Faculteit Industriële Ingenieurswetenschappen

master in de industriële wetenschappen: informatica

Masterthesis

Data management plan completion and reviewing through human-in-the-loop AI

Seppe Vandenberk

Scriptie ingediend tot het behalen van de graad van master in de industriële wetenschappen: informatica

PROMOTOR :
Prof. dr. Davy VANACKEN

Gezamenlijke opleiding UHasselt en KU Leuven

Universiteit Hasselt | Campus Diepenbeek | Faculteit Industriële Ingenieurswetenschappen | Agoralaan Gebouw H - Gebouw B | BE 3590 Diepenbeek

Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE 3590 Diepenbeek
Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE 3500 Hasselt

2024
2025

Faculteit Industriële Ingenieurswetenschappen

master in de industriële wetenschappen: informatica

Masterthesis

Data management plan completion and reviewing through human-in-the-loop AI

Seppe Vandenberk

Scriptie ingediend tot het behalen van de graad van master in de industriële wetenschappen: informatica

PROMOTOR :

Prof. dr. Davy VANACKEN

UHASSELT KU LEUVEN

Foreword

The choice for this topic stems from my ambition to begin my doctoral research in October. As a data management plan forms part of the complementary aspects of research, I considered it valuable to contribute to the advancement of current support tools. The aim was to simplify certain administrative tasks and explore how technology can assist in this process. Furthermore, the theory behind delivering context-dependent feedback and support aligns closely with the focus of my doctoral work, which investigates the integration of AI-driven digital twins to support self-regulated learning in laboratories. I also chose to complete my thesis in English in order to further strengthen my language skills and better prepare myself for future presentations and publications throughout my academic career.

I would like to express my sincere gratitude to my promotor, Prof. Dr. Davy Vanacken, for his willingness to guide me, for understanding my way of working, and for offering valuable suggestions that have elevated my research. His follow-up and insights have helped me uncover new perspectives. I would also like to thank Prof. Dr. Gustavo Rovelo Ruiz and Ing. Jarne Thys for their continuous guidance throughout this project and for their assistance in connecting me with participants with the appropriate profile for the user study. In addition, I am grateful to Nicky Daniels for the time she dedicated to introducing me to the domain of data management and to the cognitive walkthrough, which provided insights that were instrumental in the development of AIDD4DMP and the prototype. Finally, I wish to express my sincere appreciation to my parents, whose unwavering support, encouragement, and care have been invaluable throughout my studies and in bringing this thesis to completion.

In the preparation of this thesis, I utilized several AI-based tools to enhance clarity, coherence, and efficiency. Language models such as ChatGPT, Gemini, and Claude were employed to assist in refining the text, improving wording, and ensuring readability throughout the manuscript. Additionally, Claude AI played a key role in debugging and refactoring portions of the codebase, helping to streamline development processes. It also supported the organization and cleaning of my personal notes and results gathered during the user studies. These tools were valuable aids that complemented my own efforts, while all intellectual contributions and interpretations remain my own.

Nederlandse Samenvatting

0.1 Achtergrond en Literatuurstudie

Modern datagedreven onderzoek vereist efficiënt Research Data Management (RDM), waarbij Data Management Plans (DMPs) cruciale documenten vormen die beschrijven hoe onderzoeksdata verzameld, opgeslagen, gedeeld en bewaard worden gedurende de gehele onderzoekscyclus [1], [2]. Ondanks hun belang ondervinden onderzoekers vaak moeilijkheden bij het invullen van DMPs door complexe vereisten, evoluerende standaarden en beperkte ondersteuning. Deze uitdaging wordt versterkt door het tijdsintensieve reviewproces voor RDM-teams, die gepersonaliseerde feedback moeten leveren over diverse onderzoeksdomeinen terwijl ze een toenemend aantal inzendingen beheren [3].

Recente ontwikkelingen in Generatieve AI en Large Language Models (LLMs) bieden veelbelovende oplossingen voor academische ondersteuning. Transformer-gebaseerde architecturen, vooral GPT-modellen, hebben tekstgeneratie gerevolutioneerd door hun vermogen om volledige sequenties gelijktijdig te verwerken via aandachtsmechanismen [4], [5]. Deze modellen kunnen worden verbeterd door technieken zoals Retrieval-Augmented Generation (RAG), die dynamisch externe kennis integreert om antwoorden te baseren op feitelijke data [6], en prompt engineering, die inputstructuren optimaliseert om modeloutput te sturen [7]. Hoewel AI-schrijfassistenten potentieel hebben getoond in verschillende academische contexten [8], [9], brengt hun toepassing op gestructureerde formulieren zoals DMPs unieke uitdagingen met zich mee, waaronder het waarborgen van institutionele compliance, het behouden van domeinspecifieke nauwkeurigheid, en het balanceren van automatisering met onderzoekersautonomie [10].

0.2 Framework Ontwikkeling en Implementatie

Het AIDD4DMP framework werd ontwikkeld volgens een systematische methodologie gebaseerd op het IDEE-framework (Identificatie, Design, Executie, Evaluatie) [11]. De systeemarchitectuur rust op vier fundamentele pijlers:

1. **Assisteren (Assist)**: Biedt real-time AI-ondersteuning voor het genereren en verfijnen van DMP-inhoud
2. **Integreren (Integrate)**: Consolideert alle bronnen en tools binnen één verenigd platform
3. **Dialoog (Dialogue)**: Maakt multidirectionele communicatie mogelijk tussen onderzoekers, AI en RDM-teams
4. **Ontwikkelen (Develop)**: Verzekert continue verbetering door feedback en adaptieve up-

dates

De technische implementatie gebruikt een modulaire architectuur met drie hoofdcomponenten. De LLM-omgeving gebruikt lokaal geïmplementeerde modellen (Llama3 en Gemma3n) via Ollama om dataprivacy te waarborgen. De dataverwerkingspijplijn gebruikt LangChain voor documentparsing en FAISS voor vectorembedding-opslag, wat semantisch ophalen van relevante inhoud uit geüploade onderzoeksdocumenten mogelijk maakt.

Het systeem implementeert RAG met Maximum Marginal Relevance (MMR) om relevantie en diversiteit in opgehaalde documenten te balanceren [12]. Belangrijke parameters omvatten het ophalen van 30 kandidaatdocumenten en het retourneren van 12 na ranking ($\lambda = 0,5$ voor gelijke relevantie/diversiteitsweging). Drie feedbackmechanismen werden ontwikkeld: een sterbeoordelingssysteem voor eenvoudige kwaliteitsbeoordeling, een state machine-gedreven chatbot voor gedetailleerde conversationele feedback, en een hybride gelaagde aanpak die detail aanpast op basis van tevredenheidsniveaus.

0.3 Resultaten en Evaluatie

0.3.1 Bevindingen Gebruikersstudie

De evaluatie betrok elf onderzoekers met eerdere DMP-ervaring, getest over twee LLM-implementaties. Gemma3n toonde superieure nauwkeurigheid (83% correcte beoordelingen versus 40% voor Llama3) en consequent hogere pragmatische scores. Deze bevinding sluit aan bij onderzoek dat aantoont dat beknopte, taakgerichte outputs vaak effectiever zijn dan uitgebreide verklaringen [13].

De User Experience Questionnaire-Short (UEQ-S) resultaten onthulden positieve percepties voor beide modellen:

- **Pragmatische scores**: Llama3: 0,82, Gemma3n: 0,97 (beide $> 0,8$ duidt op goede bruikbaarheid);
- **Hedonische scores**: Llama3: 1,14, Gemma3n: 1,02 (beide $> 1,0$ duidt op positieve betrokkenheid).

Van de drie geteste feedbackbenaderingen:

- de **hybride gelaagde oplossing** behaalde de hoogste bruikbaarheidsscores (1,22),
- het **sterbeoordelingssysteem** (0,99) werd gewaardeerd voor eenvoud en duidelijkheid,
- de **chatbot-benadering** (0,83) toonde potentieel voor complexe tekstverbetering;

Temporele analyse onthulde dat gebruikers die recent DMPs hadden ingevuld Gemma3n's directe stijl prefereerden, terwijl degenen met oudere DMP-ervaring Llama3's verklarende aanpak waardeerden, wat suggereert dat gebruikersverwachtingen evolueren met AI-toolblootstelling [14].

0.3.2 Resultaten Cognitieve Walkthrough

De evaluatie met de RDM-databeheerder benadrukte verschillende kritische vereisten consistent met best practices in human-in-the-loop systemen [15], [16]:

- Voorkeur voor beknopte, uitvoerbare feedback boven uitgebreide verklaringen

- Behoefte aan versiebeheer en antwoordgeschiedenis-tracking
- Waarde van contextueel vlaggen boven traditionele e-mailcommunicatie
- Belang van geïntegreerde institutionele richtlijnen en sjablonen

0.4 Discussie

De bevindingen onthullen fundamentele afwegingen in AI-ondersteunde onderzoeksondersteuning. Gemma3n's succes toont aan dat beknoptheid en directheid vaak zwaarder wegen dan transparantie in praktische toepassingen, wat aannames uitdaagt over gedetailleerde uitleg die tevredenheid verbetert. De sterke prestaties van de hybride feedbackoplossing geven aan dat adaptieve interfaces die diverse gebruikersvoorkeuren accommoderen cruciaal zijn voor acceptatie. Beveiliging en privacy kwamen naar voren als fundamentele vereisten, waarbij de meerderheid van de deelnemers lokale hosting binnen het instituur voorwaardelijk accepteerde alleen met expliciete beveiligingsgaranties. Deze voorzichtige acceptatie reflecteert bredere zorgen over dataprivacy in academische contexten.

0.5 Conclusie

Het AIDD4DMP-framework toont succesvol aan hoe human-in-the-loop AI DMP-invulling kan verbeteren terwijl onderzoekersautonomie en institutionele compliance behouden blijven. De modulaire architectuur maakt aanpassing aan verschillende contexten en gebruikersvoorkeuren mogelijk, terwijl lokale implementatie dataprivacy waarborgt. Beide LLM-implementaties behaalden positieve bruikbaarheidsscores, waarbij de hybride feedbackbenadering naar voren kwam als de meest effectieve oplossing.

Het onderzoek levert verschillende belangrijke inzichten: beknopte AI-antwoorden presteren over het algemeen beter dan uitgebreide verklaringen; adaptieve interfaces die zich aanpassen aan gebruikerstevredenheidsniveaus zijn het meest effectief; en beveiligingsoverwegingen zijn van het grootste belang voor adoptie in onderzoekscontexten. Toekomstig werk zou zich moeten richten op het uitbreiden van institutionele contextualisering door verbeterde RAG-implementatie, het ontwikkelen van meer geavanceerde prompt engineering-technieken, en het uitvoeren van longitudinale studies om evoluerende gebruikersbehoeften te volgen.

Contents

List of Tables

List of Figures

Glossary

Data Management Plan (DMP)	An official document that outlines how the data will be generated, stored, accessed, and shared; what metadata and documentation will be included; and, if long-term storage is planned, how the data will be protected securely
Fine-tuning	The process of taking a pre-trained AI model and adjusting its parameters on a smaller, task-specific dataset to improve performance on a particular application or domain.
Human-in-the-loop AI	An approach to artificial intelligence in which human feedback, oversight, or decision-making is integrated into the AI workflow to improve accuracy, reliability, and usability.
MoA (Mixture of Attention)	An AI architecture that combines multiple attention mechanisms, allowing a model to focus on different aspects of input data simultaneously for improved context-awareness and performance.
MoE (Mixture of Experts)	An AI architecture that dynamically routes input to specialized "expert" sub-models, enabling efficient scaling and improved performance on diverse tasks.
Prompt Engineering	The practice of designing, refining, and optimizing input prompts for AI models to achieve desired outputs, responses, or behaviors effectively.
RAG (Retrieval-Augmented Generation)	A method that enhances generative AI models by retrieving relevant information from external data sources to inform and improve the generated output.
Research Data Management (RDM)	The organization, storage, preservation, and sharing of data collected and used during research projects, ensuring compliance with policies and facilitating reproducibility and reuse.

Abstract in Dutch

Modern, data-gedreven onderzoek vereist efficiënt en kwalitatief beheer van onderzoeksdata (Research Data Management, RDM), waarbij het plannen van gegevensverzameling, opslag en delen volgens richtlijnen centraal staat. Deze thesis onderzoekt hoe AI-technologie onderzoekers en RDM-teams kan ondersteunen bij het invullen van verplichte Data Management Plans (DMP's) via real-time, contextbewuste suggesties en continue feedbackloops.

Een human-in-the-loop-systeem werd ontwikkeld voor interacties tussen gebruikers, AI en RDM-teams in complexe situaties. De oplossing gebruikt prompt engineering en retrieval-augmented generation, gecombineerd met vector-embeddings en gelijkeniszoekopdrachten, om large language model (LLM)-antwoorden te optimaliseren. Een hybride interface met gelaagde feedbacktools en vraagvlagknoppen maakt communicatie met het LLM en het RDM-team mogelijk.

Gebruikersonderzoek met onderzoekers die eerder een DMP invulden, toonde tevredenheid: zowel de pragmatische als de hedonische dimensies werden positief beoordeeld ($> 0,8$). De hybride feedbacktool scoorde het hoogst op bruikbaarheid (1,22), beter dan sterbeoordelingen (0,99) en chatbot-gestuurde feedback (0,83). De cognitive walkthrough met een RDM-admin benadrukten het belang van overschrijfrechten, persistente antwoordlogs en vraagvlagging. Deze bevindingen suggereren dat generatieve AI niet alleen DMP-processen verbetert, maar ook breder toepasbaar is voor gestructureerde formulierondersteuning.

Abstract in English

Modern data-driven research necessitates efficient and high-quality research data management (RDM), which involves planning data collection, storage, and sharing according to applicable guidelines. The main objective of this thesis is to investigate how AI technology can support researchers and RDM teams in completing mandatory Data Management Plans (DMPs) through real-time, context-aware suggestions and continuous feedback loops.

A human-in-the-loop system was developed to enable interactions between users, AI, and RDM experts in complex scenarios. The solution combines prompt engineering and retrieval-augmented generation (RAG) with vector embeddings and similarity search to optimize large language model (LLM) responses. A hybrid interface with tailored components, such as layered feedback tools and question flag buttons, allows users to both interact with the LLM and communicate seamlessly with the RDM team.

User studies with researchers who had previously completed a DMP indicated high satisfaction, with both pragmatic and hedonic quality scoring $> 0,8$ on the UEQ-S scale. The hybrid feedback tool achieved the highest usability score (1,22), outperforming star ratings (0,99) and chatbot-driven feedback (0,83). Cognitive walkthroughs with RDM staff underscored the importance of overwrite permissions, answer logs, and question flagging. These findings suggest that generative AI can significantly enhance DMP workflows and has broader potential for supporting the completion of forms in research and beyond.

Chapter 1

Introduction

1.1 Context and Background

In recent years, substantial investments from both public and private sectors have significantly bolstered research and development (R&D) initiatives. As illustrated in Figure 1.1, higher education emerged as the second-largest sector in terms of R&D expenditure in 2022 [17]. Modern academic research is characterized by a growing volume of data and an increasing demand for robust data governance [18].

This amplifies the necessity for effective and transparent Research Data Management (RDM), which involves systematically collecting, organizing, documenting, storing, and sharing research data throughout its lifecycle to maximize its accessibility, reproducibility, and long-term preservation. RDM is more than a set of internal institutional guidelines or best practices; it also encompasses legal and regulatory frameworks, such as the General Data Protection Regulation (GDPR) [1], and the ethical review processes governed by bodies like the Social and Societal Ethics Committee (SMEC) [2]. The rules and requirements shaping RDM come not only from funding agencies, such as the Bijzonder Onderzoeksfonds (BOF) [3] and the Fonds Wetenschappelijk Onderzoek (FWO) [4], but also from research institutions, including universities (such as Hasselt University [5]), research facilities, and private companies. The final set of RDM guidelines for a project is therefore a combination of requirements from these stakeholders. Effectively managing this set is essential to ensure that all parties are aligned, obligations are met, and the research process remains efficient, compliant, and collaborative [19].

In response to this growing need for effective RDM, researchers will likely be required to set up comprehensive Data Management Plans (DMPs). A DMP outlines how the data will be generated, stored, accessed, and shared; what metadata and documentation will be included; and, if long-term storage is planned, how the data will be protected [1], [2]. Researchers at Hasselt University must complete this process within the first six months of their research project, update their DMP regularly, and at the end of their study [20]. Researchers in Flanders create

[1] https://gdpr-info.eu/

[2] https://www.uhasselt.be/en/research/responsible-research-and-innovation/social-and-societal-ethics-committee-smec

[3] https://www.uhasselt.be/en/research/research-funding/funding-programmes/bof-programmes

[4] https://www.fwo.be/en/

[5] https://www.uhasselt.be/en

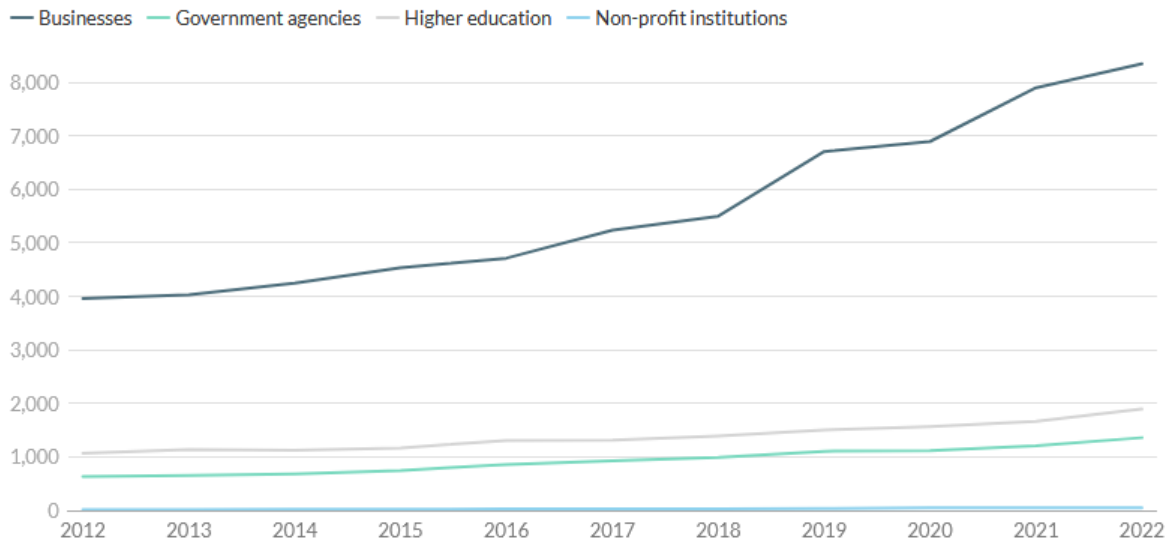


Figure 1.1: R&D expenditure according to implementation sector, for the Flemish Region (in million euros, current prices) [17].

DMPs through an online platform called DMPonline.be [20]. Hasselt University strongly advises researchers to engage with their in-house RDM team for feedback before final submission. This feedback is pivotal in validating proposed data management strategies' completeness, accuracy, and adherence.

However, DMPs can be challenging for researchers to complete effectively. Each research project requires a DMP tailored to its specific context, data types, and workflows. The process often involves navigating RDM-specific terms and interpreting evolving standards. These demands are compounded by the variety of data formats, storage solutions, and legal or ethical considerations, making it difficult for researchers, mainly those new to the process, to produce comprehensive and compliant plans without additional guidance. Consequently, support in the form of proposed answers and tailored recommendations in feedback becomes particularly valuable.

Nevertheless, providing effective feedback is not easy. Clear, constructive, and actionable feedback is essential to help researchers improve their plans, deepen their understanding of RDM practices, and adopt more effective strategies. Poorly formulated feedback (whether vague, overly critical without guidance, or delayed) can discourage improvement, foster misunderstandings, and hinder progress [21]. When applied to DMPs, such ineffective feedback can lead to plans that fail to meet requirements and regulations, overlook key aspects of data organization, or expose projects to risks such as data loss or misuse, thereby undermining the integrity of the research process.

At Hasselt University, the RDM team has encountered these operational hurdles firsthand and has proactively developed a structured checklist to systematize recurring feedback points, recognizing that DMP submissions frequently exhibit similar errors. Despite these measures, the dynamic nature of data management, combined with the growing number of DMP submissions (as it became mandatory), increases the time required for thorough evaluation, while available time decreases. As a result, turnaround times for feedback can be long, meaning researchers often wait extended periods before receiving general comments. This delays iteration and prevents them from refining their plans at their own pace, creating a bottleneck in the improvement process.

Recognizing these persistent challenges in delivering effective, timely, and consistent feedback, this thesis explores the potential of leveraging AI models to support and enhance the DMP feedback generation process. By integrating AI-driven support, the aim is to provide researchers with efficient, context-aware and real-time assistance, thereby improving the overall quality of DMPs and reducing the administrative load on the RDM team.

1.2 Research Objectives and Questions

The overarching objective is to enhance the completion and validation of DMPs by integrating human expertise with real-time AI assistance, thereby optimizing the process for researchers and RDM teams. While Generative Artificial Intelligence (GenAI) has demonstrated transformative potential in automating feedback generation and text writing, leveraging sophisticated approaches like Retrieval-Augmented Generation (RAG), model fine-tuning, and prompt engineering to deliver context-relevant information [22], [23], its direct implementation in DMP evaluation or completion support is not straightforward. This is because DMPs are highly nuanced and context-dependent. Furthermore, a critical practical constraint is data availability: due to confidentiality and privacy regulations, previous DMP submissions cannot be utilized for training or fine-tuning models. Consequently, the usable data is a set of predefined general guidance points, the questions, and their accompanying instructions.

To address this, the thesis is guided by the following main research question: **"How can a human-in-the-loop system, powered by generative AI models, effectively provide accurate, context-aware, real-time feedback and answer proposals for DMPs?"** To systematically explore this main research question, this research is structured around a set of interlinked objectives and corresponding sub-questions.

The first objective (RO1) involves a comprehensive review of both the current DMP process at Hasselt University and the existing academic literature on state-of-the-art techniques and frameworks for AI-assisted document completion and feedback. This leads to foundational design requirements and constraints for the integration of AI components into a modular DMP completion framework. Specifically, RO1 informs and partially addresses the following sub-questions: "Which AI algorithms are optimal for providing accurate and relevant feedback to applicants?" (RQ1), "How can an AI model generate meaningful feedback without relying on historical user data? Specifically, how can a Large Language Model (LLM) be effective using only general feedback principles and question formulations?" (RQ2), and "How can AI seamlessly integrate into a human-in-the-loop system to produce and refine structured feedback reports?" (RQ3).

The second objective (RO2) is developing a feedback generation and text proposal pipeline for the DMP completion framework. This involves a functional and practical pipeline for processing input and generating output to support researchers. The pipeline must comply with the RO1 design requirements and constraints, drawing on theoretical insights from similar frameworks and practical approaches and guidelines from user-centered design. Completing RO2 contributes to all three previous research sub-questions.

The third objective (RO3) is to steer the LLMs' output for context-aware DMP support. This entails integrating and directing LLMs to provide relevant, context-aware feedback and text proposals for DMPs across various research domains. The model will need access to general feedback principles and question formulations, without reliance on historical DMP answer data.

This objective directly addresses the core of RQ2.

Finally, the fourth objective (RO4) involves prototype development and user evaluation to validate the effectiveness and usability of the AI-driven approach. The functional prototype will integrate the AI-driven pipeline within one of DMP forms, allowing RDM administrators to monitor and interact with feedback and answers, and providing researchers with real-time AI-generated feedback and text proposals. This objective directly addresses the following sub-questions: "How does AI-driven support impact the applicant's experience, and what are its potential advantages and disadvantages?" (RQ4). By examining the impact on applicants' experience, complemented with expert findings, RO4 evaluates the effectiveness and impact of the system.

By pursuing these objectives and addressing the associated sub-questions, this thesis will develop a robust framework for AI-assisted feedback generation, and provide insights into the ways users and AI can interact to provide effective feedback to guide each other. This will not only contribute to maintaining the quality of feedback but also improve the efficiency of the DMP review and completion process.

1.3 Methods

This research employs an iterative design methodology, integrating theoretical research, prototyping, and evaluation to develop an AI-driven DMP form system. The first stage focuses on establishing the conceptual groundwork, beginning with a literature review to analyze existing AI-assisted feedback and evaluation tools, human-in-the-loop approaches, and AI feedback generation techniques. This review will identify gaps in current methodologies and define the requirements for the proposed system. Additionally, an interview was conducted with an RDM-team member to gain insights into the current challenges associated with DMPs. Based on the literature review and the interview, a modular framework (AIDD4DMP) will be designed to integrate AI agents into the form completion protocol, together with a structured communication pipeline to facilitate the user-AI interaction (e.g., defining the used information sources, response processing mechanisms, and integration protocols).

The second stage focuses on developing a functional prototype. A local web application will be built using Reflex[6] to simulate a platform similar to DMPonline.be, allowing for realistic user interactions with AI-supported features. The backend will be implemented in Python, using libraries such as LangChain[7] to integrate AI features. Local file storage and an event logger will be included to track user progress and generate data for analysis. A locally hosted LLM (e.g., Gemma3n[8], Llama3[9]) will run through Ollama[10]. This approach reduces development costs and enable easy plug-and-play model switching. The models will have access to general feedback guidelines, the RDM team's checklist, question phrasing, and data chunks with information from uploaded documents stored in vector embeddings.

[6]Reflex is a web framework for building reactive applications maintained by Pynecone Inc. (https://reflex.dev/)

[7]LangChain is a framework for developing applications powered by language models, developed by LangChain Inc. (https://www.langchain.com/)

[8]Gemma3n is a locally hosted large language model (Google Deepmind holds the trademark): https://deepmind.google/models/gemma/gemma-3n/.

[9]Llama3 is a large language model developed by Meta AI (Meta Platforms, Inc. holds the trademark): https://www.llama.com/models/llama-3/.

[10]Ollama is a platform for deploying and managing AI models locally, enabling flexible model switching (Ollama Inc. holds the trademark): https://ollama.com/.

The final stage evaluates the AI-enhanced DMP form system from both end-user and administrator perspectives. User studies will be conducted with researchers who completed a DMP before, and a UEQ-S survey [24], together with semi-structured interviews. This study assesses the usability, efficiency, and perceived quality of the AI-generated support. Administrators will evaluate the system through a cognitive walkthrough, demonstrating all admin and user features. During an open dialogue, feedback on the system's functionalities will be collected. The results from the evaluation phase will guide further refinements and research opportunities. The research outcomes will be disseminated through this thesis and a poster, highlighting key aspects.

1.4 Thesis Outline

This master's thesis is organized to address the research objectives in a clear and logical order. After this introduction, Chapter 2: Background provides the theoretical and empirical foundation for the study. It reviews relevant literature and presents an overview of concepts, such as generative AI, large language models, and related techniques (e.g., RAG and prompt engineering). The chapter also examines the current use of AI in academic support and identifies the gap this thesis aims to address.

Chapter 3: Framework Development and Implementation describes the design and implementation of the proposed solution. It introduces the AIDD4DMP framework, explaining its components and how it supports collaboration between human users and AI. The technical implementation is outlined, including key design decisions and system architecture. The chapter concludes with discussing the human-in-the-loop functionalities.

Chapter 4: Evaluation of AIDD4DMP: Findings and Discussion presents the evaluation methodology used to assess performance, usability, and overall impact and shows its outcome. It begins with findings from user study, followed by results from the cognitive walkthrough, providing practical validation of the system. The chapter ends with an analysis of how these results address the research questions. It concludes with reflections on the broader impact of the developed system and suggestions for future research.

Finally, Chapter 5: Conclusion summarizes the main findings and contributions of the thesis, highlighting implications for software systems engineering and research data management.

Chapter 2

Background

This research builds upon existing efforts in the fields of human-in-the-loop AI systems and generative AI. This section explores related tools to analyze their strengths, limitations, and potential gaps. This review also covers some fundamental concepts, such as LLMs and their performance enhancement techniques. This forms the basis for the design of the proposed AI form completion support, ensuring that it leverages proven techniques while addressing current challenges in the domain.

2.1 Data Management Plans in Research

Data management is important in academic research, not only to prevent data loss or the unintentional sharing of sensitive information, but also to meet institutional and funder requirements. As part of this, creating a DMP is often a mandatory step in the research process. However, the requirements of a DMP depend on the agency that mandates it, as these are shaped by their policies [3]. Still, many researchers struggle with implementation, as they are usually expected to navigate growing standards with limited support and unclear guidelines. The current support tools consist of consultation sessions with experts and participating in educational programs [1]. Data experts are increasingly recognized in the literature as key stakeholders in improving research data management. Their expertise in data organization and institutional policies positions them well to support DMP development and researcher training [1]. Given that data experts must balance limited time and resources with the quality of support they provide, GenAI offers a promising way to deliver more scalable, consistent, and accessible assistance in creating and reviewing DMPs.

2.2 Generative AI and Large Language Models

AI underwent a revolutionary advancement over the past years and gained more attention than ever before; ChatGPT, for example, emerged as the hottest topic on the Internet in 2022 [25]. Moreover, LLMs were one of the most transformative subjects. These AI models are trained on large amounts of textual data, enabling them to generate human-like text [5]. Traditional machine learning models primarily analyze structured numerical data, but textual data is considered unstructured. LLMs are trained to understand patterns, not only to relate words to

each other, but also entire sentences, so they can interact in conversations or respond to questions [26]. Tools, such as ChatGPT and Gemini, and models, such as Llama3, are based on a Generative Pre-trained Transformer (GPT) architecture, focused on generating human-like text by using next word prediction. Other architectures that can be used for language tasks include RNNs (Recurrent Neural Networks) and CNNs (Convolution Neural Networks), but transformer-based models exhibit enhanced capability in capturing long-distance dependencies within textual data [27], [28].

2.2.1 Transformer-based Architectures

Transformers are a type of neural network architecture that revolutionized natural language processing by allowing models to process an entire sequence of text at once, rather than one token at a time. The key innovation is the attention mechanism, which enables the model to focus on different parts of the input simultaneously. Each "attention head" learns to capture specific relationships between words, for example, one head might track grammar, while another detects semantic or thematic connections across sentences. By combining multiple attention heads, the model builds rich, context-aware representations of the language [4], [5].

At a high level, a transformer consists of an encoder and a decoder, shown in Figure 2.1. The encoder reads the input text and creates a contextual representation of each token. Such a representation contains information about its relation to other tokens. The decoder then uses these representations to generate an output sequence. Multi-head attention, feed-forward layers, and normalization layers are repeated in multiple layers to progressively refine these representations [4], [29].

2.2.2 The GPT Model Family and ChatGPT

GPT is a decoder-only variant of the transformer. Unlike full transformers, it does not use an encoder for separate input processing. Instead, these models undergo a pre-training phase using a large amount of data sources such as books, websites, journals, and transcripts. During this training, a technique called causal masking is employed. This ensures that the model generates text sequentially, token by token, by strictly attending only to preceding tokens and preventing it from 'peeking' at future information in the sequence, making the text more coherent, much like a human writer, who composes text progressively [30].

When interacting with a user, ChatGPT receives a prompt and treats this as the initial segment of the sequence it needs to complete. Leveraging the intricate patterns and knowledge acquired during pre-training, it then predicts the most probable next token to extend the sequence. The evolution of GPT models has seen a dramatic increase in scale and capability: from the initial GPT-1 with 117 million parameters, which demonstrated impressive performance across various natural language processing (NLP) tasks, to subsequent iterations like GPT-2 (1.5 billion parameters), GPT-3 (175 billion parameters), and GPT-4 (approximately 1 trillion parameters). The most recent iteration, based on GPT-5 (with an undisclosed parameter amount), further expands its capabilities by handling a higher input size (272,000 tokens) and output size(128,000 tokens)[31], [32].

The large pre-training phase enables GPT models to do different NLP tasks, often without fine-tuning for specific objectives. This allows the models to do zero-shot learning, where they can

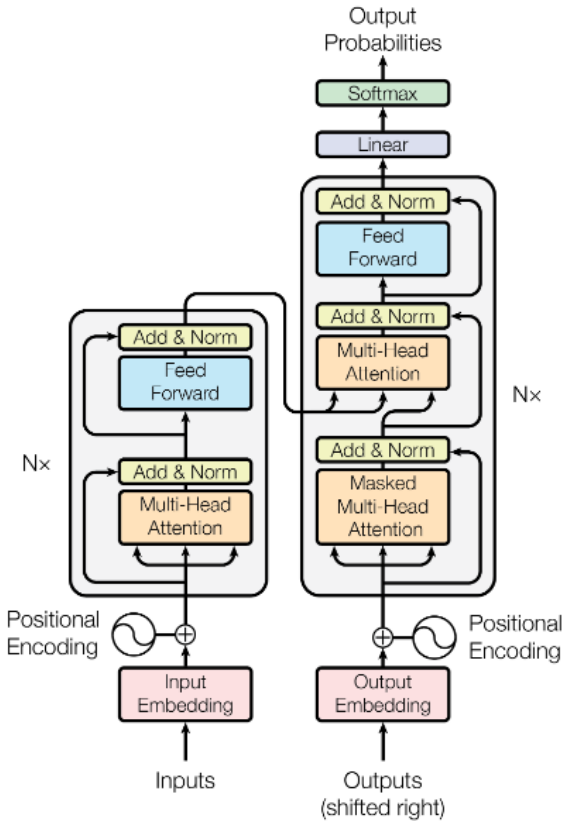


Figure 2.1: Transformer architecture, showing the encoder and decoder layers with multi-head attention, feed-forward networks, and normalization, reproduced from Vaswani, Brain, Shazeer, *et al.* [4].

tackle a task they have never seen before with just the prompt as guidance, or few-shot learning, where they achieve strong performance after being provided with only a handful of examples. Meanwhile, creating 'task experts' often requires more targeted approaches [31]. Subsections 2.2.3 and 2.2.4 will delve into methods designed to specialize these models for particular tasks.

2.2.3 LLM output enhancement techniques through structure adaptations

The most traditional approach for specialization is fine-tuning, where the pre-trained model undergoes additional training on a smaller, task-specific dataset. By doing this, the model updates its parameters to specialize in this task. While highly effective in boosting performance on specific tasks, full fine-tuning may risk 'catastrophic forgetting', where the model loses some of its broader pre-trained capabilities. In contrast, insufficient amounts will not lead to new reasoning capabilities [13], [33].

There are generally two approaches to implementing fine-tuned agents: either conducting the fine-tuning process independently using custom data and infrastructure or utilizing an existing fine-tuned model that has been trained for a similar task. The independent approach offers complete control and potential for better performance on specific use cases, but requires substantial computational resources and access to the specialized data. Using an existing fine-tuned model provides faster implementation and lower computational requirements, but offers less control and

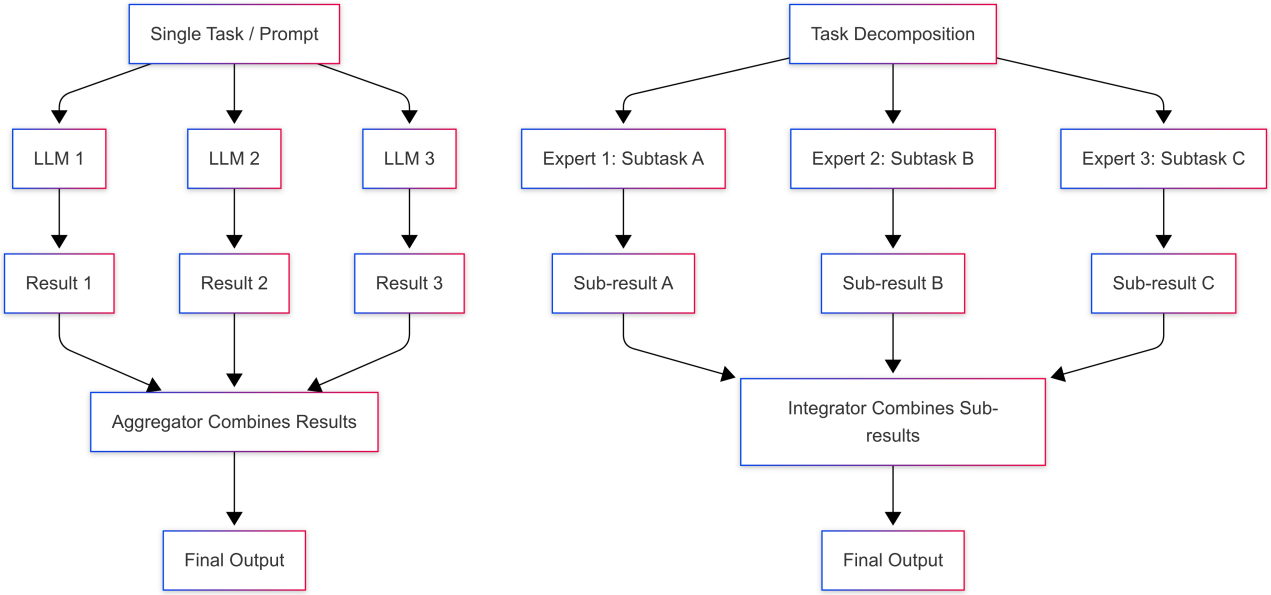


Figure 2.2: Parallel solving approach or MoA (left) vs. decomposition approach or MoE (right)

may not perfectly match specific requirements. For example, Gemma is a general model built by Google DeepMind, but some specialized public models are also available, such as CodeGemma, which has its expertise in general coding tasks, but might not be as experienced in a certain language as needed [34]. The choice between these approaches depends on available resources and specific requirements, with recent research showing that neither approach always dominates the other in terms of performance [13], [31], [33]. In the context of DMPs, historical submissions cannot be used for training due to confidentiality constraints, and compiling a dataset with detailed feedback for each question would be prohibitively time-consuming. In contrast, more advanced models, such as GPT-4, currently possess sufficient fine-tuning and domain-general capability to function effectively as feedback experts [35].

Alternatively, it is possible to combine the capabilities of multiple models. This concept is central to techniques like Mixture of Agents (MoA) and Mixture of Experts (MoE). The multi-agent systems have enhanced performance by combining intelligence, parallel decision making, and collaborative workflows, letting them tackle more complex problems [36], [37].

Collaboration among agents in multi-agent systems can be organized in various ways, typically distinguished by their strategies for task decomposition and information sharing. A widely used architecture distributes the same problem to multiple agents in parallel, followed by an aggregation of their diverse responses, as illustrated in Figure 2.2 (left). This architecture is called MoA; each agent, though receiving the identical initial problem, potentially phrased slightly differently, applies distinct internal inference mechanisms or draws upon different external knowledge sources (e.g., separate retrieval databases). Through parallel processing, it generates a range of potential solutions or insights. An 'aggregator' component then collects these multiple outputs, identifies consensus, resolves discrepancies, and synthesizes the most robust and comprehensive final answer. This approach benefits from the 'wisdom of the crowd', enhancing the confidence in the predictive outcomes and improving generalization by capturing a wider array of interpretations and solutions [38].

An alternative collaborative architecture is MoE, which involves splitting a complex problem into sub-problems and delegating these to specialized agents, which are called 'experts'. As can be seen in Figure 2.2 (right), MoE uses a central 'planner' or 'orchestrator' agent, which analyzes the bigger task and breaks it down into smaller sub-tasks. Each sub-task is then routed to a specific expert best suited for that particular domain [37] (e.g., a 'Math Agent' for calculations, or a 'Coding Agent' for code generation [38]). These specialized agents work on their assigned sub-problems, potentially accessing their own dedicated knowledge bases or tools. Their individual outputs are then returned to an 'aggregator' agent, which synthesizes these partial solutions into a comprehensive final answer. This hierarchical decomposition allows for efficient parallel processing and leverages the distinct strengths of each agent, mimicking real-world organizational structures [38], [39].

Despite the significant promise of multi-agent systems, their implementation and optimization present several notable challenges. A primary concern is the computational cost and latency associated with running multiple models, either in parallel or sequentially. While the quality of output often improves, the increased demand for concurrent inference can lead to slower response times and higher operational expenses, especially for large-scale deployments [38]. Furthermore, the complexity of coordination and communication among agents is a non-trivial problem. Designing effective communication protocols, managing information flow, and ensuring coherent collaboration across diverse agents remains an active area of research [36]. Another challenge lies in debugging and steering these complex systems. Identifying the precise point of failure or the cause of an incorrect output becomes increasingly difficult. Tools are needed that allow developers to interactively inspect agent states, rewind conversations, and experiment with interventions without restarting the entire workflow [37]. Finally, a key challenge is ensuring robust generalization across diverse domains. The range of possible domains and applications in the context of DMPs, for instance, is too vast to predefine expert agents for every scenario. Combined with the inherent unpredictability of LLMs, this limits reliability and scalability. Research into unified frameworks and adaptive agent coordination continues to address these limitations [36], [40].

2.2.4 LLM output enhancement techniques through input and retrieval strategies

Even without modifying the underlying model or the structure, improvements in output quality can be achieved by optimizing the inputs. Prompt engineering is a crucial, albeit challenging, process for optimizing LLM performance on specific tasks. It improves quality by enabling precise task communication and guiding the model towards desired outputs through detailed instructions, contextual information, and structured reasoning templates. This allows LLMs to perform complex reasoning and generate more accurate, multi-step responses [7]. However, prompt engineering is difficult because it demands sophisticated human insight to diagnose model errors and formulate effective corrections. Even attempts at automated prompt engineering face limitations, as LLMs themselves often lack sufficient guidance to perform the complex reasoning needed for optimal prompt design, thus highlighting the continued importance of human expertise in this domain [7], [13]. The best results still come from using a conversational approach, where the user and AI model send messages back and forth [13]. Also, the lack of standardization across models can lead to inefficient prompts: a prompt that works well for model A can lead to less optimal responses from model B.

Besides changing the input or changing a model's parameters through retraining, it is possible to offer extra data directly to the LLM during inference. RAG enhances an LLM's output quality by dynamically incorporating external knowledge [6]. It operates by first retrieving relevant information from an external knowledge base based on a query. This retrieved context then directly informs the LLM's response generation, effectively grounding its output in factual data from these sources. RAG itself primarily complements an LLM's pre-trained knowledge rather than constituting a direct fine-tuning technique for the generative model's core weights [41].

The initial RAG paradigm has evolved into a sophisticated discipline, as the state-of-the-art goes beyond simple, single-step retrieval. Modern RAG pipelines are often multi-staged and incorporate advanced techniques to enhance accuracy and relevance. These include pre-retrieval processing, which involves modifying or enhancing a query before it is submitted to a retrieval system to improve result quality. Common strategies include query rephrasing, where the original query is reformulated to better match the terminology and structure of the target data, and query decomposition, which breaks a complex query into smaller, more focused sub-queries to address different aspects of the information need [12]. The retrieval step itself may use hybrid search, combining keyword-based and vector-based methods to find the most relevant documents. After retrieval, post-retrieval processing employs re-ranking models to ensure the most pertinent information is presented to the LLM first, mitigating the 'lost in the middle' problem where important facts are buried within a long context window [42].

Despite its potential, implementing a robust RAG system is not without its challenges. The effectiveness of the system is highly dependent on the quality of the external knowledge base, which introduces a fundamental 'garbage in, garbage out' risk. If the source data is inaccurate, outdated, or poorly structured, the RAG system will likely produce a poor response, regardless of the LLM's capabilities [43]. Furthermore, RAG does not completely solve the inherent limitations of LLMs. Even with high-quality retrieved data, the LLM may fail to correctly synthesize the information or may still hallucinate, especially when dealing with ambiguous or contradictory context. Finally, the complexity and computational cost of building and maintaining a production-grade RAG pipeline, which includes vector databases, indexing pipelines, and advanced retrieval models, can be a significant barrier to implementation [43].

Finally, the human-in-the-loop approach involves incorporating human expertise. These represent a highly effective strategy for AI system development, recognized for combining the analytical power of AI with the contextual understanding and critical judgment of human experts. Human-in-the-loop AI involves human oversight and intervention directly within the operational workflow. This allows for continuous validation and refinement of AI outputs, ensuring reliability and safety [15]. Humans stay in control and AI acts as a helpful assistant [16]. Human-in-the-loop relates to MoA by leveraging diverse capabilities for enhanced problem-solving, but distinctively integrates human intelligence alongside artificial intelligence, with human judgment often serving as the ultimate authority. However, challenges persist, including the complexity of designing effective human-AI interaction, potential biases, and scalability limitations due to human involvement. Recent studies demonstrate the great potential of human-in-the-loop AI applications across diverse domains, including academic peer review, where LLMs support reviewers in evaluating submissions [44]; healthcare, where clinicians continuously validate generative AI outputs [15]; and knowledge-intensive settings, where human experts provide context-specific explanations to enhance the interpretability of LLM outputs [45].

2.3 AI in Academic Support and Writing Tools

The advancements in AI offer a diverse array of tools designed for academic support. These tools span from grammar and style checkers (e.g., Grammarly [1]) to sophisticated writing assistants and specialized applications for managing literature reviews and data (e.g., Elicit [2]). Specifically, AI-powered writing assistants, such as QuillBot [3], ChatGPT [4], Microsoft Bing Copilot [5], and Writefull [6], have become increasingly prevalent [8], [9].

The integration of these technologies holds potential for enhanced writing proficiency [46], more personalized learning experiences [14], instant and consistent feedback [47], and ultimately boosts overall efficiency and productivity in academic tasks. However, the reliance on AI tools can inadvertently stifle the development of core academic skills such as critical thinking, originality, and self-editing [47]. Overreliance could lead to a decline in students' and researchers' ability to perform tasks without AI support [5]. This suggests a need to balance technological efficiency with the preservation of fundamental research discipline.

This section explores how AI can enhance feedback generation for DMP forms, emphasizing contextualization, transparency, and researcher autonomy. Prior research highlights the growing potential of LLMs in delivering personalized, context-aware feedback within digital environments. For instance, Limna, Jakwatanatham, Siripipattanakul, *et al.* [48] demonstrates how automating evaluations through machine learning and cognitive modeling can enable structured AI feedback to support complex evaluative tasks. Similarly, Bany Abdelnabi, Soykan, Bhatti, *et al.* [49] shows that LLMs can enhance students' skills by providing personalized, insightful feedback that fosters critical thinking and improves clinical case analysis and presentation. These studies illustrate how LLMs and NLP techniques can be leveraged not only to comprehend textual input but also to intervene meaningfully in documentation workflows, including those involved in DMP creation.

However, integrating AI into structured academic forms such as DMPs presents specific challenges, such as:

Ethical and legal responsibility: The introduction of AI complicates the question of who is ultimately responsible for the final content of a DMP. Suppose an AI tool produces an incorrect suggestion that results in a failed grant application, it remains unclear who bears the ethical and legal accountability. This ambiguity can undermine both researcher autonomy and accountability.

Reliability and factual grounding: As Bender, Gebru, McMillan-Major, *et al.* [10] describes, LLMs can act as "stochastic parrots", skilled at generating plausible-sounding text without true understanding or factual grounding. In DMP creation, where precision and adherence to institutional or funder guidelines are critical, this risk can lead to well-written but non-compliant plans.

Inspiration can be drawn from established educational frameworks that balance automation, ethical responsibility, and human agency to successfully and meaningfully integrate AI into DMP

[1] Grammarly. Grammarly: AI Writing Assistance. https://www.grammarly.com/
[2] Elicit, Elicit: The AI Research Assistant, https://elicit.com.
[3] QuillBot, a Learneo, Inc. business. https://quillbot.com/
[4] OpenAI, ChatGPT, 2025, https://chat.openai.com.
[5] Copilot, Microsoft, 2025, https://copilot.microsoft.com/.
[6] Writefull, Writefull, 2025, https://www.writefull.com/.

completion. For example, the IDEE framework, emphasizing Identification, Design, Execution, and Evaluation of AI tools [11], provides a structured approach to building systems that are acceptable and effective. Complementing this, Kong and Yang [50] highlights the value of fostering self-regulated learning by scaffolding complex tasks into manageable steps. For DMPs, this could translate into AI tools guiding researchers through iterative revisions; for example, when multiple key concepts are missing from a response, the AI might first focus feedback on the most critical one, and address secondary elements in subsequent iterations.

Furthermore, the FIRST-ADLX framework builds on these ideas by prioritizing ethical transparency and iterative review. Its "R" (Reviewing Actively) component aligns closely with the need for researchers to revisit and validate AI-suggested improvements in DMP drafts. Likewise, its "F" (Focusing on learner behavior) principle emphasizes respecting researcher autonomy, AI tools should offer non-prescriptive suggestions, such as highlighting inconsistencies, rather than overriding the researcher's responses. Together, these frameworks demonstrate how AI support for structured form completion can be strengthened by combining structured development principles (IDEE), scaffolded task decomposition, and ethically grounded iterative refinement (FIRST-ADLX).

These approaches are increasingly complemented by explainability features, which are essential for building user trust and ensuring transparency in feedback generation. Explainable AI (XAI) enables users to understand the reasoning behind an AI's suggestions, providing insight into why a specific change was recommended rather than expecting acceptance without justification [45].

LLMs can thus deliver adaptive feedback tailored to the content and context of a DMP by identifying inconsistencies, suggesting targeted improvements, and aligning responses with institutional and funding guidelines. By embedding these principles, AI systems have the potential to evolve from intelligent grammar checkers into active collaborators in DMP workflows, enhancing efficiency without compromising scholarly rigor or researcher agency, in much the same way as they have been successfully applied in personalized education and teacher support.

2.4 Synthesis and Research Gap

The recent surge in Generative AI and LLMs, particularly transformer-based architectures like GPT, has revolutionized text generation and various NLP tasks. These models, trained on vast datasets, excel at understanding and producing human-like text, enabling applications from general conversation to specialized content creation. However, a critical challenge in applying AI to highly structured academic tasks such as DMPs lies in balancing automation with domain-specific expertise. While current tools may handle surface-level corrections effectively, they often struggle with the nuanced requirements of specialized forms, which demand strict adherence to institutional standards and methodological best practices, whilst also working in a specific research context. For instance, AI can safely reference institutional and funder policies to suggest compliant data storage locations, but adapting these recommendations to the specific needs of a research project and providing informed support on questions related to data types, associated ethical considerations, and methodological implications requires access to project-specific information. Enabling such access, however, raises concerns about potential leakage of confidential or sensitive data [9].

Moreover, the dynamic nature of research projects necessitates feedback mechanisms that can

evolve with ongoing revisions, a feature still underdeveloped in most AI writing tools. Existing solutions also fall short in offering deep integration with institutional guidelines, research context, and transparent explanations for their suggestions. Techniques such as RAG and fine-tuning techniques offer promising directions, enabling LLMs to access real-time guidelines or historical data to improve contextual accuracy [14]. Nevertheless, adaptive prompt engineering, essential for maintaining output quality under changing conditions, remains a technical challenge [7], [13].

This research addresses these gaps by developing a human-in-the-loop system for DMP completion, leveraging static prompt engineering and RAG to optimize LLM output. The goal is to provide accurate, context-aware feedback and tailored DMP suggestions while preserving confidentiality and enabling researchers to iteratively improve their plans.

Chapter 3

Framework Development and Implementation

This chapter presents the methodological approach, architecture, and technical implementation behind the AIDD4DMP framework for AI-assisted DMP completion. This work is grounded in literature, as it follows the systematic methodology for integrating AI technologies into educational and academic support contexts provided by the IDEE framework [11]. Applied to DMP completion, this framework enables the development of an AI tool that supports and enhances rather than replaces expert knowledge.

3.1 Identification: Gap Analysis in DMP Workflows

The first phase of the IDEE framework involves systematic identification of gaps and opportunities where AI integration can provide meaningful value. For this thesis, the identification phase consisted of an analysis of the existing DMP completion workflows at Hasselt University through an open discussion with the data steward for Sciences & Technology of the RDM department and analyzing related documents containing the DMP template, guidelines, and instructions, which helped in establishing and confirming the challenges of current support tools.

From this initial interview, it was observed that the existing institutional framework at Hasselt University follows a traditional two-party model where researchers complete DMP forms independently before offering them to the RDM team for review and feedback. Analysis of this workflow revealed several inefficiencies that align with the broader challenges identified in the literature regarding the integration of AI into structured academic tasks. More specific challenges are that problems in DMPs are often similar across submissions, which has led the RDM team to design a feedback checklist allowing them to select guideline points that researchers should integrate. However, due to the diversity of research topics, research-related details can be overlooked, as RDM staff possess expertise in data management rather than in every specific research domain; that is why they mainly evaluate what is written down by the researcher. This process is also time-consuming, as it is not feasible to quickly pinpoint problems without explicit communication from the researcher, highlighting a potential tool that puts focus on specific problems in the DMP.

The current DMP support process at Hasselt University reflects the challenges documented in

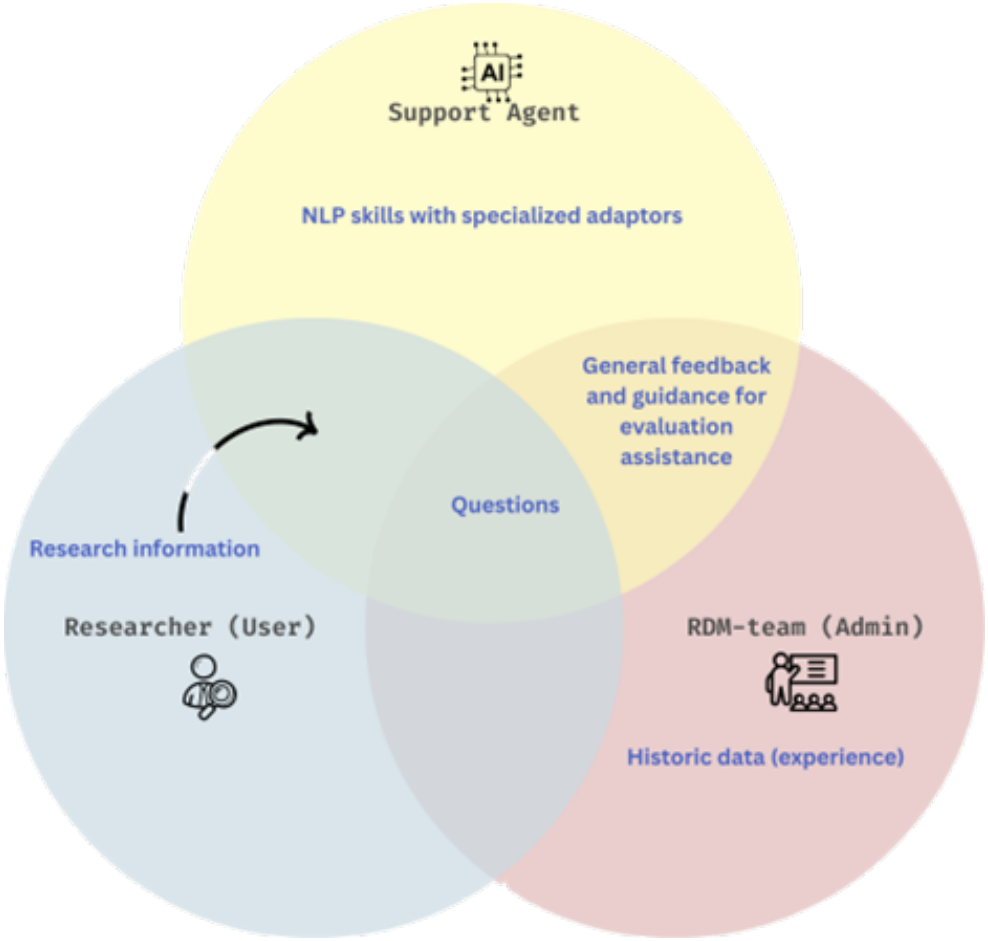


Figure 3.1: The three participating parties (the researcher, the RDM-team, and the AI-support agent) of the AIDD4DMP framework and the information they carry.

the literature (see Section 1.1), particularly balancing automation with the preservation of domain expertise and the assurance of correctness. Researchers face significant administrative burdens and struggle to navigate increasingly complex requirements with limited guidance [1], while RDM staff must deliver personalized, high-quality feedback across diverse disciplines despite constrained time and resources [21]. Addressing these constraints requires a framework that not only respects institutional compliance and best practices but also incorporates AI-driven capabilities such as context integration, real-time iterative assistance, embedded quality assurance mechanisms, and adaptive improvement through user feedback. These requirements emerged during the identification phase and form the foundation for the subsequent design process.

3.2 Design: AIDD4DMP Framework Architecture

The design phase of the IDEE framework translates the identified gaps and requirements into a coherent system architecture. The AIDD4DMP framework (for which the important relations are shown in Figure 3.1) was conceptualized around four foundational pillars that collectively address the requirements that were identified in the first phase.

Each of the four pillars of the AIDD4DMP framework addresses specific gaps and requirements while ensuring modularity and extensibility.

The **Assist** pillar focuses on providing researchers with real-time support during DMP completion. The AI system analyzes partial or complete researcher answers in light of relevant guidelines, best practices, and institutional requirements. It transforms this information into well-structured prompts, relieving researchers of the burden of sourcing the correct information or crafting effective prompts themselves. The resulting AI-generated outputs can take the form of proposed additions, context-aware feedback, or alternative phrasings. For RDM staff, this capability reduces administrative load and enables more targeted feedback in complex cases.

The **Integrate** pillar addresses the challenge of consolidating all relevant resources and tools within a single environment. Instead of using separate resources and tools, such as the main platform (e.g., dmponline.be) for completion, standalone institutional and funder portals to retrieve policies, third-party AI tools (e.g., ChatGPT) for advanced writing support, and email-based feedback requests, the aim is to unify these into one integrated, context-rich platform. By embedding support tools directly into the DMP platform, the system minimizes context-switching and delivers more effective support.

The **Dialogue** pillar operationalizes a human-in-the-loop approach by enabling multi-directional communication among researchers, RDM staff, and AI agents. This is represented by the intertwining of each party in Figure 3.1. Researchers can request guidance from both AI and RDM staff, while also providing feedback if responses are unclear or incorrect. RDM staff can intervene to steer AI outputs toward more effective guidance when necessary, ensuring that the AI remains a collaborative assistant rather than an autonomous decision-maker. This falls back on the theory behind the "F" (Focusing on learner behavior) of the FIRST-ADLX framework, where AI should trust the user. This communication framework preserves researcher autonomy, fosters collaboration, and allows each actor to contribute expertise where it is most valuable.

The **Develop** pillar emphasizes the continuous evolution of both the DMP content and the AI support. With all actors working together within a unified platform, the development of a high-quality DMP and the improvement of the underlying tools and workflows become an interwoven task. This requires ongoing performance monitoring, analysis of user feedback, and timely updates of the underlying structure (e.g., prompts, data repositories, and models). By embedding adaptability into the framework, the system remains aligned with evolving institutional policies, research practices, and AI capabilities [36], [40].

3.3 Execution I: AIDD4DMP Theoretical Implementation

The execution phase of the IDEE framework focuses on transforming the theoretical design into a functional prototype. This section begins by outlining the conceptual flow of data within the proposed system, illustrating how information is processed and routed. Section 3.4 presents the implementation details of the prototype, demonstrating how the design principles are realized in practice.

The data distribution pipeline, as shown in Figure 3.2, represents the complete process of transforming raw inputs into structured outputs for DMP support. At the start of each session,

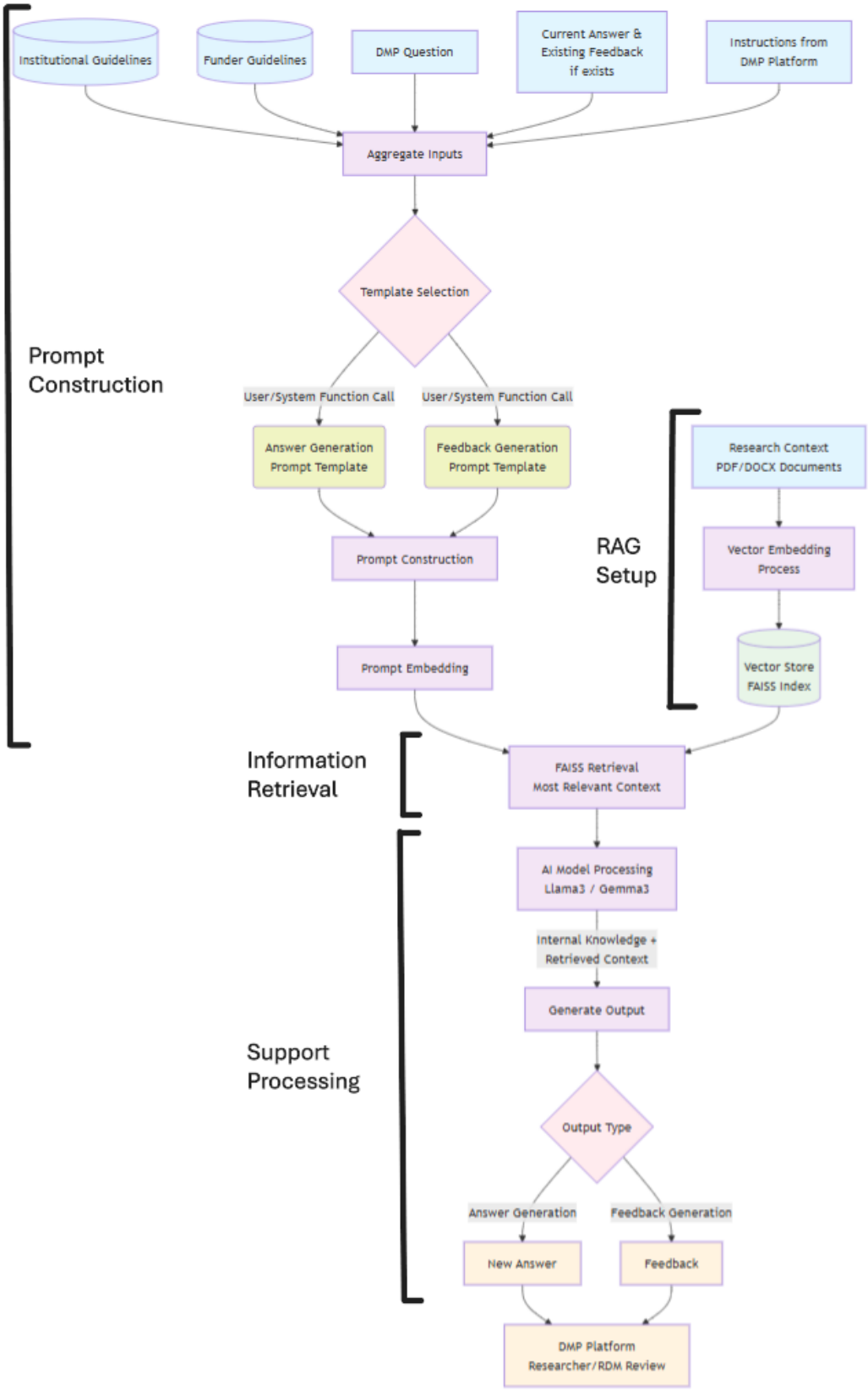


Figure 3.2: Overview of the dataflow pipeline, illustrating the full process from collecting data from different repositories to the final LLM-generated output using RAG enhanced processing

research context documents are converted into vector embeddings and stored in a vector store index, enabling efficient retrieval of relevant information during prompt execution. This stage is labeled "RAG Setup" in the diagram. On the opposite side, under "Prompt Construction", the system collects all relevant contextual inputs needed to address each DMP question.

For each DMP question, the system aggregates the following inputs:

- relevant **guidelines** from institutional repositories and funding agencies;
- the **current answer**, if one has been given already;
- any **existing feedback** from prior iterations;
- the **question formulation** itself;
- associated **instructions** from the DMP platform.

These elements are combined into one of two predefined, custom, structured prompt templates:

1. an **answer generation** prompt,
2. a **feedback generation** prompt.

The choice between these templates depends on the function call initiated by the system workflow.

Once the prompt is selected, it is embedded and processed through the "Information Retrieval" stage, where FAISS (Facebook AI Similarity Search [1]) identifies the most relevant contextual passages from the stored research data. Leveraging both this retrieved context and the model's internal knowledge, the "Support Processing" step produces either a new answer or targeted feedback. The output is then returned to the DMP platform for review by the researcher or RDM staff.

3.4 Execution II: AIDD4DMP Practical Prototype Implementation

The technical implementation of AIDD4DMP translates the needs of the tool identified in Section 3.2 and transforms the conceptual pipeline of Section 3.3 into a practical prototype for demonstration and evaluation purposes. The used approach integrates local AI deployment, sophisticated document processing, and human-in-the-loop mechanisms to create a comprehensive DMP assistance system.

3.4.1 LLM environment

The foundation of AIDD4DMP rests on local deployment using Ollama, which is a local LLM environment that enables downloading, managing, and running large language models directly on a personal machine. This method ensures that sensitive research data never leaves the institutional environment, which allows the tool to use research-related documents as an extra knowledge repository, without facing privacy concerns regarding the potential leakage of confidential research information when using cloud-based AI services. For the prototype, two LLMs, Llama3 and Gemma3n, were used. Llama3 is labeled as a general-purpose open-weight LLM optimized

[1] Faiss is a library for efficient similarity search and clustering of dense vectors, https://faiss.ai/index.html.

for broad NLP tasks, offering relatively strong reasoning and instruction-following capabilities, meaning that it can complete multi-step processes and explain what it did to get to that response, though it is not a model explicitly specialized for reasoning tasks. Gemma3n, on the other hand, is a lightweight, efficiency-focused model designed for local deployment, enabling faster inference and lower computational costs while maintaining competitive accuracy on smaller-scale tasks. These models were chosen to mitigate the performance bottlenecks inherent to local execution via Ollama, where prompt–response latency can be significant. By reducing model size and using newer models, response times were kept relatively short while maintaining sufficient output quality for system evaluation. The final implementation would benefit from higher-performance hardware to support larger, more advanced models or a multi-agent approach, which was not feasible during user testing due to the requirement to minimize delays. Furthermore, using both models also highlighted differences in output style: Gemma3n tended to produce efficient and concise responses, while Llama3 generated more detailed and explanatory outputs. Chapter 4 analyzes the difference in user experience with each LLM.

3.4.2 Data preprocessing

The beginning of the data distribution pipeline is document processing, labeled "RAG Setup" in Figure 3.2. Figure 3.3 shows where researchers can upload PDF and DOCX research-related documents (e.g., project proposals, partially written papers, previous related DMPs). AIDD4DMP employs LangChain's [2] document loaders to parse uploaded files, creating temporary file handles that ensure secure processing without persistent storage of sensitive documents. Each document undergoes semantic chunking using a recursive character text splitter (chunk size: 1000 characters, overlap: 200 characters, which are the maximum advised sizes for general tasks [51]) to maintain contextual coherence while enabling efficient retrieval. The chunking strategy addresses the challenge of preserving document structure while creating searchable segments that can be effectively embedded and retrieved during the question-answering process.

LangChain further orchestrates the pipeline by processing the extracted content through HuggingFace's[3] sentence-transformers model (all-MiniLM-L6-v2, a compact and efficient model suitable for laptop-based user studies), to generate vector embeddings that capture semantic relationships within the research documents. These embeddings are stored in a FAISS index, which provides efficient similarity search across large collections of research documents. In the context of DMP support, this allows the system to quickly retrieve relevant sections from the uploaded documents corresponding to the researcher's current question. For example, if the question is about what sensitive data is collected, the FAISS index can return previously embedded sections that specifically address data collection. The FAISS implementation also tracks metadata such as source filenames, processing timestamps, and session identifiers, enabling the system to cite the original documents or highlight the exact source of guidance in its output. This can be further illustrated with a screenshot showing a retrieved guideline linked to a specific DMP question.

Vector embedding–based retrieval offers an advantage over traditional flat text retrieval by enabling semantic rather than purely lexical matching. Whereas keyword-based approaches rely on exact word or phrase matches, embeddings represent text as high-dimensional numerical vectors

[2]LangChain is a framework for building applications with LLMs using modular components such as memory management and chains for task orchestration https://www.langchain.com/.

[3]Hugging Face provides open-source tools and models for natural language processing, including Transformers and sentence-transformers for embedding generation https://huggingface.co/.

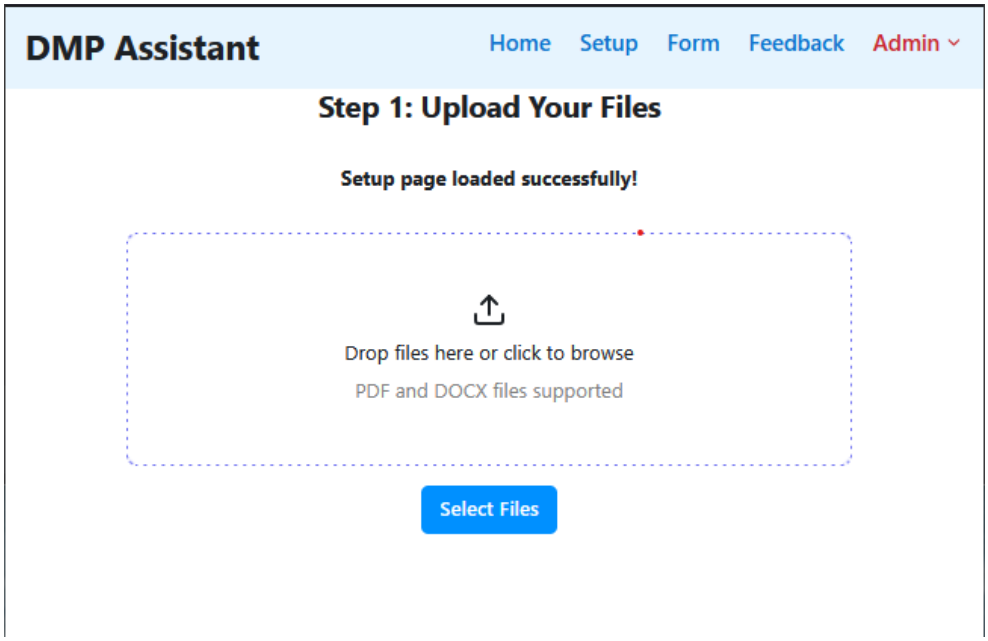


Figure 3.3: Interface for researchers to upload PDF and DOCX documents related to their research, such as project proposals, partially drafted papers, or previous DMPs

that capture contextual meaning. This allows the retrieval system to identify relevant content even when the query and the source use different wording, synonyms, or paraphrases. In the context of DMP assistance, such semantic matching ensures that institutional and funder guidelines, as well as research context extracted from project documents, can be retrieved based on conceptual similarity to a question rather than surface-level term overlap. This capability not only improves the precision of retrieved information but also reduces the risk of missing critical content due to variations in terminology, thereby addressing one of the core limitations of keyword-driven retrieval methods.

The RAG architecture uses a method called Maximum Marginal Relevance (MMR) to select the most relevant documents. MMR tries to balance two goals: finding documents that are relevant to the question and making sure the documents are diverse, so they do not repeat the same information. This way, the attention is spread across multiple sources. The parameters used with MMR are as follows: fetch_k equals 30, which is the number of candidate documents initially retrieved; k equals 12, which means it returns 12 documents after ranking. These parameters balance coverage and computational efficiency, retrieving too few candidates risks missing relevant content, while too many increases processing time without a significant gain. This approach is particularly important because LangChain splits each page of a PDF into a separate document, meaning that multiple documents in the storage can actually originate from the same original document, thus potentially containing a lot of the same information. And from my own experience a FWO proposal is 12 pages, so add some more documents such as old DMPs or other research descriptions that are available at the beginning of a research, they probably fall under 30 and are then all considered based on diversity and relevancy, and twelve was more of a random selection, because a proposal is 12 pages and contain all data of research very summarized and interlinked. Furthermore, lambda is set to 0.75, which means relevance and diversity are weighted equally. This setup helps AIDD4DMP pull information from multiple documents instead of just one, leading to more comprehensive answers based on a broader set of sources. This approach is particularly important because LangChain splits each

page of a PDF into a separate document, meaning that multiple documents in the storage can actually originate from the same original document, thus potentially containing a lot of the same information. Additionally, in real use cases, there may be multiple documents of different types, such as research proposals, previous DMPs, or (partial) publications.

The RAG architecture uses Maximum Marginal Relevance (MMR) to select the most relevant documents. MMR balances two goals: identifying documents that are highly relevant to the query while ensuring diversity, so that similar information is avoided and attention is spread across multiple sources. In this setup, the parameters are defined as fetch_k = 30, which represents the number of candidate documents initially retrieved, and $k = 12$, the number of documents returned after ranking. The parameter λ is set to 0.75, giving equal weight to relevance and diversity.

These values balance coverage and computational efficiency: retrieving too few candidates risks missing relevant content, while retrieving too many increases processing time without a substantial gain. This is particularly important in the context of DMPs, because LangChain splits each page of a PDF into a separate document, so multiple stored documents may originate from the same source. In practice, $k = 12$ is largely a heuristic choice, reflecting the typical length of an FWO research proposal (12 pages), which usually summarizes and interlinks all relevant research data. By considering additional documents, such as prior DMPs or other research descriptions available at the start of a project, up to fetch_k = 30 documents/pages are evaluated for relevance and diversity, with $k = 12$ selected. This ensures that AIDD4DMP pulls information from multiple sources, producing comprehensive answers that integrate knowledge from different document types, including proposals, previous DMPs, and partial publications.

3.4.3 DMP form environment

AIDD4DMP's external design begins with the user interface built with the Reflex tool-kit. The interface has two perspectives, one for the researchers and one for the RDM team. For both perspectives, a component factory pattern [52] ensures consistent generation of a card-based layout, where each question gets its own card customized to its type: single answers (Figure 3.4), multi-part responses (Figure 3.5), and table-based inputs(Figure 3.6). This consistency is reinforced by centralized state synchronization, which ensures that changes made in one part of the application are immediately reflected in the backend, supporting dynamic adaptations within the current session. For example, if a researcher modifies an answer and then returns to the RAG setup phase to add or update files, the previously entered answers remain intact. Additionally, each component includes "blur event" handling that automatically saves progress to a local data repository for subsequent sessions, preserving work without requiring explicit save actions and minimizing the risk of lost input.

Besides the AI-driven support tools found in Subsection 3.4.4, as shown in Figure 3.5, the interface includes information icons to display institute-specific guidelines and instructions (shown in cursive) for each question. This integrated information helps researchers evaluate their own answers. When additional support is needed, researchers can fall back on the human expertise of the RDM team through the flagging functionality, visible in Figure 3.9 to 3.11. A flag can be created with an accompanying comment, allowing researchers to ask a question, provide feedback, or highlight a specific issue. These flagged items generate a notification on the RDM team's dashboard, ensuring timely follow-up. Researchers can also flag without adding a com-

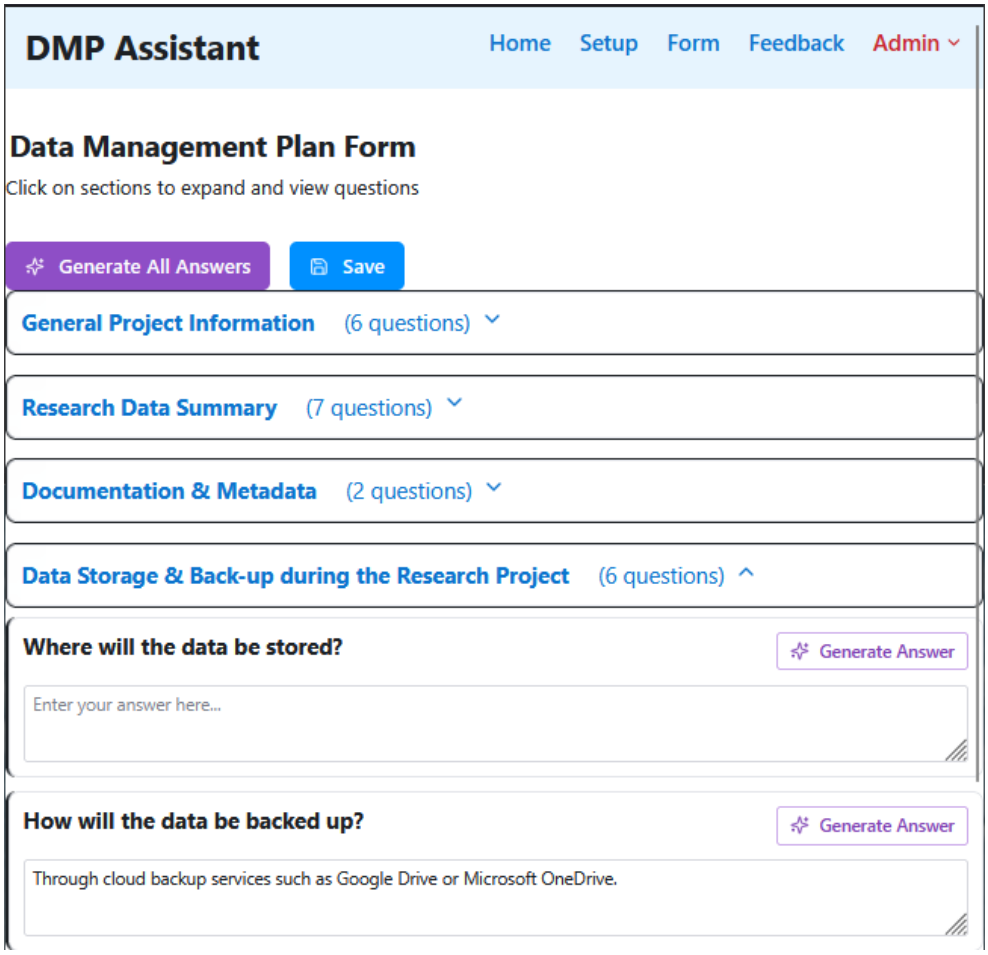


Figure 3.4: Base view during initial completion with the section about data storage and back-up expanded, and the simple one-answer questions on display, and the generate answer button on the right

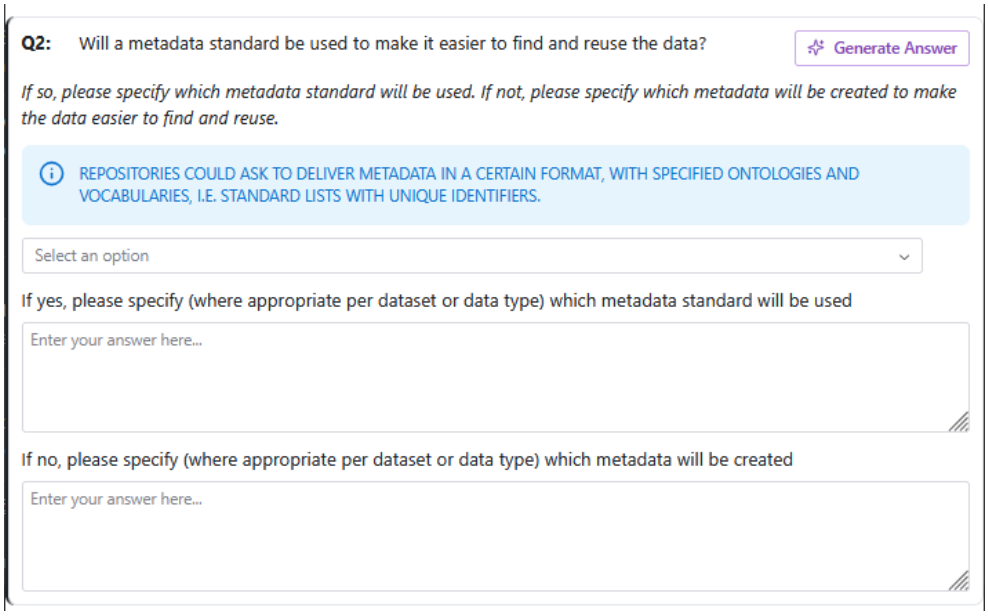


Figure 3.5: Card view of a question within the initial completion phase, featuring a multi-part answer comprising one selection response and two open-ended responses

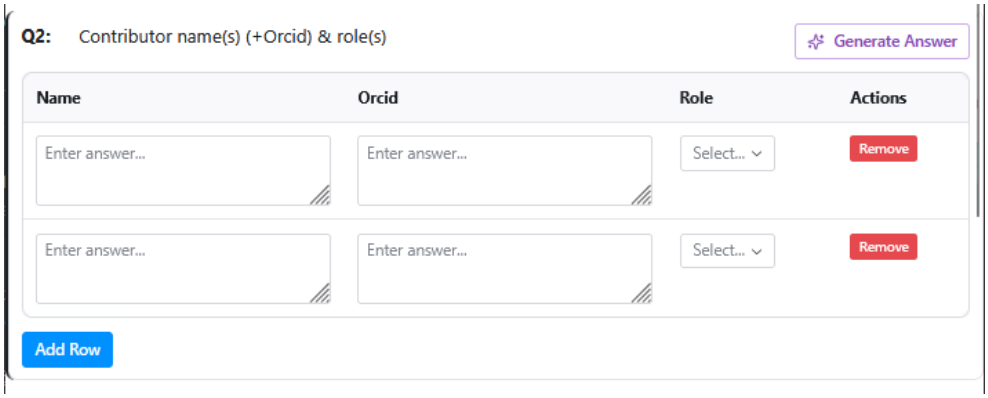


Figure 3.6: Card view of a question within the initial completion phase, featuring a table-based input featuring two open-answer columns and one selection-based column. Each row has a "delete row" button, and each table has an "add row" button.

ment, which does not trigger a notification but still marks the question for later review; these silent flags remain visible to RDM staff when they open the DMP.

The RDM side consists of two main views: a dashboard and an environment that mirrors the researcher's view, but with extended capabilities. The dashboard provides an overview of all DMPs under management. The dashboard is shown in Figure 3.7. In the individual form view (Figure 3.8), they have both read and write access to the answer and feedback fields, enabling them to overwrite either the researcher's answer or the AI-generated answer when necessary. For every question, they can view the full history of changes, including past answers, the feedback provided, and the ratings that feedback received. This shared, transparent workspace ensures both parties can collaborate effectively while maintaining a clear record of how each answer evolved over time.

3.4.4 Human-in-the-loop feedback mechanisms

Building on this foundation, AIDD4DMP incorporates a feedback loop designed for continuous improvement of both the DMP and the underlying AI models. As shown in Figure 3.9, researchers can toggle the feedback section for context-dependent guidance on each question. The "Regenerate" button allows them to request more guidance until it meets their needs before proceeding. If they want to rate the AI-generated feedback or highlight concerns, researchers can provide their own feedback. Two distinct approaches were developed, each based on different underlying theories, along with a third hybrid solution that combines elements of both.

The first approach uses a regular star-based rating system (Figure 3.10, where researchers assign between one and five stars to the feedback. Low ratings of one or two stars automatically trigger the regeneration of feedback, with the added instruction to the system that the previous feedback was unsatisfactory. Higher ratings (three to five stars) do not alter the feedback for the researcher in real time but are still recorded for later analysis. This approach intentionally makes little effort for the user and leverages a well-known visual metaphor for expressing satisfaction. All ratings, along with the corresponding question, current answer, and the rated feedback, are stored in an event logging repository. By extracting these ratings as logs, without retaining the full DMP or underlying research information, the tool produces training data that reflects what is considered good or bad feedback in specific contexts, maintaining privacy while supporting potential future

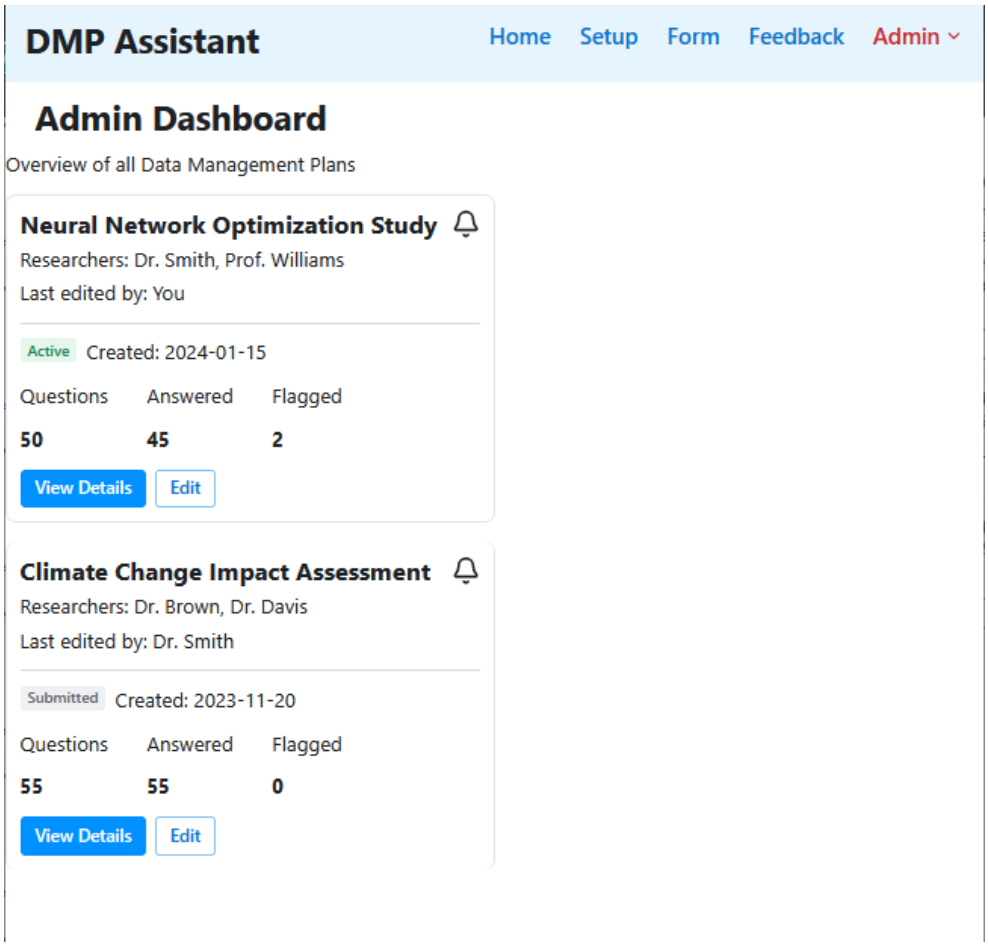


Figure 3.7: RDM team dashboard providing an overview of all DMPs, showing metrics such as flagged questions, unanswered questions, and overall progress for efficient monitoring and management.

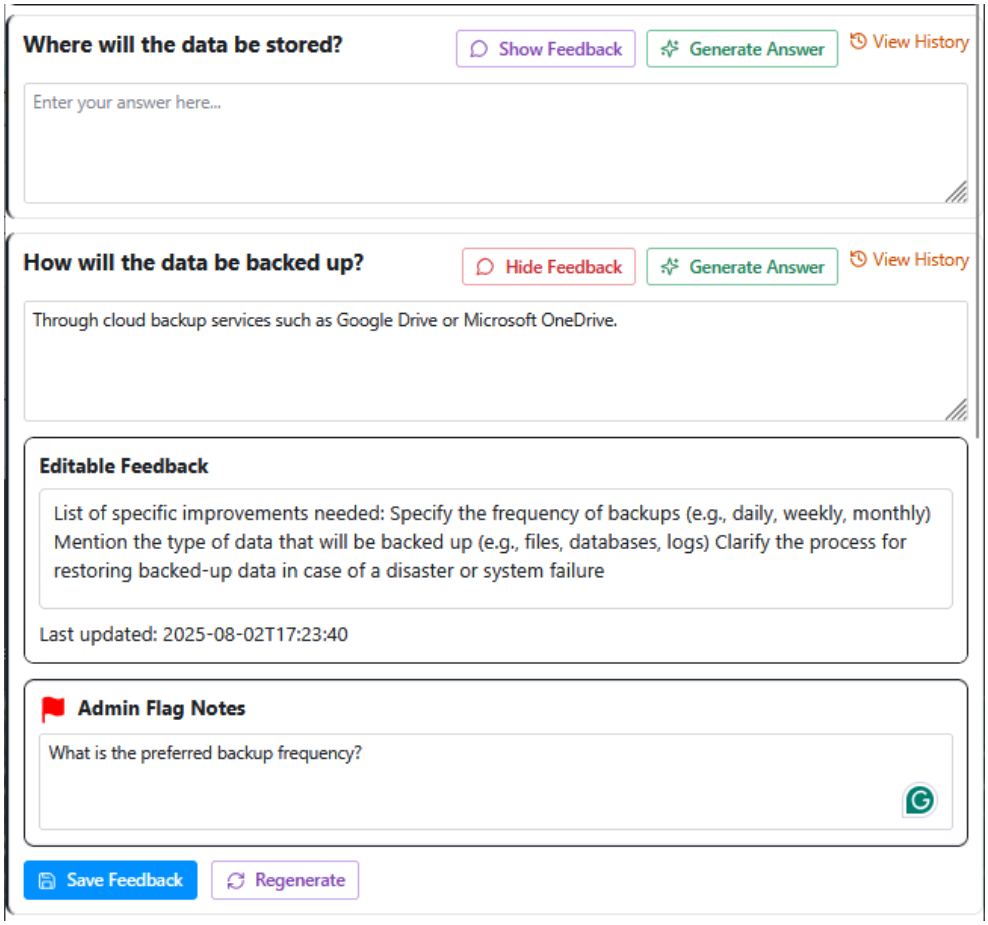


Figure 3.8: Feedback loop view from the RDM perspective, showing the button that leads to the full history of answers, feedback, and ratings for each question, with the ability to edit responses or provide additional guidance.

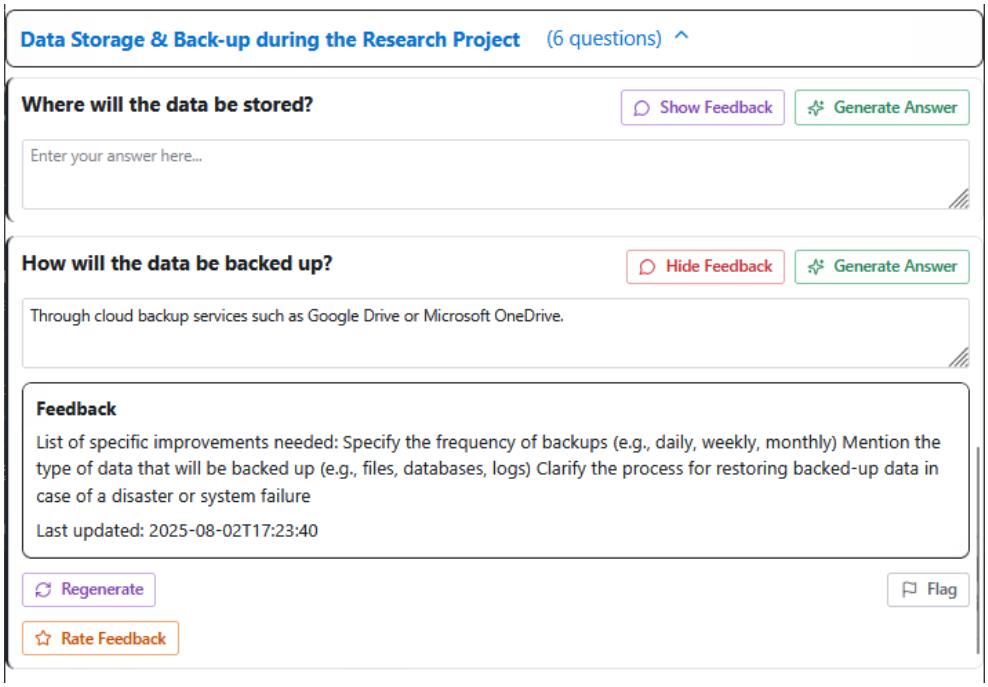


Figure 3.9: View during the feedback loop phase, showing two question cards. Each card includes a "Generate Answer" button, consistent with the initial fill phase, and a "Toggle Feedback View" button. On the second card, the feedback view is enabled, revealing the feedback field along with a "Regenerate" button (to regenerate the AI feedback) and a "Rate Feedback" button. The "Flag Question" button is also visible at the bottom left, to leave comments on the question.

LLM fine-tuning.

The second approach addresses the limitation of the first, namely its lack of precise steering, by introducing a state machine–driven conversational chatbot (shown in Figure 3.11) that orchestrates the feedback process rather than directly generating responses through AI. The state machine identifies specific keywords in user comments and responds accordingly, asking clarifying questions (e.g., extra details regarding inaccuracies). By framing feedback as a natural conversation rather than a static summary, the chatbot lowers the barrier for researchers to provide useful input. Once enough information is gathered, which is capped by a set conversation length to avoid excessive delay, AIDD4DMP compiles the data into a structured prompt, sends it to the LLM, and logs the exchange in a repository. While this approach is more demanding for the user, it has the potential to produce more targeted improvements.

The hybrid solution combines the star rating with a layered approach that adjusts the level of detail based on the user's rating. Positive responses of five stars are simply logged, while lower ratings trigger specialized workflows. For ratings of one or two stars, the system initiates chat-based feedback collection to gather as much detail as possible. A rating of three stars opens an annotation interface that allows users to highlight specific good or problematic text segments (shown in Figure 3.12), and a four-star rating presents a simpler comment section for targeted improvement suggestions. This layered structure ensures that feedback remains actionable while keeping the process lightweight for the user.

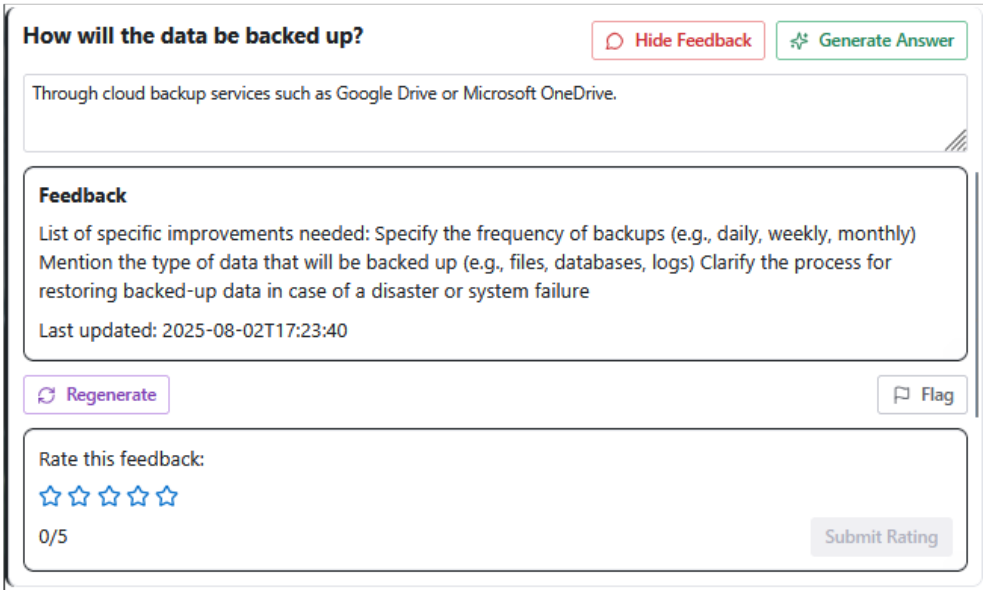


Figure 3.10: One of the DMP's question cards implementing the star-based feedback system, researchers rate feedback from one to five stars, triggering automatic feedback regeneration for low ratings and recording all responses for future analysis.

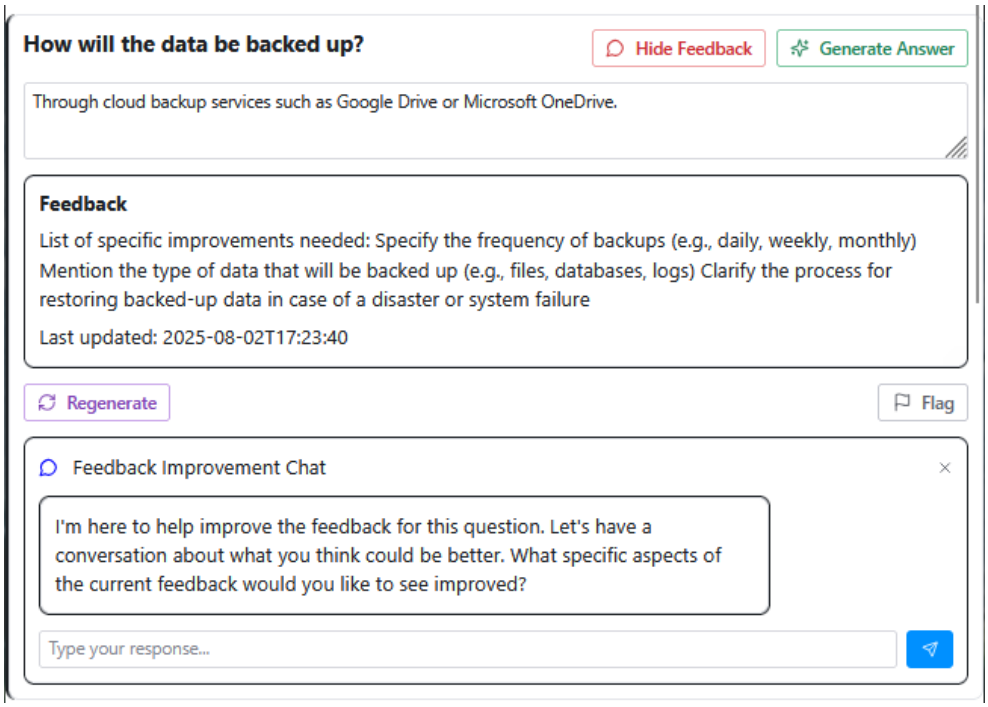


Figure 3.11: Example of a DMP question card with the state machine–driven conversational chatbot for feedback. The chatbot guides the researcher through clarifying questions, collects input in a structured manner, and compiles it into a prompt for the LLM while logging the interaction for later analysis.

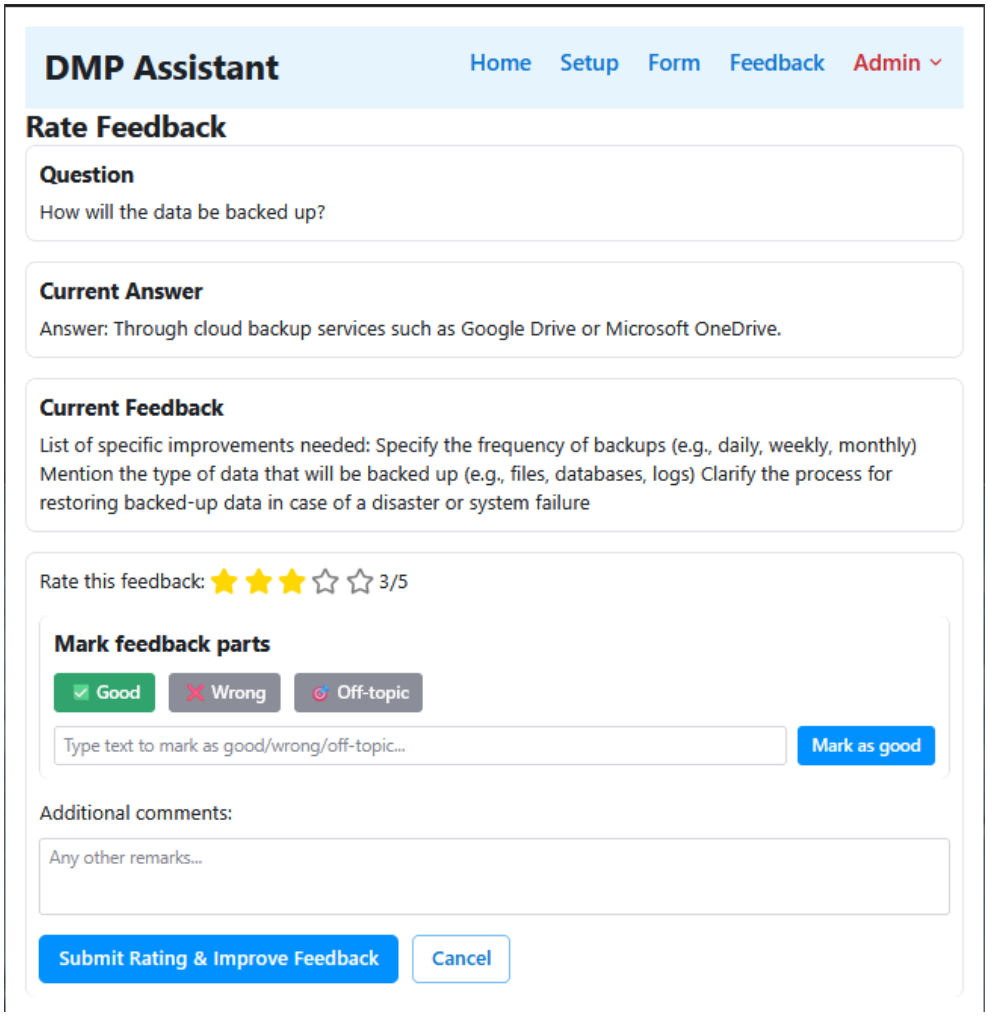


Figure 3.12: Three-star feedback interface in the hybrid system, showing the annotation view where users can highlight specific good or problematic text segments for targeted improvement.

Chapter 4

Evaluation of AIDD4DMP: Findings and Discussion

4.1 Multi-dimensional Assessment Methodology

The evaluation phase of the IDEE framework employs a comprehensive mixed-methods approach designed to assess technical effectiveness and practical utility within real-world workflows. In this thesis, the researcher-facing side of the system was evaluated through a user study involving active researchers. A cognitive walkthrough with an RDM team member was conducted to assess the overall functionality of the system, including both researcher and RDM perspectives.

4.1.1 User Study with Researchers: Study Design and Protocol

The user study employed a between-subjects experimental design to evaluate various aspects of the AIDD4DMP framework. Eleven participants were recruited from the target user population of academic researchers with prior experience completing DMPs, ensuring that feedback reflected realistic usage scenarios and informed user expectations. Five participants used a version of the tool interfacing with Llama3 (Llama3), while the remaining six used a version based on Gemma3n (Gemma3n).

The experimental protocol simulated authentic DMP completion scenarios, enabling systematic evaluation of different system components. Participants completed structured tasks including context extraction, AI-assisted question answering, feedback evaluation, and testing of multiple interaction modalities. User perceptions were assessed at multiple points using the User Experience Questionnaire-Short (UEQ-S), chosen over the full UEQ+ due to its focused evaluation of pragmatic and hedonic quality dimensions. This provided sufficient detail while minimizing participant burden during four repeated administrations.

To maintain ecological validity, all participants worked within the standardized context of a pseudo-research project investigating "the impact of social media on the mental health of adolescents aged 13 to 17, conducted through surveys and data collection over a 12-month period." This scenario ensured consistent evaluation conditions while offering sufficient complexity to test the framework across various DMP sections and question types.

The structured evaluation protocol comprised seven distinct phases, each targeting specific as-

pects of the framework.

1. **Upload Phase:** Participants uploaded research documents and were asked about their comfort with local versus cloud-based document processing. They were also asked if they used AI tools during their latest DMP completion.

2. **Initial Fill Phase:** Participants manually responded to questions in Section 4 of the DMP, "Data Storage and Backup During the Research Project." Specifically, they addressed Question 1, "Where will data be stored?" and Question 2, "Where will data be backed up?" This approach allowed participants to use their own reasoning, enabling them to more effectively evaluate the feedback they received later.

3. **Generation Testing Phase:** In Section 2 of the DMP: "Research Data Summary," participants had to describe what the "Generate Answer" button would do and use it for Question 3, which is related to ethical issues. After generating the answer, they were asked to review and adapt the AI-generated response as needed. This phase evaluated user expectations for AI-generated answers, satisfaction with the quality, and the need to modify the response.

4. **Feedback Evaluation Phase:** Participants navigated to the Feedback page and used the "Regenerate" to generate feedback for the two questions they had answered manually. They then reviewed the generated feedback for each question. Protocol questions focused on their expectations for the feedback system, the quality of the feedback, its relevance, and how actionable and helpful it was for improving their responses.

5. **Rating and Feedback Mechanisms Phase:** Three feedback improvement approaches were tested: a simple star rating system, a chatbot-based suggestion system, and a hybrid solution. After experiencing each approach, participants completed a UEQ-S to evaluate that specific component. Furthermore, for the star rating, they were asked about its perceived purpose; for the chatbot, whether they considered it an appropriate approach for each situation; and for the hybrid system, they were invited to provide additional comments.

6. **Human-in-the-Loop Support Phase:** Participants evaluated the flagging system by identifying a piece of poor or problematic feedback and using the flag functionality to report it. The protocol questions addressed the visibility of the flag feature, the reason for flagging, and the intuitiveness of the process, as well as whether they preferred this system over the existing comment, message, or email-based approaches.

7. **Integrated Policy Review Phase:** Finally, participants were introduced to the info icons and instructions shown in Figure 3.5, if they had not seen them previously. They were asked what types of information they would like to see integrated into the tool, such as guidelines, instructions, templates, or other relevant resources.

Data collection combined quantitative and qualitative methods. The UEQ-S was administered at four key evaluation points: post-star rating, post-chatbot interaction, post-hybrid feedback experience, and after full system evaluation. Semi-structured interviews consisted of the questions asked during the entire protocol; these captured participant reflections, expectations, and reactions in real time. Behavioral observations and interaction logging complemented self-reported data, enabling analysis of actual usage patterns alongside subjective feedback. The UEQ-S

methodology was particularly suited for this academic tool evaluation due to its focus on core pragmatic dimensions, "efficiency" and "usability" and hedonic dimensions, "stimulation", "novelty", and "attractiveness", while avoiding less relevant scales.

4.1.2 Cognitive Walkthrough with Data Steward: Methodology

The cognitive walkthrough was designed to evaluate AIDD4DMP from the perspective of RDM administrators responsible for oversight and ongoing support. The session aimed to simulate real-world use by examining both the researcher-facing interface and the administrative dashboard within a single evaluation. This structure allowed the administrator to first experience the tool as researchers would during DMP preparation, and then transition into the RDM role to assess how the administrative dashboard could be used to review, guide, and manage those plans.

The walkthrough followed a structured presentation to ensure that every functionality received attention, while also incorporating open conversation to enable both the steward and me to ask questions and exchange comments. The first part of the session provided a theoretical overview of the framework's goals, while the second part focused on hands-on exploration of the prototype.

The evaluation plan specifically targeted:

1. **Usability**: navigation, clarity of controls, and ease of accessing relevant information;
2. **Integration potential**: how the system could complement current review processes, communication channels, and institutional policies;
3. **Feedback loop mechanisms**: ways in which feedback and ratings could be used to maintain and improve DMP quality;
4. **Version control**: identifying which events are feasible to log for both researcher and LLM assessment.

4.2 User Study with Researchers: Findings

This section presents the findings from the user study, which evaluated the researcher side of the DMP support tool through a multi-phase protocol.

4.2.1 User Demographics, Acceptance and Trust

The eleven participants ranged in age from 24 to 45 years old, with the majority (8 out of 11) in the range of 20 to 30 years old. Most participants had completed a Master's in Computer Science (8 participants), while one held a Master's in Software Engineering Technology, and two completed a PhD in Computer Science. The recency of their last DMP completion varied, with dates ranging from September 2022 to June 2025. This indicates that all users have experience with DMPonline, offering a relevant demographic for evaluating the support tool.

Regarding prior engagement with AI tools, the participant group showed a moderate level of experience. Seven out of eleven participants had previously utilized AI assistance for Data Management Plan (DMP) creation. The nature of this usage varied considerably, ranging from comprehensive DMP development to more specific tasks, such as data cleaning and answers that were based on content from previous DMPs. In contrast, four of the participants reported no

prior AI experience in this domain. Three of these four individuals remarked that AI tools were "not yet booming" at the time of their last DMP creation, which suggests a rapid acceleration in the adoption of AI for research support tasks.

Participants in the study exhibited a cautious approach to data handling, particularly concerning security and privacy. A significant majority, specifically eight of the eleven participants, indicated conditional acceptance of internally hosted support systems that needed access to their research documents. This acceptance was contingent upon explicit security guarantees, such as data encryption and assurances of institutional control over data storage. This observation underscores a general concern regarding data privacy within the research community. Two participants explicitly articulated reservations about AI systems potentially releasing sensitive research data.

4.2.2 System Functionality and Performance Evaluation

This user study had five participants who used the tool that received output from Llama3, which incorporated a significant amount of descriptive and narrative text, and the other six participants had a tool based on Gemma3n, which delivered the answers and feedback more directly and concisely. It revealed notable differences in performance. Gemma3n consistently demonstrated superior accuracy; five out of six participants said the generated answer was correct and complete. In contrast, Llama3 received correct ratings from only two participants.

User expectations for the "Generate Answer" feature were diverse. Approximately six participants anticipated complete responses based on the provided documents, while two others expected template-based suggestions. Others anticipated generic suggestions derived from institutional and funder policies. Llama3 frequently failed to meet these expectations due to its verbose "flavor text", as one participant described, which often obscured the core content. One of its responses to the question about how data will be backed up was:

> Based on the provided context, I found no specific information regarding data backup procedures in the uploaded documents [Title.docx]. However, I can suggest general best practices for data storage and backup during the research project:
>
> - Store digital data on a secure server or cloud storage service that is regularly backed up.
> - Use an external hard drive or USB drive to store non-digital data (e.g., journal entries) and back it up regularly.
> - Implement a version control system for any collaborative files or documents.
> - Refer to institution-specific policies regarding backup procedures when appropriate, such as guidelines on backup frequency, acceptable storage media, and assigned responsibilities.
>
> Please note that this answer does not provide specific details about the data backup procedures used in the study, as this information was not available in the provided context.

Llama3 also received negative comments regarding several recurring issues, including problematic Markdown formatting, repeating the question within its generated responses, use of the third-person perspective (e.g., "the researcher should ..."), and lack of fact-checking (e.g., whether

1417 year-olds are considered adolescents).

Gemma3n, however, aligned more closely with user preferences by providing concise and direct responses. Despite its better performance, Gemma3n still received feedback for omitting institution-specific details, such as references to local ethics committees (SMEC) and storage guidelines (password-protected shared Google Drive), indicating a need for greater contextual awareness in generated content.

4.2.3 Interaction Preferences

Besides the preferred wording style from the LLM models, the interviews provided insight into preferences for communication with RDM staff, and AI models became evident. Feedback designed for researchers should identify errors in the answer coupled with actionable instructions, as opposed to what is correct about the answer, generic suggestions, or unsolicited examples. Returning feedback to the AI is more discussed in Subsection 4.2.4, but several participants reported difficulties in accurately evaluating the quality of feedback when they lacked sufficient domain expertise to assess correctness. Furthermore, participants indicated that integrated guidelines through info icons can improve clarity and reduce the amount of back-and-forth communication needed.

Participant preferences for communication channels to RDM clearly favored contextual tools, with seven researchers opting for the flagging system over traditional email communication. The advantages cited for flagging included its direct contextual relevance, reduced ambiguity regarding the location of specific issues, and lower effort requirements compared to drafting structured emails. The minority of two researchers preferred email valued the personal nature of direct communication, with one participant even mentioning a preference for telephone contact in complex scenarios. The other two did not disclose a preference. However, the dual flagging system, which offered options for flagging with or without comments, received universal positive reception from all participants. Participants, however, emphasized the necessity for more comprehensive documentation to clarify the functional distinctions between the different flag types. The inclusion of an 'unflag' option was specifically highlighted by two participants as a feature that lowered the psychological barrier to using the system, suggesting that features allowing for reversibility enhance user confidence in interactive tools.

4.2.4 Feedback Preferences

When looking at the technologies that provide feedback to the AI model, the star rating system demonstrated high conceptual clarity among users. A majority of nine participants correctly understood its primary purpose as contributing to long-term model improvement rather than merely regenerating immediate responses. This understanding suggests that participants generally possessed accurate mental models regarding the system's underlying learning mechanisms. Nevertheless, some ambiguity persisted concerning whether these ratings served personal, institutional, or broader research community objectives (a participant mentioned that this is unnecessary for the final product and will only have a use case in early stages).

The effectiveness of the chatbot varied significantly between the two model implementations. Llama3 exhibited greater transparency by explicitly detailing changes made during interactions, although this often led to information overload and excessively verbose responses. Gemma3n,

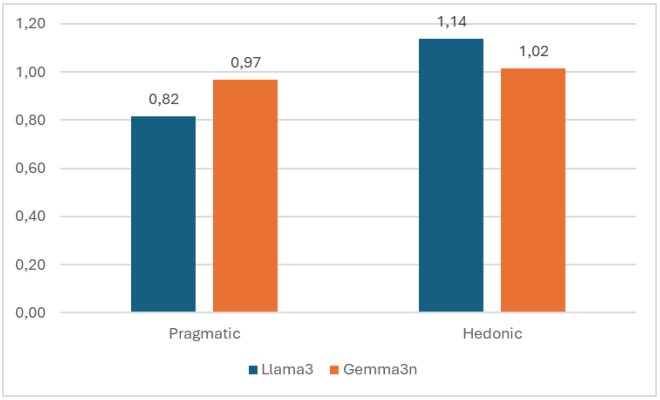


Figure 4.1: Comparison of average pragmatic and hedonic UEQ-S scores for AI-driven support tool, with Llama3 and Gemma3n as LLM. Since pragmatic scores rise above 0.8, both models were perceived as useful and functional, and hedonic scores are above 1.0, reflecting an even more positive perception of the tool's engaging and innovative aspects.

while offering a more balanced feedback style, occasionally provided vague responses and demonstrated inconsistency between different interaction rounds. In the end, users identified the most effective use cases for the chatbot as processing larger blocks of text, facilitating iterative refinement through extended dialogue.

The hybrid solution was perceived as the most effective out of the three solutions, which aligns with feedback patterns seen on popular consumer platforms. For example, participants compared it to Airbnb and Amazon, where they noted an inverse correlation between satisfaction levels and the detail provided in feedback: highly satisfied users tended to offer minimal additional information, whereas dissatisfied users provided extensive explanatory comments. This was one of the theories in mind when developing this solution. A particularly notable comment came right after using the chatbot: one participant suggested a layered approach in which more detail is requested as the rating decreases, similar to the implementation tested here. In contrast, another participant cautioned that altering the feedback method based on the star rating might influence the rating itself, as users could deliberately adjust their score to access a preferred approach.

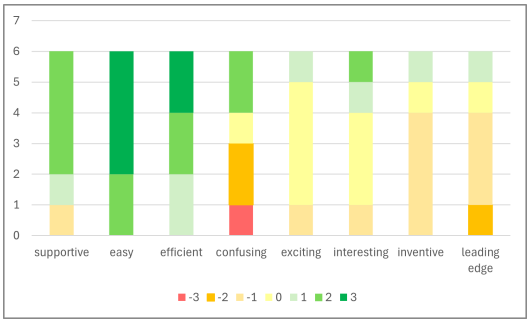
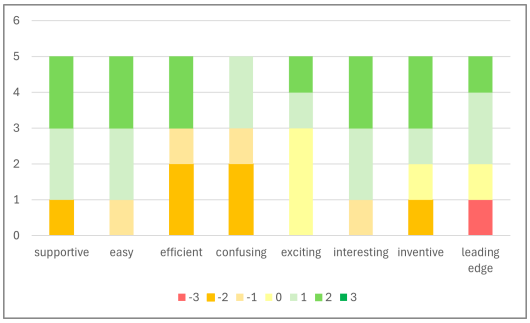
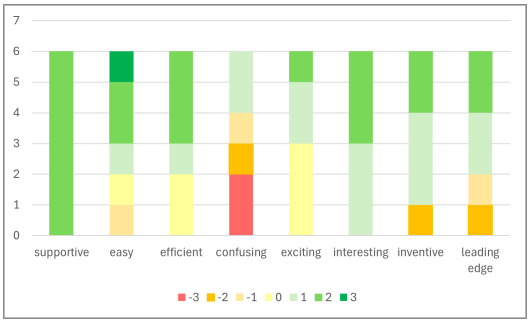
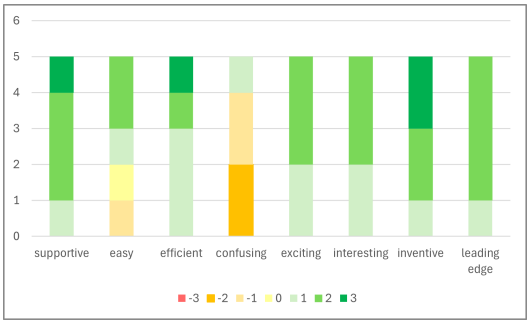
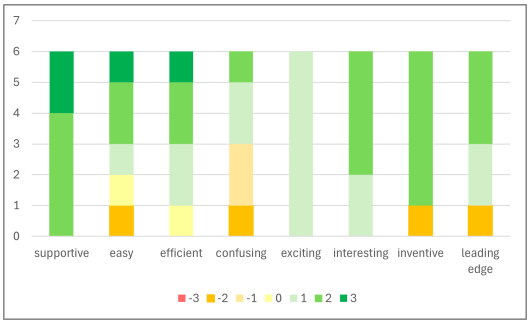
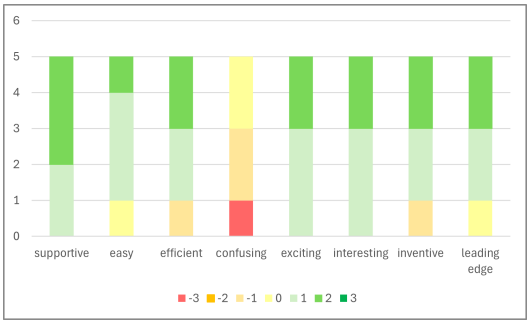
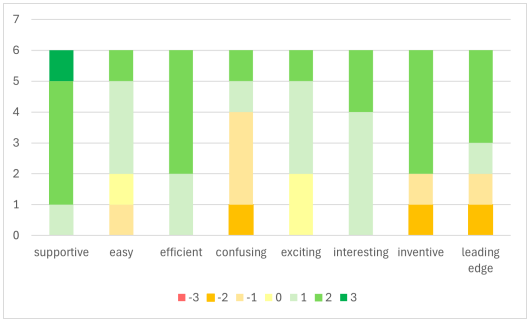
4.2.5 Quantitative Performance Metrics

The UEQ-S results indicate overall user satisfaction, as seen in Figure 4.1: Llama3 scored an average pragmatic score of 0.82 and a hedonic score of 1.14, while Gemma3n scored 0.97 and 1.02. Because both models scored above 0.8 on the pragmatic scale, AIDD4DMP was rated useful and functional. With the hedonic scores above 1.0, the ratings indicate a distinctly positive perception of the tool's innovative and engaging aspects.

Further analysis of the solution–model combinations, as summarized in Table 4.1, reveals distinct performance patterns:

- **Solution A - Star Rating:** Llama3's version had mixed results. While it performed well in *easy* and *confusing*, it rated poorly in *interesting*, *inventive*, and *leading-edge* dimensions, indicating a perception of being conventional. Gemma3n significantly outperformed Llama3 in nearly all categories, particularly in *supportive*, *efficient*, *interesting*, and *inventive* dimensions, although *leading-edge* and *inventive* scores remained moderate.

Table 4.1: Score Distributions for UEQ-S Across Models, Feedback Approaches, and Overall Functionality.

	Llama3	Gemma3n
Star Rating		
Chatbot		
Hybrid Solution		
Overall		

The categories on the x-axis are: *supportive, easy, efficient, confusing, exciting, interesting, inventive,* and *leading edge.* The y-axis represents the frequency of responses. Note that the *confusing* dimension is inversely scaled: red indicates lower confusion and is therefore more positive, while green indicates higher confusion and is less desirable.

- **Solution B - Chatbot:** Gemma3n received predominantly positive ratings across *supportive*, *easy*, *efficient*, *exciting*, and *interesting* dimensions, with excellent clarity in the *confusing* category. However, *inventive* and *leading-edge* ratings were lower. Llama3's chatbot resulted in lower *efficiency* and higher *confusion*, alongside weaker scores in *supportive* and *leading-edge* dimensions.
- **Hybrid Solution – Layered Feedback:** For Llama3, the hybrid solution received predominantly positive ratings across all dimensions, particularly in *supportive*, *efficient*, *exciting*, *interesting*, *inventive*, and *leading-edge.* The *confusing* category shows mostly moderate clarity. For Gemma3n, performance was less uniformly positive, with *confusing* slightly higher than optimal, but still strong clarity. This solution stands out for balancing strong functional performance with broad appeal.
- **Overall Functionality:** In the Llama3 implementation, Overall Functionality received mainly positive ratings for both pragmatic dimensions (first four) as well as hedonic (last four). Notably, the *confusing* category received the best possible clarity score. Gemma3n also achieved strong results with predominantly positive ratings in *supportive*, *easy*, *efficient*, *exciting*, and *interesting* dimensions. However, *confusing* for Gemma3n indicates moderate clarity rather than the exceptional clarity seen in Llama3. Both models show solid innovation and engagement levels in this solution type.

We further examined pragmatic and hedonic scores relative to the time since participants' last DMP completion (see Annex A for all graphs). Overall, a clear temporal trend emerges across models and tools: participants who had completed a DMP recently tended to rate the tools less positively. Across all feedback approaches, Gemma3n consistently outperformed Llama3 in perceived usefulness and engagement, except for the star rating.

The star rating tool exhibited an overall decline in both pragmatic and hedonic scores as DMP recency increased for both models, though Llama3 generally maintained a higher pragmatic rating than Gemma3n. For the chatbot interface, Llama3's pragmatic and hedonic scores declined sharply among participants with recent DMP experience, whereas Gemma3n's scores increased steadily, suggesting that Gemma3n better supports both novice and experienced users. In the hybrid feedback solution, Gemma3n again shows rising hedonic ratings over time, reflecting strong appreciation for its adaptive feedback, while Llama3 scores decline for more recent completions. Notably, participants who completed their DMP prior to 2023 (around the time AI in the form of ChatGPT became a hype, as mentioned by the participants) rated the hybrid solution particularly highly for its flexibility. These patterns indicate that the perceived effectiveness and enjoyment of the tools are influenced by both the underlying model and the user's familiarity with DMP processes, with Gemma3n providing a more consistently positive experience.

4.3 Cognitive Walkthrough With Data Steward: Findings

A cognitive walkthrough session was conducted with an experienced RDM data steward to evaluate expert usability, review workflows, and identify gaps between the AI-driven support tool and professional review. The following key findings were identified:

1. Preference for shorter, more direct answers, with the suggestion that the tool automatically

trim verbose or redundant content provided by researchers.

2. Recognition that providing per-answer feedback and suggestions is both cost-efficient and effective, as it reduces the need for comprehensive full-document reviews; proximity of feedback to the relevant answer improves clarity and usability.
3. Emphasis on the value of real-time feedback, contrasting with traditional RDM review processes that often suffer from lengthy turnaround times.
4. Recommendation to include a glossary of terms, external references, and interactive help fields, such as a chatbot, where researchers could ask questions to enhance understanding.
5. Emphasis on the need for version control, including the ability to track changes and maintain answer history.
6. Reconition of the importance of an initial completion phase allowing users to fill out the entire DMP before receiving improvement suggestions, thereby saving time; for example, a user may skip refining a response if the following question requests further detail.
7. Interest in implementing a dynamic, responsive information panel displaying institution-specific guidelines.

4.4 Discussion

This section synthesizes the findings from the user study and cognitive walkthrough to examine the relationships between different results, identify underlying patterns, and provide theoretical rationales for the observations across model performance, interface design, and user experience factors.

4.4.1 Model Performance and Interaction Methods

The comparative analysis between Llama3 and Gemma3n reveals a fundamental trade-off between transparency and efficiency in AI-assisted research support tools. Gemma3n's superior accuracy and consistently higher pragmatic scores for the chatbot, hybrid approach, and overall functionality can be attributed to its concise, direct response style that aligns with users' preference for actionable feedback over explanatory narrative. This finding challenges the common assumption that more detailed explanations inherently improve user satisfaction.

However, the relationship between model performance and user experience proves more nuanced when examining specific contexts. Llama3's inclusion of narrative feedback, while sometimes perceived as verbose flavor text, demonstrated superior clarity in certain applications, particularly within the Overall Functionality solutions. This suggests that the optimal balance between transparency and conciseness is context-dependent: simple feedback benefits from concise instructions, and having the extra explanation after returning feedback helps understand the model's rationale.

The consistently strong performance of the hybrid layered feedback solution across both models indicates that user control and adaptability are critical factors in tool acceptance. Users with higher satisfaction provide minimal feedback, while those requiring more support can access detailed explanatory tools, creating a self-regulating system that scales complexity to need. This

approach's positive ratings can be attributed to its accommodation of diverse user preferences and expertise levels. Meanwhile, Llama3 provided more text to analyze and respond to, which could make giving feedback easier since there was more content to agree or disagree with. This sometimes led to slightly higher perceived usability, whereas Gemma3n achieved similar usability through the quality of its answers.

4.4.2 User Experience Trends

The observed temporal trends reveal patterns in user adaptation with potential implications for tool design. Users favored the direct approach of Gemma3n, whose concise feedback reduced the need to rely on the rating tool. In terms of hedonic experience, Gemma3n felt more stimulating to recent DMP completers due to its straightforward style that resembles current AI tools. Meanwhile, those who completed their DMP longer ago appreciated Llama3's narrative feedback more, likely because they recall past frustrations such as scattered information and long feedback turnaround times.

The divergent performance patterns between models can be attributed to differences in feedback style. Llama3's narrative approach, while providing transparency, often included redundant information, markdown formatting issues, and third-person perspective usage. This may explain why users with recent DMP experience — and presumably greater familiarity with streamlined AI tools — showed decreasing satisfaction with Llama3 over time. Gemma3n's consistent ratings across user groups suggest its approach effectively delivers essential information without unnecessary elaboration.

4.4.3 User Experience Trends

The observed temporal trends reveal patterns in user adaptation with potential implications for tool design. Users favored the direct approach of Gemma3n, whose concise feedback reduced the need to rely on the rating tool. In terms of hedonic experience, Gemma3n felt more stimulating to recent DMP completers due to its straightforward style that resembles current AI tools. Meanwhile, those who completed their DMP longer ago appreciated Llama3's narrative feedback more, likely because they recall past frustrations such as scattered information and long feedback turnaround times.

4.4.4 Security and Privacy

The finding that 73% of participants conditionally accepted internal hosting with explicit security guarantees reveals the complex relationship between functionality and trust in research environments. This cautious acceptance, coupled with the observation that data security emerged as one of the biggest deciders if researchers want to use this tool, indicates that technical capabilities alone are insufficient for adoption.

4.4.5 General Takeaways for Data Management Support

The evaluation with both researchers and a data steward revealed several guidelines for designing AI-assisted research support tools:

Feedback Architecture: Effective feedback should clearly identify errors and provide concise, actionable instructions. Examples and proposed solutions should be implemented as separate, optional functions rather than integrated into primary feedback streams, allowing users to access elaboration when needed without overwhelming the primary interaction.

Answer Generation Strategy: The tool should leverage existing resources to provide more context-aware answers, focusing not only on the specific research content but also on the relevant parties involved. This can be achieved by analyzing previous DMP answers to identify applicable approaches (e.g., data storage strategies), provided that researchers consent to upload their prior DMPs. Additionally, answers can draw from institutional and funder-specific guidelines or answer repositories. This approach minimizes user effort while maintaining personalization and ensuring compliance with institutional requirements.

Adaptive Interface Design: A layered feedback approach emerged as the most preferred solution, but users expect access to explanatory tools regardless of satisfaction level, suggesting that interface restrictions based on ratings may create bias.

Contextual Integration: Bringing all necessary information and tools within the primary interface context obviously outperforms multi-platform approaches. Users strongly preferred integrated solutions that eliminated ambiguity and reduced the need to navigate between different systems or communication platforms.

Security and Privacy as a Foundational Requirement: Data security and privacy considerations emerged as a primary determinant of tool adoption willingness. The conditional acceptance patterns observed indicate that security and privacy features must be prominently communicated and actively demonstrated rather than assumed or buried in documentation.

These findings collectively indicate that successful AI-assisted RDM tools must balance multiple competing demands: efficiency versus transparency, simplicity versus capability, and innovation versus security. The temporal patterns suggest that these balances can shift over time and with experience, requiring adaptive approaches that can accommodate evolving user expectations while maintaining the fundamental principles of clarity, control, and security.

4.5 Limitations and Future Work

This research offers valuable insights but also has several limitations that suggest directions for future research and development.

First, the participant group consisted of only eleven researchers and one data steward. While sufficient for qualitative insights and initial quantitative trends, this limited sample size reduces the generalizability of the findings. Future studies should include larger and more diverse participant pools to validate and expand upon these results, as well as longitudinal studies to monitor evolving user adoption patterns and perceptions over time. Including expert validation loops to review critical AI-generated content could formalize the human-in-the-loop process into a structured quality assurance mechanism, balancing AI efficiency with necessary expert oversight.

Second, the research was conducted within one particular research context and institutional setting (UHasselt). Consequently, some findings may be context-dependent and might not fully translate to different environments or alternative RDM frameworks. Future work should explore

the approach in other institutional contexts to test its broader applicability.

Another limitation concerns temporal analysis: this study focused on how user opinions varied with time since their last DMP completion, but did not compare perceptions between the existing platform and the newly developed AIDD4DMP. Future research should include baseline measurements of the current platform to better evaluate improvements.

Regarding AI performance, Llama3's narrative style reduced user satisfaction due to overly verbose textual outputs, whereas Gemma3n's concise approach was preferred. Future tool development should focus on enhancing AI precision through advanced fine-tuning of LLMs, leveraging well-documented institutional guidelines or template answers, potentially using researcher feedback by adding it to the data repositories to ensure AI responses remain relevant and accurate. Further system enhancement can come from using dynamic prompts that adapt to each DMP context, and applying post-processing to extract concise, actionable feedback (e.g., removing markdown formatting and introductory sentences). Prioritizing the direct, clear communication style exemplified by Gemma3n is expected to increase pragmatic utility. Furthermore, the chatbot interface shows promise; evolving it into a fully interactive LLM-driven system could enable conversational prompt engineering, potentially improving usability. Integration of Mixture of Experts (MoE) or Mixture of Attention (MoA) models could further enhance output quality, though this would require higher-performing infrastructure or acceptance of longer response times.

Finally, trust, privacy, and security emerged as crucial factors influencing adoption. Future work must strengthen transparency around data handling, AI training, and security and privacy protocols. Clear user documentation for features like the dual flagging and "unflag" options will further empower users and foster confidence.

Chapter 5

Conclusion

This thesis presents the design and implementation of the AIDD4DMP framework that supports researchers and RDM teams in their task to complete DMPs through human-in-the-loop AI. The framework is grounded in literature and an analysis of current DMP workflows, and rests on four core pillars: Assist, Integrate, Dialogue, and Develop. The Assist pillar focuses on providing AI-driven support to help researchers generate, refine, and validate DMP content more efficiently, reducing repetitive work while ensuring alignment with institutional and funding requirements. The Integrate pillar emphasizes seamless incorporation of the framework into existing research workflows and tools, enabling compatibility with institutional systems and minimizing disruption to established practices, while also bringing together all relevant information sources, such as institutional guidelines and research-related documents, into one place. The Dialogue pillar centers on creating dynamic, bidirectional communication channels between researchers, AI systems, and RDM teams, fostering more targeted feedback and collaborative problem-solving. Finally, the Develop pillar encompasses continuous improvement through user feedback, analytics, and adaptive model updates, ensuring that the framework evolves alongside changing research needs and technological capabilities.

The AIDD4DMP framework was designed around a modular architecture, a principle carried into the prototype to ensure adaptability and scalability. A user study with researchers compared different feedback approaches (star rating, chatbot, and hybrid layered feedback) across two LLMs (Llama3 and Gemma3n) to explore how interaction style and model characteristics influence user experience. Results showed that the hybrid solution, which adjusts the amount of detail in the feedback based on the rating, was perceived as the most usable and innovative. The star rating was valued for its clarity and simplicity, while the chatbot was considered more supportive and engaging, particularly for refining large, complex, or multidisciplinary text segments.

Across models, Gemma3n generally scored higher on pragmatic measures, offering concise, direct responses that reduced cognitive load. Llama3's more narrative style sometimes improved clarity, but could also lead to information overload, especially for users who had recently completed a DMP and could compare it to their previous experiences with tools such as ChatGPT. Quantitative UEQ-S scores confirmed strong functional performance for both models (pragmatic scores $> 0,8$), with hedonic scores indicating positive perceptions of the AIDD4DMP's novelty and engagement. The cognitive walkthrough with an RDM data steward further highlighted the importance of concise feedback, integration of institutional guidelines, and features such as version control and contextual help.

When evaluating the overall approach, researchers emphasized the value of combining AI efficiency with human oversight, adaptive feedback mechanisms, and seamless integration of institutional resources. In conclusion, these findings suggest that the AIDD4DMP framework can improve both efficiency and satisfaction in DMP creation, while providing a scalable, secure foundation for future adoption across research institutions.

In summary, future work should aim to validate the AIDD4DMP framework across larger and more diverse participant groups and institutional contexts, incorporate structured expert review loops to enhance the human-in-the-loop process, and include baseline comparisons with existing platforms to quantify improvements. From a technical perspective, continued refinement of AI models—through fine-tuning, dynamic prompts, post-processing, and exploration of MoE or MoA architectures—will be critical to provide concise, context-aware, and actionable feedback. Expanding the chatbot into a fully interactive conversational system offers additional opportunities for usability improvements. Finally, ensuring robust trust, privacy, and security mechanisms, alongside clear documentation and user empowerment features, will be essential to facilitate widespread adoption and sustainable use of AI-supported DMP tools.

Bibliography

[1] M. H. Burnette, S. C. Williams, and H. J. Imker, "From Plan to Action: Successful Data Management Plan Implementation in a Multidisciplinary Project," *Journal of eScience Librarianship*, vol. 5, no. 1, e1101, Sep. 2016, ISSN: 2161-3974. DOI: 10.7191/JESLIB.2016.1101. [Online]. Available: https://publishing.escholarship.umassmed.edu/jeslib/article/id/409/.

[2] Ghent University, *Preparing a Data Management Plan (DMP)*. [Online]. Available: https://www.ugent.be/en/research/openscience/datamanagement/before-research/datamanagementplan.htm.

[3] Belnet, *DMPonline.be*, 2025. [Online]. Available: https://dmponline.be/.

[4] A. Vaswani, G. Brain, N. Shazeer, *et al.*, "Attention Is All You Need," Tech. Rep., 2023.

[5] E. Kasneci, K. Sessler, S. Küchemann, *et al.*, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102 274, Apr. 2023, ISSN: 1041-6080. DOI: 10.1016/J.LINDIF.2023.102274.

[6] P. Lewis, E. Perez, A. Piktus, *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems*, vol. 2020-December, May 2020, ISSN: 10495258. [Online]. Available: https://arxiv.org/abs/2005.11401v4.

[7] Q. Ye, M. Axmed, R. Pryzant, and F. Khani, "Prompt Engineering a Prompt Engineer," *ACL*, Nov. 2023, ISSN: 0736587X. [Online]. Available: https://arxiv.org/abs/2311.05661v3.

[8] B. Bordalejo, D. Pafumi, F. Onuh, A. K. Khalid, M. S. Pearce, and D. P. O'Donnell, ""Scarlet Cloak and the Forest Adventure": a preliminary study of the impact of AI on commonly used writing tools," *International Journal of Educational Technology in Higher Education*, vol. 22, no. 1, pp. 1–25, Dec. 2025, ISSN: 23659440. DOI: 10.1186/S41239-025-00505-5/FIGURES/12. [Online]. Available: https://link.springer.com/articles/10.1186/s41239-025-00505-5%20https://link.springer.com/article/10.1186/s41239-025-00505-5.

[9] E. Nurchurifiani, A. Maximilian, G. D. Ajeng, P. Wiratno, T. Hastomo, and A. Wicaksono, "Leveraging AI-Powered Tools in Academic Writing and Research: Insights from English Faculty Members in Indonesia," DOI: 10.18178/ijiet.2025.15.2.2244.

[10] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623, Mar. 2021. DOI: 10.1145/3442188.3445922. [Online]. Available: https://dl.acm.org/doi/10.1145/3442188.3445922.

[11] J. Su and W. Yang, "Unlocking the Power of ChatGPT: A Framework for Applying Generative AI in Education," *ECNU Review of Education*, vol. 6, no. 3, pp. 355–366, Aug. 2023, ISSN: 26321742. DOI: 10.1177/20965311231168423/ASSET/6A33C26A-6D57-4B88-8430-BA3BDC4C1B6B/ASSETS/IMAGES/LARGE/10.1177{_}20965311231168423-FIG1.JPG. [Online]. Available: https://journals.sagepub.com/doi/10.1177/20965311231168423.

[12] C.-M. Chan, C. Xu, R. Yuan, *et al.*, "RQ-RAG: Learning to Refine Queries for Retrieval Augmented Generation," Mar. 2024. [Online]. Available: https://arxiv.org/abs/2404.00610v1.

[13] J. Shin, C. Tang, T. Mohati, M. Nayebi, S. Wang, and H. Hemmati, "Prompt Engineering or Fine-Tuning: An Empirical Assessment of LLMs for Code," Oct. 2023. [Online]. Available: http://arxiv.org/abs/2310.10508.

[14] A. Mohebbi, "Enabling learner independence and self-regulation in language education using AI tools: a systematic review," *Cogent Education*, vol. 12, no. 1, Dec. 2025, ISSN: 2331186X. DOI: 10.1080/2331186X.2024.2433814. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/2331186X.2024.2433814.

[15] A. Fahad and C. Z. Huang, "Human-in-the-Loop AI: A Framework for Continuous Validation of Generative AI Outputs in Healthcare," DOI: 10.1186/s12911-020-01332-6. [Online]. Available: https://doi.org/10.1186/s12911-020-01332-6.

[16] S. Natarajan, S. Mathur, S. Sidheekh, W. Stammer, and K. Kersting, "Human-in-the-loop or AI-in-the-loop? Automate or Collaborate?" *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 27, pp. 28 594–28 600, Apr. 2025, ISSN: 2374-3468. DOI: 10.1609/AAAI.V39I27.35083. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/35083.

[17] Flanders.be, *R&D intensity*, Jul. 2024. [Online]. Available: https://www.vlaanderen.be/en/statistics-flanders/science-and-innovation/rd-intensity.

[18] S. Leonelli, "Scientific Research and Big Data," in *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta, Ed., Summer 2020, Metaphysics Research Lab, Stanford University, 2020. [Online]. Available: https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=science-big-data.

[19] *Research Data Management*. [Online]. Available: https://scienceeurope.org/our-priorities/open-science/research-data-management/.

[20] Research Data Management, *Data Management*. [Online]. Available: https://www.uhasselt.be/en/university-library/research/research-data-management/data-management.

[21] J. Rogito and M. Makabe, "The Art and Act of Providing Feedback at the Workplace: Effective Feedback for Positive Results," *Pan-African Journal of Education and Social Sciences (PAJES)*, vol. 4, no. 1, pp. 49–56, Jan. 2023. [Online]. Available: https://www.researchgate.net/publication/372104151_The_Art_and_Act_of_Providing_Feedback_at_the_Workplace_Effective_Feedback_for_Positive_Results.

[22] R. Merrit, *What Is Retrieval-Augmented Generation aka RAG — NVIDIA Blogs*, Jan. 2025. [Online]. Available: https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/.

[23] I. Fischer, "Evaluating the ethics of machines assessing humans The case of AQA: An assessment organisation and exam board in England," *Journal of Information Technology Teaching Cases*, Nov. 2023, ISSN: 20438869. DOI: 10.1177/20438869231178844.

[24] M. Schrepp, A. Hinderks, and J. Thomaschewski, "Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S)," *International Journal of Interactive*

Multimedia and Artificial Intelligence, vol. 4, no. 6, p. 103, 2017. DOI: 10.9781/IJIMAI.2017.09.001.

[25] T. Teubner, C. M. Flath, C. Weinhardt, W. van der Aalst, and O. Hinz, "Welcome to the Era of ChatGPT et al.: The Prospects of Large Language Models," *Business and Information Systems Engineering*, vol. 65, no. 2, pp. 95–101, Apr. 2023, ISSN: 18670202. DOI: 10.1007/S12599-023-00795-X/METRICS. [Online]. Available: https://link.springer.com/article/10.1007/s12599-023-00795-x.

[26] Pecan, *The Role of LLMs in AI Innovation*. [Online]. Available: https://www.pecan.ai/blog/role-of-llm-ai-innovation/.

[27] A. Jain, *Journey LLM 12: A Deep Dive into Convolutional Network Architectures — Explanation, Implementation, and Comparison*, Oct. 2024. [Online]. Available: https://medium.com/@akshayush007/journey-llm-12-a-deep-dive-into-convolutional-network-architectures-explanation-implementation-62f90eb20eea.

[28] H. Narasimhan, *RNN to Transformers: The principle behind LLMs*, Dec. 2023. [Online]. Available: https://www.linkedin.com/pulse/rnn-transformers-principle-behind-llms-harini-narasimhan-1xncc.

[29] J. Ferrer, *How Transformers Work: A Detailed Exploration of Transformer Architecture*, Jan. 2024. [Online]. Available: https://www.datacamp.com/tutorial/how-transformers-work.

[30] M. Hoque, *A Comprehensive Overview of Transformer-Based Models: Encoders, Decoders, and More*, Apr. 2023. [Online]. Available: https://medium.com/@minh.hoque/a-comprehensive-overview-of-transformer-based-models-encoders-decoders-and-more-e9bc0644a4e5.

[31] P. P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet of Things and Cyber-Physical Systems*, vol. 3, pp. 121–154, Jan. 2023, ISSN: 2667-3452. DOI: 10.1016/J.IOTCPS.2023.04.003.

[32] S. Willison, *GPT-5: Key characteristics, pricing and model card*, Aug. 2025. [Online]. Available: https://simonwillison.net/2025/Aug/7/gpt-5/.

[33] A. Ramachandran, "Fine-Tuning Advanced Reasoning Models Methodologies, Empirical Insights, and Strategic Implications for OpenAI o3, LLaMA 3.3, Claude 3.7, and Gemini 2.0," Tech. Rep., Mar. 2025. [Online]. Available: https://www.researchgate.net/publication/389944901_Fine-Tuning_Advanced_Reasoning_Models_Methodologies_Empirical_Insights_and_Strategic_Implications_for_OpenAI_o3_LLaMA_33_Claude_37_and_Gemini_20.

[34] *Gemma models overview — Google AI for Developers*, Mar. 2025. [Online]. Available: https://ai.google.dev/gemma/docs.

[35] J. Li, L. Gui, Y. Zhou, D. West, C. Aloisi, and Y. He, "Distilling ChatGPT for Explainable Automated Student Answer Assessment," [Online]. Available: https://github.com/lijiazheng99/aera..

[36] Y. Yang, H. Chai, S. Shao, *et al.*, "AgentNet: Decentralized Evolutionary Coordination for LLM-based Multi-Agent Systems," Apr. 2025. [Online]. Available: http://arxiv.org/abs/2504.00587.

[37] W. Epperson, G. Bansal, V. C. Dibia, *et al.*, "Interactive Debugging and Steering of Multi-Agent AI Systems," 2025. DOI: 10.1145/3706598.3713581. [Online]. Available: https://doi.org/10.1145/3706598.3713581.

[38] S. Chen, L. Zeng, A. Raghunathan, F. Huang, and T. C. Kim, "MoA is All You Need: Building LLM Research Team using Mixture of Agents," Sep. 2024. [Online]. Available: https://arxiv.org/abs/2409.07487v2.

[39] J. Liao, M. Wen, J. Wang, and W. Zhang, "MARFT: Multi-Agent Reinforcement Fine-Tuning," Apr. 2025. [Online]. Available: https://arxiv.org/abs/2504.16129v3.

[40] S. Hu, M. A. Hady, J. Qiao, J. Cao, M. Pratama, and R. Kowalczyk, "Adaptability in Multi-Agent Reinforcement Learning: A Framework and Unified Review,"

[41] Y. Gao, Y. Xiong, X. Gao, *et al.*, "Retrieval-Augmented Generation for Large Language Models: A Survey," [Online]. Available: https://github.com/Tongji-KGLLM/.

[42] N. F. Liu, K. Lin, J. Hewitt, *et al.*, "Lost in the Middle: How Language Models Use Long Contexts," Nov. 2023. [Online]. Available: http://arxiv.org/abs/2307.03172.

[43] M. Galushko, "Navigating the RAG Landscape: Key Challenges in AI-Powered Information Retrieval," *Devrain*, Sep. 2024. [Online]. Available: https://devrain.com/blog/navigating-the-rag-landscape-key-challenges-in-ai-powered-information-retrieval.

[44] I. Drori and D. Te'eni, "Human-in-the-Loop AI Reviewing: Feasibility, Opportunities, and Risks," *Journal of the Association for Information Systems*, vol. 25, no. 1, pp. 98–109, 2024, ISSN: 15369323. DOI: 10.17705/1JAIS.00867.

[45] S. Vanbrabant, G. Alberto, and R. Ruiz, "ECHO: Enhancing Conversational Explainable AI through Tool-Augmented Language Models," *Proc. ACM Hum.-Comput. Interact*, vol. 9, no. 4, DOI: 10.1145/3734191. [Online]. Available: https://doi.org/10.1145/3734191.

[46] D. Zhao, "The impact of AI-enhanced natural language processing tools on writing proficiency: an analysis of language precision, content summarization, and creative writing facilitation," *Education and Information Technologies*, vol. 30, no. 6, pp. 8055–8086, Apr. 2024, ISSN: 15737608. DOI: 10.1007/S10639-024-13145-5/TABLES/4. [Online]. Available: https://link.springer.com/article/10.1007/s10639-024-13145-5.

[47] L. K. Nhan, N. T. M. Hoa, and L. V. N. Quang, "Leveraging AI for Writing Instruction in EFL Classrooms: Opportunities and Challenges," *Educational Process: International Journal*, vol. 15, 2025, ISSN: 25648020. DOI: 10.22521/EDUPIJ.2025.15.158.

[48] P. Limna, S. Jakwatanatham, S. Siripipattanakul, P. Kaewpuang, and P. Sriboonruang, "A Review of Artificial Intelligence (AI) in Education during the Digital Era," *Technology in Society*, vol. 74, Jul. 2022, ISSN: 0160791X. DOI: 10.1016/J.TECHSOC.2023.102279. [Online]. Available: https://papers.ssrn.com/abstract=4160798.

[49] A. A. Bany Abdelnabi, B. Soykan, D. Bhatti, and G. Rabadi, "Usefulness of Large Language Models (LLMs) for Student Feedback on H&P During Clerkship: Artificial Intelligence for Personalized Learning," *Trans Comput Healthcare*, 2025. DOI: 10.1145/3712298.

[50] S. C. Kong and Y. Yang, "A Human-Centered Learning and Teaching Framework Using Generative Artificial Intelligence for Self-Regulated Learning Development Through Domain Knowledge Learning in K-12 Settings," *IEEE Transactions on Learning Technologies*, vol. 17, pp. 1588–1599, 2024, ISSN: 19391382. DOI: 10.1109/TLT.2024.3392830.

[51] K. Kutumbe, *Comprehensive Guide to Chunking in LLM and RAG Systems*, Sep. 2024. [Online]. Available: https://kshitijkutumbe.medium.com/comprehensive-guide-to-chunking-in-llm-and-rag-systems-c579a11ce6e2.

[52] T. Cohen, J. (Yossi, and). Gil, “Better Construction with Factories ”The Factory-Owning Class Controls the Means of Production.” K. Marx [14],” vol. 0, no. 0, [Online]. Available: http://www.springframework.org.

Appendix A

Annex - Score Distributions for UEQ-S in function of the time since last DMP completion

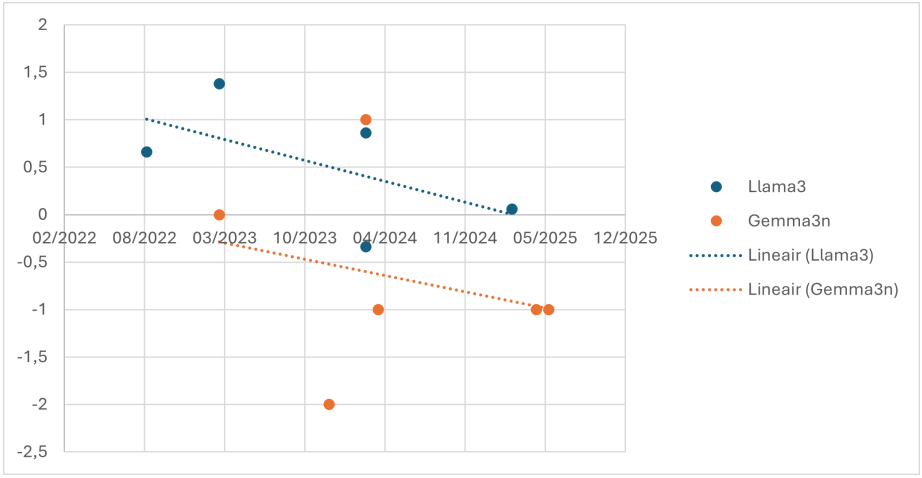


Figure A.1: Pragmatic scores for the star rating tool as a function of time since last DMP completion.

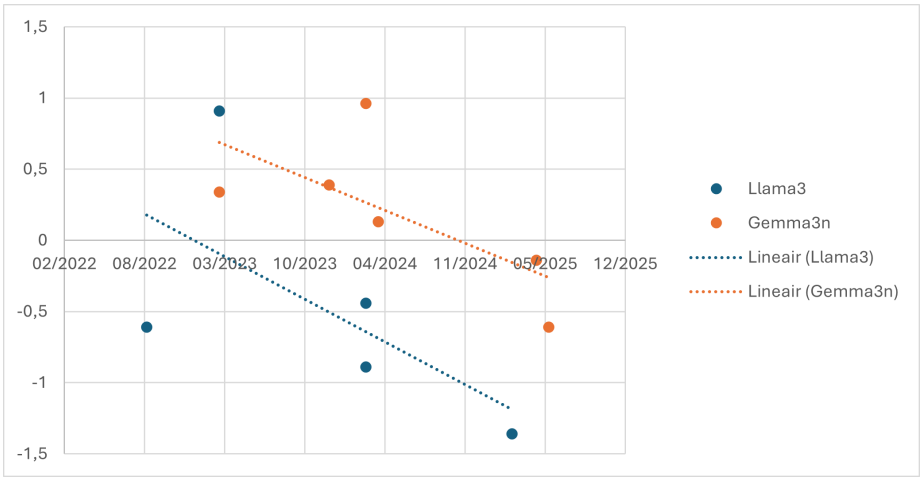


Figure A.2: Hedonic scores for the star rating tool as a function of time since last DMP completion.

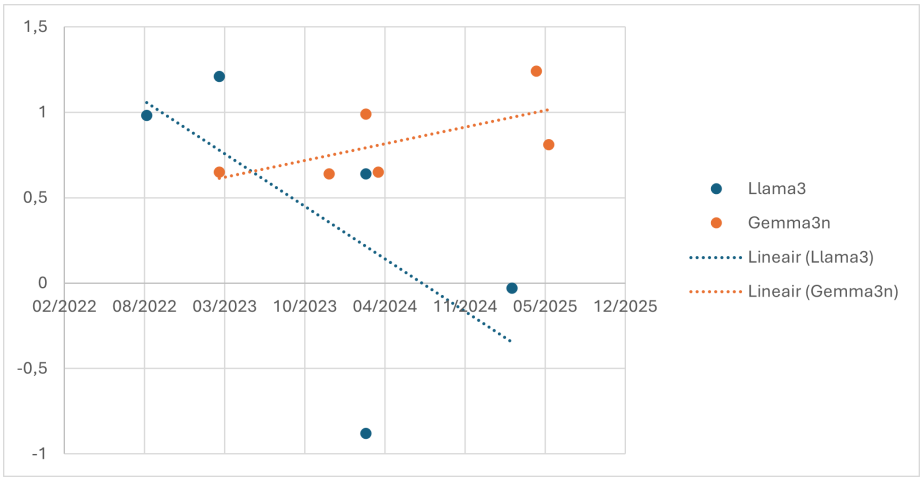


Figure A.3: Pragmatic scores for the chatbot tool as a function of time since last DMP completion.

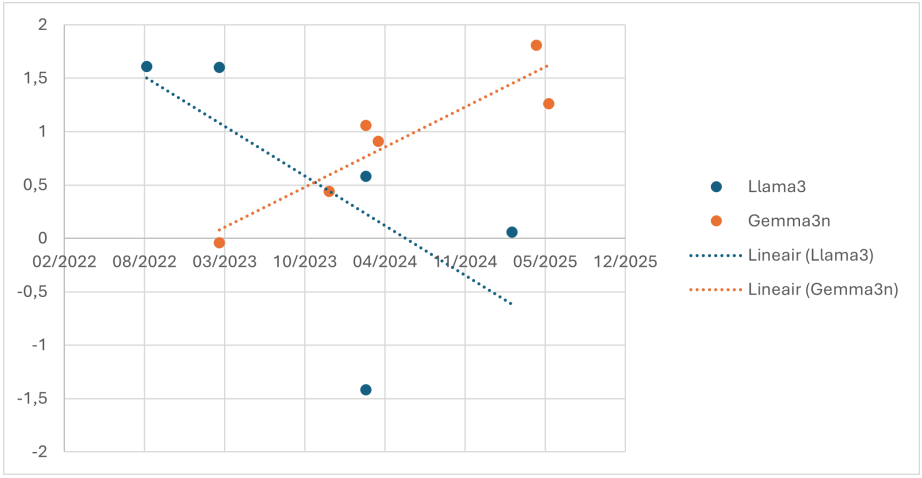


Figure A.4: Hedonic scores for the chatbot tool as a function of time since last DMP completion.

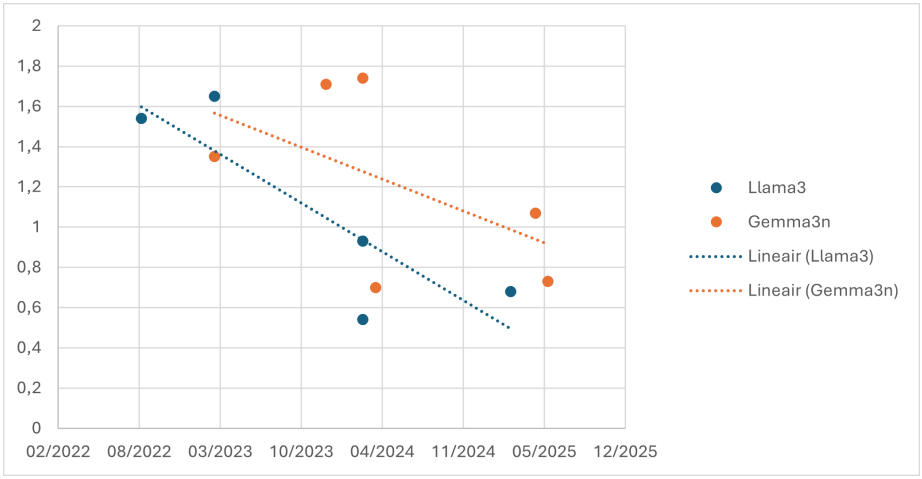


Figure A.5: Pragmatic scores for the hybrid feedback solution as a function of time since last DMP completion.

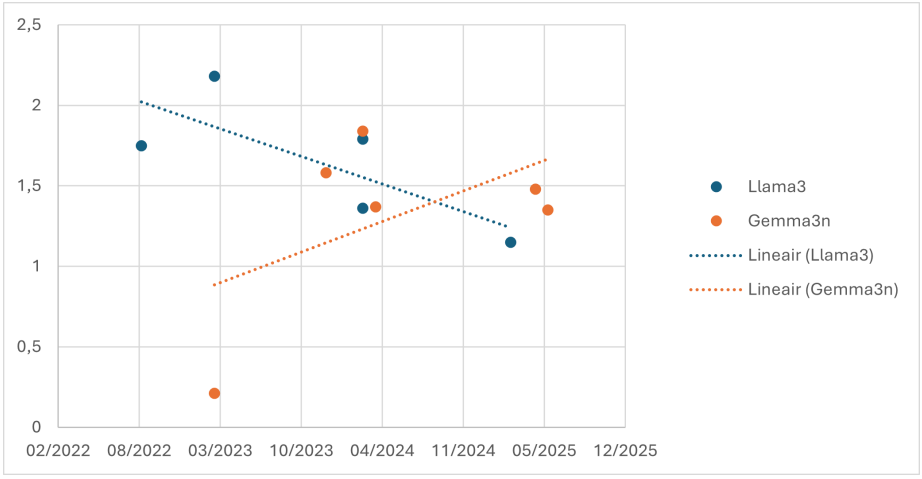


Figure A.6: Hedonic scores for the hybrid feedback solution as a function of time since last DMP completion.

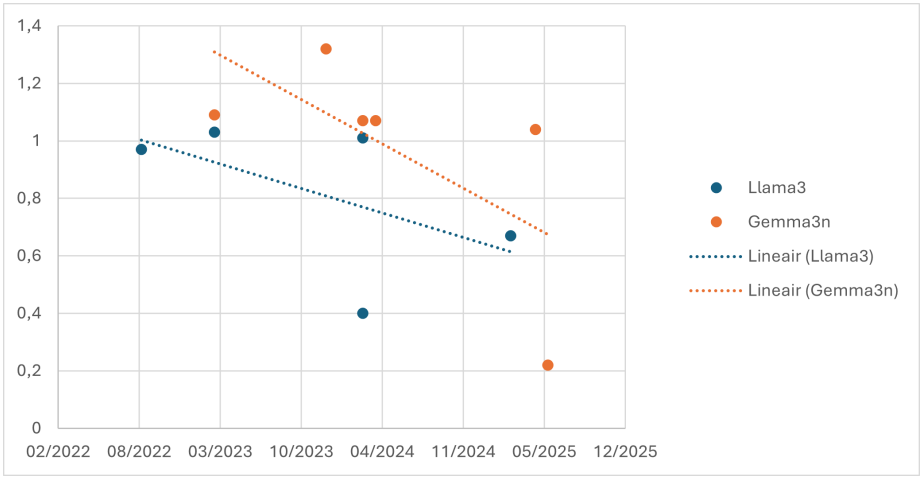


Figure A.7: Overall pragmatic scores as a function of time since last DMP completion.

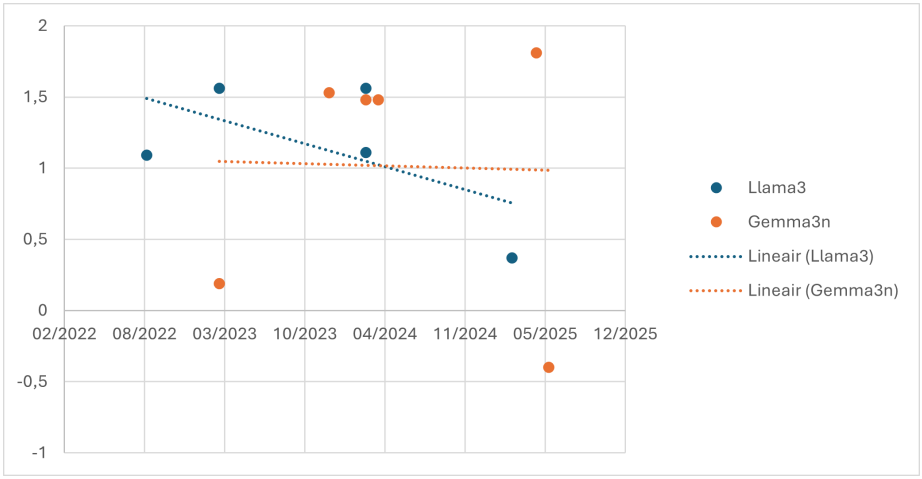


Figure A.8: Overall hedonic scores as a function of time since last DMP completion.