for example, if it is a high-level risk sector such as medical or finance, as defined in the EU AI Act, then the application must comply with certain standards. For this reason, we made a design choice of not providing weights to the assessment dimensions and let this flexible for consultants when applying the framework to twist it depending on their project, as it usually already happens in due diligence engagements.

Moreover, the "so what" implications also impacted how the strategy could align with a specific investment thesis and ultimately help clients define this along their growth objectives. For instance, Does the company want to be a disruptor, an optimizer, or an enhancer with their GenAI usage? Does the company want to enhance its traditional business model with GenAI or be an AI-driven business? In order to use the framework in a use-case way in TDD, it needs real-world application throughout diverse engagements, so that, with time, the consulting team could recognize patterns, know what dimensions are best applied to each use case, and be able to apply the framework more effectively for each specific use case. That capability can only be developed with time, observation, and applied expertise using the GenAI assessment framework.

With applicability restrictions in mind, our ultimate intention was to cover as much as possible of technical concepts found in literature in the framework content. This approach allowed us to structure the framework in way that consultants could assess GenAI solutions with a long-term strategy lens, since it covers the three categories content (Technology, Quality, and Oversight), looking at fundamental aspects that can define a technology that is robust, enveloped in best practices possible, and that will

avoid structural flaws (e.g., re-word and re-coding in the future), and also mitigate risks that, with regulations, come at a price.

Following the usability testing, we performed the last two iterations on the framework continuing to evaluate the framework using the “why would an investor care about this assessment?” and "so what?" perspectives. This refinement also used co-design sessions to evaluate the framework from a top-down perspective. We dissected each individual dimension of the framework, asking repeatedly what was the "so what?" of each subdimension, and "why an investor should care about this?" until we were able to identify and define the most important aspects and only keep in the framework content that had a direct business-technical connection and could also be translated into financial or regulatory implication. This exercise helped us not only revise the framework content, but also ensure alignment with strategic requirements, ensuring that every subdimension had a strong investment question linked to it that could help support its importance. During the second refinement iteration, subsection 7.4.2 we were able to further visualize the content of the framework as lego-bricks and decide the final organization of its categories, dimensions, and subdimensions content. This was an extensive work that consisted of finding the balance between what was the optimal way to organize the dimensions, while also keeping an understandable and user-friendly storyline in which each topic built on each other. In this last refinement, MECE compliance was extremely important. Since during testing there was some overlap between subdimensions content that led users to confusion, we also looked at every word in entire framework and scoring maturity levels descriptions and ensured that there were no repetitions and that everything fit into the assessment storyline seamlessly.

8.3 Framework methodology approach

From the beginning of the research, we looked at the framework as a Lego-brick tool, in which ideally, one consultant could choose which dimensions applied to its specific due diligence project and fit each target's application. However, we encountered challenges in designing the framework in a way that supported this approach while also complying with all the requirements, and also had an easy storyline and could be used by consultants with varying level of GenAI knowledge and expertise.

The design approach taken was to design the framework, with the three-tier structure (i.e., category/dimension/subdimension) and a standard scoring method (i.e., 1-3-5 maturity levels) that standardised the framework and could provide a TDD assessment for the GenAI system. However, how to interconnect and how do different assessment dimensions affect each other? How do we explain what are the implications of dimensions in the GenAI quality? How to support consultants in understanding those things directly embed them into the design of the framework?

In the framework, conceptually, dimensions impact each other, if a GenAI model has bad data then probably it will have also bad performance metrics, bad explainability and trustworthiness, all of which bring regulatory consequences. In order to translate this into the scoring method, it would bring too many use-cases-score-possibilities, so we decided to leave it to consultants' interpretation on how to understand the results and its interdependencies. The solution was to explain the impact and the interconnection through supporting material, in the management questions we are able to show the implication of why a topic is important but it needs work from the consultant side to learn how to link this, that its usually the case in normal due diligences project (e.g., such as how cybersecurity affects risks). Since the framework needs to work for multiple use cases, we decided not to give weights to its scoring but rather translate these implications through the investment and management questions, and engage consultants in analysing and interpreting the scoring results.

During the refinement process of the framework, Section 7.4, we had to make final design choices regarding the technical depth of the framework content. While performing the last two refinement iterations, we revised the structure of the framework (category, dimension and subdimentions) and its content, including associated maturity levels and descriptions. In those activities, we had to decide how much technical depth we could keep in the content while maintaining the user-friendly language necessary for the end-users.

The challenge was aggravated because we did not have a TDD project with GenAI during the research period, and were unable to test the framework in a real due diligence. Consequently, we could only hypothesize how the framework would be used, using the normal TDD process map as reference, but not predict exactly what depth of information target companies would provide. According to feedback from consultants, the scope of the framework exceed the normal technical depth of typical TDDs, particularly in the Performance, AI Security subdimensions, and in the Quality dimension.

To account for the different use cases possibilities, while translating the technical importance, was to keep the framework with the established technical depth and the 39 subdimensions. However, we reinforced the Lego-brick modular approach, encouraging consultants to apply the framework through use-case lens, even though we could not design its structure explicitly for this flexibility. This design trade-off allowed us to preserve the technical depth that Generative AI assessment requires while aligning with TDD applicability and usability in future projects by motivating consultants to apply each subdimension as needed, depending on project scope and GenAI characteristics.

8.4 Comparison with previous research

Compared to previous research, our goal was to create a complete assessment that covered multiple topics that we found most relevant in the context of a technology due diligence of GenAI, with a level of depth that was adequate for the due diligence work, and in line with the users usual assessment strategy and process. The literature review revealed that existing works propose methodologies for evaluating Generative AI based on different aspects including: model architecture classification and evaluation [Bandi et al., 2023], input-output modalities [Kim et al., 2024, Gozalo-Brizuela and Garrido-Merchan, 2023, Feuerriegel et al., 2024], precision-based metrics [Singh et al., 2025], trustworthiness concepts [Huang et al., 2025], risk-based evaluation [Raut and Singh, 2024], and user experience design [Kim et al., 2024]. These related works propose assessment tools, or similar scoring methodologies, but they lack the holistic view and a standardised scoring method that allows for the same scoring system across all assessment dimensions and one approach to do it, resulting in a systematic assessment of Generative AI systems.

8.5 Contributions

Our research addresses the identified research gap with an assessment framework that provides a standardised and objective methodology to answer clients questions in due diligence projects, and defined criteria for evaluation through its dimensions and subdimensions. Specifically, it provides consistency and quality for the output of due diligence assessments even with varied project scopes, investment theses and strategic considerations.

The assessment framework can establish the foundation for GenAI assessment that can evolve with the technology, its structure enables future research additions, including Agentic AI dimensions or specific industry concepts.

This research provides seven key contributions: comprehensive assessment framework, standard evaluation methodology, holistic integration, GenAI parameters, knowledge accessibility, implementation guidelines, and practical due diligence toolkit. Described in detail below:

- **Comprehensive assessment framework**: A GenAI-specific assessment framework, containing 9 dimensions and 39 subdimensions across three categories (Technology, Quality, Oversight), with standardised 1-3-5 maturity level scoring model providing maturity definitions of non-existent, intermediate and expert;
- **Standardised evaluation methodology**: A scoring approach with MECE-compliant dimensions that enables objective evaluation of target GenAI systems across diverse implementations and use cases;
- **GenAI parameters**: Definition of key parameters and criteria that determine GenAI models quality based on data, architecture, performance, trustworthiness, explainability, human-AI, security, governance and compliance dimensions, synthesised from of 35 reviewed academic sources;
- **Holistic integration**: A comprehensive and unified assessment approach that integrates different works from existing literature and frameworks that evaluated generative AI, creating a holistic assessment that considers breadth and depth of evaluation scope;
- **Knowledge accessibility**: Framework design that bridges technical concepts with business implications, that supports practical understanding and assessment of GenAI technological capabilities accessible to consultants with varying levels of GenAI expertise;

- **GenAI implementation guidelines**: Best practices and guidelines regarding GenAI development and usability that serve as reference material to support consultants in their evaluations and technology implementation across different use cases;
- **Practical due diligence toolkit**: Supporting materials specifically designed for due diligence context including management questions, data request list, investment questions, scoring calculation, and presentation templates.

In addition to the academic contribution, the GenAI assessment framework constitutes a valuable consulting asset for EY-Parthenon in its assessments of software companies that claim to be advanced in AI technologies. EY-Parthenon has indicated that the framework is ready to be proposed to clients and used in future technology due diligence. The framework enables consultants to discern whether these companies genuinely utilise AI and adhere to best practices, thereby enhancing the quality of their evaluations. This internal acceptance and explicit intention to use the tool demonstrates its practical relevance for future GenAI engagements.

8.6 Unexpected results

During the prototyping process, we expected to be able to concentrate all of the framework material in one document, directly use the original form of from the reference work, selected in Subsection 5.2, and have a different structural approach to the scoring methodology. However, over the course of the project, in collaboration with the mentors, we realised the initial strategy was neither possible nor would it provide the best usability for the final users.

This resulted in unexpected structural changes throughout the different versions of the framework prototype, which in its last version resulted in a satisfactory assessment framework that complied with the requirements, was aligned to the company work methodology, had the standard layout for its materials, and provided value to its users.

8.7 Limitations

The main limitation of this research was the impossibility of validating the assessment framework in a real project. During the course of the nine-month thesis internship at EY-P, there were no technology due diligence projects with generative AI. To inform our approach, we received input from senior consultants and a senior manager, who explained what is typically assessable and what information targets usually provide. They indicated that in most due diligence projects, the scope of the TDD can be much narrower compared to our assessment coverage, and that targets sometimes do not provide sufficient documentation as would be necessary to complete the assessment. Considering these constraints, rather than modify our framework based on what previous experience suggested could be assessed in a TDD, we preferred to deliver a comprehensive assessment covering 39 subdimensions designed specifically for GenAI, aligning with our research objectives.

The second limitation of this research is the scope of application of the assessment framework, that is shaped directly by the type of work the partner company does, its clients and work methodology. So, we were unable to test and validate how the framework could be adapted to different scenarios and domains of the economy, that could've provided different angles to the framework content. Other limitations include the testing time that was limited to one month, and the number of users (8) who participated in those sessions, and the refinement iterations made on the prototype were limited to the time available to conclude the project.

Furthermore, since the field of Generative AI and its related literature is growing rapidly, we cite our limitation in covering all existing literature and focusing on the most closely related literature chosen through our research methodology, explained in Chapter 4, with a highlight for Agentic AI considerations, as, at the beginning of this project, this technology was not widespread. Throughout the research we observed that with the rise of Agents, new concerns, especially regarding security and access emerged and would require further research.

Although these limitations constrained the results of the current research, they also revealed open issues from not being able to test the framework in real due diligence projects.

8.8 Open issues

Three open issues were identified regarding the usage and usefulness of the assessment framework in real due diligence projects.

- **Open issue 1:** Will consultants with low level of GenAI knowledge be able to carry out this assessment alone relying only on the framework scoring, the maturity level descriptions, or will they need additional knowledge and support?;
- **Open issue 2:** How feasible this assessment is in a real life scenario, and throughout different kinds of generative AI assets?;
- **Open issue 3:** What would be the level of documentation access the company would get from clients to be able to carry the assessment, and working with limited documentation and relying on the management questions, would it be enough to complete it?.

These open issues are dependent on testing the assessment framework in real due diligence projects and observe and quantify the impact of consultants using the tool through projects as well as the efficacy in evaluating different kinds of GenAI use cases, and the interaction with the client. All of these issues could be answered by conducting tests with GenAI due diligence projects and observing the interaction of users, consultants, in using the material throughout the technology diligence process.

Beyond these implementation issues, the framework's usability underwent a separate review from a strategic and business perspective, that revealed gaps for future work.

8.9 Future work

Following project completion at EY-P, there was a final review with a business, strategy, and investor lens. This revision was done by the EY-P mentors to identify possible gaps in the framework that were less technical and more connected to possible investor theses and business questions. Afterwards, senior consultants and management also reviewed the framework, posing questions about the translation of the technical content into TDD practicalities.

The revision of this research identified several opportunities for extending the framework, including content gaps and research questions about practical application.

First, we identified the following content gaps:

- Guardrails;
- Financial cost including people FTE (full-time equivalent);
- Onboarding AI products;
- Product assessment metrics;
- Model theft and user access;
- ROI assessment.

Second, questions emerged about how to translate the technical aspects of the assessment into the TDD decision-making process:

- **Q1**: How should subdimensions be prioritized for different application types?
- **Q2**: Which framework subdimensions are most important for which use case?
- **Q3**: How can importance be explained to investors, either by revenue impact or compliance necessity?*
- **Q4**: How can the cost to fix issues found during TDD be calculated?
- **Q5**: How can we assess GenAI models' unique value proposition characteristics while understanding their technical moats?

*Aside from the defined investment questions perspective.

The identified gaps represent an opportunity to extend the framework's applicability and interpretation by connecting scoring results to business impact. First, the content gaps identified are subject to the TDD and strategy context, and not all items necessarily would fit the assessment framework *per se*, but they could be added to the supporting project materials.

Addressing Q1 and Q2 requires work to analyse and classify the framework content and create a methodological guide based on use cases and GenAI applications types, and they should be a medium-term activity. These questions should take into account TDD time constraints (projects may be as short as one week or less) and the length of assessment, in order to define a prioritization that considers what is most important for each type of GenAI assessment.

Questions Q3, Q4, and Q5 represent long-term activities and would require deep research. The question Q3 requires analysis to define and quantify, for each subdimension of the framework, its impact and importance, which could then be translated into TDD materials. Q4 would require a benchmark of issue costs, organized by use case, application, size of the application, and other factors. Finally, Q5, which is related to the Sub-research questions 4 and 5, needs application of the assessment framework into real TDD scenarios, and observation of results.

The future work, areas range from short-term additions to long-term research questions, their work would further enhance the applicability and tailor the framework to specific concerns of technology due diligence projects.

Chapter 9

Conclusion

This thesis aimed to develop a comprehensive, standardized assessment framework for evaluating Generative AI systems in technology due diligence projects. It addresses a literature gap about the lack of an unified framework that integrates multiple concepts and assessment perspectives that are applicable to diverse Generative AI implementations and use cases, particularly for the due diligence transactions evaluations.

We used the *research through design* method, as the artefact creation was the research, and its outcome was the GenAI assessment framework, combined with design thinking project phases of immersion, ideation, prototyping, and testing. This approach provides an effective methodology to our research problem, particularly for translating technical literature knowledge into an assessment tool that bridges theoretical concepts with practical application.

The main result of the research is the GenAI-specific assessment framework, containing three categories, nine assessment dimensions, and 39 subdimensions, featuring a standardised 1-3-5 maturity level scoring model, not currently found in existing literature or publicly available market solutions to the author's knowledge. The framework was specifically created for assessing GenAI systems during technology due diligence engagements. The testing of the framework revealed that even consultants with little knowledge of GenAI were able to assess a GenAI case by using the framework, and we were able to comply with all the defined requirements. The TAM questionnaire results supported the framework acceptance, with a mean score of 3.06 out of 4 indicating overall agreement of the users with the framework.

The created GenAI assessment framework establish a foundation for GenAI assessment that can evolve with the technology, contributing to a standardised evaluation methodology, defining key parameters and criteria that determine GenAI models quality. The framework features a design that bridges technical elements with business implications, enabling future integration of emerging concepts while maintaining its core structure. The research produced materials specifically created to support technology due diligence project process. Its main limitation was the lack of validation of the assessment framework in a due diligence project, leading to open issues about feasibility, documentation access, and usability of the framework in real contexts.

Future work includes short-term content additions and long-term research questions (e.g., prioritisation of subdimensions and translation of technical aspects into specific TDD considerations), which would further enhance the applicability and tailor the framework to specific concerns of technology due diligence projects.

Bibliography

Matthew O. Ayemowa, Roliana Ibrahim, and Muhammad Murad Khan. Analysis of recommender system using generative artificial intelligence: A systematic literature review. *IEEE Access*, 12:87742–87766, 2024. doi: 10.1109/ACCESS.2024.3416962.

Ajay Bandi, Pydi Venkata Satya Ramesh Adapa, and Yudu Eswar Vinay Pratap Kumar Kuchi. The power of generative AI: A review of requirements, models, input–output formats, evaluation metrics, and challenges. *Future Internet*, 15(8):260, 2023.

Jeffrey W Berkman. *Due Diligence and the Business Transaction: Getting a Deal Done*. Apress, 2013.

Rishi Bommasani, Kevin Klyman, Shayne Longpre, Betty Xiong, Sayash Kapoor, Nestor Maslej, Arvind Narayanan, and Percy Liang. Foundation model transparency reports. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):181–195, Oct. 2024. doi: 10.1609/aies.v7i1.31628. URL `https://ojs.aaai.org/index.php/AIES/article/view/31628`.

Le Cheng and Xiuli Liu. Unravelling power of the unseen: Towards an interdisciplinary synthesis of generative AI regulation. *International Journal of Digital Law and Governance*, 1(1):29–51, 2024. doi: doi:10.1515/ijdlg-2024-0008. URL `https://doi.org/10.1515/ijdlg-2024-0008`.

Fred D Davis et al. Technology acceptance model: TAM. *Al-Suqri, MN, Al-Aufi, AS: Information Seeking Behavior and Technology Adoption*, 205(219):5, 1989.

Deloitte. State of generative AI in the enterprise: Wave four survey results, January 2025. URL `https://www2.deloitte.com/content/dam/Deloitte/us/Documents/consulting/us-state-of-gen-ai-q4.pdf`.

EY-Parthenon. Five forces redefining the CTO mandate in software companies: Insights from 350 European software CTOs, 2025. URL `https://www.potloc.com/hubfs/EY-Parthenon%20-%20Five%20forces%20redefining%20the%20CTO%20mandate%20in%20software%20companies%20report.pdf`. Research conducted in partnership with Potloc.

F. Farhat, E. S. Silva, H. Hassani, D. Ø. Madsen, S. S. Sohail, Y. Himeur, M. A. Alam, and A. Zafar. The scholarly footprint of ChatGPT: A bibliometric analysis of the early outbreak phase. *Frontiers in Artificial Intelligence*, 6:1270749, 2024. doi: 10.3389/frai.2023.1270749.

Georgios Feretzakis, Konstantinos Papaspyridis, Aris Gkoulalas-Divanis, and Vassilios S. Verykios. Privacy-preserving techniques in generative AI and large language models: A narrative review. *Information*, 15(11), 2024. ISSN 2078-2489. doi: 10.3390/info15110697. URL `https://www.mdpi.com/2078-2489/15/11/697`.

Stefan Feuerriegel, Jochen Hartmann, Christian Janiesch, and Patrick Zschech. Generative AI. *Business & Information Systems Engineering*, 66(1):111–126, 2024.

Abenezer Golda, Kidus Mekonen, Amit Pandey, Anushka Singh, Vikas Hassija, Vinay Chamola, and Biplab Sikdar. Privacy and security concerns in generative AI: A comprehensive survey. *IEEE Access*, 12:48126–48144, 2024. doi: 10.1109/ACCESS.2024.3381611.

Mandeep Goyal and Qusay H. Mahmoud. A systematic review of synthetic data generation techniques using generative AI. *Electronics*, 13(17), 2024. ISSN 2079-9292. doi: 10.3390/electronics13173509. URL `https://www.mdpi.com/2079-9292/13/17/3509`.

Roberto Gozalo-Brizuela and Eduardo C. Garrido-Merchan. ChatGPT is not all you need. a state of the art review of large generative AI models, 2023. URL https://arxiv.org/abs/2301.04655.

Philipp Hacker, Andreas Engel, and Marco Mauer. Regulating ChatGPT and other large generative AI models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1112–1123, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594067. URL https://doi.org/10.1145/3593013.3594067.

Elise Han. What is design thinking?, 2021. URL https://online.hbs.edu/blog/post/what-is-design-thinking. Accessed on: March 4, 2025.

Yue Huang, Chujie Gao, Siyuan Wu, Haoran Wang, Xiangqi Wang, Yujun Zhou, Yanbo Wang, Jiayi Ye, Jiawen Shi, Qihui Zhang, Yuan Li, Han Bao, Zhaoyi Liu, Tianrui Guan, Dongping Chen, Ruoxi Chen, Kehan Guo, Andy Zou, Bryan Hooi Kuen-Yew, Caiming Xiong, Elias Stengel-Eskin, Hongyang Zhang, Hongzhi Yin, Huan Zhang, Huaxiu Yao, Jaehong Yoon, Jieyu Zhang, Kai Shu, Kaijie Zhu, Ranjay Krishna, Swabha Swayamdipta, Taiwei Shi, Weijia Shi, Xiang Li, Yiwei Li, Yuexing Hao, Zhihao Jia, Zhize Li, Xiuying Chen, Zhengzhong Tu, Xiyang Hu, Tianyi Zhou, Jieyu Zhao, Lichao Sun, Furong Huang, Or Cohen Sasson, Prasanna Sattigeri, Anka Reuel, Max Lamparth, Yue Zhao, Nouha Dziri, Yu Su, Huan Sun, Heng Ji, Chaowei Xiao, Mohit Bansal, Nitesh V. Chawla, Jian Pei, Jianfeng Gao, Michael Backes, Philip S. Yu, Neil Zhenqiang Gong, Pin-Yu Chen, Bo Li, and Xiangliang Zhang. On the trustworthiness of generative foundation models: Guideline, assessment, and perspective. *arXiv preprint arXiv:2502.14296*, 2025.

Faisal Kalota. A primer on generative artificial intelligence. *Education Sciences*, 14(2):172, 2024.

Dominik K. Kanbach, Lena Heiduk, Gregor Blueher, and Stefan Wilde. The GenAI is out of the bottle: generative artificial intelligence from a business model innovation perspective. *Review of Managerial Science*, 18(4):1189–1220, 2024. doi: 10.1007/s11846-023-00696-z. URL https://doi.org/10.1007/s11846-023-00696-z.

Tae-seok Kim, Marvin John Ignacio, Seounghee Yu, Hulin Jin, and Yong-guk Kim. UI/UX for generative AI: Taxonomy, trend, and challenge. *IEEE Access*, 2024.

Sascha Kraus, Dominik K. Kanbach, Patrycja M. Krysta, Maximilian M. Steinhoff, and Nicoletta Tomini. Facebook and the creation of the metaverse: radical business model innovation or incremental transformation? *International Journal of Entrepreneurial Behavior & Research*, 28(9):52–77, 2022. doi: 10.1108/IJEBR-12-2021-0984. URL https://doi.org/10.1108/IJEBR-12-2021-0984.

Tiffany H W Kung et al. The ChatGPT (generative artificial intelligence) revolution has made artificial intelligence approachable for medical professionals. *Journal of Medical Internet Research*, 25(1):e48392, 2023. doi: 10.2196/48392.

Clayton H. Lewis. Using the "thinking aloud" method in cognitive interface design. Technical Report RC-9265, IBM T.J. Watson Research Center, Yorktown Heights, NY, 1982.

Andreas Liesenfeld and Mark Dingemanse. Rethinking open source generative AI: open-washing and the EU AI act. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1774–1787, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3659005. URL https://doi.org/10.1145/3630106.3659005.

Rajkumar Modake and Dipali Patil. Evaluating generative AI applications, November 2024. Available at SSRN: https://ssrn.com/abstract=5032389 or http://dx.doi.org/10.2139/ssrn.5032389.

Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Ramona Comanescu, Canfer Akbulut, Tom Stepleton, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, William Isaac, and Laura Weidinger. Gaps in the safety evaluation of generative AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1): 1200–1217, Oct. 2024. doi: 10.1609/aies.v7i1.31717. URL https://ojs.aaai.org/index.php/AIES/article/view/31717.

Gaurav Raut and Apoorv Singh. Generative AI in vision: A survey on models, metrics and applications. *arXiv preprint arXiv:2402.16369*, 2024.

Raphael Ronge, Markus Maier, and Benjamin Rathgeber. Towards a definition of generative artificial intelligence. *Philosophy & Technology*, 38(1):1–25, 2025. doi: 10.1007/s13347-025-00863-y.

Johannes Schneider. Explainable generative AI (GenXAI): a survey, conceptualization, and research agenda. *Artificial Intelligence Review*, 57(11):289, September 2024. ISSN 1573-7462. doi: 10.1007/s10462-024-10916-x. URL https://doi.org/10.1007/s10462-024-10916-x.

Rajendra Singh, Ji Yeon Kim, Eric F Glassy, Rajesh C Dash, Victor Brodsky, Jansen Seheult, ME de Baca, Qiangqiang Gu, Shannon Hoekstra, and Bobbi S Pritt. Introduction to generative artificial intelligence: Contextualizing the future. *Archives of pathology & laboratory medicine*, 149(2): 112–122, 2025.

Alex Singla, Alexander Sukharevsky, Bryce Hall, Lareina Yee, Michael Chui, and Tara Balakrishnan. The state of AI in 2025: Agents, innovation, and transformation. Survey, McKinsey & Company, November 2025. URL https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai. Online survey conducted June 25–July 29, 2025, with 1,993 participants from 105 nations.

Pieter Jan Stappers and Elisa Giaccardi. Research through design. In Mads Soegaard and Rikke Friis-Dam, editors, *The Encyclopedia of Human-Computer Interaction*. Interaction Design Foundation, Aarhus, Denmark, 2nd edition, 2017. URL https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/research-through-design.

Sebastian Stober. Possibilities of source documentation and disclosure for generative AI systems. *Available at SSRN*, 2025.

C. T. Ungureanu and A. E. Amironesei. Legal issues concerning generative AI technologies. *Eastern Journal of European Studies*, 14(2):45–75, 2023. doi: 10.47743/ejes-2023-0203.

Justin D Weisz, Jessica He, Michael Muller, Gabriela Hoefer, Rachel Miles, and Werner Geyer. Design principles for generative AI applications. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–22, 2024.

Shangyu Wu, Ying Xiong, Yufei Cui, Haolun Wu, Can Chen, Ye Yuan, Lianming Huang, Xue Liu, Tei-Wei Kuo, Nan Guan, and Chun Jason Xue. Retrieval-augmented generation for natural language processing: A survey, 2025. URL https://arxiv.org/abs/2407.13193.

Alice Xiang. Fairness &; privacy in an age of generative AI. *Science and Technology Law Review*, 25(2), Jun. 2024. doi: 10.52214/stlr.v25i2.12765. URL https://journals.library.columbia.edu/index.php/stlr/article/view/12765.

Ning Yu, Vladislav Skripniuk, Dingfan Chen, Larry Davis, and Mario Fritz. Responsible disclosure of generative models using scalable fingerprinting. *arXiv preprint arXiv:2012.08726*, 2020.

Y. Zeng, K. Klyman, A. Zhou, Y. Yang, M. Pan, R. Jia, and B. Li. AI risk categorization decoded (AIR 2024): From government regulations to corporate policies. *not yet released*, 2024. URL https://arxiv.org/abs/2406.17864. arXiv preprint arXiv:2406.17864.

John Zimmerman, Jodi Forlizzi, and Shelley Evenson. Research through design as a method for interaction design research in hci. pages 493–502, 04 2007. doi: 10.1145/1240624.1240704.

John Zimmerman, Erik Stolterman, and Jodi Forlizzi. An analysis and critique of research through design: Towards a formalization of a research approach. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems*, DIS '10, pages 310–319, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0103-9. doi: 10.1145/1858171.1858228. URL https://dl.acm.org/doi/10.1145/1858171.1858228.

Appendix A

Project timeline

Phase	**Period**	**Key Activities**
Pre-Immersion	Feb-Mar 2025	Literature search, foundational reading
Immersion	April-May 2025	Internal research, shadowing TDD, questionnaire, conversations
Ideation	May-June 2025	Process mapping, material organization, co-design session
Prototyping	June-Sept 2025	Framework development (V1→V5), weekly co-design iterations
Testing	Sept-Oct 2025	Usability testing (8 participants), TAM questionnaire
Refinement	Oct-Dec 2025	Post-testing iterations, final framework, thesis writing

Table A.1: Research project timeline (2025)

Appendix B

Immersion material: GenAI knowledge and technology due diligence experience questionnaire

Demographics

1. What is your role?
    - Intern
    - Consultant
    - Sr. Consultant
    - Manager
    - Sr. Manager
    - Director
    - Partner
2. Age Group
    - 18 - 24
    - 25 - 34
    - 35 - 44
    - 45 +

Generative AI Knowledge
The next few questions are designed to quantify technical aspects regarding Generative AI technology.

3. For the following statements respond in terms of Strongly Disagree to Strongly agree:

    a) I can explain in simple terms what generative AI is and how it differs from other types of AI (e.g., discriminative models)

    b) I understand the basic concepts of large language models (LLMs) and how they function

    c) I understand the concepts of pre-training and fine-tuning in the generative AI training process

    d) I can identify the types of data sources required for train generative AI models

    e) I am familiar with different techniques used in training generative AI models, such as supervised, unsupervised and reinforcement learning

    f) I can distinguish between different generative AI architectures (e.g., transformer-based models, diffusion models, GANs, VAEs, etc.)

    g) I understand how to evaluate generative AI performance using appropriate metrics

    h) I understand the concept of hallucinations in generative AI and strategies to mitigate them

i) I can explain how Retrieval-Augmented Generation (RAG) works and its benefits over standalone generative AI

j) I understand how to implement guardrails and safety measures when deploying generative AI solutions. (e.g., content filtering and flag, human-in-the-loop)

k) I understand the computational and environmental impact associated with training and running generative AI models

l) I understand the data privacy and security implications of deploying generative AI in enterprise environments

m) I can identify and assess the data privacy and security risks when integrating third-party generative AI tools into business applications

n) I am aware of ethical considerations and potential biases in generative AI implementations

o) I am familiar with specific regulations governing AI/generative AI in Europe, (e.g., EU AI Act, GDPR)

p) I can identify business processes where generative AI would provide significant value versus where it would be ineffective

q) I understand how training data quality affects model performance

(Scale for all items: Strongly Disagree / Disagree / Neutral / Agree / Strongly Agree)

Technical Evaluation Process

4. How many Technical DD's have you contributed on?
   - 0
   - 1-5
   - 6-10
   - 11-15
   - 16-20
   - 20+
5. What aspects do you focus on a technical evaluation?
   - _____
6. Have you ever performed a technical (maturity) assessment of (Gen)AI-native companies?
   - Yes
   - No

   Note: here the questionnaire branches, if the answer is yes go to section "Technical Evaluation Process [Cont]", if the answer is no go to section "GenAI/AI Evaluation Process".

Technical Evaluation Process [Cont]

7. What were the tools used in the evaluation process? (Eg. framework/excel template, benchmarks, google search, GPTs, past projects, own experience)
   - _____
8. What are the most challenging steps and main pain points you encountered during the technology due diligence and evaluation process?
   - _____
9. What did you miss during the technical assessment, or what additional information or resources would have benefited your evaluation process?
   - _____

10. Would you feel confident conducting a GenAI maturity assessment of a target organization with your own experience and knowledge?
    - Yes
    - No
    - Maybe
11. Can you briefly outline your hypothetical approach to this assessment?
    - _____

GenAI/AI Evaluation Process

12. What model of GenAI or AI did you evaluate? e.g., Transformers, Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), etc.
    - _____
13. In the evaluation, what were the main priorities when assessing the GenAI/AI model?
    - _____
14. Were there any differences between a GenAI/AI technical DD and other technologies evaluations?
    - _____
15. What were the tools used in the evaluation process? (Eg. framework/excel template, benchmarks, google search, GPTs, past projects, own experience)
    - _____
16. What are the most challenging steps and main pain points you encountered during the GenAI/AI due diligence and evaluation process?
    - _____
17. What did you miss during the GenAI assessment, or what additional information or resources would have benefited your evaluation process?
    - _____

Appendix C

Testing material: mock case

Techflow AI Solutions B.V.
Mock investment due diligence case, created with Perplexity AI

COMPANY PROFILE
Founded: 2019 | **Location**: Amsterdam, Netherlands
Industry: Logistics Technology (B2B SaaS)
Target Market: Mid-sized European logistics companies (50-500 employees)

TechFlow AI Solutions provides AI-powered logistics optimization platform serving 180+ companies across Europe. The company transformed from traditional rule-based software to generative AI-driven solutions in 2024.

CORE AI IMPLEMENTATION
Primary Use Case: Intelligent Route Optimization
TechFlow's platform uses generative AI to create optimal delivery routes by processing real-time traffic data, weather conditions, vehicle capacity constraints, and customer preferences. The system generates natural language explanations for route decisions and provides predictive insights about potential delays.

AI Technology Stack:

- Core Model: GPT-4 Turbo fine-tuned on logistics data
- Implementation: Cloud-based API integration with existing customer systems
- Training Data: 2.5 million historical delivery records, weather patterns, traffic data
- Update Frequency: Real-time optimization with daily model retraining

System Characteristics:
The AI system provides detailed explanations for its routing decisions through:

- Natural Language Summaries: Plain text explanations like "Route modified due to construction on A4 highway, adding 12 minutes but avoiding 45-minute delay"
- Visual Decision Trees: Interactive maps showing why specific routes were chosen over alternatives
- Confidence Scoring: Each route recommendation includes confidence percentages (typically 85-95%)
- Factor Attribution: System identifies and ranks the top 3-5 factors influencing each decision (traffic: 40%, weather: 25%, fuel efficiency: 20%, etc.)
- Historical Comparison: Shows how current recommendations compare to similar past scenarios with outcome data
- API Response Time: Average system response time is 1.2 seconds for route calculations, with 99.5% uptime recorded over the past 12 months

Human Oversight: Logistics managers can override AI recommendations with one-click alternatives, and the system learns from these interventions to improve future suggestions.

KEY FINANCIAL METRICS

Revenue Performance:

- Annual Recurring Revenue (ARR): €8.2M (2024)
- Growth Rate: 145% year-over-year
- Client Count: 180 active customers
- Average Contract Value: €45,600 per year
- Customer Retention Rate: 94%

Investment Context:

- Funding Sought: €15M Series B
- Use of Funds: 60% product development, 25% market expansion, 15% team scaling
- Current Runway: 18 months

ADDITIONAL INFORMATION

- The system processes 50,000+ route optimizations daily
- Customer satisfaction scores average 4.2/5.0 specifically for AI-generated explanations
- 12% of AI recommendations are overridden by human operators
- Average time savings per route: 18 minutes
- Fuel cost reduction: 15% average across customer base

Disclaimer: This case is designed for assessment using your generative AI evaluation framework. Focus on applying your methodology to evaluate the technical and business aspects of TechFlow's AI implementation.

Appendix D

Testing material: TAM questionnaire

Question 1 - Perceived Usefulness
To what extent would this generative AI assessment framework improve your ability to evaluate AI implementations effectively in consulting projects?
Scale: 1 = Not at all | 2 = Slightly | 3 = Moderately| 4 = Significantly
Question 2 - Perceived Usefulness
How much would using this assessment framework enhance your confidence when making AI investment recommendations to clients?"*
Scale: 1 = Not at all | 2 = Slightly | 3 = Moderately| 4 = Significantly
Question 3 - Perceived Ease of Use
How easy did you find it to navigate and apply the assessment framework during the evaluation process?
Scale: 1 = Very Difficult | 2 = Difficult | 3 = Easy | 4 = Very Easy
Question 4 - Output Quality
How would you rate the quality and reliability of the assessment insights generated using this framework?
Scale: 1 = Poor Quality | 2 = Fair Quality | 3 = Good Quality | 4 = Excellent Quality
Question 5 - Result Demonstrability
How clearly could you communicate and justify your assessment findings to stakeholders using this framework?
Scale: 1 = Very Unclear | 2 = Somewhat Clear | 3 = Clear | 4 = Very Clear
Question 6 - Behavioural Intention
How likely are you to use this generative AI assessment framework in your future consulting assignments?
Scale: 1 = Very Unlikely | 2 = Unlikely | 3 = Likely | 4 = Very Likely

Appendix E

Framework related work components

This appendix documents the academic sources used to construct each assessment dimension in the final framework, described in the Table 7.6. For each dimension, we identify which papers contributed content (subdimensions, metrics, definitions, or guidelines) and briefly note what was adapted.

This appendix has the objective of providing clear accountability and explain what exact sources we used from reference material to develop our framework, giving due credit to the works.

The descriptions bellow reflect the final version of the assessment guide, not explaining the framework prototyping evolution process versus the use of the reference materials. The prototyping framework versioning is described in detail in Subsection 6.2 and the last two framework iterations are described in Subsection 7.4.

E.1 Technology

E.1.1 Data

- **Ungureanu and Amironesei [2023]:** Legal data type classification (open source, voluntary, protected, synthetic);
- **Goyal and Mahmoud [2024]:** Synthetic data generation techniques;
- **Xiang [2024]:** Data quality, curation process, bias considerations;
- **Stober [2025]:** Taxonomy of data sources.

E.1.2 Architecture

- **Kalota [2024]:** Transformer, GPT architectures;
- **Bandi et al. [2023]** Taxonomy of GenAI architectures;
- **Wu et al. [2025]:** RAG architecture explanation;
- **Ronge et al. [2025]** GenAI-AI specific caracteristics;
- **Feuerriegel et al. [2024]** Multi-level classification (technical-business application).

E.1.3 Performance

- **Modake and Patil [2024]:** Quantitative metrics (text/image-based, diversity, computational);
- **Ayemowa et al. [2024]:** GAN/VAE evaluation metrics (probabilistic, qualitative, ranking);
- **Singh et al. [2025]** Foundational model-evaluation approaches;
- **Bandi et al. [2023]** Task-specific metrics.

E.2 Quality

E.2.1 Trustworthiness

- **Huang et al. [2025]:** Trustworthiness subdimensions framework (hallucination, sycophancy, honesty, jailbreak, toxicity, bias, robustness, ethics, privacy).

E.2.2 Explainability

- **Schneider [2024]:** Explainable AI (XAI) desiderata, white box vs. black box, GenXAI techniques;
- **Kalota [2024]:** XAI definition and techniques.
- **Modake and Patil [2024]:** XAI metrics and explanation-quality evaluation;
- **Bommasani et al. [2024]:** Foundational model transparency.

E.2.3 Human-AI

- **Weisz et al. [2024]:** Six design principles for GenAI applications;
- **Cheng and Liu [2024]:** Five risk categories and best practices.
- **Raut and Singh [2024]:** Human autonomy and integrity harms taxonomy

E.3 Oversight

E.3.1 AI Security

- **Feretzakis et al. [2024]:** Privacy risks in LLMs, data privacy preserving techniques;
- **Golda et al. [2024]:** Privacy and security implications across 5 perspectives;
- **Xiang [2024]:** Bias and privacy of generated content;
- **Stober [2025]:** Data provenance and lawful data use;
- **Ayemowa et al. [2024]:** Prompt security and input-manipulation safeguards;
- **Ungureanu and Amironesei [2023]:** Documentation of sources and legal use of data.

E.3.2 AI Governance

- **Zeng et al. [2024]:** AIR Taxonomy (314 risks), EU AI Act mandatory requirements;
- **Rauh et al. [2024]:** Taxonomy of GenAI harms;
- **Stober [2025]:** Data sources and documentation requirements;
- **Bommasani et al. [2024]:** Transparency requirements across 6 government policies;
- **Golda et al. [2024]** Data handling procedures, governance structure, and risk mitigation and reversal;
- **Hacker et al. [2023]** Incidents and mitigation strategies.

E.3.3 AI Compliance

- **Zeng et al. [2024]:** Risk taxonomy;
- **Cheng and Liu [2024]:** Legal and ethical risk categories;
- **Hacker et al. [2023]:** Three policy layers for LGAIMs derived from the EU AI Act;
- **Golda et al. [2024]:** Overview of data protection laws.

E.4 Supporting Materials

E.4.1 Usability Assessment

- **Kim et al. [2024]:** UI/UX usability heuristics (10 heuristics, cognitive walk-through, evaluation criteria), used in the original form.

E.4.2 Development Guide

- **Bandi et al. [2023]:** Development guidelines;
- **Singh et al. [2025]:** Implementation best practices.

E.4.3 Reference Materials

- **Ronge et al. [2025]:** GenAI definition (four characteristics);
- **Hacker et al. [2023]:** Policy standards for LGAIML regulation;
- **Stober [2025]:** Data source recognition techniques;
- **Yu et al. [2020]:** Fingerprint mechanism for AI-generated images.

Appendix F

GenAI prompts repository

This appendix describes all the prompts used during the project, providing transparency, accountability around the prompt process, and enabling other researchers to replicate and critically access how Generative AI to support the development and testing of the framework.

F.1 Prompts used for the prototyping of the framework

As cited in Chapter 6, Subsection 6.3, Generative AI was used to support the creation of framework based on the content of the selected reference work. For applicability reasons, we provide only the prompts that were used during the prototyping of the version five and later in the refinement done in framework the after the testing.

Additionally, we provide the MECE analysis prompt that we used during the testing, Chapter 7, especially in the refinement of the prototype interactions Subsection 7.4.

Several prompts were tested in this phase, however the one below was the one that provided the best results for our framework.

F.1.1 Creation of the framework's maturity descriptions

- Tool: Perplexity AI
- Model: Claude Sonet 4.5
- Use: Those two prompt were used for all the subdimensions of the framework

Creating the maturity level ranging from 1 to 5

Prompt 1: "Based on these domains and subdomains with its own metrics create a table with maturity levels score description from 1 to 5 for each subdomain + metric what it means for a generative AI model to have a score of 1 vs 3 vs 5 in [insert: topic]. In the domains descriptions, include the subdomains topics.

Each maturity score description needs to be descriptive in a user friendly language that help consultants, that may or may have not a technical background, know what maturity level looks like when scoring and assessment the target company generative AI, consider maturity 5 like state of the art / top notch.

Use the academic paper(s) attached to create the metrics and cite all the quote that justifies the levels.

Expected output: row with domain/ subdomain / metric / maturity level 1 / maturity level 2 / maturity level 3 / maturity level 4 / maturity level 5

Include also 3 - 4 key indicators for each level"

Creation of the maturity level only for 1-3-5

Prompt 2: "I am creating a generative AI assessment framework, containing [insert: topic] as an assessment domain. I need to create maturity levels descriptions, based on the content of the academic papers provided [attachment: academic papers pdf files]. This assessment is to be used by consultants, with

technical and without technical knowledge, during technical due diligence assessments, so the language needs to be simple and user-friendly.

The maturity levels are to be maturity 1 - non existent / maturity 3 - intermediate / maturity 5 - expert

Create this descriptions providing detail and key indicators for each maturity level

Use the correct English grammar

Start every sentence of the maturity level descriptions with "- target has", or "-target has not" or similar wording

Expected output: table with rows organised by subdomain / maturity 1/ maturity 3/ maturity 5 "

F.1.2 MECE analysis

- Tool: Perplexity AI
- Model: Claude Sonet 4.5
- Use: Perform MECE analysis throughout the framework content

Prompt for MECE analysis: "Make a complete detailed MECE analysis of the [insert: category] category including [insert:subdomains names] subdomains content, considering also the maturity levels against all the other categories/domains/subdmomains of this framework. My goal is to identify the content from the [insert: category] category that is already elsewhere in the assessment framework for GenAI. Provide me the output with detailed quotes showing where the framework is not MECE compliant"

F.2 Prompts used in the testing phase

As mentioned in 7, perplexity was used to support the creation of some aspects of the testing material, including the mock case, the questionnaire and the testing agenda. Here are the prompts for the last two items.

F.2.1 Interactions to create the first long mock case

- Tool: Perplexity AI
- Model: Claude Sonet 4.5
- Mode: Deep research
- Use: Diverse interactions were needed to result in the final version of the long mock case

First interaction

Input: "You are a consultant writing a mock case to test out a assessment framework guide for generative AI in due diligence transaction projects. The assessment framework guide in is ppt format, the goal of the slide deck is: "This guide is used to evaluate Generative AI during a technology due diligence.

In a technical DD, when a company uses generative AI is necessary to understand what kind of model and use case it actually has. Then evaluate the model operation, quality and compliance." The assessment guide agenda is:

- Usage guide – how to apply this framework
- Detailed assessment dimensions
- Operational
- Quality
- Compliance
- Usability assessment
- Development guide

- Reference materials
- Appendix

The goal for the mock case is to be used as a reference case when testing out the guide material with a group of consultants (aprox. 6), the dynamic is to go through the assessment guide sections. The requirements for the case:

- Tech due diligence mock case scenario that is realistic
- Buy side tech due diligence
- It needs to have the technology of generative AI.
- Closer to the practice → normal company with technology → application layer
- Country of reference: Netherlands market (can be mixed with other European countries such as France, Italy, Belgium etc)

core thing → find out what their testing subject attitude is towards the instrument, and with the help of the mock case help them visualize and apply the assessment and guide the discussion for the subjects to be able to evaluate the assessment and provide feedback and improvements"

Second interaction

Input: "Modify this mock use case document [attachment: mock case 1 - v2 - deep research]. Delete the sections "Technical Challenges & Risks and Competitive Landscape

Modify: Make the beginning of the case more narrative in the executive summary and company provide sections. one page of actual context of the company before the next section

Keep" the Generative AI Technology Stack as it is"

Third interaction

Input: "Provide a second version [attachment: mock case 1 - v2 - deep research], where the AI platform is one of the offerings of the company and the main service, but the company needed to move all the data and business into this AI platform, "" migrating"" the original solution. Now they are finished moving the client database and are fine tuning the AI platform to support the original database, also considering the 2025 technology developments.

Modify and provide a new text not a pdf"

Based on the original mock case, that was more than 10 pages long, we created the smaller mock case version to be used in the usability tests.

F.2.2 Interactions to crate the second small mock case

- Tool: Perplexity AI
- Model: Claude Sonet 4.0
- Mode: Pro research
- Use: Diverse interactions were needed to result in the final version of the small mock case

First interaction

Input: "I need a new version of this mock case, create a new case with this new structure and providing textual descriptive information about the core AI implementation explainability characteristics [attachment: mock case 1 - v3 - deep research]"

Second interaction

Input: "Consider that these are the subdimensions for generative AI explanability evaluation [insert: subdimensions]. Does the case provide aligned but not direct information that allows the user to complete this assessment? The case does not need to provide all of these assessment dimensions themes, but a few because not having those things could also be a realistic possibility.

analyse the current match of information and what is currently in the mock case"

Third interaction **Input**: "Provide me quotes from the use case document that matches the subdimensions themes so I can see where the information could be found in the document"

Fourth interaction

Input: "Make modifications on the streamlined_mock_case_techflow.md -> take out explainability characteristics and put only system characteristics: with all the bullets but include another additional bullet with another information that is not explainability related. take out the assessment challenge part put the available information part as additional information take out decision point part"

The output of this last and final iteration is documented in AppendixC.

F.2.3 Iterations to create the TAM questionnaire

- Tool: Perplexity AI
- Model: Claude 4.0 Sonet
- Title: TAM questionnaire for test
- Reference paper: [Davis et al., 1989]

List of iteration messages to create the TAM questionnaire:

- **Message 1:** Based on Davis tam, and the case on the paper create a 5 question, tam model, to test a generative ai assessment guide in ppt format;
- **Message 2:** Refactor the 5 questions considering that this guide would be used during technology due diligence assessment;
- **Message 3:** Recap the tam questionnaire;
- **Message 4:** Provide a new version, addressing the concerns on question 2, that it should focus on purely on easy of use, and 5, that had mixed concepts. and using a 4 level scale instead of 7;
- **Message 5:** Recap the new tam questionaire version with all questions;
- **Message 6:** What are the tam3 version constructs? what would be a new questionnaire version that incorporates more constructs?;
- **Message 7:** For the mock use case and assessment subdimensions content, what are the most relevant constructs of TAM?;
- **Message 8:** Create a new TAM questionaire, with max 10 questions but ideally 5, comprehensive of those constrains of TAM for this specific use test. Perceived Usefulness; Perceived Ease of Use; Behavioural Intention; Output quality and Result demonstrability;
- **Message 9:** Create a section before the questions with the academic definition of each construct, base it on the original academic papers of tam, tam2 and tam3.

The output of this last and final iteration is documented in AppendixD.

F.2.4 Iterations to create the agenda for testing session

- Tool: Perplexity AI
- Model: Claude 4.0 Sonet
- Title: Agenda for test
- **Message 1:** Provide me a feedback on the agenda for the user test, designed for 1 hour or 1.5 hour testing agenda: explain the material -> slide deck + spreadsheet go through the usage guide -> present on the slide deck provide investment questions (2) -> to guide the test provide a mock case scenario with factual information -> to test the usability of the slide deck material with a actual case mimicking a real life scenario ask to perform one dimensions assessment and calculate the score -> using the spreadsheet ask the TAM questions -> will provide them later
- **Message 2:** So the flow would be to present the mock case + the materials with specific questions and see how users would use it naturally?
- **Message 3:** But if i want to focus on the TAM questionnaire validation with questions (draft of the questions not final version)-> how the testing flow needs to be adjusted? or by providing a task + the materials and then asking the tam-based survey questions is the structure to have? + TAM questionnaire from previous section
- **Message 4** Could you create a instructions sheet or agenda recapping the structure of the user testing session?
- **Message 5:** I am going to be using this case and this tam questionnaire [inputted last version of tam questionnaire] in the use test, recap the agenda for the whole use test

The output of this last and final iteration is documented in Chapter 7.2.1.