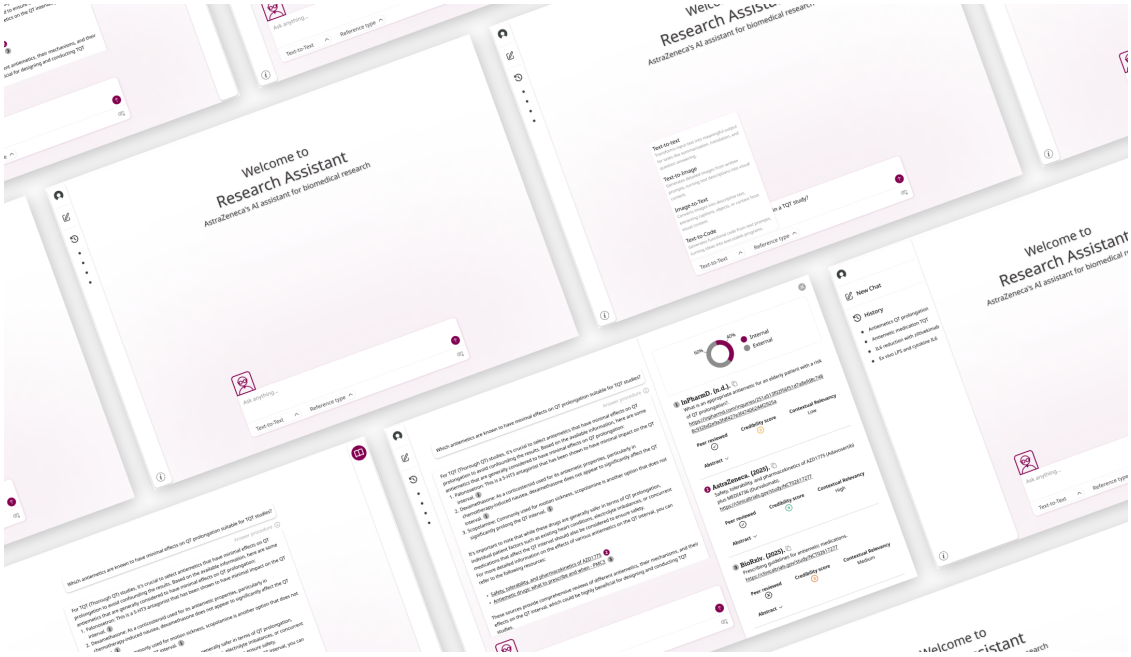
CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG



Designing Conversational Agents – Important Factors for Enhancing User Satisfaction Among Field-Specific Experts

Master's thesis in Computer science and engineering

SARA BÖRJESSON
KARIN ÖRN ANDERSSON

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2025

Master's thesis 2025

Designing Conversational Agents – Important Factors for Enhancing User Satisfaction Among Field-Specific Experts

SARA BÖRJESSON
KARIN ÖRN ANDERSSON

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2025

Designing Conversational Agents – Important Factors for Enhancing User Satisfaction Among Field-Specific Experts
SARA BÖRJESSON
KARIN ÖRN ANDERSSON

Academic supervisor: Swen Gaudl, Department of Applied IT
Industry supervisor: Sara Stegemann, AstraZeneca
Examiner: Michael Heron, Department of Computer Science and Engineering

Master's Thesis 2025
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2025

Designing Conversational Agents – Important Factors for Enhancing User Satisfaction Among Field-Specific Experts
SARA BÖRJESSON
KARIN ÖRN ANDERSSON
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

This thesis investigates the use of conversational agents, particularly how scientists use an AI assistant for biomedical research. The research aim revolves around increasing the user satisfaction, by finding important factors that influence it. The research utilizes both qualitative and quantitative methods from human-computer interaction, including external tool analysis, surveys, interviews, prototyping, user testing, and evaluations. By iteratively involving users and maintaining a reflective mindset while utilizing design methods, a Research through Design approach has been adopted. The result consists of a new design of the application, focusing on references and prompt settings such as models and reference types. It also resulted in important factors to consider for conversational agents, including the themes research foundations, trust, usability, time efficiency, transparency, and motivation. The factors have the purpose to gather the insights from every used method, created with a high-level perspective to make the aspects applicable for all conversational agents designed for internal use by field-specific experts.

Keywords: Conversational agent, Design recommendations, AI chatbot, User Experience, User Satisfaction, Interface, Computer science, Engineering, Project, Thesis.

Acknowledgements

We would like to express our gratitude to all those who have supported us throughout this project.

First and foremost, we wish to thank Swen Gaudl at University of Gothenburg for his exceptional guidance and support. Your academic insights and thoughtful feedback have been crucial in shaping our research.

We are also deeply grateful to AstraZeneca for providing the opportunity and necessary resources to conduct this research. It has been an invaluable experience for our professional growth.

Finally, our sincere thanks go to Sara Stegemann at AstraZeneca for her invaluable advice and encouragement. Your expertise and dedication have greatly contributed to the successful completion of this project.

Sara Börjesson and Karin Örn Andersson, Gothenburg, May 2025

Contents

List of Figures

List of Tables

1

Introduction

AstraZeneca is constantly looking for ways to improve their work and research. In recent years, several new applications have been launched to support their scientists in their daily work and to make their processes more efficient and thorough. One of the application types that has increased is conversational agents. The use of conversational agents has increased both internally at AstraZeneca and externally [1]. A reason for this increase could be that they have a wide range of potential applications, and when combining machine learning and natural language processing, they get more powerful and can solve more complex problems.

The increase of conversational agents used internally has led to several applications being used, where it is uncertain how the users use them, and how efficient the applications actually are. Therefore, there is a need for further evaluation of design of the tools, to ensure that they are adapted to the user needs.

1.1 Aim & Scope

This thesis aims to explore the conversational agent, CA, paradigm, to examine the design structure and interaction of applications using artificial intelligence with the purpose to find information. This includes to analyze how they function, interact with users and fulfill their intended purpose. Furthermore, it will be investigated how the existing CAs are perceived from a user point of view and how they are used by scientists. The purpose is to identify areas of improvements related to the design of existing CAs. This is done to increase the value of the product and the identified factors will serve as a guideline when designing CAs. The thesis will focus on increasing the user satisfaction levels through user research and design, where the expected outcome will be presented in a high-fidelity prototype.

To the best of our knowledge, only a limited number of studies have explored the use of CAs by professionals in their work. Our study will address a gap in the research by targeting a specific user group, where the focus is on scientists and their use of CAs in an internal and professional context. The thesis strives to result in a development of important factors and a prototype tailored to their needs and work conditions.

1.2 Research Question

The research question is as follows:

1. *What are important factors for designing conversational agents that influence user satisfaction in the context of internal field-specific experts, and how can these factors be realized?*

1.3 Demarcations

This thesis is limited to focus on internal use of conversational agents, by field-specific experts. External available conversational agents will be analyzed for inspiration purposes, however, the results will address the ones for internal use. The project is based on an interaction design viewpoint, with a user-focused perspective, while not addressing developmental considerations such as coding or technical infrastructure.

2

Background

In this chapter, the technique behind conversational agents will be described. The application which the thesis centers around will be introduced. Concepts such as artificial intelligence, machine learning, deep learning, natural language processing, generative AI, large language models, agentic AI, and subfields will be described. Furthermore, external conversational agents will be presented, as well as its history and related work. Lastly, AI applications terms will be compared.

2.1 Internal Conversational Agent – Research Assistant

Within research and development, R&D, at AstraZeneca, several AI applications have been developed. One of them is Research Assistant, RA, which is a conversational agent using natural language processing (see section 2.7), and large language models (see section 2.5) for processing prompts and generating text. A prototype was created in February of 2023 and the current version was created and released in April of 2024. The targeted user group of RA is people working within R&D, primarily targeted for scientists that can use it to ask scientific inquiries and expect a scientific response.

The data is collected from the open web, literature, and internal databases. One of the main benefits of RA is the feature that provides references for each new type of information that is presented. With the verifiable sources, the application becomes more reliable since the scientist can verify that the information is correct, which is of great importance within the science field, where great rigor is required.

2.2 Artificial intelligence

Artificial intelligence, AI, is a trending topic, yet it was discussed already in the 1950s [2]. A decade later, logical methods that could create mathematical proof were to be found in computer programs. To exemplify, this made it possible for computer programs to beat humans in chess. However, it was extremely difficult, if not impossible, to create performances such as recognizing objects in pictures of understanding language. To do this, a different approach was needed, which did not have a breakthrough until the twenty-first century because of the need of large data

sets. Today, conversational agents are a central part of the AI evolution, developed with natural language processing and machine learning [3].

2.3 Machine Learning

Machine learning is generally categorized into three main types: supervised learning, unsupervised learning, and reinforcement learning [4]. Supervised learning requires labeled input data, allowing the model to learn from training examples and generalize to make accurate classifications beyond the training set. Unsupervised learning, on the other hand, does not use labeled data. Instead, the algorithm groups similar items together based on shared features. This method provides less insight into the decision-making process compared to supervised learning. Reinforcement learning involves rewarding desirable actions and penalizing less preferred ones to train the model. The algorithm is not explicitly told how to achieve the correct action but learns from feedback, either through rewards or penalties, indicating whether the action taken was right or wrong.

2.4 Deep Learning

Deep learning is inspired by the human brain in processing information [5]. While both machine learning and deep learning learn from data, the difference is that deep learning uses neural networks, which are generally larger with many hidden layers. It uses these layers of neural networks to go through large amounts of data, being self-capable of extracting information, which allows it to handle various kinds of unstructured data effectively. It can be used to solve complicated problems, for computer vision, natural language processing, cybersecurity, and more.

2.5 Large Language Models

Large Language Models, LLMs, are designed to perform NLP tasks [6]. Built on the foundation of artificial intelligence, they can process and transform vast amounts of data. The ability to handle these large datasets is largely attributed to the rise of the transformer architecture. Transformers can have billions of parameters and are the most commonly used architecture for LLMs. This architecture involves extensive preprocessing and is trained on massive datasets. During this process, the model learns statistical patterns to understand the complex interconnections between components that make up a text segment. LLMs typically utilizes statistical models. This mean that once the model is trained, its parameters remains unchanged in response to new data unless it's explicitly being updated.

2.6 Generative AI

Generative AI is the field in artificial intelligence focusing on generating new content [7]. Based on deep learning techniques, it can generate new material that closely

resembles the original input data it has been trained on. It can generate material such as images, text and audio. Generative AI differs from traditional AI in its ability to create new and original data, rather than strictly following predefined instructions. It utilizes both supervised and unsupervised learning methods and is primarily based on neural networks that excel at pattern recognition. By learning from existing data, generative AI can produce novel outputs such as text, images, or audio.

2.6.1 Retrieval-Augmented Generation

Retrieval-Augmented Generation, RAG, is a specific architecture often integrated with LLM to improve their abilities [6]. A recurrent problem with statistical models, such as LLMs, is that it has no possibility to learn and adapt to new real time information. This may result in outdated or inaccurate information [8]. RAG has surfaced as a solution to this problem by integrating external sources from where it can retrieve real-time data [9].

2.7 Natural Language Processing

Natural language processing, NLP, consists of several computational techniques for analyzing and representing data with the purpose to achieve human-like processing for different tasks or applications [10]. The language analysis made can be done through multiple techniques and the represented data can be of any language, type or mode, both oral and written. NLP is used for several types of applications, including spam detection, auto-complete functions, recruitment processes, smart assistants and conversational agents. For NLP applications, pre-processing is needed to prepare the data into an appropriate form for the understanding unit [3]. This includes data cleaning where text is converted to lowercase and punctuation is removed, tokenization where sentences, words, and characters are separated and normalization where stemming trims a word into its origin form.

2.7.1 Natural Language Understanding

Natural language understanding, NLU, is a subfield of NLP that focuses on spoken or written human language and the interpretation of it [11]. It enables the human-machine communication by building a comprehensive understanding of semantics. An example of systems using NLU is call centers, that uses conversational agents to solve frequently occurring problems, which allows their personnel to answer and solve more complicated problems, leading to more users getting help in less time. Applications using NLU are those with a large range of semantics, with multiple concepts that can be combined into a more complex request. It can handle longer user utterances by extracting smaller pieces of it.

2.7.2 Natural Language Generation

Natural language generation is a subfield of AI and an essential part of natural language processing, that involves the process of generating natural language text into an understandable format for humans [12]. The aim is to transform data from a non-language format into a human-readable format to simplify the communication between humans and machines. Its tasks can be divided into four categories, which are text-to-text, data-to-text, image-to-text, and video-to-text.

2.8 Agentic AI

The definition of Agentic AI is that it should successfully perform a series of advanced assignments for a longer time frame without any human involvement or supervision [13]. Agentic AI primarily differs from traditional AI in the sense that it can perform more advanced reasoning with several steps and has more flexibility to move within. The flexibility, compared to traditional AI which has more definitions and restrictions to fit, also makes the agentic AI more interdependent. Similar to LLMs, the implementation of Agentic AI require large amount of training data and robust data pipelines.

2.9 External Conversational agents

There are several conversational agents available for both individuals and companies. Below are three examples of available CAs, all based on large language models as a base technology.

2.9.1 ChatGPT

ChatGPT is a model that interacts in a conversational way [14]. It can answer follow-up questions, admit mistakes and reject requests that are inappropriate. It is trained using Reinforcement Learning from Human Feedback, RLHF, which is a variant of Reinforcement learning, RL [15]. RLHF is similar to RL, however, it does not require an engineered reward function, but instead learns from feedback from humans. The users provide feedback to the agent and behaves in accordance with the feedback.

2.9.2 Claude

Claude, made by Anthropic, consists of a model family with the versions ”Haiku”, ”Sonnet”, and ”Opus” [16]. The differences between the models are that one is light and fast, one is described as hardworking and the last is the most powerful. It is able to analyze images, generate code, and translate between languages, among other features. It focuses on integrity, misuse prevention and accuracy.

2.9.3 Copilot

Microsoft Copilot is described as an AI-powered productivity tool that provides real-time intelligence [17]. It is paired to Microsoft Graph and Microsoft 365, which offers a connection to emails, documents, Word, Excel, and other applications offered by Microsoft. This includes being able to create drafts in Word, add slides in PowerPoint, and summarize emails.

2.10 History of CAs

Conversational agents, CAs, originated from chatbots. While the term ”chatbot” is broader and encompasses CAs within its scope, a chatbot is defined as a system designed to digitally simulate human conversation. This term applies to systems regardless of the underlying technology, whether powered by advanced AI or more simple rule-based with predetermined actions [18]. While multiple definitions of conversational agents exist, this thesis will use the definition by Laban and Araujo, and thereby define a CA as a virtual entity that interacts with users using artificial intelligence [19]. Five waves of conversational agent development have occurred, leading up to the current version [20]. The first wave featured simplistic applications with predetermined rules, operating under specific conditions. This wave began with the creation of ELIZA, the first-ever chatbot, developed by Weizenbaum in 1966 [21]. ELIZA was limited in terms of usage since it could only operate within specific conditions with its predetermined rules. For instance, it faced problems in generating a response when keywords to which it had been programmed to respond were lacking. However, there is no doubt that ELIZA has had a profound impact on its field [22]. The term ELIZA-effect was coined in 1970, which is the tendency to devote human attributes such as empathy to a responsive computer. The second wave in 1995 introduced AI for the first time, allowing systems to recognize emotions via scripted dialogue while the first embodied CAs appeared [20]. However, the embodiment of CAs also raised issues related to anthropomorphism and the uncanny valley. The third wave saw the real-world application of CAs, such as IBM’s Watson. The fourth wave brought voice assistants like Alexa, and the fifth wave, where we are today, is marked by innovations such as OpenAIs ChatGPT.

2.11 Related work and possible aspects to consider when designing CAs

A meta-analysis of 60 different studies found that attitudes, effort expectancy, performance expectancy, perceived usefulness, and trust are strong indicators of user acceptance of AI technologies [23].

The role of trust in technology adoption is critical, and studies have identified several dimensions of trust that influence user acceptance. For instance, a study on trust in the digital workplace categorized trust into three dimensions: cognitive, emotional, and organizational [24]. Cognitive trust is based on knowledge

and can change quickly depending on performance. However, emotional and organizational trust are more resilient, often remaining stable even after errors. When interactions with a conversational agent include social cues, they can form long-lasting trust, supporting continued product use.

Other research highlights the importance of managing user expectations of the CAs capabilities [25]. Regardless of whether the CA can perform the desired action, it is essential to clearly communicate the capabilities to set appropriate expectations. Other factors contributing to a positive user experience include handling unsuccessful dialogues gracefully, so the CA does not negatively affect the user's perception. It's also important to enhance contextual understanding to improve conversation flow, responsiveness, and adding human-like conversational features to contribute to a more engaging experience.

Building on this idea, additional research emphasizes that employees' expectations and assumptions about technology play a significant role in its successful adoption [26]. Employees' prior experiences with similar technologies shape their expectations and influence how they evaluate and embrace new tools. This underscores the importance of cognitive factors such as prior knowledge or experience in forming initial trust and acceptance. In line with this, research has found that users' awareness of a CA's capabilities impacts how competent they feel when interacting with it [27]. Additionally, flexibility, personalization, and control over data contribute to users' sense of autonomy, further reinforcing trust and positive engagement.

In contrast, a study on AI adoption in office environments found that increased productivity and efficiency were the strongest motivators for adopting CA's [28]. Although hedonic qualities such as enjoyment and social acceptance played a role in user adoption, they were less important than practical factors such as increased productivity and efficiency.

These aspects can be summarized as:

- Trust and perceived usefulness are crucial for AI acceptance
- Clear communication of CA capabilities manages expectations.
- Employee expectations and past experiences influence adoption, with productivity and efficiency being key drivers.

These insights emphasize the importance of trust and managing user expectations, guiding the design of CAs to enhance user satisfaction and adoption in the workplace.

2.12 AI application terms

Recently, the use of AI chatbots, also known as conversational agents, have increased, which are equipped with AI [1].

The AI applications use machine learning, ML, and natural language processing, NLP, to carry on the human-like conversations [29]. The growth has led to artificial intelligence, AI, being used by both individuals and companies, where companies strive to increase the integration of AI in both products and services. Internally, companies can use AI applications to ease with coding tasks, offer clearer communications and to assist with creative work [30]. The organizational version of an AI tool is often equipped with data and security protection, enabling the use of encrypted conversations.

The growth of AI applications has led to several terms being used for tools with a similar, or even identical, structure. A chatbot simulates human conversations and often use NLP to understand and automate responses [18]. Another AI tool is AI agents, which are programs capable of performing tasks, either when asked by a user or connected to another system [31]. These can have many functionalities beyond NLP, including interacting with external environments and executing actions. Conversational AI is another term, which refers to technologies that users can communicate with, including both chatbots and virtual agents [32]. In contrast to chatbots, conversational AIs sometimes lack the embodiment aspect, such as an avatar. They use machine learning and NLP to imitate human language and interactions, where the machine learning is used to process the inputs to improve the AI algorithms. In order to keep a consistent designation throughout the report, the term Conversational Agent, CA, will be used hereinafter.

3

Theory

The following chapter explorer the theory considered relevant for the project, including human-computer interaction, research through design, cognitive load theory, working memory, computers are social actors, theories for user engagement, and design theory. Lastly, ethical considerations are described.

3.1 Human-computer Interaction

Human-computer interaction, HCI, is the study of the interaction between the user and a computer system, by investigating the efficiency and communication between them and how the systems are designed and used [33]. The purpose is to enhance computer systems' usability and user experience, by adapting to people with different preferences, backgrounds and abilities.

3.2 Research through Design

Research through Design, RtD, is an approach where new knowledge is generated by using design methods [34]. It uses the strengths of design as a reflective practice by reframing and reinterpreting problems through a design process, with a speculative mindset investigating what the world could be. However, it is important to be cautious when establishing norms and outcomes from RtD [35]. This is because the outcome might not be applicable in every case, but rather flexible and evolving for different design cases. To reflect on the result, annotated portfolios can be used to visualize and describe the process, since not only the result is of value, but the process as whole.

3.3 Cognitive load theory

Cognitive load theory focuses on the amount of effort needed by working memory to conduct a task, and by this structure a design that optimizes learning. [36]. Cognitive load theory is based on the hypothesis that the working memory has limited capabilities, while long term memory has potentially infinite. Cognitive load is based on an architecture of three different types of cognitive load, which is intrinsic, extraneous load and germane load. The intrinsic load is related to the

complexity of the subject you are learning, extraneous load is how the information is presented, while germane load involves constructing schemas that aids learning.

3.4 Working memory

The humans' working memory is a system that temporarily stores and manipulates information needed for complex tasks like reasoning, comprehension, learning, and problem-solving [37]. Working memory has limited capacity, containing short-term memory, while actively engaging with information.

3.4.1 Short-term memory

Short-term memory is active in working memory, and refers to the short-term holding and handling of information that is quickly accessible, usually lasting only a few seconds [38]. Short-term memory (STM) contains three different systems with a finite amount of capabilities, Phonological STM, Visual STM and Spacial STM. Phonological STM refers to the temporary stored auditive material, visual STM refers to the temporary stored visual recognition memory for objects and spatial STM refers to the temporary information about locations. The systems operate independently. Two key components of STM is encoding to register information and consolidation to contain the representation in STM. The limited capacity of short-term memory is evident in the suggested number of items a person can store, which is approximately seven, plus or minus two.

3.5 Computers are social actors

The theory *Computers are social actors* states that humans interact with computers as a social entity [39]. This means that although there is an understanding that computers are not human, humans tend to interact with computers as they are social and generates social responses in their interaction with the computer. The theory is described as an experimental paradigm of the human-computer interaction field of study. From five experiments made, it became clear that humans interact with computers in a social manner. One of the results was that uniformity of interfaces is double-edged, meaning that creating consistent and standardized interfaces has both positive and negative effects. A positive outcome could be that users understands the interface and easily interacts with it, while a negative implications could be that creativity gets restricted and the interface is not as adapted to different user needs.

Connected to the social actors aspect is anthropomorphism, meaning that anything can be perceived as characterized by human traits [40]. It can be made both intentional and unintentional and often appears in AI functionalities. Both media and entertainment has influenced how AI is perceived, leading to an idea that AIs are being like humans emotionally, cognitively and morally. This reflects

how limited the understanding of AI can be, which can lead to false expectations of its capabilities.

3.6 Theories for user engagement

Below, the theories technology acceptance model and task technology fit theory are described. These theories can be used to provide a comprehensive framework for understanding user engagement.

3.6.1 Technology Acceptance Model

The Technology Acceptance Model, TAM, developed by Fred Davis in 1989, was created to address the limited tools available to predict the acceptance of information systems, specifically computer-based systems, at the time [41]. Over time, TAM has expanded beyond its initial focus and become one of the most widely used frameworks for understanding how users accept and use a broad range of emerging technologies [42]. TAM proposes that two main predictors determine the probable adoption of a new technology: perceived usefulness and perceived ease of use.

Perceived usefulness refers to the extent to which a user believes that using a specific tool or technology will enhance their performance or help them achieve their goals. It focuses on the potential benefits that the user expects to gain from using the tool. Perceived ease of use, on the other hand, is the degree to which the user perceives the technology as easy to use, intuitive, and requiring minimal effort to learn or operate. This concept focuses on the user's initial experience with the tool and their perception of how effortless it is to interact with it. It's theorized that if the tool is perceived as both useful and easy to use, the user is more likely to adopt the new tool.

3.6.2 Task Technology Fit theory

Task Technology Fit theory is about how well the fit is between the technology and its use case [43]. If the features of the technology aligns well with the intended use case and the technology can help the user complete the task, it is more probable that the user would want to use the technology.

3.7 Design theory

The design theories explored during the project are usability, user experience, user interface, and human-centered design, which are described below.

3.7.1 Usability

Usability is a term describing how well a system, product or service can be used, which is described with the ISO-standard as:

The extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use. [44]

3.7.1.1 Usability principles

Four usability principles are described below.

Mental models
Mental models are a type of conceptual models that represent the understanding of how things work [45]. They are created of the people themselves, other people, the environment, as well as of things they interact with. Of the same item, different people can have different mental models. It is therefore important to consider what mental models the users of an application have, to be able to ensure that they understand how to interact with the application.

Affordance
Affordance is the relationship between a person and a physical object [45]. It is determined by the properties of the object and the capabilities of the person, influencing how the object can be used. By taking this aspect into consideration, possible interactions can be clearly indicated, which makes the application more user-friendly.

Signifiers
Signifiers are clues of how objects function [45]. They signal how and where to interact with the object, and must be perceivable to not fail to function. They are important to include to ensure user-friendliness, since the user needs to understand how to use the object.

Consistency
Consistency is virtuous, meaning that when a user has learned how to use one system, they will transfer that knowledge when learning a new system [45]. It addresses uniformity across various element and interactions in the interface. With consistency, the user can easier form a mental model and easier build an understanding of how the interface works. This in turn enhances usability and a better user experience.

3.7.2 User Experience

User Experience, UX, is focused on the perceived experience of the user when interacting with a product, system or device [46]. The experience includes factors such as seeing, feeling, or doing, and is connected to and affected by the user's previous

knowledge, assumptions, and motivation. UX design therefore focuses on creating a positive experience and fulfilling their needs during each part of the interaction.

3.7.3 User Interface

A user interface, UI, is an significant component between a group of users and a system and is the systems' primary component that the user interacts with [47]. It enables the users to interact with the system in a efficient way, which is done with different strategies depending on the purpose of the system or interaction. An effective UI-design strategy can lead to rich user experience, by making it user-friendly, while the opposite leads to unsatisfied user-attitude and performance.

There are guidelines that can be followed when creating UIs. For the navigation, one examples is to standardize the task sequences to allow the user to perform tasks in similar ways across the interface [47]. With data display, it is important to be consistent with terminology, colors, abbreviations, capitalization, and formats when possible. Prioritization of information is important to ease the usage, by making the most important information the most visible and by keeping related information close together.

3.7.3.1 Posture

The posture of an application can be transient or sovereign [48]. When only using the application for short periods of time, with the purpose to achieve a larger goals elsewhere, it has a transient posture. When an application instead hold the users' attention for extended periods of time, the posture is sovereign.

3.7.3.2 Hierarchy

For visual interface design, different elements supports specific tasks, which requires a clear hierarchy of the elements to provide understanding and minimize cognitive load [48]. The visual hierarchy can be created with visual properties, such as use of color, typography, size, whitespace, position, and shape.

3.7.3.3 Details-on-demand

When designing advanced user interfaces, a good starting-point is the principle of Overview first, zoom and filter, then details-on-demand [49]. It means to only give the user details when needed, by offering a button to the user that opens up more information. By implementing the details-on-demand principle, the user is less likely to get overwhelmed by the interface, and can use it more efficiently.

3.7.3.4 Saliency

Saliency refers to the attribute of an object which stands out in comparison to other surrounding objects, and grabs the attention of the viewer [50]. Saliency can be achieved with differences in color, contrast, brightness, motion, size or other visual features.

3.7.4 Social Desirability bias

Social Desirability bias is when people portray themselves in a better light by highlighting behaviors that are viewed positively and downplaying those that are viewed negatively [51]. This can lead to that people alter what they think in order to please the researcher, or behave in a way they would think is more favorable, than their true and initial feelings.

3.7.5 Evaluation Apprehension

Evaluation apprehension refers to the anxiety that occurs due to the fear of being judged or evaluated negatively by others. Factors such as gender and educational level, self-esteem and self-assessment as well as fear of negative evaluation are all factors that can create evaluation apprehension. This may result in lower levels of participation, academic performance and can infer social anxiety, depression and low self-esteem [52].

3.8 Ethical Considerations

There are some ethical considerations that need to be taken into account. Below, ethical aspects of artificial intelligence is listed, as well as climate concerns to keep in mind.

3.8.1 Ethics of AI

There are several points to consider when creating AI regarding ethical concerns. Leslie discusses potential harm of AI, some of which are listed below [53]:

Bias and Discrimination
Since the creators of AI systems choose features, metrics, and analytic structures of how data should be accessed by the system, AI can potentially be influenced by the creators' biases and preconceptions [53]. The system then gains insights and trains its algorithm from existing sources and user inputs, which can lead to reinforcement and amplification of inequality and discrimination that exists in the society. These aspects are therefore important to keep in mind when creating AI systems.

Non-transparent, Unexplainable, or Unjustifiable Outcomes
Generated results are often done with high dimensional correlations that are beyond capabilities of human scale reasoning [53]. These outcomes are not transparent or understandable, leading to the user being influenced by an answer but not fully comprehending the reasoning behind it. In some cases, this lack of explainability and opaqueness is unethical.

Invasions of Privacy
Since AI systems are based on processing data, the creation of AI technologies will often require the use of personal data [53]. By sometimes being extracted

without consent or handled in a way that reveals personal information, this is a threat of privacy. AI systems that target or influence users without their consent or awareness can be seen as a violation of their ability to maintain a private life.

3.8.2 Climate concerns

There have been several attempts to calculate the carbon footprint of AI using NLP and although these efforts have resulted in varying estimates, the calculations consistently indicate that the carbon footprint can be a significant concern [2]. The electrical consumption of AI models is sensitive to the size of AI models, and when wanting to create more accurate inference, larger AI models have been required and developed based on deep neural network architecture.

4

Methodology

The following section describes relevant methods for the project, including literature review, user analysis, survey, ideation methods, creation methods, and evaluation methods.

4.1 Literature Review

A literature review can be done to get an overview of areas in which more research is needed [54]. It can also be used to find research that is interdisciplinary or different, which both help creating theoretical frameworks for processes. The approaches can be both quantitative and qualitative, depending on the wanted outcome.

4.2 User analysis

To get an understanding of the users and who they are, a user analysis can be made. This can be done by collecting quantitative data, which can be summarized into statistics [55]. It can also be done by gathering qualitative data, which is more in depth data from fewer users. Both approaches leads to an understanding of the user, by finding patterns in the results.

4.3 Survey

A survey is a method for collecting self-reported information about characteristics, thoughts, behaviors, or attitudes [56]. It is an efficient tool to collect a lot of data quickly, usually done with a large sample of respondents. By being self-reported, they do not reflect true thoughts or perceptions accurately. Therefore, careful administration is needed, and it is recommended to complement the survey with observations or other methods.

4.3.1 Questionnaire

A questionnaire is a data collection instrument commonly used within surveys [56]. It can include either open-ended or closed-ended questions. Open-ended questions allow respondents to elaborate on their thoughts and opinions, providing more detailed insights. In contrast, closed-ended questions offer predefined response options,

making analysis and comparison more straightforward. While questionnaires are time-efficient, careful consideration should be given to both wording and response options, as they significantly influence the quality and accuracy of the responses.

4.3.2 Net promoter score

Net Promoter Score is a method for investigating and measuring loyalty and user satisfaction with the product, and how likely it is that they would recommend the product to others [57]. A score of 0-6 makes the person a detractor, 7-8 is considered passive, and 9-10 is regarded as promoters of the product.

4.4 Interviews

Interviews are a valuable method for gathering rich qualitative data [56], allowing interviewees to express themselves not only through words but also through body language and facial expressions. The level of structure in an interview varies depending on its purpose. Structured interviews follow a predefined set of questions, ensuring consistency across multiple interviews, which is particularly useful when standardization is required. Semi-structured interviews offer a balance between structure and flexibility, combining predetermined questions with opportunities for spontaneous discussion. In contrast, unstructured interviews do not follow a fixed format, allowing for a more open-ended and personal interaction. Additionally, interviews can be categorized based on the target group, such as expert or stakeholder interviews.

4.4.1 Pilot test

A pilot test is a small-scale trial conducted with one or a few participants before the main testing begins, allowing researchers to evaluate the test's effectiveness [58]. This approach helps identify potential issues and refine the test design without risking the loss of valuable participants due to ineffective data collection. Pilot testing can be applied in both qualitative and quantitative research.

4.5 Ideation

Ideation is the process of generating novel ideas, an essential entity for innovation [56]. There are numerous methods for the creative process of ideation. There are collaborative methods such as brainstorming, visual methods such as story boarding and analytical methods such as reverse brainstorming among others.

4.5.1 Brainstorming

Brainstorming is a widely used method for generating a large number of ideas within a group setting [59]. It is guided by key principles, including prioritizing quantity

over quality, withholding criticism, encouraging creative and unconventional thinking, and building upon others' ideas to further develop them.

Crazy 8

Crazy 8 is a brainstorming technique with the purpose to generate a lot of ideas, fast [60]. The process is to sketch one idea in one minute, and do this for eight consecutive times. When one minute has passed and the timer is ringing, no further action is allowed on the sketch, just move on and start on next sketch.

4.5.2 Dotvoting

Dotvoting is a prioritization technique to rank and prioritize different ideas, features, to know what to design. The basis is democracy, where each participant is allowed to vote on ideas they find important [61].

4.5.3 Mood board

Mood board is a method used as a source of information, to visually communicate ideas and create a visual entity that allows people to align in the same direction of the project [62]. The mood board offers a blueprint to the direction of the project in communicating design concepts or ideas. Mood boards are made up of visual entities, where the purpose is to represent a certain style to embody in the artifact or represent a certain emotion the visual entity originates.

4.6 Creation

During a creation phase several methods can be used. The methods sketching, wireframing, mood board, and prototyping are described below.

4.6.1 Sketching

Sketching is a method to externalize ideas, to create them into a tangible format to make them permanent [63]. Ideas can easily be remade several times for improvement. This process not only aids in clarifying ones own thinking but also facilitates communication and collaboration with others.

4.6.2 Wireframing

Wireframing represents the basic structure and layout of a design, serving as a blueprint for the final interface [64]. It is a valuable tool for visualizing the foundation of a design in a simple and cost-effective manner before any coding begins. Wireframes can be created by using pen and paper, whiteboards, or digital tools. They are typically made up of simple shapes, such as boxes and lines, to represent key elements, such as buttons, text areas, images, or navigation menus. They focus on the structure and arrangement, and usually avoid the use of colors.

4.6.3 Prototyping

Prototyping is a valuable method for transforming implicit or intangible knowledge into a more tangible form [64]. This knowledge-generation process enhances communication and understanding among stakeholders in a project. Prototypes can take various forms, including digital or physical representations, with their level of fidelity typically increasing as the project progresses. The primary purpose of prototyping is to support product development by providing a shared reference point for stakeholders, facilitating decision-making, and enabling user testing.

4.7 Evaluation & Analysis

To evaluate an idea, concept or prototype, different analyzing methods can be used. Mixed methods research, word clouds, thematic analysis, affinity diagramming, and user testing are described in this section.

4.7.1 Mixed Methods Research

Mixed methods research encompasses both qualitative and quantitative data [65]. By leveraging the benefits of both quantitative and qualitative data, mixed methods can enable a more comprehensive and nuanced understanding of a problem. Insights can be made by validating findings in utilizing different types of data, as well as help to minimize bias of the researcher.

4.7.2 Word clouds

Word clouds are a method of information visualization [66], used to display the most frequent words that emerge from a survey. In a word cloud, the size of each word corresponds to how often it has been mentioned. The more frequently a word appears, the larger it is displayed. This method provides a clear overview of the number and types of topics, as well as the relevance of each topic based on how many respondents have mentioned it.

4.7.3 Thematic analysis

Thematic analysis is a method used to identify patterns in data and derive insights from qualitative material [67]. It is widely applied, particularly in academic research. Braun and Clarke outline six phases of thematic analysis, beginning with familiarization with the data and coding. These two phases include reading the data and making initial notes, as well as to generating brief labels of important features. Phase three and four include generating initial themes and then developing and reviewing the them. Broader patterns are developed and structured, and then each potential theme is reviewed with both the coded data and entire data set. When further developed, the themes can be split, combined, or discarded. The last phases are to phase five: refining, defining and naming themes, and six: Writing up. During these phases, a detailed analysis of each theme is made and each theme gets

a determined name, and then the analytic narrative and data extracts are weaved together.

4.7.4 User testing

User testing is a key component of successful product development, helping to increase future user adoption [68]. It evaluates usability, user experience, and overall functionality of the product.

4.7.5 A/B testing

There are different versions of A/B testing. When participants in A/B test are only exposed to one of the versions, it is called between-subjects design [69]. When participants see all versions it is known as within-subjects design. To avoid the ordering effect playing a role in affecting the result, counterbalancing is often deployed in within-subjects design. Counterbalancing is when the order of the versions are mixed for all participants, which means that all participants are shown each version in a different order.

4.7.6 Think aloud method

The think-aloud method is a research technique where participants are encouraged to articulate their thoughts aloud while performing a task [70]. This approach provides immediate insights into their cognitive processes, and can provide knowledge into their thought patterns, emotions, and problem-solving strategies. The method also involves transcription of the verbalization to be able to review and decode the thought process.

4.7.7 Usability Metric for User Experience

Usability Metric for User Experience, UMUX, was developed to measure user experience related to usability, with the primary aim to be more time-efficient, and simple but also more flexible compared with the widely used System Usability Scale [71]. UMUX consists of four questions, where two of the questions are positively framed and two of the questions are negatively framed. The questions are the following:

- [This system]'s capabilities meet my requirements.
- Using [this system] is a frustrating experience.
- [This system] is easy to use.
- I have to spend too much time correcting things with [this system].

The questions are accompanied with a five or seven point Likert scale, where the choice alternatives ranges from "Strongly agree" to "Strongly disagree" [72].

4.7.7.1 Calculate Usability Metric for User Experience Score

To be able to compare UMUX scores with the System Usability Scale, the formula for Usability Metric for User Experience (UMUX) scale can be used:

$$\text{UMUX} = \frac{((\text{UMUXitem1} - 1) + (\text{UMUXitem3} - 1) + (7 - \text{UMUXitem2}) + (7 - \text{UMUXitem4})) \times 100}{24} \tag{4.1}$$

4.7.7.2 Interpretation of UMUX score

The framework developed by Bangor evaluates and interprets the UMUX questionnaire scores for each participant, providing a understanding of their usability implications [73]. The interpretation can be seen in figure 4.1.

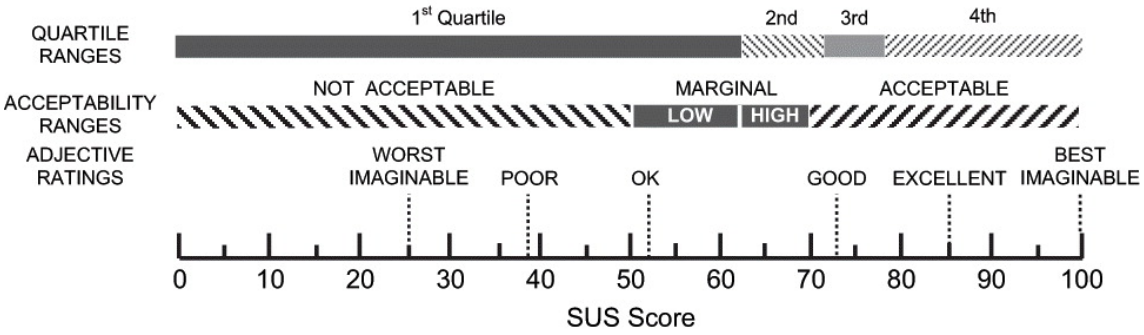


Figure 4.1: SUS Score scale.
Copyright 2008 [73]. Reprinted by permission of Informa UK Limited, trading as Taylor & Francis Group.

5

Planning

In this chapter, the initial planning and structure of the project are described.

5.1 Timeline

The execution of the project were planned to follow the initial timeline seen in the Gantt-chart in figure 5.1.

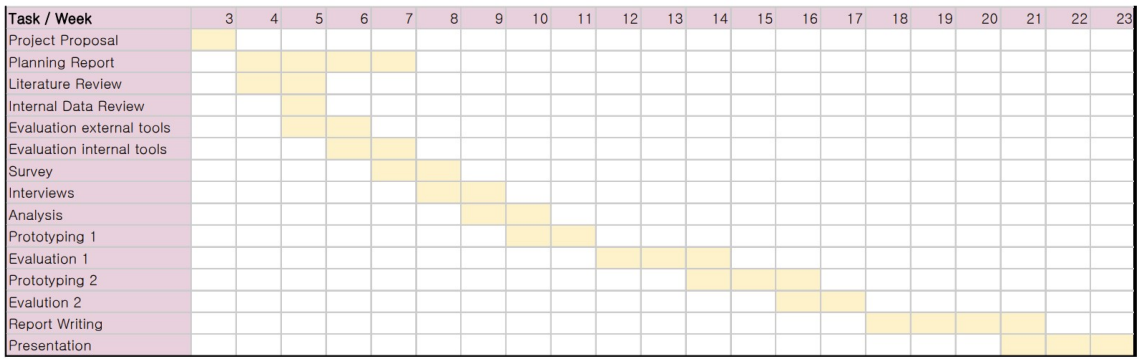


Figure 5.1: Gantt-chart with initial timeline

The purpose of the timeline was to create a visual overview of the project, divided into weeks. It gave an understanding of how much time was reasonable for each part of the process and to adapt the parts accordingly. Each task were then further defined and divided into smaller parts over the course of the project.

The literature review, internal data review, and evaluation of tools were planned to get an overview of the project and to define its scope. A survey, interviews, and analysis were then planned with the purpose to get a deeper understanding of the current application and users. Later, the creation part, including prototyping and evaluations, were planned to visualize the concept, which then could be finalized into a report and presentation.

5.2 Milestones

Related to the timeline, milestones for the project were set as the following, in table 5.1. Report writing were planned to be executed during the whole project and the

time scheduled were for finalizing, as well as to create buffer time if some parts would take longer than expected.

Table 5.1: Planned milestones

Milestone	Week
Project proposal	3
Planning report	4-7
Literature review & Internal data review	4-5
Evaluation of tools	5-8
Survey, Interviews & Analysis	7-10
Prototyping & Evaluation 1	10-14
Prototyping & Evaluation 2	14-17
Report writing	18-21
Presentation	21-23

6

Process & Execution

In this chapter, the process and execution of all methods are described, including the process description, internal data review and external tool analysis, survey, interviews, mood board, sketching, wireframing, prototyping, and evaluation. During this project, an iterative approach has been adopted, allowing for multiple revisions and refinements. The users have been included throughout the design process, involving their perspective and values. The steps of iteration will be described for each section it covers.

6.1 Process

The process of the project had initial set milestones, which during the execution of the project were adapted and turned into the list seen in table 6.1.

Table 6.1: Milestones

Milestone	Week
Project proposal	3
Planning report	4-7
Literature review & Internal data review	4-5
Evaluation of tools	5-7
Survey	7-8
Ideation & Interviews	9-11
Analysis	10-13
Insights	14
Sketching & Wireframing	15
Prototyping	16-17
User testing & Evaluation	18-19
Report writing	19-21
Presentation	21-23

6.2 Internal data review

To understand how Research Assistant is structured and what technology lies behind it, internal documentation was reviewed.

Internally at the company, earlier research of Research Assistant had been conducted. This included fault reports of bugs and written feedback from users that had been collected earlier, which were read through. Some of the information had been gathered in the form of bug reports, where users voluntarily submitted written feedback through the use of the application.

Previous interviews had also been conducted, but they were scattered in different formats without any coherent conclusion. The documentation about the previous research could be found in 9 Figma files, 14 Miro boards, and several reports, as well as in videos.

The insights that could be extracted from the documentation were that the users expect comprehensive responses, that they were struggling with understanding application modes, and that they benefit from being guided on how to structure their searches. However, since the documentation was done previously and did not show a coherent conclusion, the outcome of the internal data review was that new exploration and research were needed, where it was essential to approach the issue with an open mind and approach the problem from an unbiased and foundational standpoint, to objectively analyze the problem. This approach facilitated the exploration of new pathways without preconceived notions.

6.2.1 Original interface

Below is the current interface of Research Assistant. The homepage consists of a top bar with several buttons, some leading to a new tab, while some opens up a pop-up in the application, see figure 6.1. It also has a text box, three suggestions for questions, three modes to choose between, and a few warnings about the applications, its limitations, and what the user needs to know.

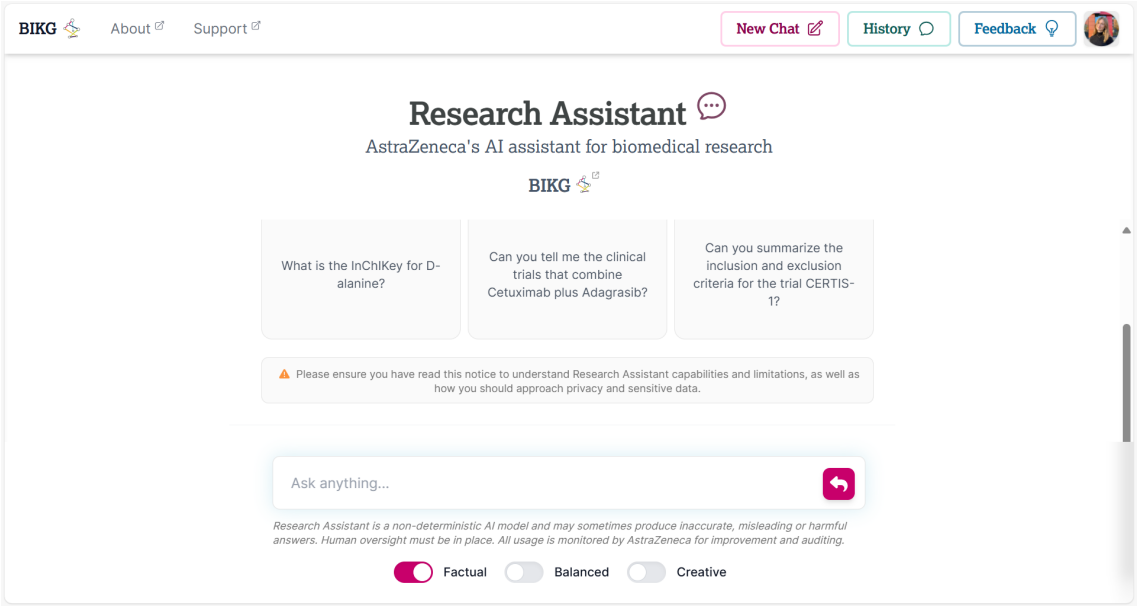


Figure 6.1: Current Homepage of Research Assistant

When receiving an answer, the current interface looks as below, see figure 6.2. In the response, there are numbers which leads to a new tab with the reference if clicked. There is also a list of the references at the bottom of the response, which consists of a heading with the publisher, as well as the start of an abstract of the article, see figure 6.3.

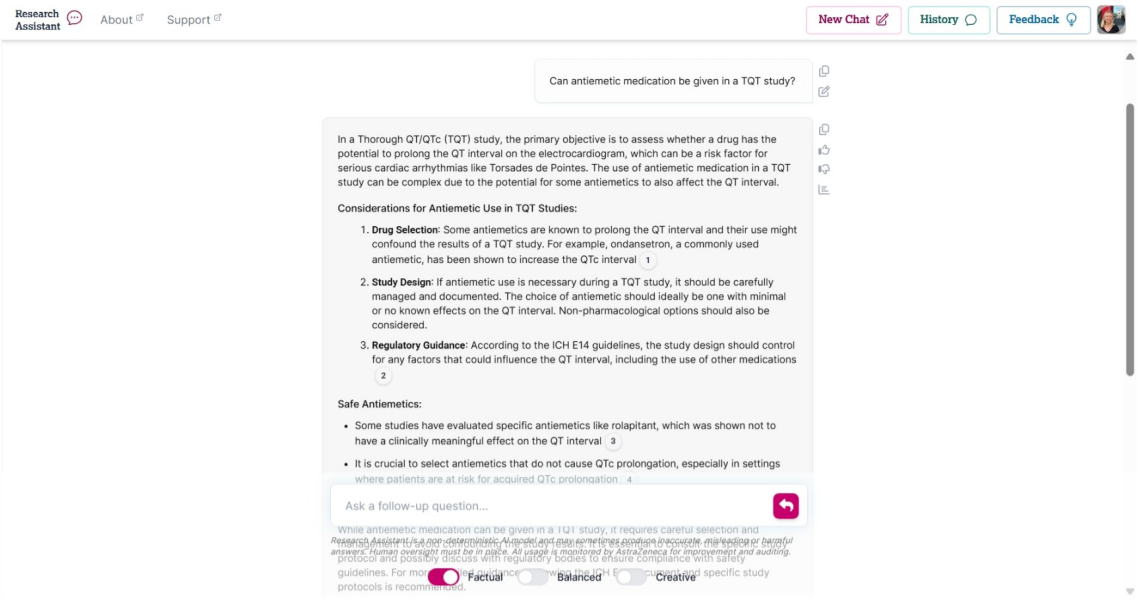


Figure 6.2: Current layout of response in Research Assistant

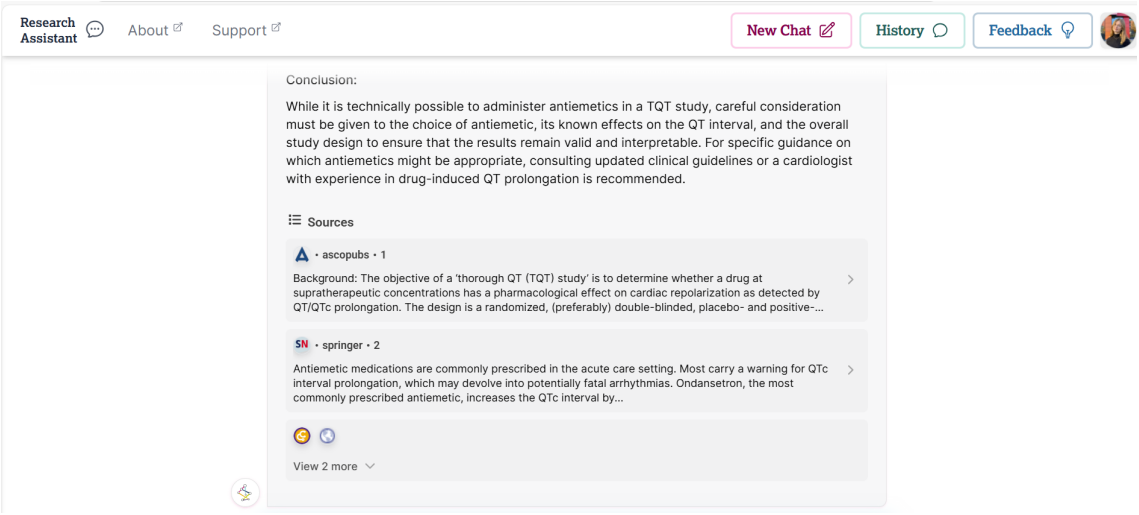


Figure 6.3: Current references in Research Assistant

6.3 Survey

A survey was created with the purpose to gain an overview of RA, where the answers would be used to find areas to explore further, see method in section 4.3. The survey was created in Microsoft Forms to make the survey accessible to all employees, and to be able to get as many users as possible to participate in the survey. Furthermore, a survey was chosen to reach out to many participants in an effective time-manner, as RA is an application that is used in several countries. The aim with the survey was to get a broad understanding of how RA is used, what is working well, and what the users are missing with the application. Therefore, several questions were kept open-ended and focused on the usage of Research Assistant, their user experience, and the interface.

Four iterations were made from draft to finished structure, which included going through the questions with the supervisor at the company and two more people within the team. The reason was to ensure that the questions created a broad understanding of the usage of the application, with benefits and pain points.

The survey gathered demographics of the participant such as gender, age, department and role, to see what kind of distribution the respondents had and if it was possible to see any specific patterns. The survey also included a section where non-users could respond, to get more information about why people did not use Research Assistant, to potentially possibly address this in our solution.

The survey included the following quantitative aspects that were measured:

- Frequency of usage
- Usage of certain functions
- The satisfaction with the current interface and the application in general

It also included the following qualitative aspects:

- Benefits and pain points
- Workflow
- Features and desired additions

The survey was published internally within the company, on internal channels with over 9000 people. The survey was shared by the product lead of the application, sent directly to 13 people working in different departments asking them to share the link with their team, and shared with a QR-code on 23 screens in 9 different buildings. Since it is an internal tool with a specific internal target group, scientists, it was not considered reasonable to open the survey for responses from external parties. It was also no additional ways to share the survey internally. For all survey questions, see appendix A.

6.4 Interviews

Semi-structured interviews were conducted for research purposes, see method in section 4.4. The interviews also included user tests with low-fidelity prototypes, see section 4.6.3. The interview questions template was iterated six times, with input from three different people, and a list with arguments and purpose descriptions for all questions was created to ensure that the interviews covered all aspects from the survey that were considered important. A pilot interview was also conducted with a person from the team, to test the structure and to estimate how long the interviews would take. The initial interviews were scheduled on different days, providing an opportunity to modify the manuscript if any issues with the interview questions were identified.

All interviews were held online to keep consistency of the interview structure. All participants received prior information to ensure they were well informed and prepared for their participation. The information contained important aspects such as time duration, request to record and if the participants could share their screen to allow for the researchers to observe their initial expressions. The full email template can be found in appendix B.

A total of 8 participants from the survey signed up for an interview, which had different level of knowledge of the application. There was also a variation in age among them, as presented in table 6.2. The duration of the interviews ranged between 25 minutes to 50 minutes. One participant was not a user of the application, while the other seven had varying experience, from only having tried it once to using it almost every day. The non-user interview included other questions than the users' interviews, and included questions about their workflow when doing literature searches, their opinion on AI tools, and how they find new tools and information internally at the company. The interviews were recorded on Teams, allowing us to transcribe and review the interview in a later stage. The transcription was made by an internal software.

Table 6.2: Age distribution of interview participants

Interview Participant	**Age [Years]**
1	30-39
2	30-39
3	40-49
4	30-39
5	50-59
6	30-39
7	50-59
8	18-29

The first part of the user interview guide included open-ended questions and depending on the answers, sub-questions could be asked. The second part of the interview included three low-fidelity prototypes of different functions, chosen from the survey findings, to inspire the interviewees about what is possible and to encourage them to share suggestions for improvement. For all interview questions, see appendix C for the script used for users of RA and appendix D for the non-user of RA.

6.4.1 Low Fidelity User Test

The low fidelity prototypes were created by using the current interface as a template. The creation of the prototypes included modes, references, and examples of wanted functions and these themes were found and chosen from the survey findings. More elements were added onto pictures of the current interface, and minor changes were made by erasing some elements and then adding new ones above it. Thereby, the low fidelity prototypes have a similar design to the current interface. The manuscript for the questions asked for the low fidelity prototypes tests during the interviews can be found in appendix C. The non-user of Research Assistant was not included in this low fidelity user test, since it was deemed to be non-fruitful. Below are the prototypes shown for the user during the interviews, see figures 6.4, 6.5, and 6.6.

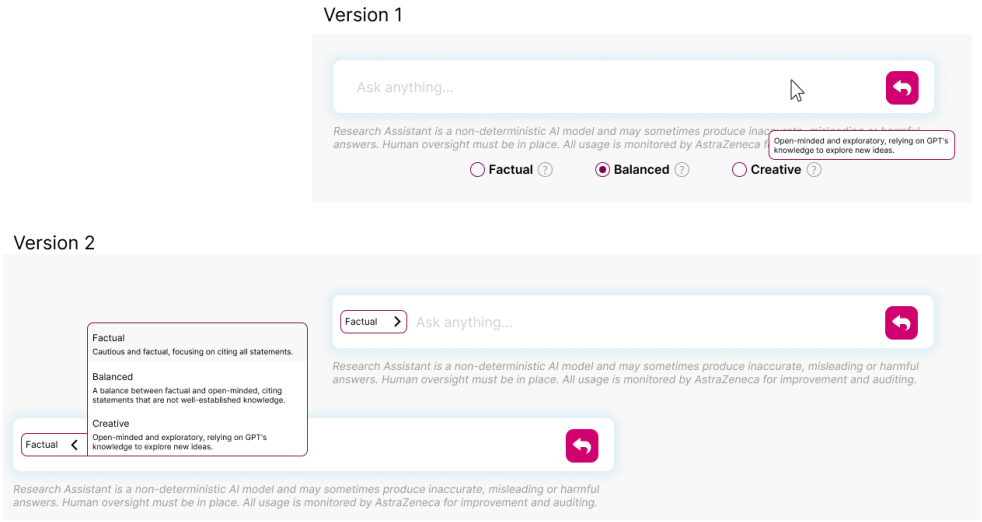


Figure 6.4: Low fidelity Modes

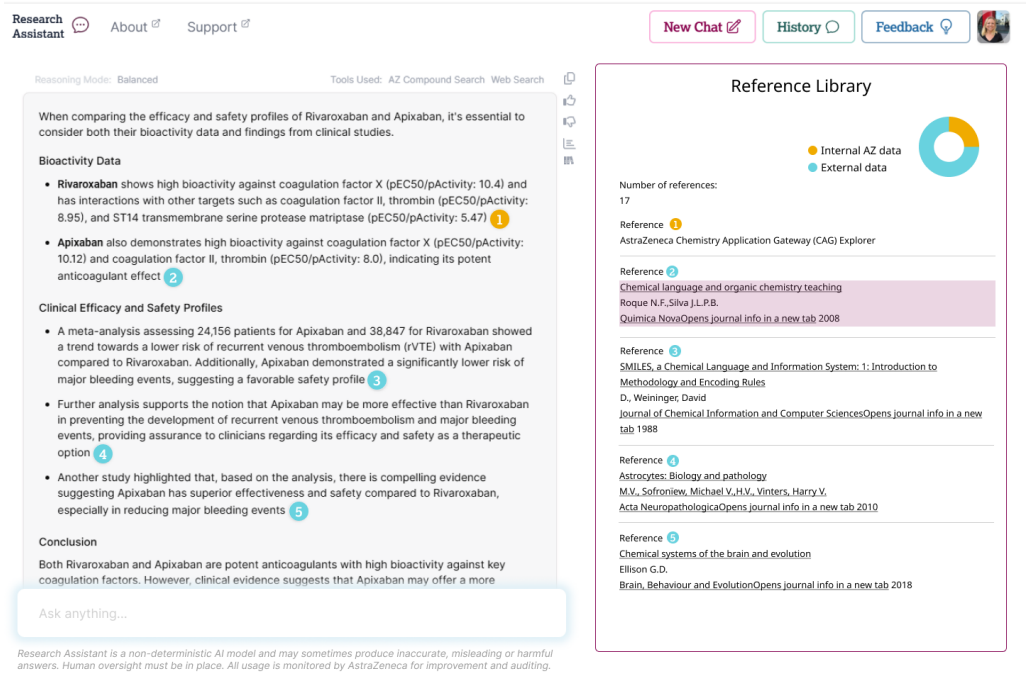


Figure 6.5: Low fidelity References

Figure 6.6: Low fidelity Functions

The insights from the low fidelity prototypes can be found in section 7.2.1 and served as a basis for the wireframing and higher fidelity prototype.

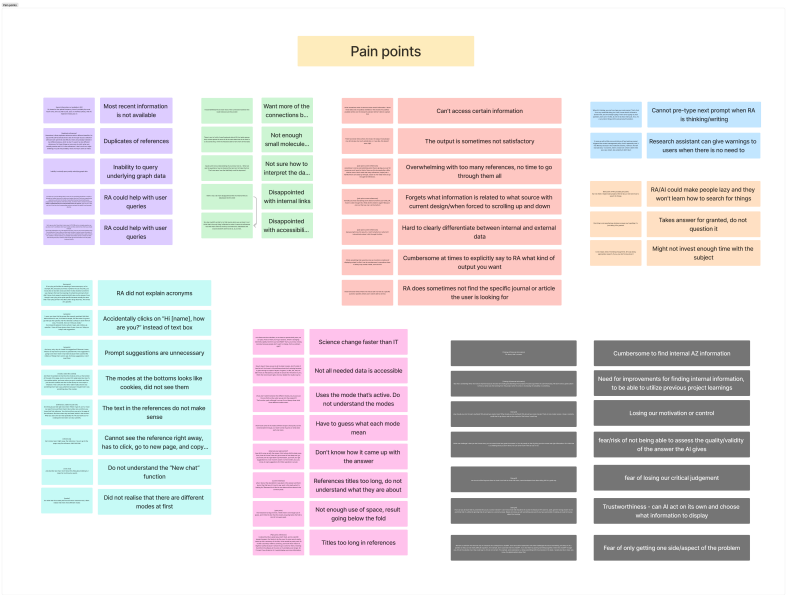
6.4.2 Interview analysis

For the interview analysis, a reflexive thematic analysis was made to structure the answers and to find insights with needs, see method in section 4.7.3. Following the Reflexive TA process, the first step was to get familiarized with the dataset, which included going through the transcript of each interview and making notes on initial observations. For each interview, coding was then done to group different quotations. These were then given initial themes, see example in figure 6.7. An deductive approach was adopted for the first stage, where it was predetermined what kind of themes the data would be categorized into. The purpose of this was to merge each interview and its individual thematic analysis into a uniform analysis in which all participants interview responses would be gathered. Without this deductive approach, it would be very difficult to compare the answers and find shared similarities across the interviews.

Figure 6.7: Interview 4 Thematic Analysis

Data from all interviews were then gathered and divided into the initial themes, to get an overview of the data, see example of structure in figure 6.8.

(a) Use cases



(b) Pain points

Figure 6.8: Use cases and Pain points - initial themes

The sticky notes were then divided into smaller groups and initial subthemes. An emergent approach was adopted, allowing codes and themes to naturally arise from the data rather than relying on preassigned ones. Allowing for emerging codes and themes, enabled a data-driven, flexible, and explorative approach. It was data-

driven because the codes were grounded in data, and the flexibility allowed to remain open for unexpected themes that did not align with initial expectations, reducing researcher bias. The sticky notes kept the color of each participant to give an overview of how many people mentioned the same aspects, meaning that each color represents a certain participant, see figure 6.9. By going through all sticky notes and themes iteratively, the themes were refined and defined into a final version. An overview of the final version can be seen in figure 6.10 and more details about the themes can be found in section 7.2. Additional figures of the final version can be found in appendix E.

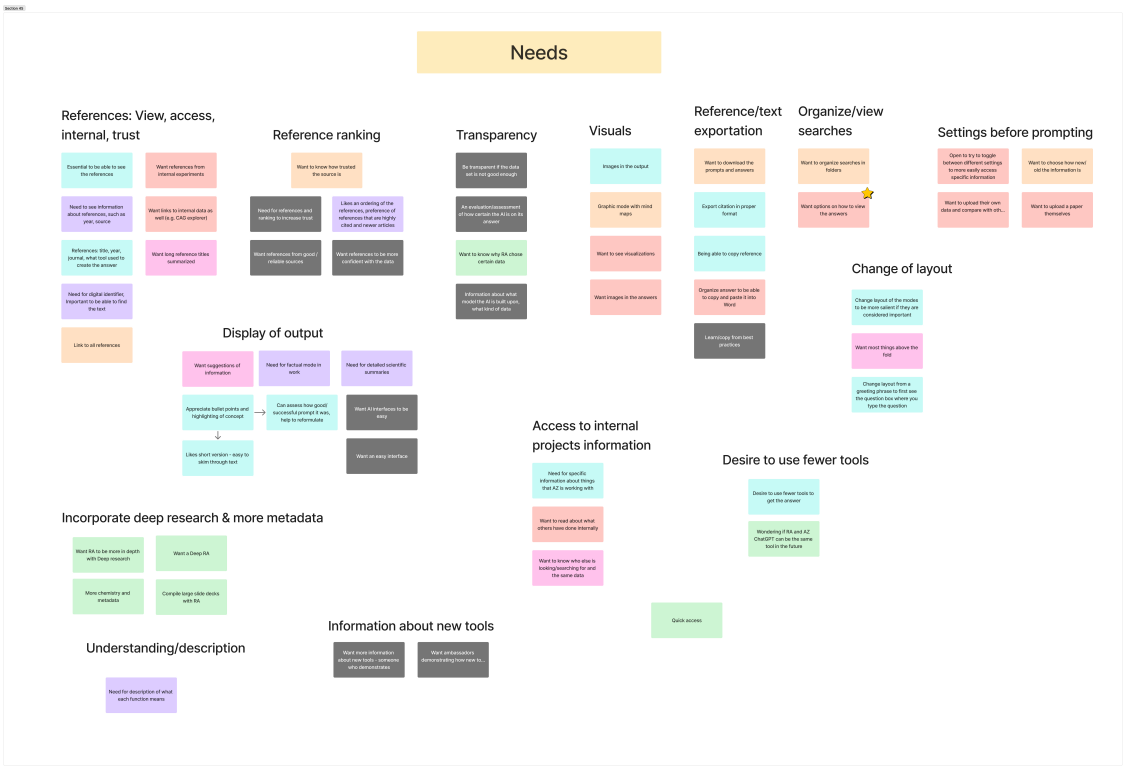


Figure 6.9: Step 3: Reviewing of themes and subcategorization of Needs.

Figure 6.10: Overview of the thematic analysis of interviews

To understand the user needs in a structured way and to clarify why they have certain needs, a list was created to summarize them in a comprehensive way. By going through the sticky notes and looking for what the user need, categories could be set up in the list. From the sticky notes, it could also be identified why the users have that need. See part of the list in figure 6.11 and the full list in appendix F. The list consisted of 14 needs categories, including the following:

- Tools for gaining more knowledge
- To get summaries
- Writing aid
- References / Verification
- Up to date information
- Clear communication
- Visuals

- Exportation
- Organization
- Prompt settings
- Save time - fast tool
- Modes
- Minimalistic layout
- Level of depth

What does the user need?	Why?
Tool for gaining more knowledge	To understand the project better
	To get background information
	Find new angles for projects
	How things are interconnected
	Explore targets
	Get additional guidance beyond the summaries
To get summaries	Save time
	Get inspiration for literature
	Get familiar with topics
	Not have to read large documents
Writing aid	
References / Verification	See if information is reasonable
	See if information makes sense
	To find additional information
	To find previous references by explaining what it was about
	Too many references can be overwhelming
	Reference titles are sometimes too long
	Want to distinguish internal and external data
	Want reference right away, not having to click and copy in new d
	To see the reference without scrolling down
	To see links to internal information as well
	Links to internal data
	Title, year, journal
	Digital identifier to find the text/info
	Ranking of references
	Highly cited references to increase trust
	Newer articles to increase trust
	Reliable sources to increase trust

Figure 6.11: List of Needs and Why's

6.5 Insight cards

To have a basis to start with for the beginning of the design phase, insights were made from the thematic analysis of the interviews, as well as from the survey, in the form of insight cards. The purpose was to get clarity on what to design, to fulfill the users needs based on what had been found from the interviews and survey. The insight cards stated what the user needs, a list describing why, gave a design recommendation on how to satisfy these needs, and had quotes to strengthen them, see example of structure in figure 6.12. The structure was decided to align with an internal company structure and to give a clear summary and overview of the insights from the interviews and survey. The quotes were added to show how the insight was formed from user opinions, the why was added to strengthen and support the need, and the design recommendations were added to provide clear guidance for the next steps in an open manner. For all insight cards, see appendix G.

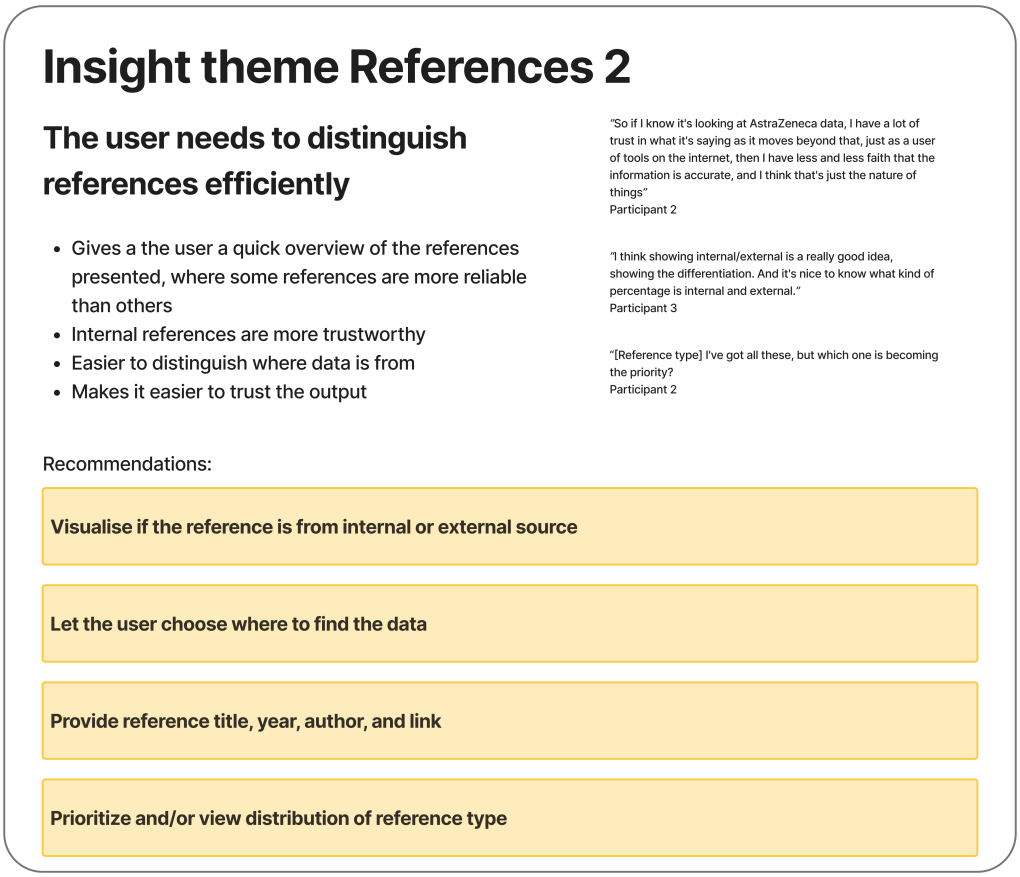


Figure 6.12: Insight References 2

6.6 External tool analysis

An external tool analysis was conducted in order to understand what currently is on the market in terms of conversational agents. By reading about the applications, trying out different functions, and inserting pictures of the different functions together with short descriptions in Figma, inspiration was received in terms of what features are in use and how these have been implemented in different applications. Six different conversational agents were examined, chosen for that they were either industry leaders, there was a buzz about them, or they were made by large tech-leaders, together with the requirement that they should be free to use. The conversational agents that were examined were ChatGPT, Claude, Copilot, Gemini, Notebook LM, and Perplexity. See example of external tool analysis in figure 6.13.

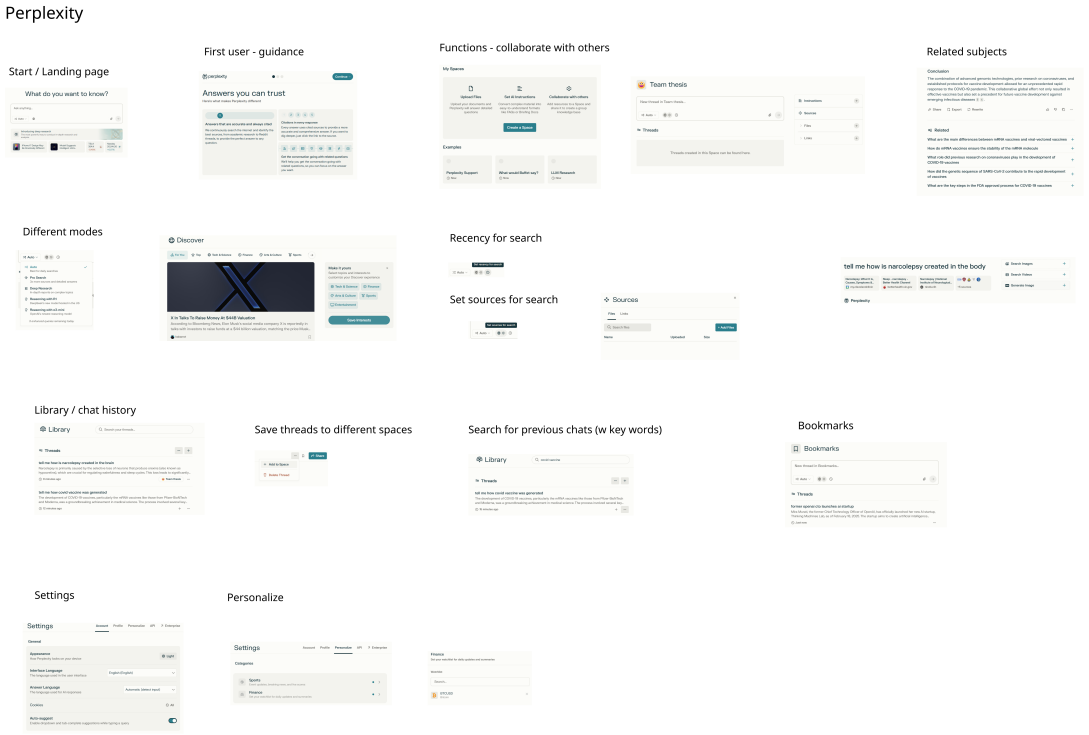


Figure 6.13: External tool analysis of Perplexity

6.7 Mood board

A mood board was created to communicate the style of the upcoming design, see method in section 4.5.3. It was also used to align the design with the company's and the department's visual identity. The mood board was created by browsing through different external sources, and by finding inspiration from the external tool analysis. Several entities were added to create an overview of the vision, including colors, inspiration from other applications, and descriptions of functions. The mood board can be seen in figure 6.14.

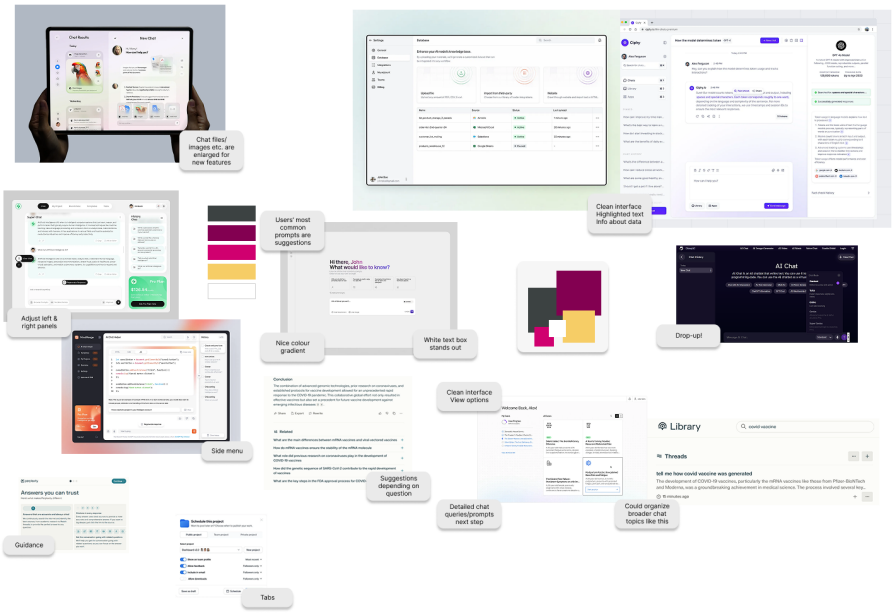


Figure 6.14: Mood board

6.8 Prototyping

The insights from the survey, interviews and thematic analysis, and the mood board served as the basis for the start of the prototyping phase, which is presented below.

6.8.1 Sketching

With the design recommendations on the insights cards, the phase of sketching started, see method in section 4.6.1. The design recommendation guided the sketching of the design to set some boundaries of what to include in each design solution. Sketches were made of all of the 10 insights cards and their design recommendation with the Crazy 8 method, see section 4.5.1. If some design recommendations could be combined, they were explored simultaneously, which turned the insights card into sketches divided into six categories:

- References 1
- References 2
- Modes 1 + 2
- Transparency 1
- Transparency 2
- Communication 1 + 2 + 3

This created 96 sketches that explored solutions for the different insights. An example of sketches for references 1 can be seen in figure 6.15, and for references 2 in

figure 6.16. They explored different solutions, for example, on how to interact with buttons, how to structure a list of references, and how to view both the response and references.

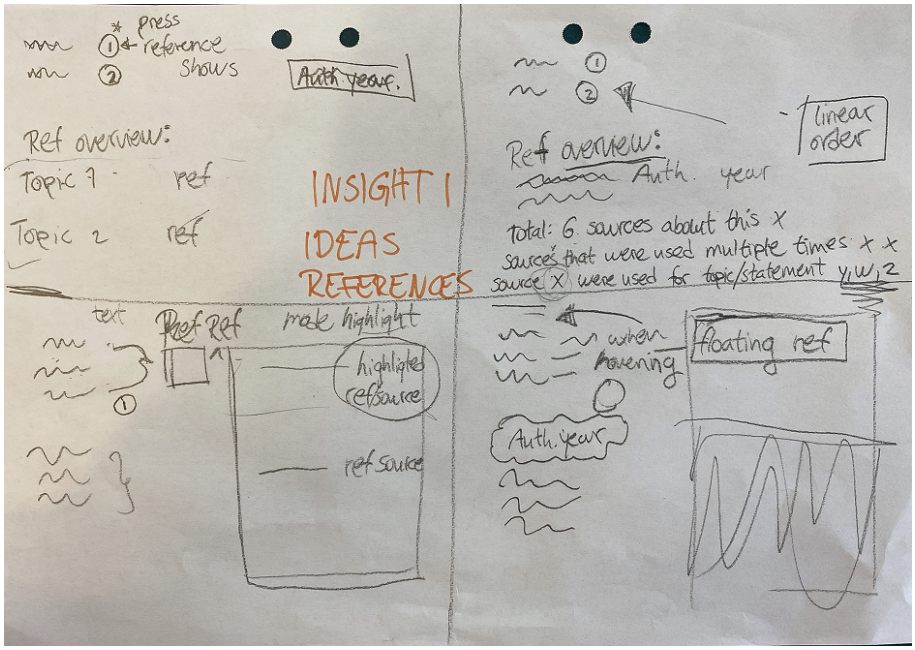


Figure 6.15: Sketches of References 1 insight

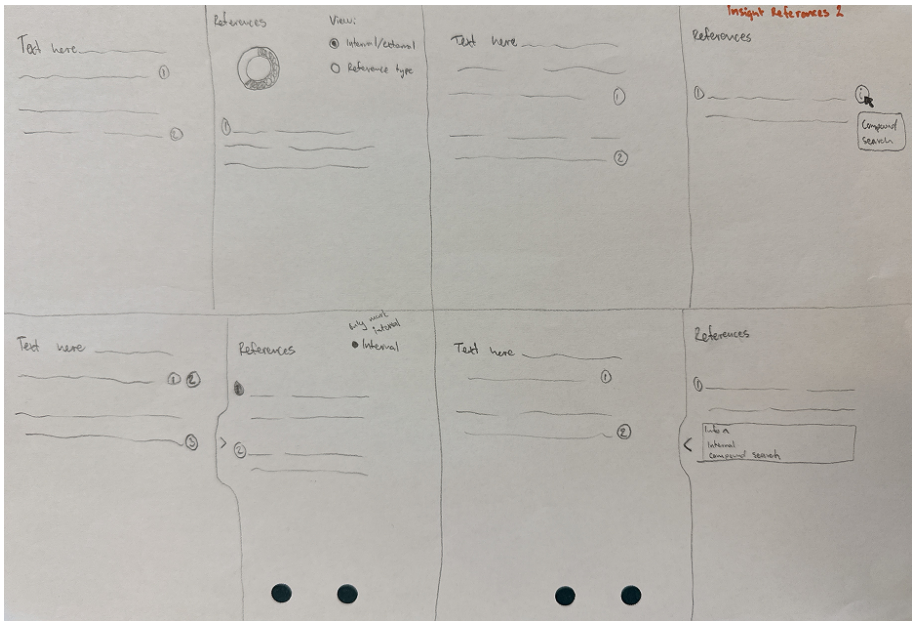


Figure 6.16: Sketches of References 2 insight

In figure 6.17, an example of sketches for insight Modes 1 and 2 can be seen. Solutions for how to visualize, for example, current modes were explored, as well as the option to not have modes, have other types of settings.

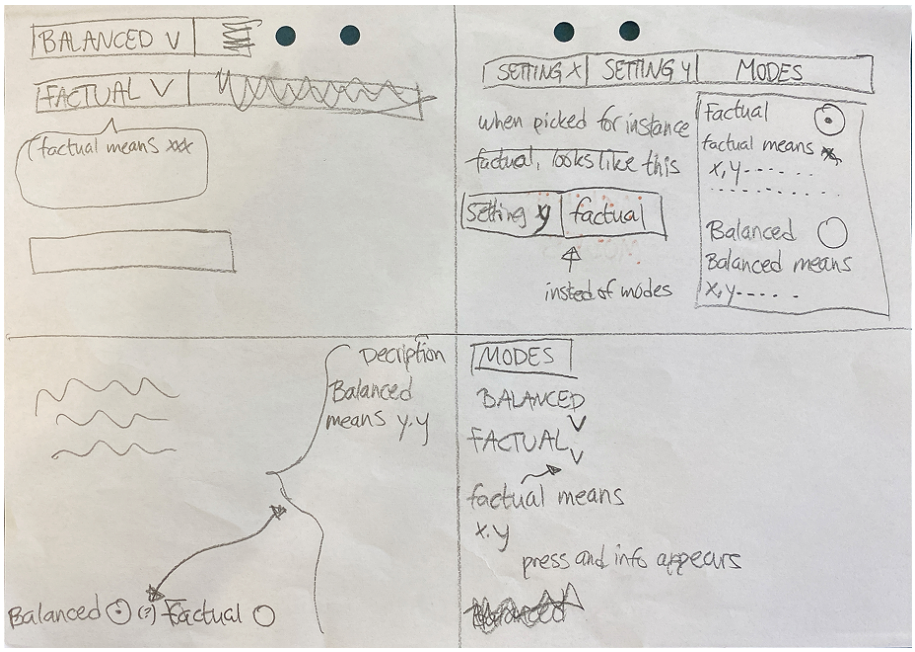


Figure 6.17: Sketches of Modes 1 + 2 insights

For the two transparency insights, it was explored how to inform the user about limitations and possibilities, how to give the user an understanding of the answer procedure, and how it can provide examples of follow-up questions. An example of transparency 1 sketches can be seen in figure 6.18 and of transparency 2 in figure 6.19.

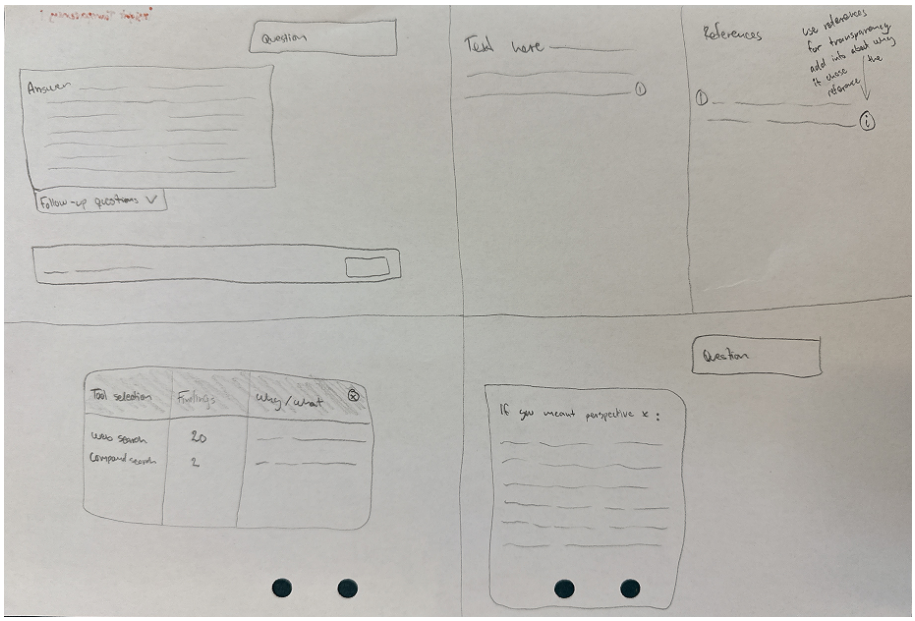


Figure 6.18: Sketches of Transparency 1 insight

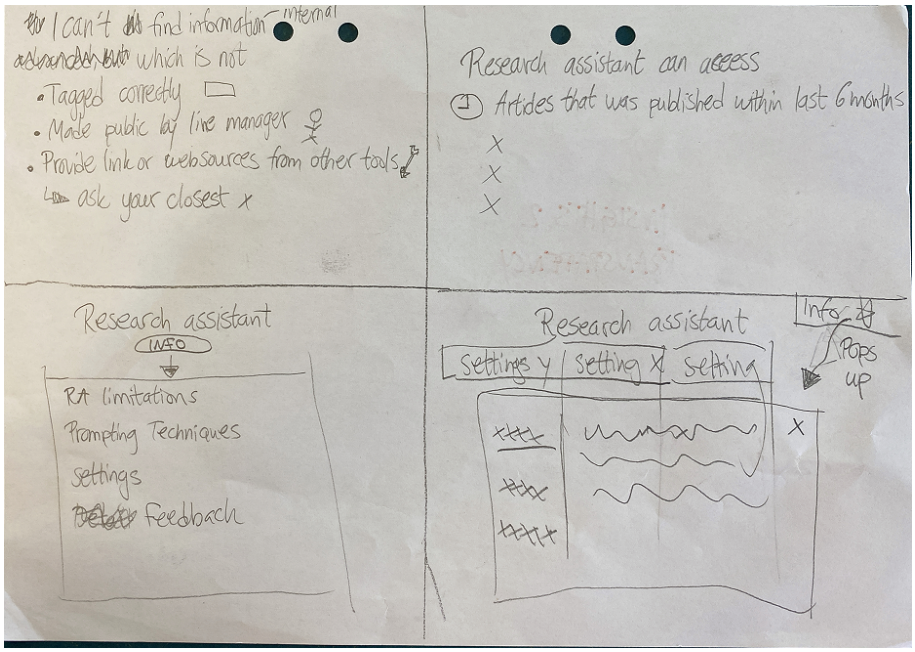


Figure 6.19: Sketches of Transparency 2 insight

The three communication insights were put together to explore solutions regarding coherency with buttons, pop-ups, and other elements, visuals responses from the application, and clearly provided data. In figure 6.20, sketches of communication insights can be seen. For all sketches, see appendix H.

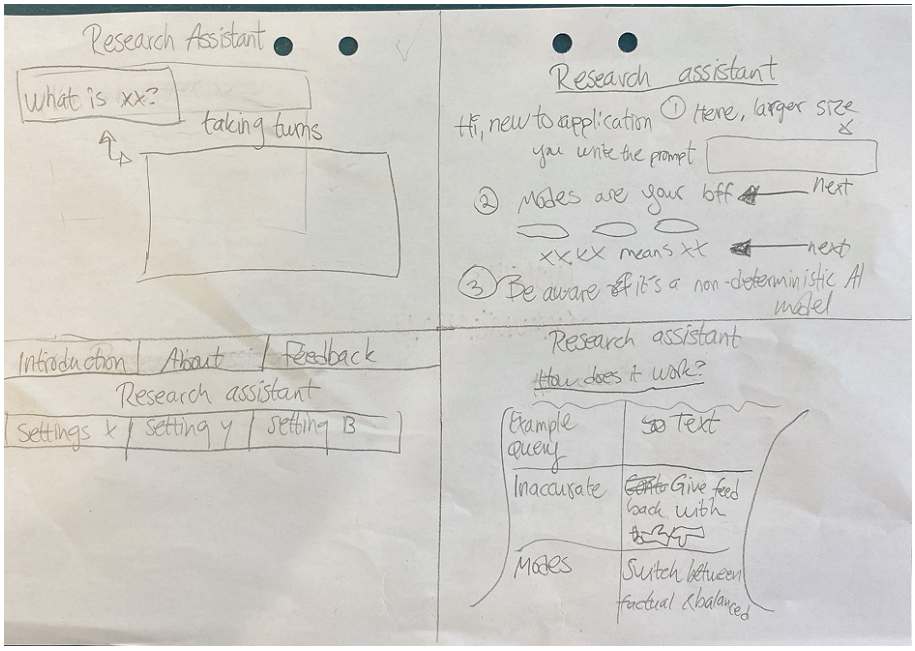


Figure 6.20: Sketches of Communication 1 + 2 + 3 insights

6.8.2 Wireframing

In order to choose what sketches to make wireframes of, the method dotvoting was utilized, see methods in section 4.6.2 and 4.5.2. The ideas that were considered to be a suitable solution for the insight cards got a star, and the ideas with stars were developed into simple wireframes in Figma. This resulted in 60 separate wireframes, all connected to the sketches and insights to various extents, where some were several versions of the same idea but a slightly different way to visualize it. The wireframes that were chosen to turn into the design were chosen together with the department team, by seeing which ones that explored and combined the most recommendations from the insight cards. An example of the wireframes can be seen in figure 6.21. More wireframes can be found in appendix I.

Figure 6.21: Wireframing Transparency 2

6.8.3 Design of New workflow

The design of the application was prototyped in Figma from the wireframes, see method in section 4.6.3. The wireframes laid out the structure and layout of the design, and which features to implement. The mood board and external tool analysis served as inspiration to the higher fidelity design, which was created for the user testing.

The new Homepage can be seen in figure 6.22. To see information about the application, an information-button is added in the side panel, which opens up as

seen in figure 6.23.

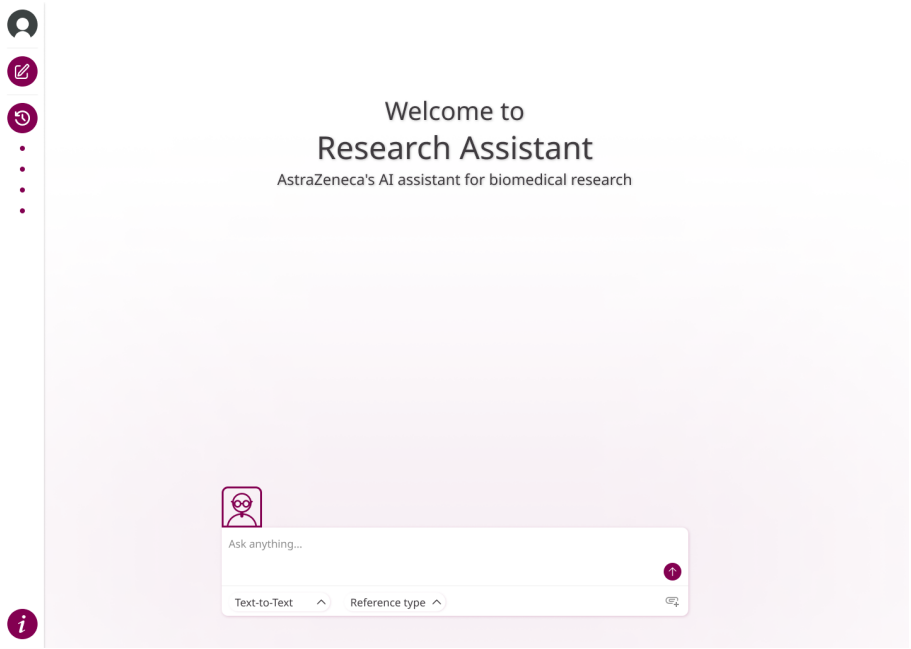


Figure 6.22: New Homepage of RA

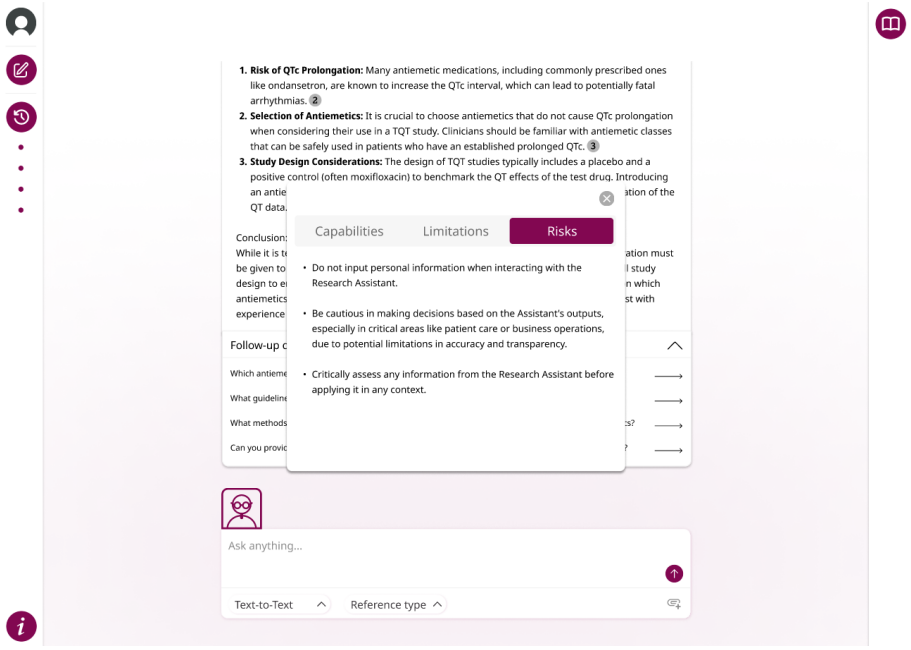


Figure 6.23: Information pop-up

The new response when asking a question can be seen in figure 6.24. The right panel can be opened up to see the references side by side. Follow-up questions have been added as well, seen in figure 6.25.

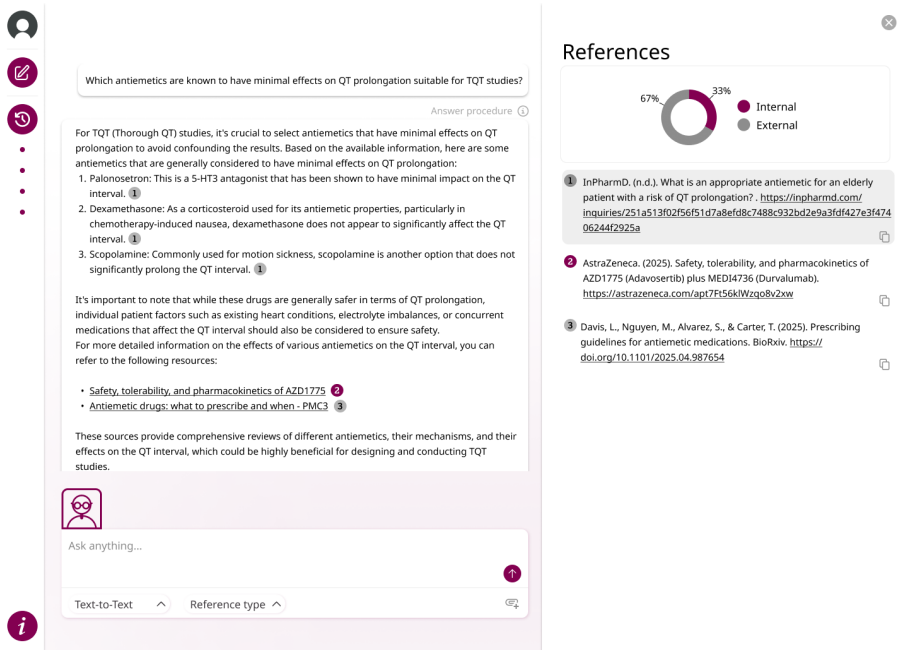


Figure 6.24: Response to prompt with references

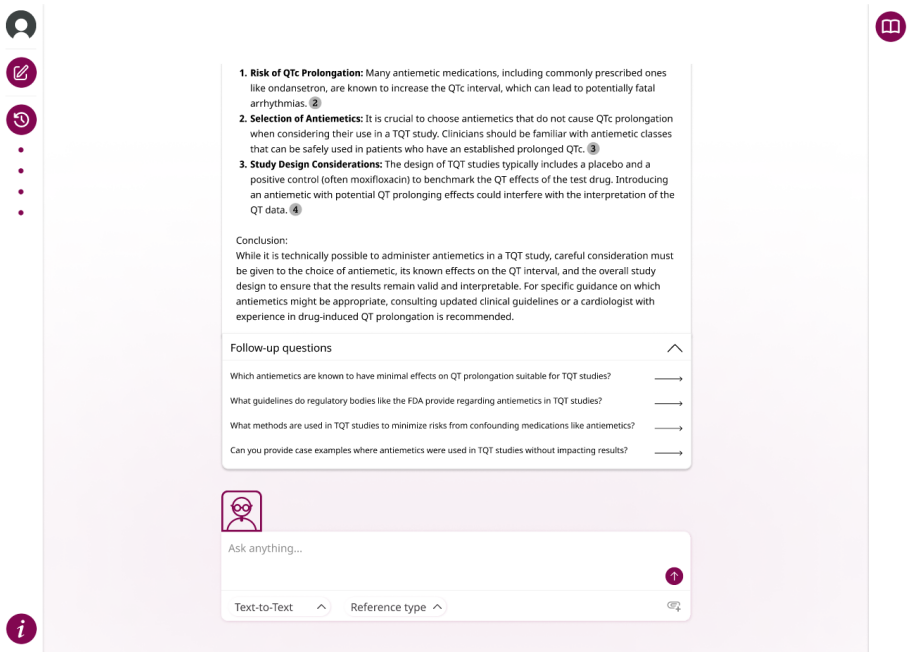


Figure 6.25: Follow up questions

Models were added instead of modes and text-to-image, as well as an attachment, can be seen in figure 6.26. The visual response can be seen in figure 6.27.

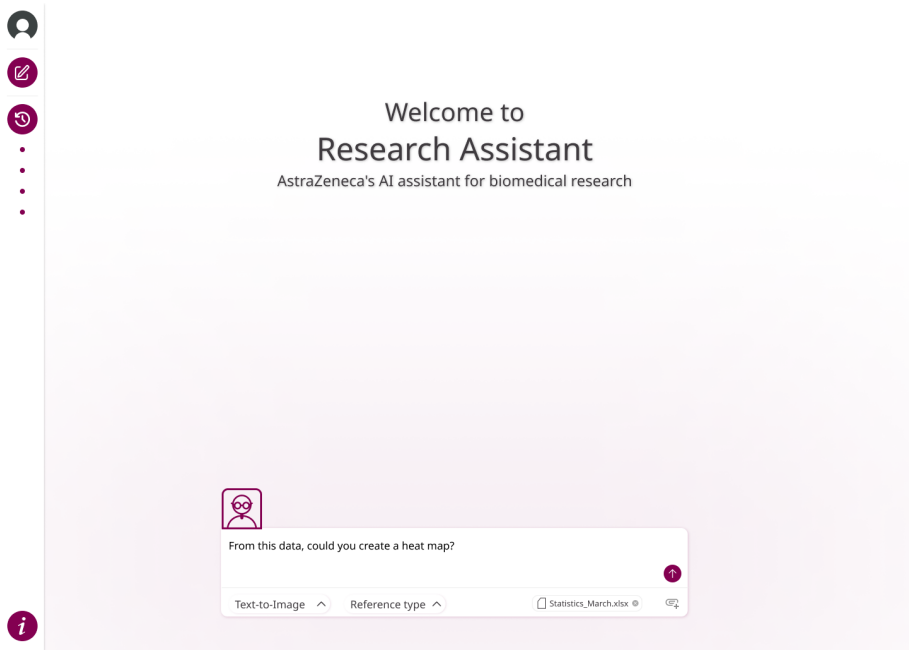


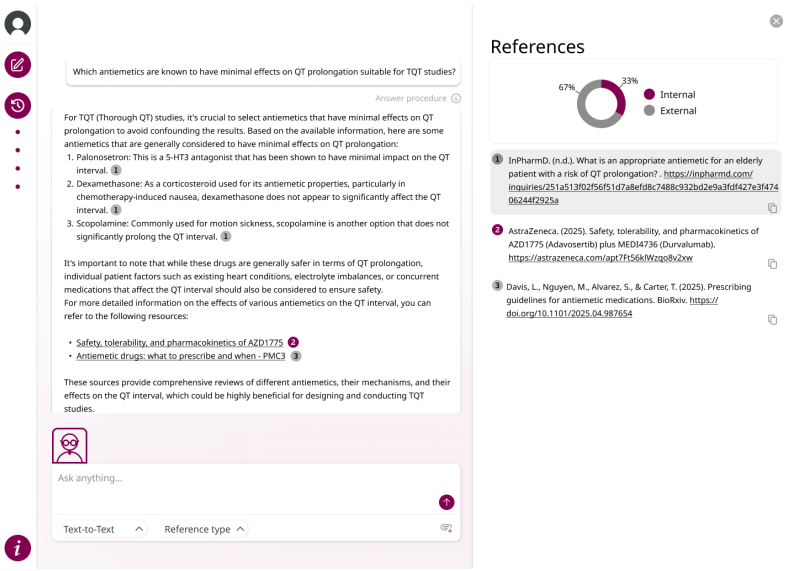
Figure 6.26: Text to image

Figure 6.27: Text to image answer

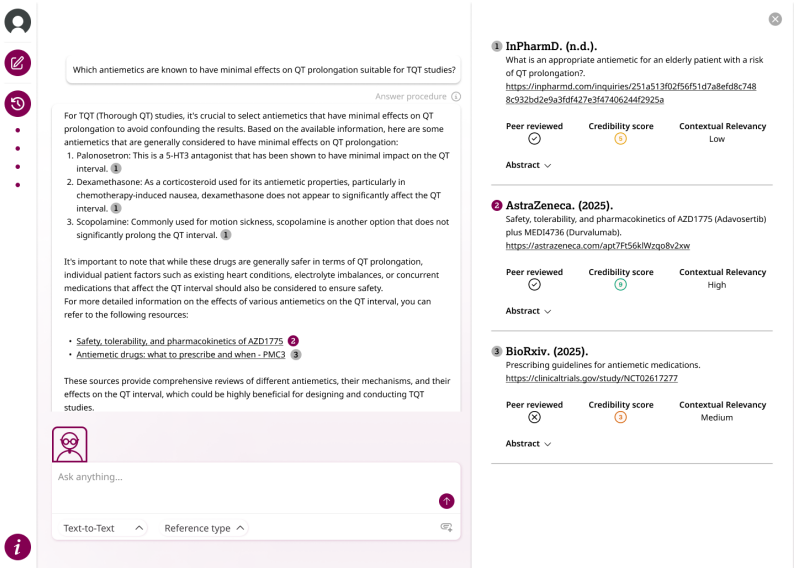
6.8.4 Design of A/B-tests

The prototype also consisted of two versions of references and modes, which were used for A/B-testing. A/B-testing was chosen to do to be able to compare several alternatives and to evaluate as many functions as possible. This was due to the participants in the interviews being very positive, and therefore there was a

need to explore broadly to be able to find what areas of improvement that could generate most user satisfaction. To choose which solutions to compare during the A/B-testing, a meeting with the department team was held, where initial alternatives were discussed and iterated into the chosen alternatives, see method in section 4.7.5. The A/B-testing for references that was selected can be seen in figure 6.28.



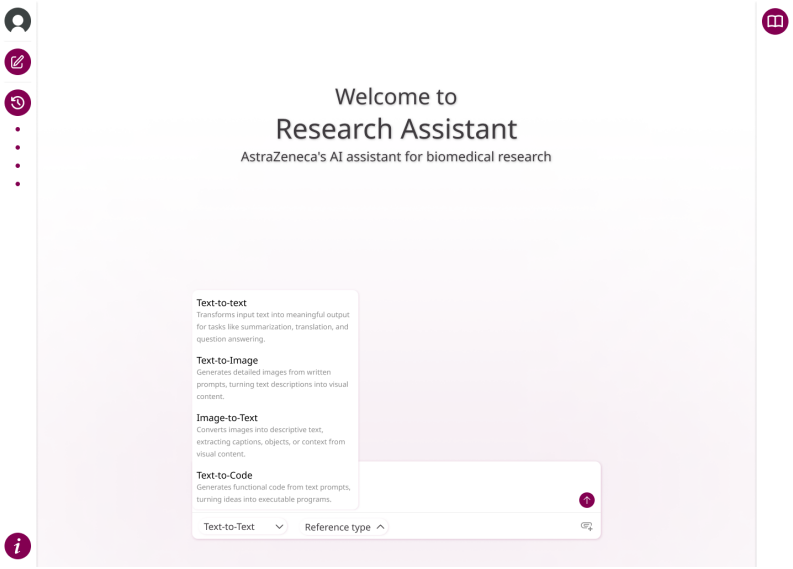
(a) Reference A



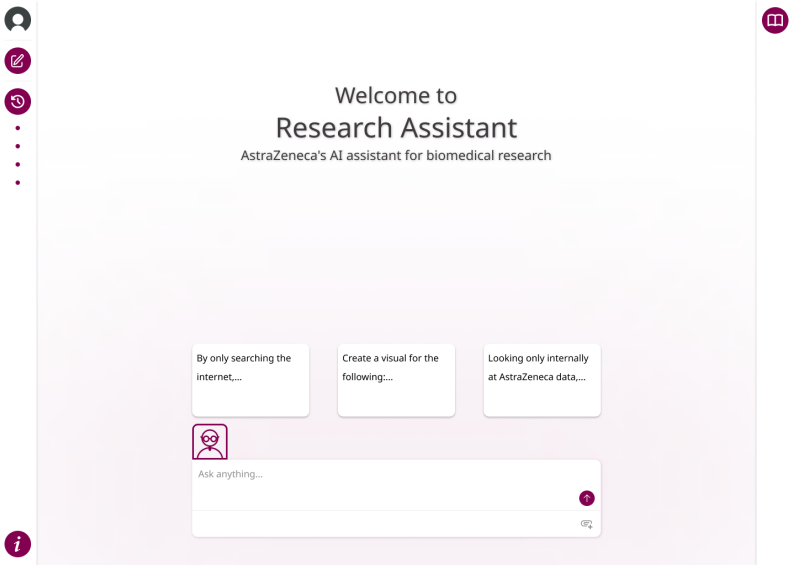
(b) Reference B

Figure 6.28: A/B-testing of References

A/B-testing for modes consisted of alternative A, with models and reference types, and alternative B, with prompting and suggestions of frequently asked prompt-functions. In figure 6.29, the models of alternative A is seen, together with alternative B. In figure 6.30, the reference types of Modes alternative A can be seen.



(a) Modes A - Models



(b) Modes B

Figure 6.29: A/B-testing of Modes

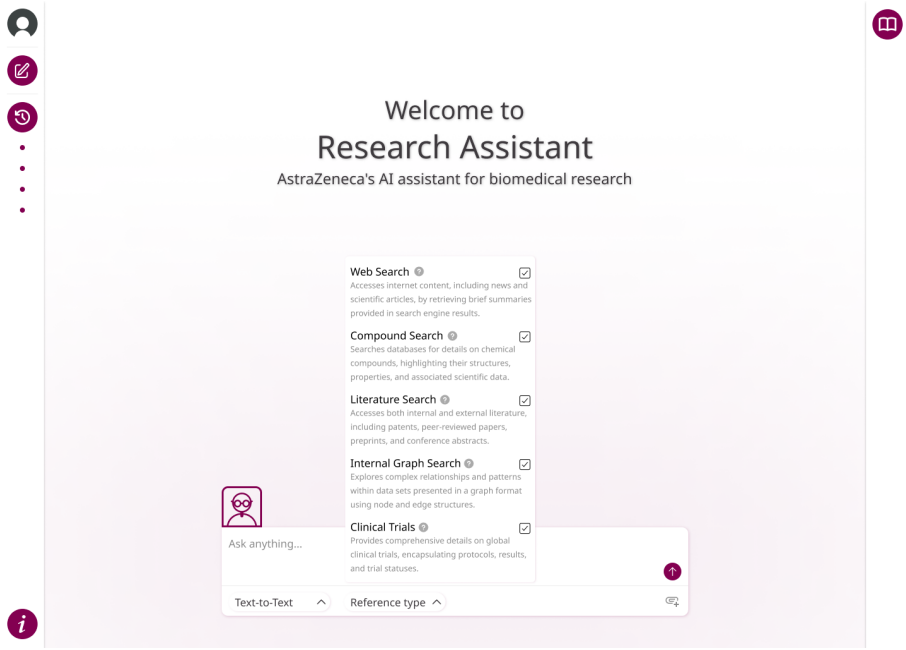


Figure 6.30: Modes A - Reference type

6.8.5 Evaluation - User tests

User testing was conducted to evaluate the design, see method in section 4.7.4. The user tests structure were iterated before they were conducted, with focus on ensuring that both qualitative and quantitative measures would be included, and that it was possible to compare the data to the survey and previous interviews during the evaluation. The structure of the user tests was developed and iterated in collaboration with the team at the company. The requests to participate in the user testing was sent out to the seven people which previously had been interviewed as RA-users. Five out of these seven requests was accepted, the other two declined. The user tests script that were used can be found in appendix K. The user tests were all held digitally, with the reason that some of the participants were located in other countries and the structure and conditions were wanted to be coherent for all participants.

To keep anonymity of participants in both the interviews and user tests, they got assigned a number. The numbering from the interviews and the corresponding participant from the user tests can be seen in table 6.3.

Table 6.3: Participants numbers for Interviews and User Tests

Interview Participant	**User Test Participant**
Participant 1	Participant 5
Participant 2	-
Participant 3	Participant 2
Participant 4	Participant 3
Participant 5	Participant 1
Participant 6	Participant 4
Participant 7	-
Participant 8	-

The user testing consisted of a mixed method approach, including both quantitative and qualitative measurements. A/B testing was applied, with a within-subjects design where all participant was shown and tested all the versions.

The user tests evaluated six different elements, where the participant was given a task to perform for each element. The participant was asked questions either during or after the task had been completed. After testing each element, the participant was asked to fill out the questionnaire Usability Metric for User Experience, UMUX. UMUX was chosen as a method to evaluate user experience since one of the factors is focusing on perceived ease of use, which was one of the essential factors for AI acceptance, see section 2.11. Furthermore, it is a time efficient method to evaluate user satisfaction. It was essential to make use of a time efficient method for the participant to have time to conduct all the tasks in each prototype and complete a questionnaire together with each prototype tested.

Outlined below is a summary of the user test, and the full script for the user test can be found in appendix K.

The structure of the user tests consisted of:

- Section 1: Conduct tasks in the current application
- Section 2: Conduct tasks in the (4) A/B versions
- Section 3: Conduct tasks in a workflow with the new design

Firstly, the participant had to go through two tasks in the current application to immerse themselves in the current interface of the application. After this, they filled out the UMUX questionnaire. The task served as a reminder for the participant, compensating for differences in their application usage frequency. It was also used to avoid the scoring being affected by how well they remembered the interface. The following four elements were the different A/B versions, which was laid out in different order for each participant to avoid the ordering effect could affect the result. The prototypes shown for the A/B test can be seen in section 6.8.4.

Order of the different A/B versions for each participant:

- Participant 1: Reference A, Reference B, Modes A, Modes B
- Participant 2: Modes A, Modes B, Reference A, Reference B
- Participant 3: Reference B, Reference A, Modes B, Modes A
- Participant 4: Modes B, Modes A, Reference B, Reference A
- Participant 5: Reference A, Reference B, Modes B, Modes A

Participants completed identical tasks when working with reference A and B, and similarly with modes A and B. However, the nature of the task varied between the references and the modes. To have identical tasks was to ensure a fair comparison between the different versions of references and modes. By keeping tasks consistent, it helps to isolate the effect of the independent variable (the interface). This makes it easier to see whether if it was the variable being tested was the aspect causing the effect rather than the change in the task.

After the A/B test was conducted, the test ended with participants completing tasks in the new workflow. The prototypes shown of the new workflow can be seen in 6.8.3.

To evaluate the quantitative UMUX scores that had been received during the user tests, the formula for Usability Metric for User Experience, UMUX, scale was applied, see section 4.7.7.1.

Each item in the questionnaire corresponds to a particular usability aspect. Specifically, Item 1 refers to Question 1 of the UMUX questionnaire, Item 2 corresponds to Question 2, and this pattern follows for Items 3 and 4.

6.9 Important factors

The important factors were created with the purpose to answer the research question. They are based on insights from the survey, interviews, and user tests, as well as design theory (see section 3.7) to find important aspects that were connected. The survey was broad and used to gain an overview of the application, and the interviews were semi-structured and in-depth. The interview analysis consisted of an in-depth thematic analysis, with insight cards as the outcome, which were iterated as well to cover all important aspects. The conducted user tests consisted of both quantitative and qualitative aspects, and by so collecting broad insights and a thorough understanding. All these mentioned methods laid the foundation for the important factors, creating an outcome based on thorough work, were each part had been iterated to ensure that the purpose was fulfilled. The outcome of each method was reviewed, to include insights from all results. By iterating this step, it could be ensured that the important insights from all of the methods were included in the list of important factors. The factors were created with a high-level perceptive, with the aim to make the aspects applicable for all conversational agents designed for internal company use.

7

Results

The following chapter contains the outcome of the project. The results from the survey, interviews, and user tests are presented, followed by important factors, the final design, and the evaluation of it.

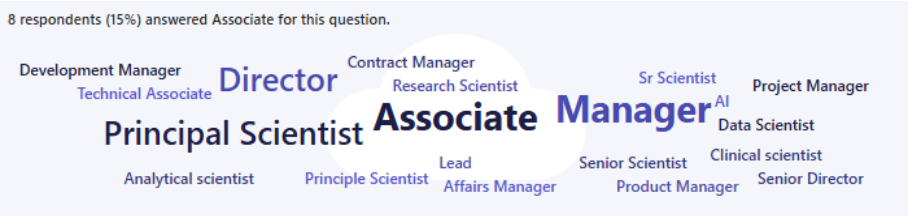
7.1 Survey

The survey gathered 54 responses from over 20 departments. The first section includes the demographics of the respondents. The respondents were almost evenly split, with 58% identifying as women and 42% as men. Participants spanned all age groups and represented various divisions and roles within the company, see figure 7.1 and figure 7.2.

(a) Gender

(b) Age

Figure 7.1: Gender and age distribution



(a) Role

(b) Department

Figure 7.2: Role and department of respondents

Non-users of RA answers

Half of the survey respondents reported not having used the Research Assistant. Of these non-users, 21% reported that they did not have a need for the tool. The majority of the non-users, 79% reported ”other”. In the text box where respondents could voluntarily choose to write to further explain the choice of ”other”, almost all of the non-users reported that they did not know that the tool existed. Furthermore, when asking when in your work the non-users could include Research assistant, the majority of the non-users were uncertain of potential use cases. The last questions included what other tools, such as AI tools, the non-users used where the majority 96% of users reported using AstraZenecas internal ChatGPT.

Users of RA answers

Respondents had varying levels of experience, ranging from never having tried it to using it almost daily. The questions with pre-written alternatives showed that 8% of RA users have only tried it once, 50% use it 1-5 times a month, 35% use it 1-5 times a week, and 8% use it almost every day. Information gathering was seen as the primary task for the application. A majority (85%) of the users identified pain points with the application, where the main pain point included lack of access to references or literature the user reported existed. The primary benefit with the application were that it saves time. Other benefits were that it increases productivity, is fast, uses scientific evidence, that the answers include internal information, and that the information is summarized, among other benefits. More than half of the respondents 58% reported not using the chat history.

There are three modes the users can switch between (Factual, Balanced, and Creative), and this function is used by 50% of the users, 19% do not use them, and 31% do not understand them. 58% of users want more functions included in the application. These included among others to be able to favorite chats, have chats organized by topic, see a metric of how reliable the sources are, and more

knowledge about its limitations.

The questions related to the interface of RA showed that respondents evaluated RA to be easy to use and were satisfied with their experience, see figure 7.3. The average rating (on a scale of 1 to 5) for how well users were able to use the application on their first attempt was 4.96.

(a) First time usage of RA

(b) User experience of RA

Figure 7.3: Interface of RA

Respondents were also asked to rate how likely they are to recommend Research Assistant to a colleague. Out of the 28 respondents of this question, 16 are marked as promoters, 8 as passives, and 2 as detractors.

Microsoft Forms created word clouds from the free text survey answers. These were put into FigJam together with the answers from the multiple choice questions, which Microsoft forms had turned into pie charts. By going through all questions, by both looking at the word clouds and by looking at and comparing specific answers, an overview of the survey results was created. The survey answers then resulted in a list of topics to explore further during the interviews:

- Workflow
- Benefits and Pain points
- Trust
- Modes
- References

- Functions including level of depth, answer length, and reference type

7.2 Interviews

From the interviews, insights were gathered through an thematic analysis. The thematic analysis resulted in 8 themes, divided further into sub themes, as well as the connected insight, see tables 7.1, 7.2, 7.3, 7.4, 7.5, 7.6, 7.7, and 7.8.

Table 7.1: Use cases theme from thematic analysis of interviews

Theme: Use cases	**Insights**
Understanding	Learn new subjects Explore and understand targets Investigate new angles to a subject
Find information	Checking if information is available Find information about prior projects
Writing Aid	Email creation with suitable tone Write documents
Summaries	Create summary tables Literature summary, get familiar with topic Inspiration tool for literature Summarization of subject, key points

Table 7.2: User workflow theme from thematic analysis of interviews

Theme: User workflow	**Insights**
References & Verification	Double checks if answer makes sense Verifies that the information is reasonable Scrolls down and uses references Verifies with other scientists
Query writing	Uses key words for context specifics Optimizing the query based on wanted aspects
Knowledge search	Checks literature and databases to find information thats unlikely to exist Follow up on the citations it provides, learning more about science

Table 7.3: Benefits theme from thematic analysis of interviews

Theme: Benefits	**Insights**
Time saving	Quick & saves time with summaries Quick & saves time with text and grammar Instant access to knowledge Quick access
Cuts down on manual work	Gives more time to compile what they have learned from literature
Understanding	Understands science Can explain in a simple way
Internal data importance	Important to access internal data Trust because of AstraZeneca data Access to internal data deemed most important
Access to sources	Satisfied with the web links Can validate the results based on the references
References	Numbers the references and lists them Efficient & reliable source information Easy access and display of references

Table 7.4: Pain points theme from thematic analysis of interviews

Theme: Pain points	**Insights**
Wanted information not accessible	Most recent information is not available Cannot access certain information Not enough small molecule data
Display of references	Forgets what information is related to what source when forced to scrolling up and down Overwhelming with too many references Titles too long in references Hard to differentiate internal and external data The text in the references do not make sense
General frustrations	Cannot pre-type next prompt when RA is thinking Prompt suggestions are unnecessary
Unclear output	Uncertain in how to interpret the data Uncertain in how to interpret the answers
Modes are unclear	Have to guess what each mode means Do not understand the modes, uses default

Table 7.5: Prompting theme from thematic analysis of interviews

Theme: Prompting	**Insights**
Common to reprompt	Rewrites prompt sometimes Gets right answer after some reprompting
How they prompt	Writes the gene and disease to get an answer
Simple prompting	Keeps it simple, 2-3 words
Being specific	As detailed as possible

Table 7.6: Attitude theme from thematic analysis of interviews

Theme: Attitude	**Insights**
Positive	Interest in learning new things
Mindset	Learning curve with using AI Don't want to combine creative AI and science
Settled mindset	Cannot ask for more Satisfied with current solution
Perception of RA	Treats it as a human Want it more as only a search tool Users treating RA as a human = creepy

Table 7.7: Needs theme from thematic analysis of interviews

Theme: Needs	**Insights**
References - View	Want references from internal experiments Want long reference titles summarized References: title, year, journal What tool used to create the answer
References - Ranking	Want to know how trusted the source is Want references from reliable sources
Transparency	Want to know why RA chose certain data
Visuals	Graphic mode with mind maps Want to see visualizations Images in the output
Exportation	Want to download the prompts and answers Export citation in proper format Be able to copy reference
Organize searches	Want options on how to view the answers
Prompt settings	Want to upload their own data Want to choose recency of information Want to upload a paper themselves
Change of layout	Make modes layout more salient
Display of output	Want suggestions of information Need for detailed scientific summaries Appreciate bullet points and highlighting of concept Likes short version - easy to skim through text
Access to internal projects information	Read about what others have done internally Who else is searching for the same data
Desire to use fewer tools	Use fewer tools to get the answer Wondering if RA and AZchatGPT can be the same tool
More data	More chemistry and metadata
Understanding	Need description of what each function means

Table 7.8: Trust theme from thematic analysis of interviews

Theme: Trust	**Insights**
References	References makes it easier to trust the output Trusts internal data more than external Trusts answer if verifying the references Need for references to be able to check every fact and thus, trust it Trust it 60%, therefore always questions it looks at the references
Verification of information	Verifies all data, since it is good practice Check if the answer seems reliable Asks other experts if the answer seems accurate Verifies if the information is accurate The quality of the information is determined by how reliable the output is

7.2.1 Results from Low Fidelity User Test

To see the prototypes shown in low fidelity user test, see section 6.4.1 and to see the full results from the thematic analysis of the low fidelity user test see appendix E.

The first prototype was modes, see figure 6.4. The feedback was that participants appreciated being able to read the descriptions of modes, which enhanced their understanding of what the mode entailed. Most of the participants preferred the drop down-version over hovering.

The second prototype shown was references, see figure 6.5. The participants were very positive to the side-by-side view, stating it was a way better solution than the current one.

The last prototype shown was functions, see figure 6.6. The participants were not fond of the level of depth nor the length of answer, but showed curiosity towards reference type.

7.2.2 Insight cards

The created insight cards included three cards for the theme References, two for Modes, two for Transparency, and three for Communication. The themes were chosen to cover all insights from the interviews and survey in a summarized and structured manner. For all themes and descriptions, see table 7.9.

Table 7.9: Categories and descriptions from the insight cards

Category	Description
References	The user needs references The user needs to distinguish references efficiently The user needs to be able to reference the source in other tools
Modes	The user needs guidance in how to interpret the modes The user needs guidance in how to use the modes
Transparency	The user needs information about how it created its answer and guidance in how to interpret it The user needs to understand what limitations there are with the application
Communication	The user needs clear communication of the interface The user needs a visual way to interpret the information The user needs clear communication of the information

7.3 User tests

The user tests consisted of a quantitative and a qualitative part, and the results from them are presented below.

7.3.1 Quantitative part

The following illustrations (figures 7.4, 7.5, 7.6, and 7.7) depict the scores obtained from participants who completed the Usability Metric for User Experience (UMUX) questionnaires. Each illustration highlights both the mean (indicated by an ”X”) and the distribution of scores.

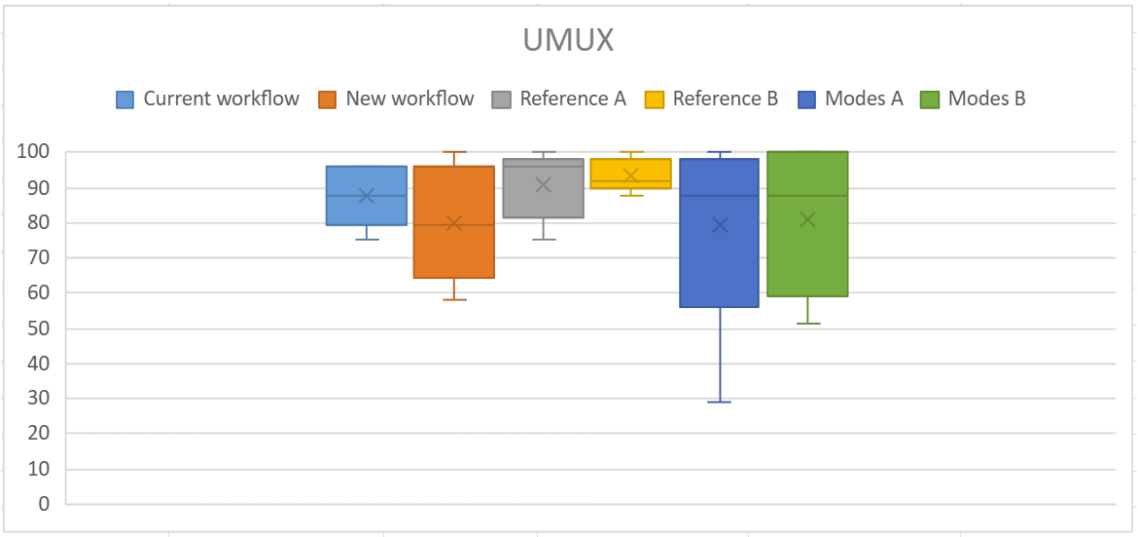


Figure 7.4: UMUX scores for all items

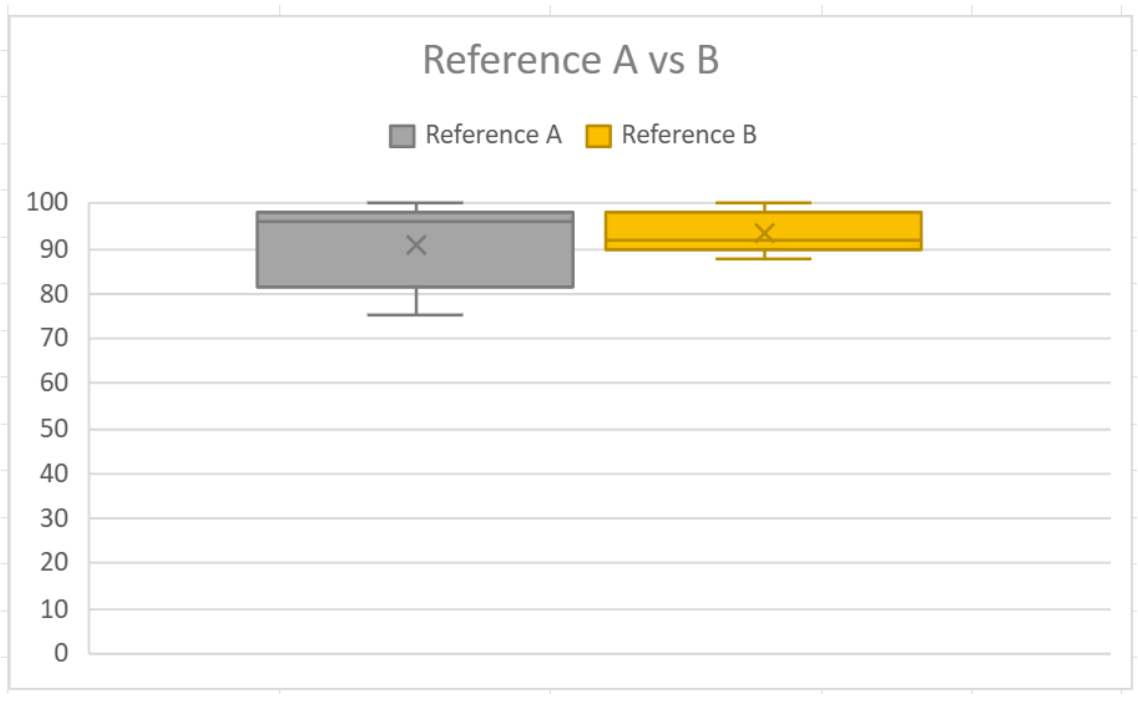


Figure 7.5: UMUX scores for references

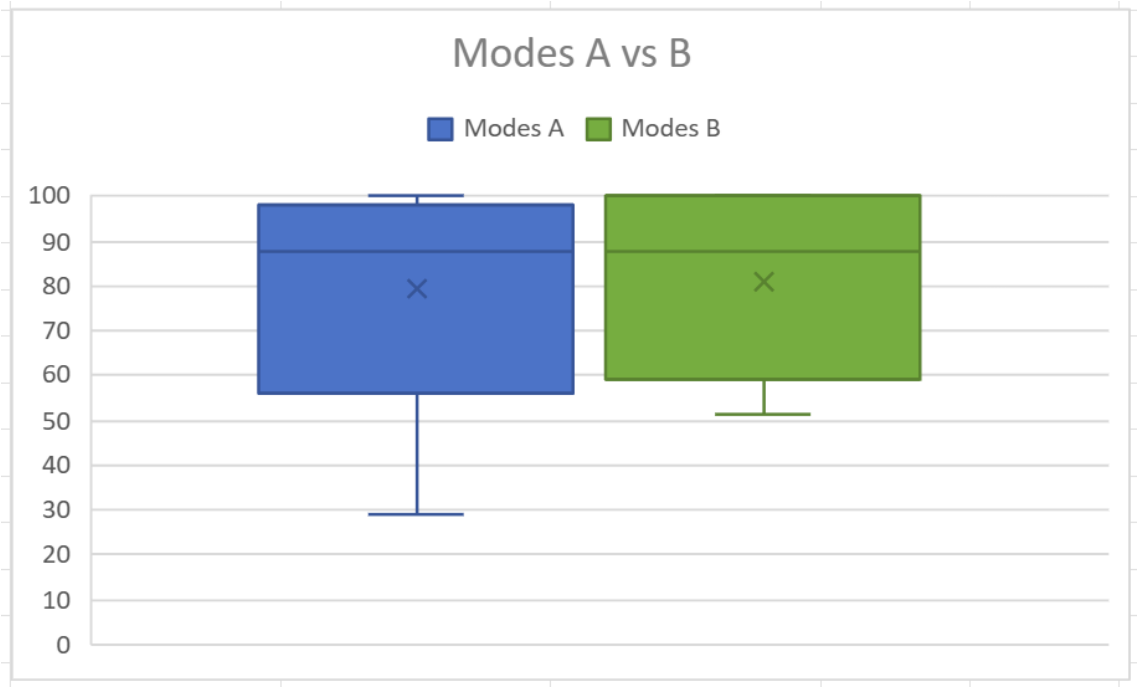


Figure 7.6: UMUX scores for modes

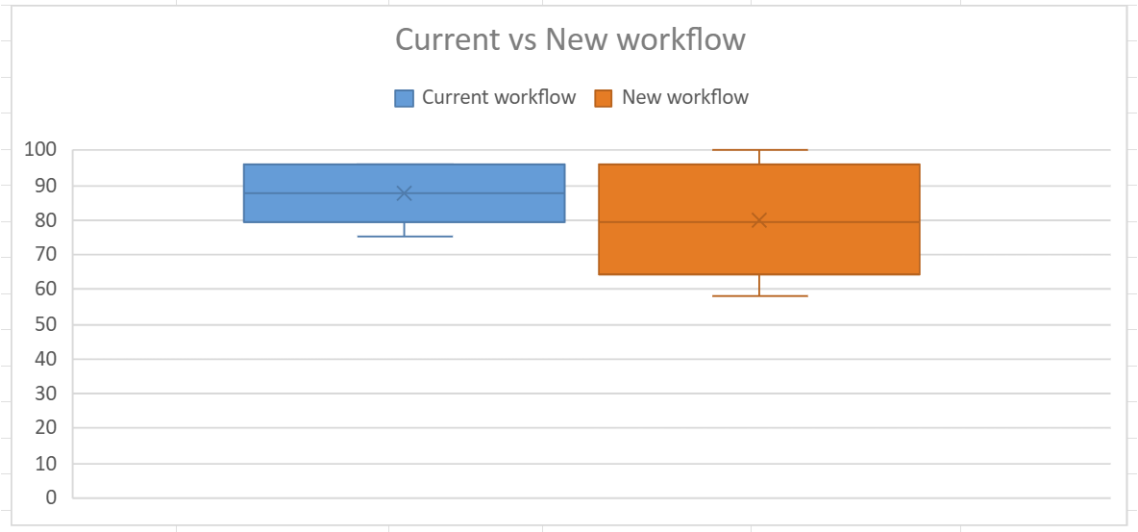


Figure 7.7: UMUX scores for Current vs New workflow

Table 7.10 and table 7.11 present the precise numeric values for each participant across all items, as well as the mean.

Table 7.10: Scores for participants(P) across A/B testing for references & modes

P	Reference A	Reference B	Modes A	Modes B
P1	87.5	91.67	87.5	51.17
P2	100	100	95.83	100
P3	75	91.67	29.17	87.5
P4	95.83	95.83	83.33	66.67
P5	95.83	87.5	100	100
Mean	**90.83**	**93.33**	**79.16**	**81.06**

Table 7.11: Scores for participants(P) across new & current workflow

P	Current workflow	New workflow
P1	83.33	91.67
P2	95.83	79.17
P3	87.5	58.33
P4	75	100
P5	95.83	70.83
Mean	**87.49**	**80**

These UMUX scores were evaluated using Bangor's framework, as outlined in section 4.7.7. According to the framework, a UMUX score above 50 is seen as *OK*, a score above 70 is seen as *good*, a score of 85 or more is seen as *excellent*, and 100 is seen as *best imaginable*.

In table 7.10, the mean for all of the six versions can be seen. The scoring shows that participants were generally very satisfied, with all of the items scoring 79 or above, meaning all items had good usability. A difference can be seen between the items, where the two versions of References have the highest score with a score above 85, indicating excellent usability. The two versions of Modes scored notably lower than references, with a drop of around 10 points in comparison. The current and new workflow has the highest spread between the scores, around 7 points. This can be contrasted with around 2 points difference between the different versions for both references and modes.

In figure 7.4 the distribution of values can be seen. Reference B is the one with the smallest variation in values in total, and also the item that had the highest mean, with a score of 93, indicating a very strong satisfaction. Modes A is the one with the biggest variation of values.

7.3.1.1 Net Promoter Score of New workflow

Participants responded to the question *How likely are you to recommend this new version of Research Assistant to a friend or colleague?* after conducting a series of tasks in the new workflow. Four participants scored 10 out of 10 on this question and is therefore considered promoters. One participant scored 8, and is therefore consider a passive according to the Net Promoter Score, see section 4.3.2.

7.3.2 Qualitative part

A thematic analysis was made from the transcriptions of all the user tests. The themes seen in table 7.12 were found.

Table 7.12: Thematic analysis of user tests

Theme	**Sub theme**
Needs clear descriptions	Metrics (contextual relevancy & credibility score) How does selection of models or reference type affect the response Color coding
Importance of information structure	Reference accessibility Organization and clarity Interactive features Visual differentiations Content presentation
Personal judgment and experience	Users have created their own judgment of what is considered trustworthy Users want to make the final call in assessing credibility instead of the algorithm An objective metric is preferred such as peer reviewed (yes/no)
Items are evaluated in terms of perceived usefulness	Assesses in terms of utility, if utility is low it is evaluated as unnecessary Visual figure of the proportion of references is nice to look at but doubts whether useful Participants questioned the utility and time saved for pre-written prompts for tasks Reference type
Values visual representations	Users are visual learners Visual aids analysis, in terms of accessibility and speed
Features that aid information searching are beneficial	Features or visualizations that help discrimination of internal information are appreciated Follow-up question is seen as a good source of inspiration

7.4 Important factors

Designing a conversational agent for internal company use, with scientists as the user group, requires several factors to consider. The following factors have been identified by conducting a survey, interviews, and user tests, divided into six different categories. The factors are important for influencing user satisfaction and directly address the research question.

7.4.1 Research Foundations

A factor to consider when designing conversational agents is the research foundations, see table 7.13. The mental models of the users should be considered, which relates to the usability principles by Norman, see section 3.7.1.1. The other aspect is to offer a visual alternative during design research, to elicit the participants' feedback, which were found during the thematic analysis of the interviews, see section 6.4.2.

Table 7.13: Research Foundations

1	**Research Foundations**
1.1	Consider the experts' mental model
1.2	Offer visual aids to elicit participant feedback during design research

7.4.2 Trust

Trust is a factor that was found to be important. For conversational agents made for internal company use, the trust is affected by the references, where the information is published and if it is internal or external, see table 7.14. Trust was mentioned during the interviews, where it was explained how the users use references to increase the trust for the answer, as well as how they trust AI in general, see section 6.4.2.

The interviews showed that all participants use the references to evaluate the quality of the response, see section 6.4.2. It was also brought up during the interviews that some publications are considered less reliable. The opportunity to distinguish internal and external data was also found to be appreciated during the interviews and was confirmed as an important aspect during the user tests, where positive feedback of the color coding was mentioned, see section 7.3 and table 7.12.

Table 7.14: Trust

2	**Trust**
2.1	Offer references for all information
2.2	Clarify the source of the publication
2.3	Facilitate how to distinguish between internal and external data

7.4.3 Usability

Usability is an important factor to consider, to ensure that the application can be used with effectiveness, efficiency, and satisfaction, see section 3.7.1. This factor consists of five aspects, seen in table 7.15. It should use a visual hierarchy of elements and information, which is based on the theory of hierarchy, see section 3.7.3.2. During the interviews, it was found that several users consider themselves to be visual learners, expressing the need for visual responses from the application, which is why incorporation for a visual response was included as an aspect, see section 6.4.2. That was also confirmed during the thematic analysis of the qualitative part of the user tests, see table 7.12. It was found during the survey, see section 7.1, that time saving is important for the user, which during the interviews, see section 6.4.2, was found to be connected to ease of use. The interviews also showed that too many functions could create confusion and ineffectiveness, leading to the aspect of ensuring a clean interface.

Table 7.15: Usability

3	**Usability**
3.1	Set visual hierarchy between elements
3.2	Incorporate visual response when possible, to aid understanding and learning
3.3	Prioritize ease of use
3.4	Ensure a clean interface

7.4.4 Time Efficiency

Connected to the usability and its effectiveness, time efficiency is a factor, see table 7.16. This factor is about easing the daily work and the users' workflow, by cutting down the manual work that is needed. To achieve this goal, the application should only have a few functions to minimize the learning curve. The first time that time efficiency was brought to attention was in the survey, see section 7.1. Speeding up the users' workflow and cutting down on manual work were confirmed during the interviews, see section 6.4.2. It was also brought up during the interviews that too many functions or settings made the users overwhelmed, and that the users preferred to only have a few functions, with the purpose to give responses and information efficiently.

Table 7.16: Time Efficiency

4	**Time Efficiency**
4.1	Ensure the application accelerates the users' workflow
4.2	Integrate only the most essential functions

7.4.5 Transparency

The transparency category is about explaining how different selections and functions affect the response, see table 7.17. During the user tests, were the user compared solutions with A/B-testing, the participants mentioned how they got confused with how different selections and settings would affect the response, see section 7.3, leading to the aspect of providing explanations for it. During the survey, see section 7.1, several participants described how they do not understand what data, functions, or capabilities that the applications has, leading to the aspect of explaining limitations. This was also confirmed during the interviews, see section 6.4.2.

5	**Transparency**
5.1	Provide explanations to how selections affect the response
5.2	Use clear communication and explain the limitations of the application for enhanced understanding

Table 7.17: Transparency

7.4.6 Engagement

Engagement is a factor to consider for conversational agents, which is about making the user want to use the application, see table 7.18. It includes having follow-up questions as a source for inspiration for related topics and questions, and to encourage the user to act on the response and use it for their work. The follow-up questions were found as a need during the interviews, see section 6.4.2, and was later confirmed to be appreciated during the user tests, see section 7.3. The follow-up questions together with the copy-button that were tested during the user tests, showed that the user got encouraged to make use of the response and to further explore, which led to the aspect of encouraging to interact with the application.

Table 7.18: Engagement

6	**Engagement**
6.1	Offer follow-up questions
6.2	Encourage to interact with the application

7.4.7 Alignment and Integration of Important Factors

Important factors aligning with previous research

A few of the important factors outlined align with previous research, as described in section 2.11. Trust has been identified as a crucial factor in earlier studies, a finding that aligns with the insights presented in table 7.14. Similarly, effective communication regarding the capabilities of the conversational agent, as highlighted in prior research, resonates with *Use clear communication and explain the limitations of the application for enhanced understanding* outlined in table 7.17. Furthermore, past

research indicates that productivity and efficiency play a significant role in user adoption, echoing the aspect *Ensure the application accelerates the users workflow* in table 7.16.

Integrated important factors
One of the main benefits and primary reasons for using Research Assistant, as identified in sections 6.4.2 and 7.1, is its ability to accelerate users workflow. Achieving this benefit relies on the integration of several important factors factors: for example, the aspect *Integrate only the most essential functions* (see table 7.16) works in conjunction with *Prioritize ease of use* and *Ensure a clean interface* (both in table 7.15). These interconnected aspects collectively contribute to streamlined time efficiency and enhanced usability, which are important for influencing user satisfaction, addressing the research question.

7.5 Design

The final design is a high-fidelity interactive prototype created in Figma. The Research Assistant is a specialized, conversational AI tool designed to ease and enhance the daily workflow and research activities of scientists. The prototype, in line with the factors above, is based on the survey, interviews, user tests, and design theory.

7.5.1 Rationale for new features added

The following features have been added in the design:

Side-by-side view of references
Relates to important factor: *Usability. 3.3: Prioritize ease of use.*
From the results of the interviews, see section 7.4, it was discovered that the display of references was an area of improvement. The workflow of reviewing references entailed forcing the user to scroll multiple times up and down, since the references were displayed below the text response from Research Assistant. This put some cognitive load on the user to remember both the information and the reference it related to, as participant 8 said ”Normally you’ll see everything at the bottom and then you’re like, OK, I have to click through this. What did this relate to again? Because one is at the top, one is at the bottom”. This was solved by creating a side-by-side view of references, see figure 7.21. With a side-by-side view, the references can be reviewed simultaneously as reading, which makes use of available side space but most and foremost was hypothesized to alleviate working memory (see section 3.4) and short-term memory (see section 3.4.1) by reducing cognitive load (see section 3.3) and makes use of the side space. The side-by-side view of references can be seen in figure 7.8

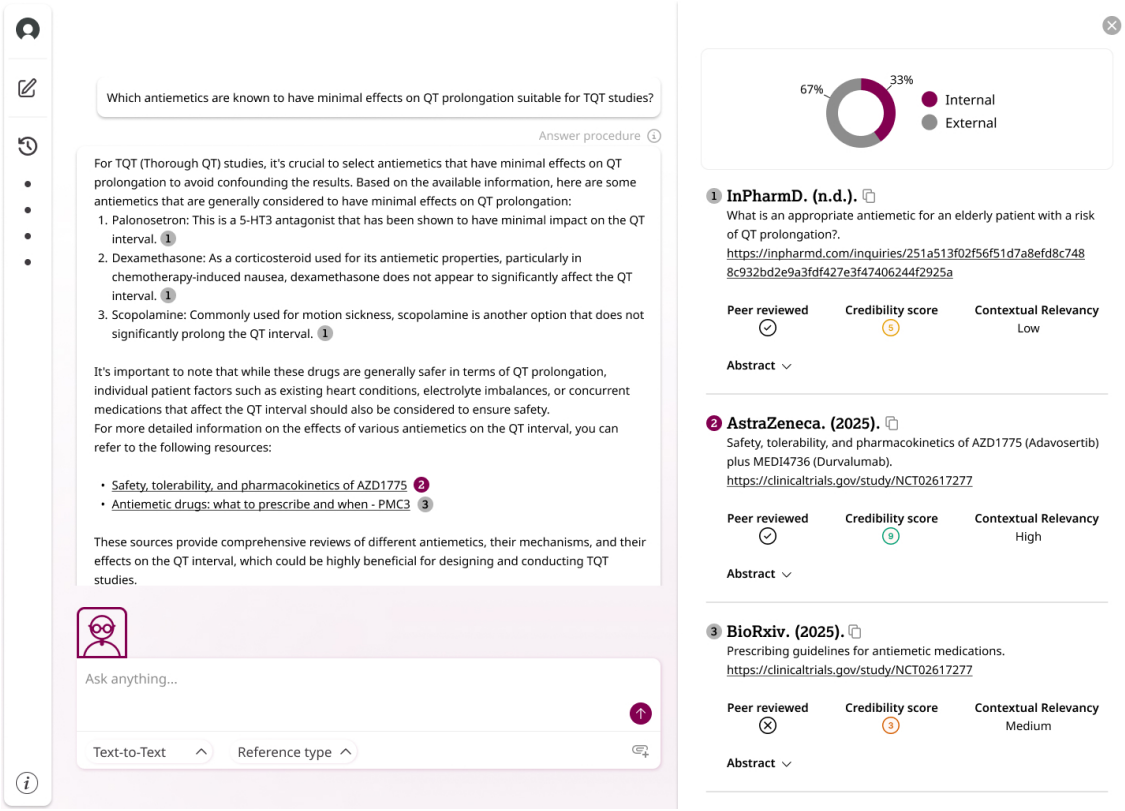


Figure 7.8: Display of references

Information structure of references

Relates to important factor: *Trust. 2.2: Clarify the source of the publication & Usability. 3.1: Set visual hierarchy between elements.*

Additional improvements for the display of references was the information structure. During the interviews, it was noted that it took time for the users to identify the most important aspects, such as origin and year of journal affects trust with the users. Solution to this was to make use of visual hierarchy. Examples of this was to use another font and larger text size to indicate the most important parts, and icons for the user to quickly be able to scan assessment of references. Can be seen above in figure 7.8.

Color coded references

Relates to important factor: *Trust. 2.3: Facilitate how to distinguish between internal and external data.*

The interview results, see table 7.8 revealed that users perceive internal references as more trustworthy than external ones. To access the credibility of a reference is something users do daily, but since internal references have inherent trust with users, internal references require less thorough examination and ultimately saving users time. However, one of the pain points was that it was hard to differentiate internal and external data in the application, see 7.4. To make it easier for the user to distinguish between internal and external references, color coding was applied to indicate whether the reference is internal or external, see figure 7.21. Internal

references are marked with company's internal color mulberry, while the rest of the references are marked with gray. The trusted internal sources' mulberry color creates a contrast against the gray references, allowing users to easily identify them and save time by requiring less thorough examination, thereby including the effect of saliency, see section 3.7.3.4. Can be seen above in figure 7.8.

Metrics

Relates to important factor: *Usability. 3.3: Prioritize ease of use & Time Efficiency. 4.1 Ensure the application accelerates the users workflow.*

Metrics was added in one of the versions of references. The interview results, see section 6.4.2, showed that it was essential for the users to verify that the reference is reliable and the information is accurate, for them to be able to trust the output from Research Assistant. Since the user reviews multiple of references every day, it was seen as an important area to explore if it could be improved in order to improve their workflow with the application. Different solutions were discussed in how to alleviate and streamline the workflow of how the user review references and assess them to be credible. The idea of metrics that assess the references was formed, to push some of the work to the algorithm instead of on the user. The metric peer reviewed (Yes/No), credibility score (1-10), and contextual relevancy (Low, Medium, High) were created, see figure 7.9.

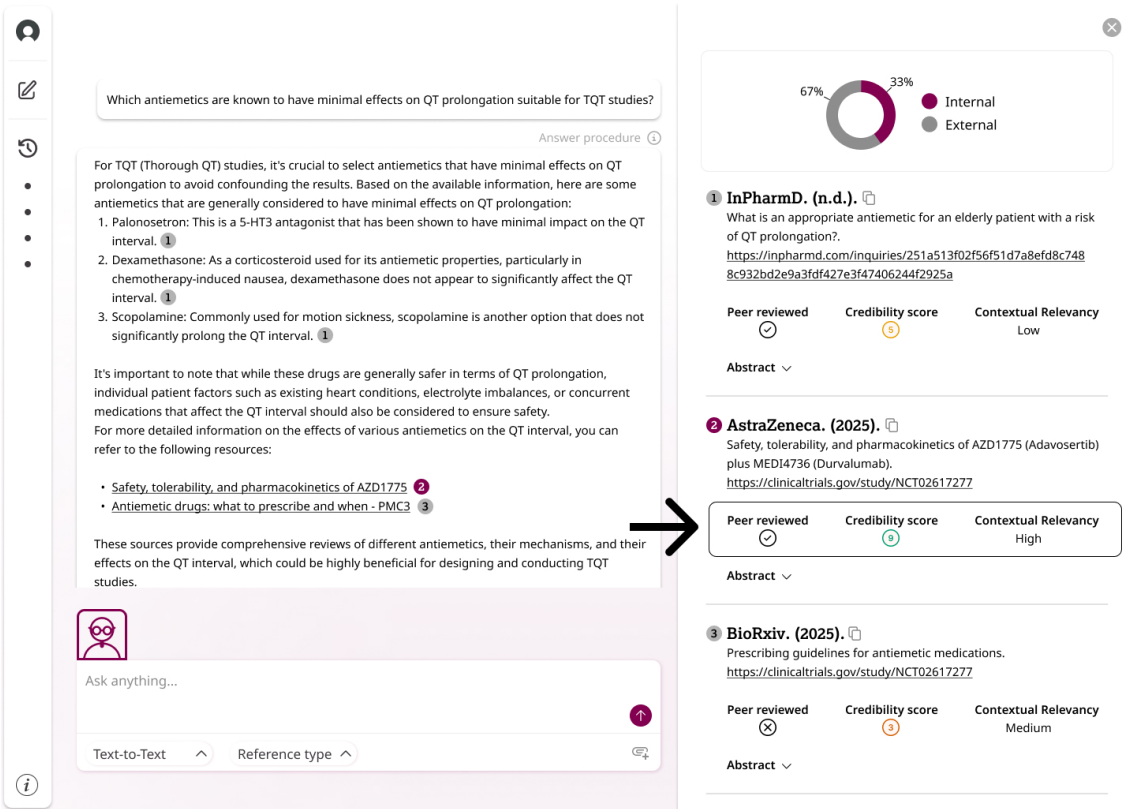


Figure 7.9: Metrics

Models replacing modes

Relates to important factor: *Time Efficiency. 4.1: Ensure the application acceler-*

ates the users workflow.
Modes was a previous solution in the interface to tailor the type of response in terms of phrasing the user could receive in the application, see section 6.2.1 and figure 6.1. However, from the results of the survey it became clear that this function is an area that could be improved. The survey results, see section 7.1, revealed that it was only 50% of the respondents that used the modes, 31% did not understand the modes and 19% responded that they did not need the them.

From the interview results and the low fidelity user testing, it was discovered that participants did typically not switch between the modes - they were strong advocates for the one mode they used. Moreover, the factual mode was the primary one used whereas the creative mode was not used at all. Since participants mainly only used one mode, it was decided to remove modes from the interface.

Additionally, it was discovered that Research assistant is an application used in several different types of subjects and the tasks can therefore differ, even though the primary usage is information gathering. It can be used for tasks such as providing summaries, interpreting a particular dataset, or coding. The idea of models were formed where the user could select how the response should be presented to fit their specific task better. It was thought that users would gain more benefit from tailoring the response in terms of data, beyond just the specific phrasing and what sources it references, which was the former case for the modes solution.

The new design solution of the added function models can be seen in figure 7.10.

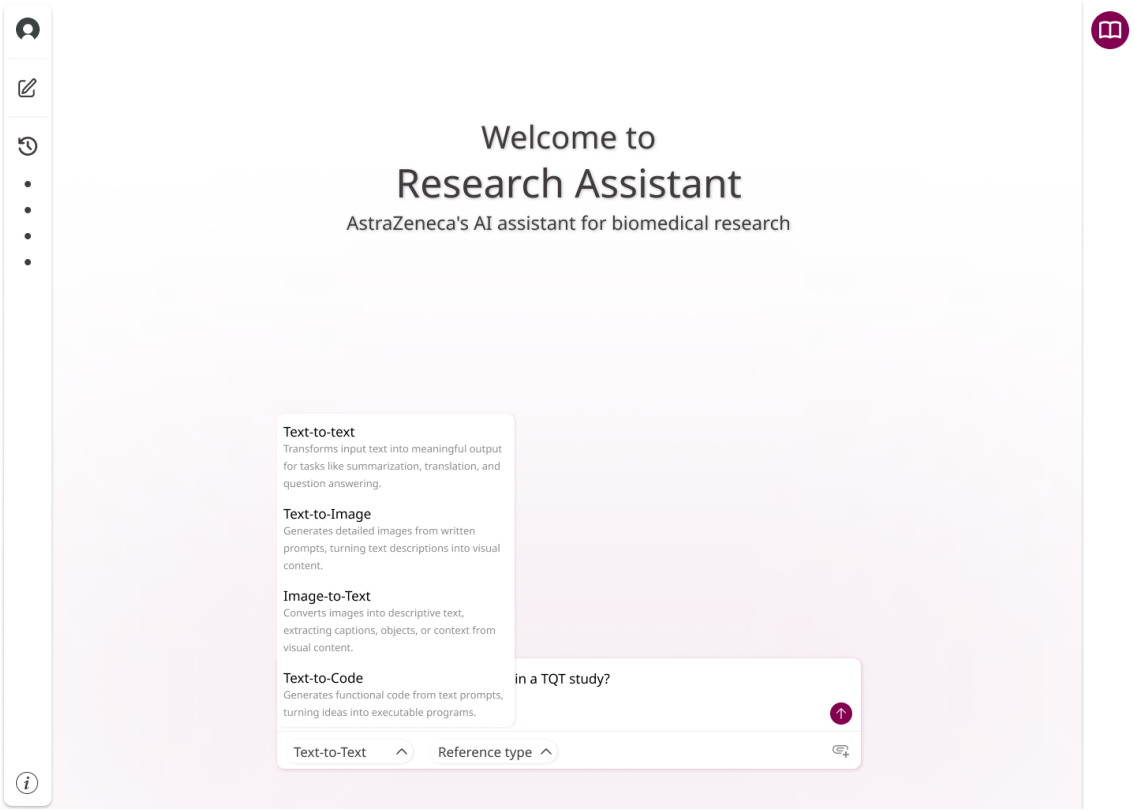


Figure 7.10: Model type

Reference type

Relates to important factor: *Time Efficiency. 4.1: Ensure the application accelerates the users workflow.*

As mentioned previously, from the survey and the interview results, (see sections 7.1 and 7.2), it was discovered that Research Assistant has a lot of users from different departments that use the application for various different tasks and different topics. Therefore, which type of information that needs to be retrieved can vary significantly. Therefore, it was theorized that it could be useful to tailor the response towards your specific needs. With selection of reference type, the user can select what kind of data points they want the conversational agent to search for. Furthermore, considering the ethics of the environmental aspect, see section 3.8.2, if only searching for a specific type of data point, the amount of electricity needed to conduct the search should be lower, than if all data points are used. Therefore, the ability for the user to select which reference type the Research Assistant should gather was deployed, see figure 7.11.

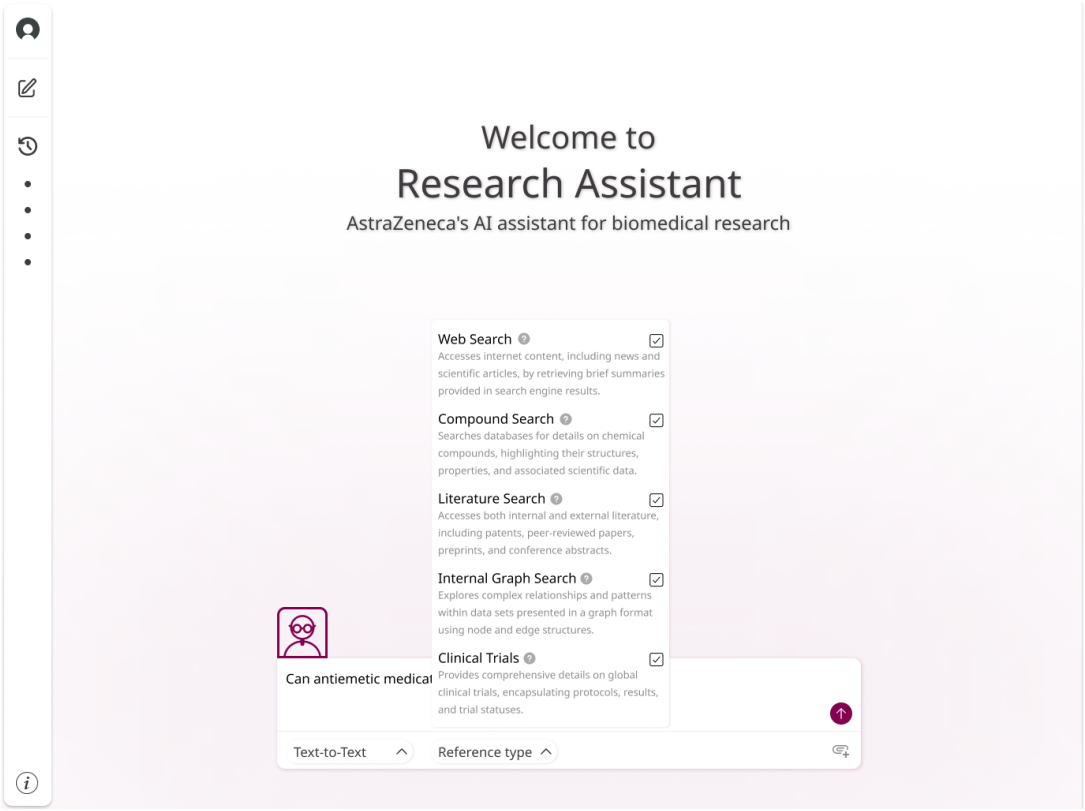


Figure 7.11: Reference type

Visual response

Relates to important factor: *Usability. 3.2: Incorporate visual response when possible, to aid understanding and learning.*

All of the participants from the interviews said that they appreciated the ability to have a visual alternative. ”I would love instead of only getting text. Also to get images of that, for example, that metabolic pathway... Im a visual learner” (Participant 4). ”I don’t know if it would be possible to have like a graphic mode, so to get your answer in some sort of mind map kind of thing, that would be really cool” (Participant 6). ”I like to see a visualization of a path as a visual learner myself, words can sometimes get confusing for me, so I like to see pictures, so like, oh, X interacts with this and it does not like this, like kind of visual things” (Participant 8). Therefore, it was decided to add a visual alternative as a way of visualizing the response. The visual response can be seen in figure 7.12

Figure 7.12: Visual response

Information button

Relates to important factor: *Transparency. 5.2: Use clear communication and explain the limitations of the application for enhanced understanding.*

From the survey results, see section 7.1, many users reported that a major pain point was not having access to literature they knew existed, and the same phenomena could be observed during the interviews, see table 7.4. This is a data integration problem, but the user experience was theorized to be improved if the user could be better informed about the limitations of the Research Assistant. In the current solution, information was scattered around in several different places in the application. To solve this problem, an information button was incorporated in the design where the user could find information about the capabilities, limitations and risks of the application, see figure 7.13. In the limitation tab the user is informed of exactly when the different type of reference sources are updated, to avoid frustration from the user why they can not access certain information. By giving all buttons a similar design, the user can easily identify them and understand how to interact with it, which incorporates the usability principle affordance, see section 3.7.1.1. It also incorporates signifiers, since the user can understand where actions take place, see section 3.7.1.1.

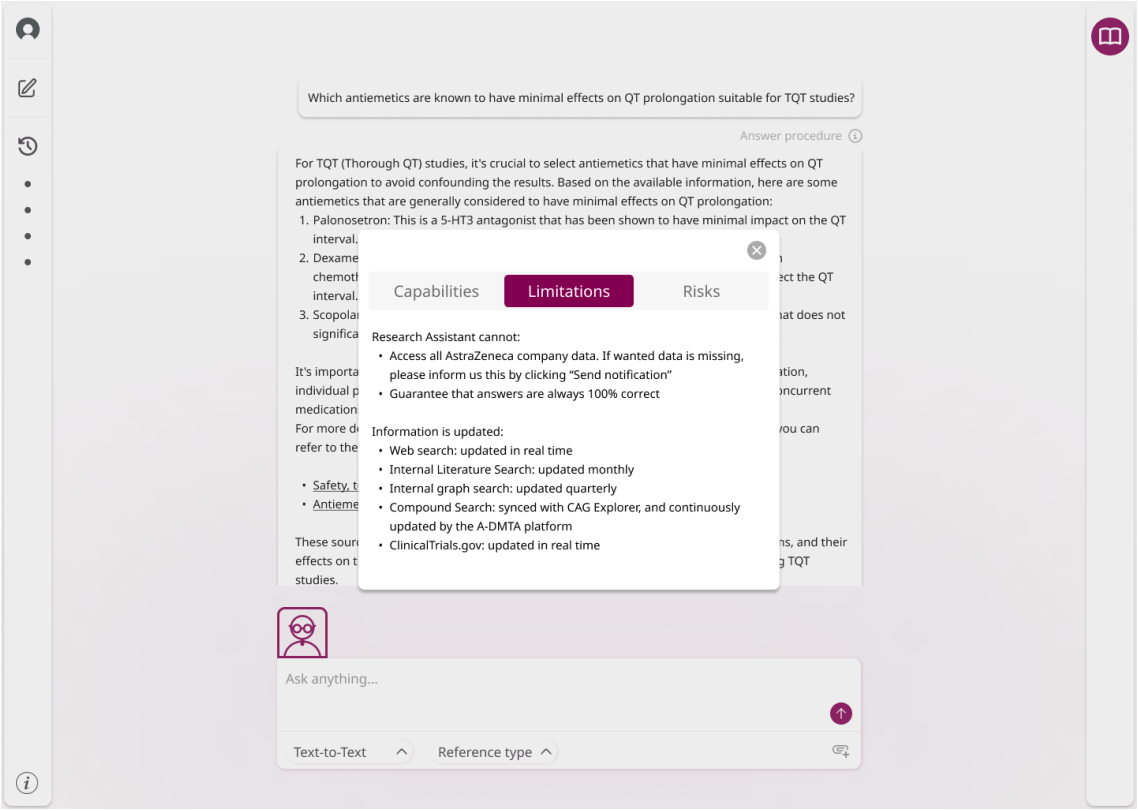


Figure 7.13: Information pop-up

Follow-up questions

Relates to important factor: *Engagement. 6.1: Offer follow-up questions.*

Research assistant is among others used to explore and understand targets, and investigate new angles to a subject, see table 7.1. To enhance and improve this workflow, follow-up questions related to the prompt was integrated. The purpose was to give the user a helpful source of inspiration, when providing suggestions of aspects to look further into. The follow up questions can be seen in figure 7.14

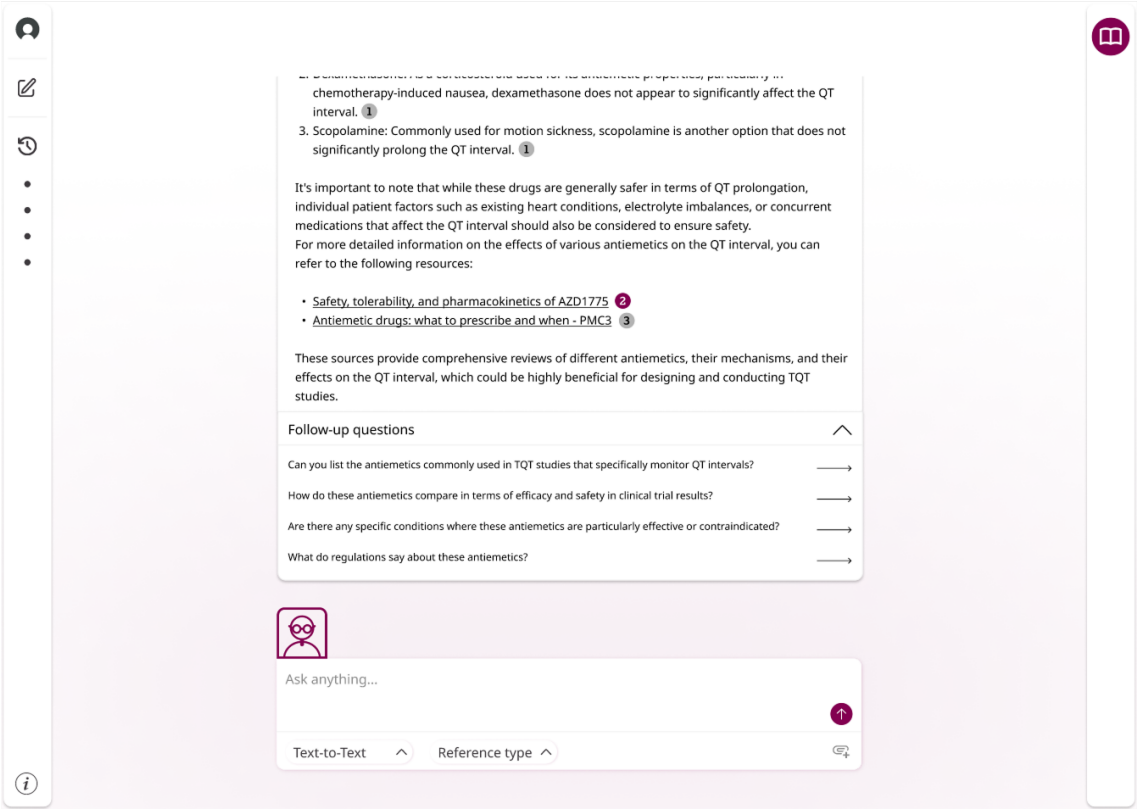


Figure 7.14: Follow-up questions function

Copy Reference

Relates to important factor: *Time Efficiency. 4.1:Ensure the application accelerates the users workflow.*

To reference to different sources and references is a common task that the users perform daily. It was found to be a need during the interviews to be able to copy the references, in order to save time and avoid doing it manually. The decision to implement a copy reference function was straightforward, since that is a task many users do manually each time they use the application and therefore has a large probability that the majority of users would find useful. The copy reference can be seen in figure 7.15.

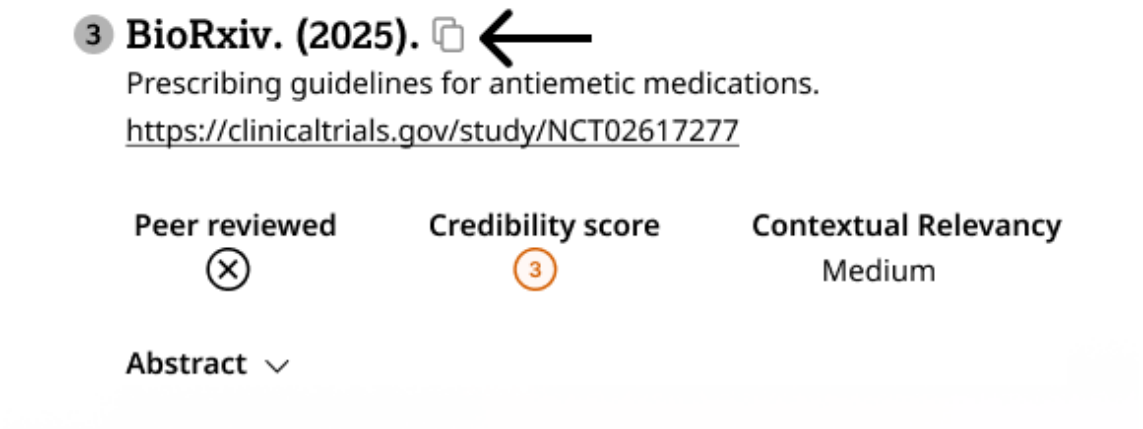


Figure 7.15: Copy reference

Add Attachment
Relates to important factor: *Time Efficiency. 4.1:Ensure the application accelerates the users workflow.*
Another need that was expressed in the interviews was the ability to add attachments. Common tasks of Research Assistant were to provide summaries of internal material, to provide presentation material and data presented in forms of Excel, where users expressed that a lot of time could be saved if the option to attach their own file was available. Lastly, with the rationale that this function already exists in another internal application and therefore the technology is ready and available, it was decided to include this feature. can be seen in figure 7.16.

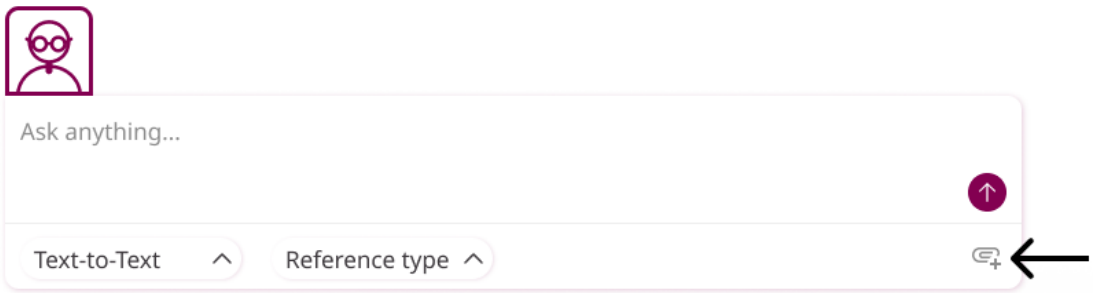


Figure 7.16: Add attachment

7.5.2 Rationale for design choices

The following design choices have been made:

Color palette
Mulberry was chosen to use as a accent color as it is one of the two colors representing the company's brand identity and the aim was to align with this

identity. The company has many diverse applications, where another internally used application, also a conversational agent, uses this type of mulberry color. By sharing a similar design language in terms of the colors, it was hypothesized to create a sense of unity. It also incorporates consistency between internally used applications, see section 3.7.1.1, as well as takes the users' mental model into consideration, by adapting the interface to what the users are used to, see section 3.7.1.1.

Gray color is used to indicate external references. Gray is often used to indicate inactive parts, however, since the current interface has gray references, it was argued that the mental model of the user is that they have learned to associate gray with active references. Other colors that is used is black and white. A white background with a black text font was used to create a contrast that ensures the best readability for the user.

To avoid overwhelming the user, it was decided to not add in additional colors. Research Assistant is an application that has users that use it for short durations of time daily, and for these users the usage of several different colors in the interface could mean too many impressions, and thereby become overwhelming.

The idea with different color behind the metrics for the credibility score was that the users can go through the references and with the help of colors quickly be able to scan through multiple references and determine which references are the most credible. The colors used to creating the metrics for the credibility score has been tested with the help of a plug-in in Figma called Colorblind. To meet accessibility standards, the contrast ratio between text and background should be at least 4.5:1 [74]. The different gradients ensure that people with color deficits are able to distinguish between the three colors as well, and with the numeric value serves as an addition to ensure a correct assessment.

Text size and typography
The text size used was determined by internal standards set by the department, to be able to align with rest of the company's design language. For the body text, size 13, 15 and 17 were used, and size 33 was used for headings. The text sizes where set to create a clear visual hierarchy of the information, referring to the important usability factor in table 7.15. The text font was also determined by internal standards set by the department, Noto Sans. Noto Sans was used across the full interface, to ensure consistency, see section 3.7.1.1. The only exception is in the A/B testing for Reference B, where the title of the journal is written with font Lexia for the user to easier differentiate the name of the journal, which is a font that is also used by the company.

Icons
Icons are visual objects that can be processed multiple times faster than regular text and is less prone to errors [75]. The icons included in this interface were selected because they are standard representations for specific functions, and therefore well known by the standard user. Icons can also communicate the action behind the

object more space efficiently than regular text.

7.5.3 Overview of Interface Functionality

The prototype were done based on the important factors, to reflect and show a visual representation of what the important factors could lead to. The design of the prototype aligns with the purpose of the conducted research, as well as the research question.

The design features a homepage, seen in figure 7.17. To the left is a side bar, that opens up to give the user more information about the functions. The opened side bar can be seen in figure 7.18. In the side bar, Profile, New Chat, History, and Information can be found. The icons are used to make it possible to minimize the side bar. When opening the side bar, the text is added to give the user more information about the functions, as well as summarized titles of the recent history.

By focusing on a few important functions, the user can quickly write a prompt and receive an answer, which incorporates the important factor time efficiency, see table 7.16. It also creates a clear transient posture of the application, by making it clear in its communication and focusing on important tasks, see section 3.7.3.1. By having the side bar closed in the homepage, the user can focus on the text box, which signifies that there is where the actions take place, which is an example of use of a signifier, see section 3.7.1.1.

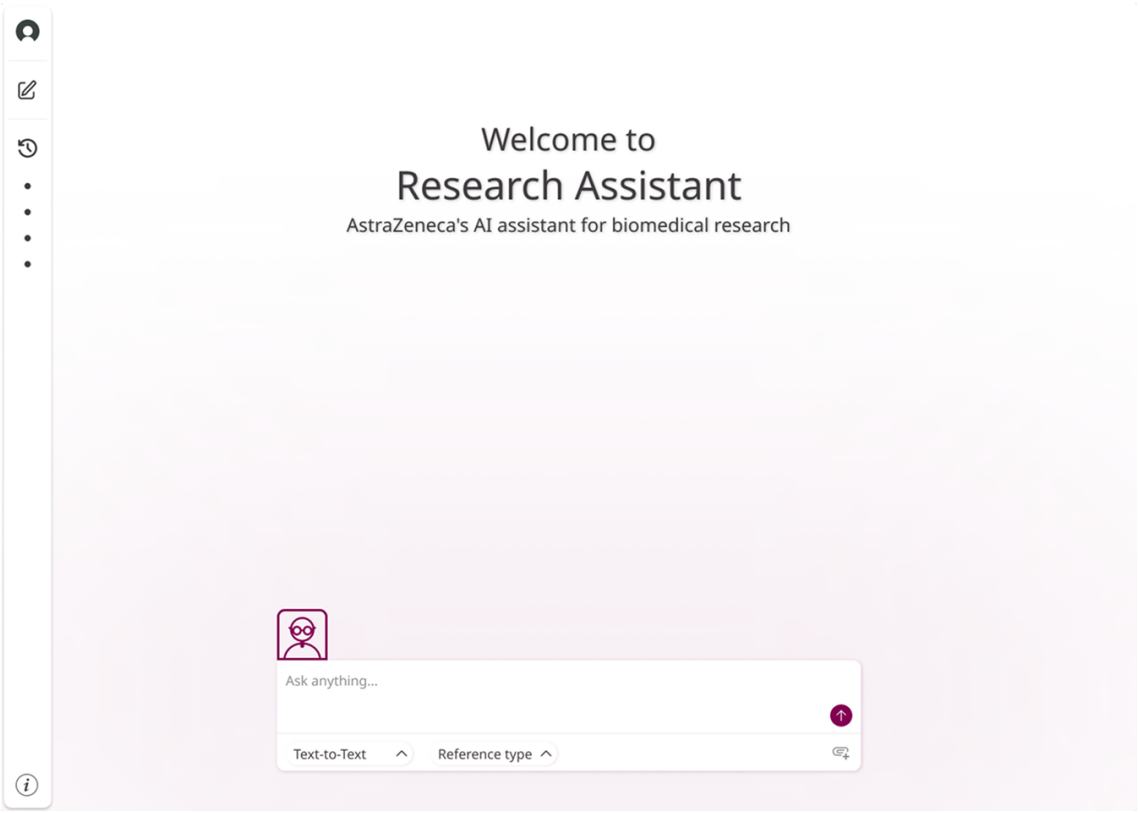


Figure 7.17: Homepage of new Research Assistant

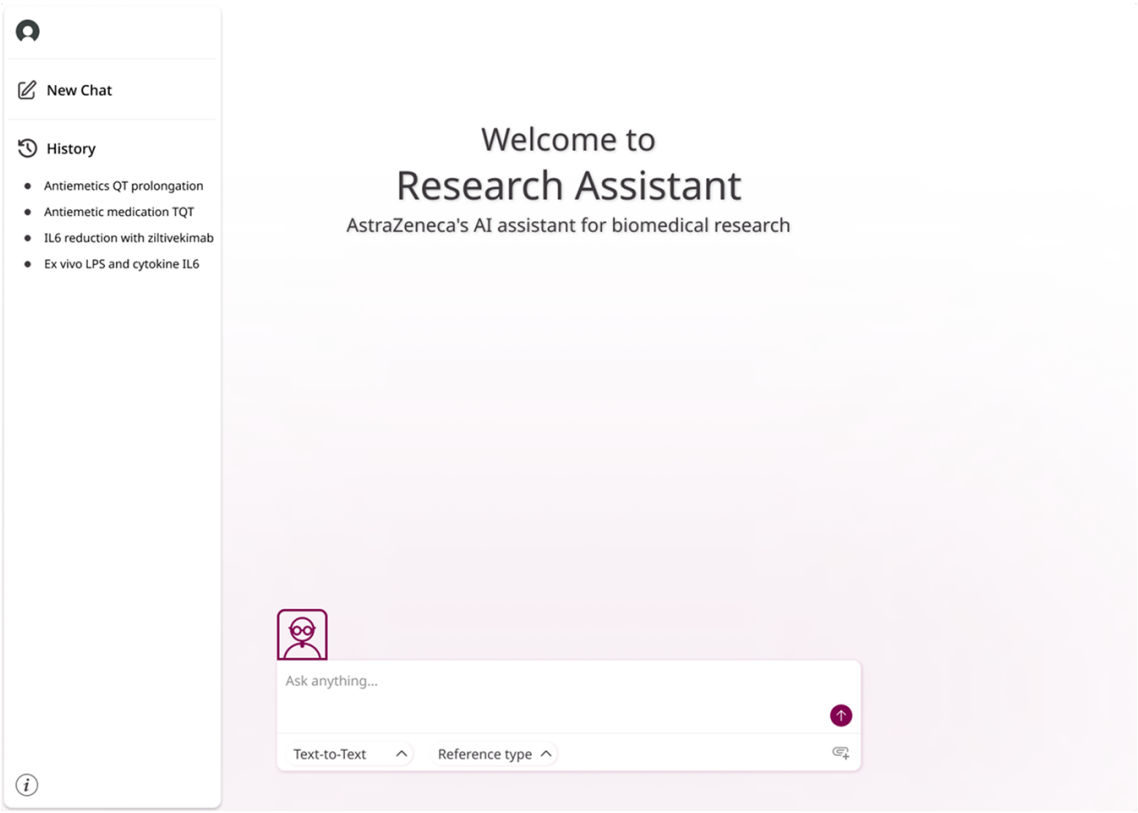
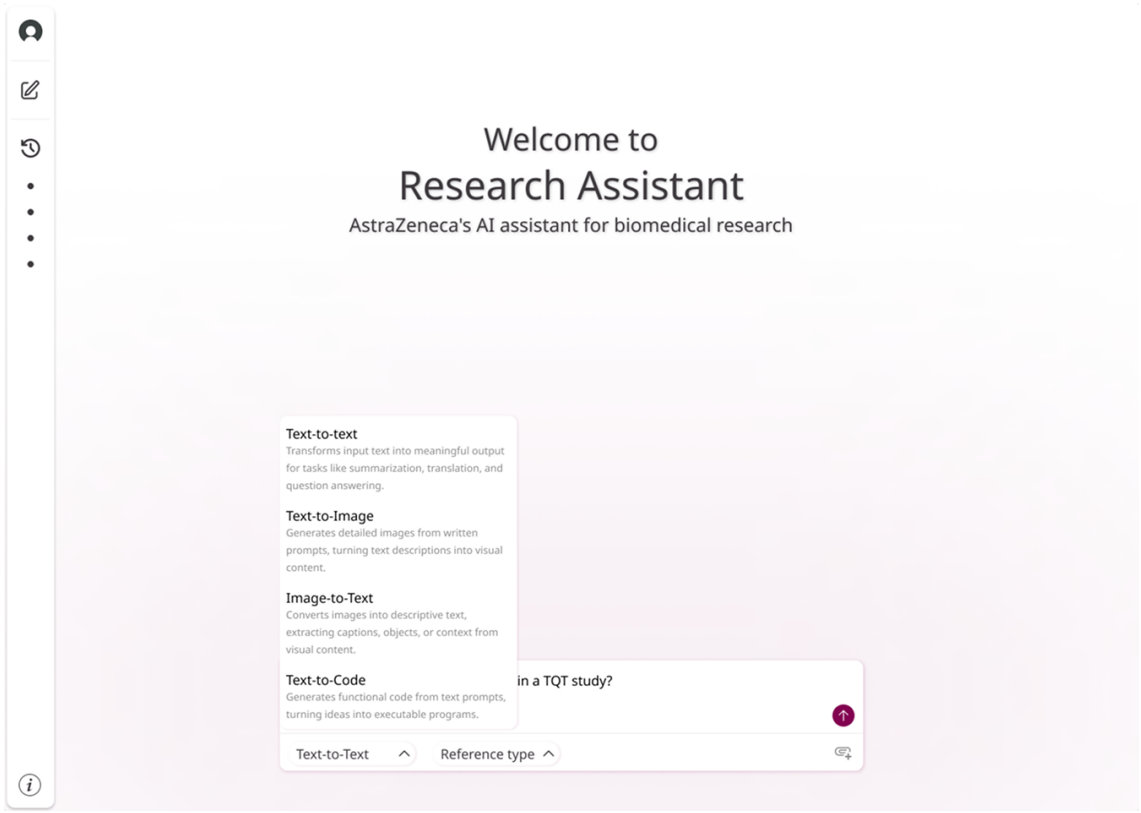


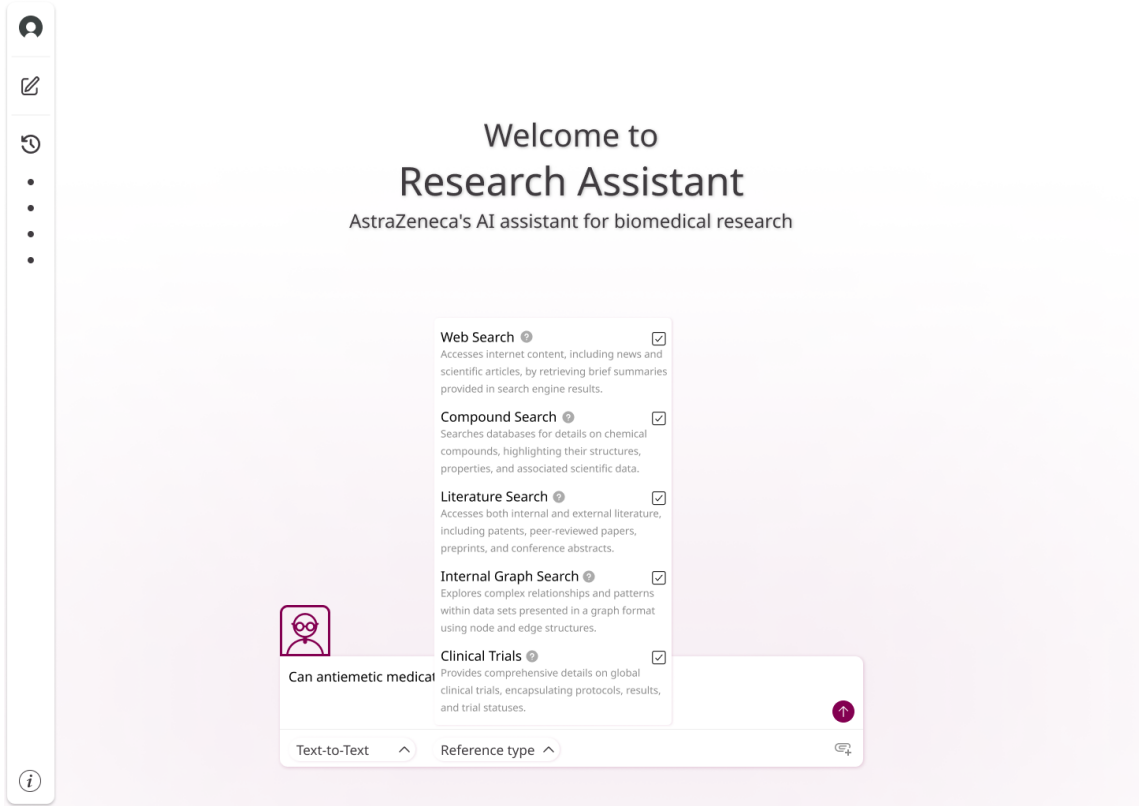
Figure 7.18: Side panel

The text box in the middle consists of Models, Reference type, an Attachment button and an avatar, see figure 7.17. The avatar was incorporated to fit with a long-term vision for the department. From the insights, it was clear that the possibility to use visuals was wanted, some mentioned being able to get help with coding, and all interview participants described the need to write prompts quick and receiving a response. Therefore, four models were included, seen in figure 7.19a. They all have descriptions, since one of the pain points included not understanding all functions in the application.

In figure 7.19b, five reference types with descriptions can be seen. The alternatives were chosen from a list of knowledge sources that the application uses today, but currently without the option for the user to choose which ones to use. By working with a size and color difference of the models and the descriptions, a visual hierarchy is set, see section 3.7.3.2.



(a) Models



(b) Reference type

Figure 7.19: Models and Reference type functions

In figure 7.20, follow-up questions can be seen. The function is a drop-down menu, to allow for the user to see and use the function if wanted, while minimizing the risk of making the user overwhelmed with information if not wanting to use it, following the details-on-demand principle, see section 3.7.3.3. The follow-up questions are based on the prompt, to let the user know what other similar prompts are common when searching for a specific topic, providing a way of letting the user aware of important topic aspects.

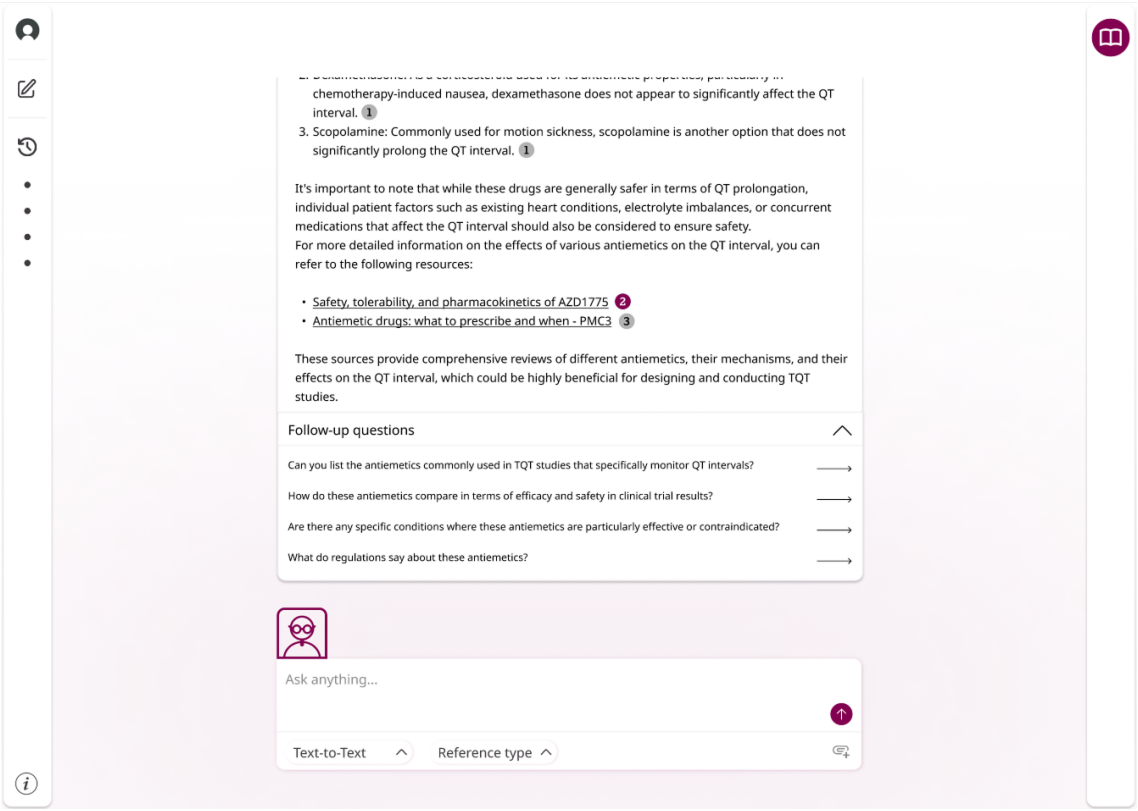


Figure 7.20: Follow-up questions function

When receiving a response, a references-button can be seen in the upper corner the right (the book icon), see figure 7.20. When clicking on it, or by clicking on one of the reference numbers in the response text, the references panel is opened in a side-by-side view, see figure 7.21. With a side-by-side view, the references can be reviewed simultaneously as reading, which makes use of available side space but most and foremost was hypothesized to alleviate working memory (see section 3.4 and 3.4.1) by reducing cognitive load (see section 3.3). The different text sizes, use of bold, and colors are used to create a clear visual hierarchy (see section 3.7.3.2), which incorporates the important factor usability seen in table 7.15.

The visual graph enables the users to get a fast overview of the range of external and internal references. The color indicates if the reference is internal or external, and the function was added because the users trust internal sources to a larger extent than external. The trusted internal sources ”pop”, allowing users to

easily identify them and save time by requiring less thorough examination, thereby including the effect of saliency, see section 3.7.3.4.

A copy-button for the references can also be seen, which was added because the users often copies references to paste in other documents and the copy-function ensures that the reference is in the correct format.

Adding the abstract in the references streamlines workflow by reducing the need to open multiple windows. By adding the abstract as a drop-down, the details-on-demand principle is added (see section 3.7.3.3), which allows for the user to get more information if wanted but without cluttering the overview.

The metrics ”Peer reviewed”, ”Credibility score”, and ”Contextual relevancy” were added to increase trust of the references.

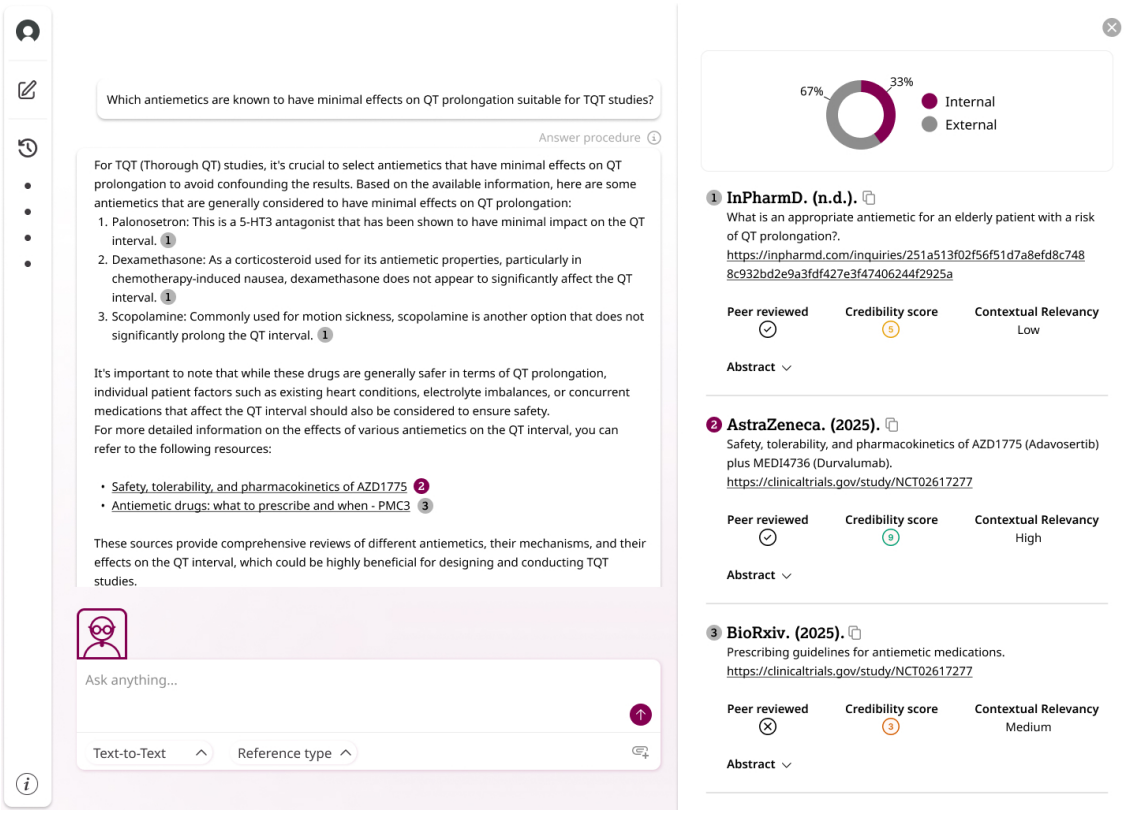


Figure 7.21: References panel

To gather all information about the application in one place, an information pop-up was added, which can be found in the side bar, see figure 7.22. The users were confused about where to find information, since there is information about the application and messages in five different places, making it hard for the user to find it and read it all.

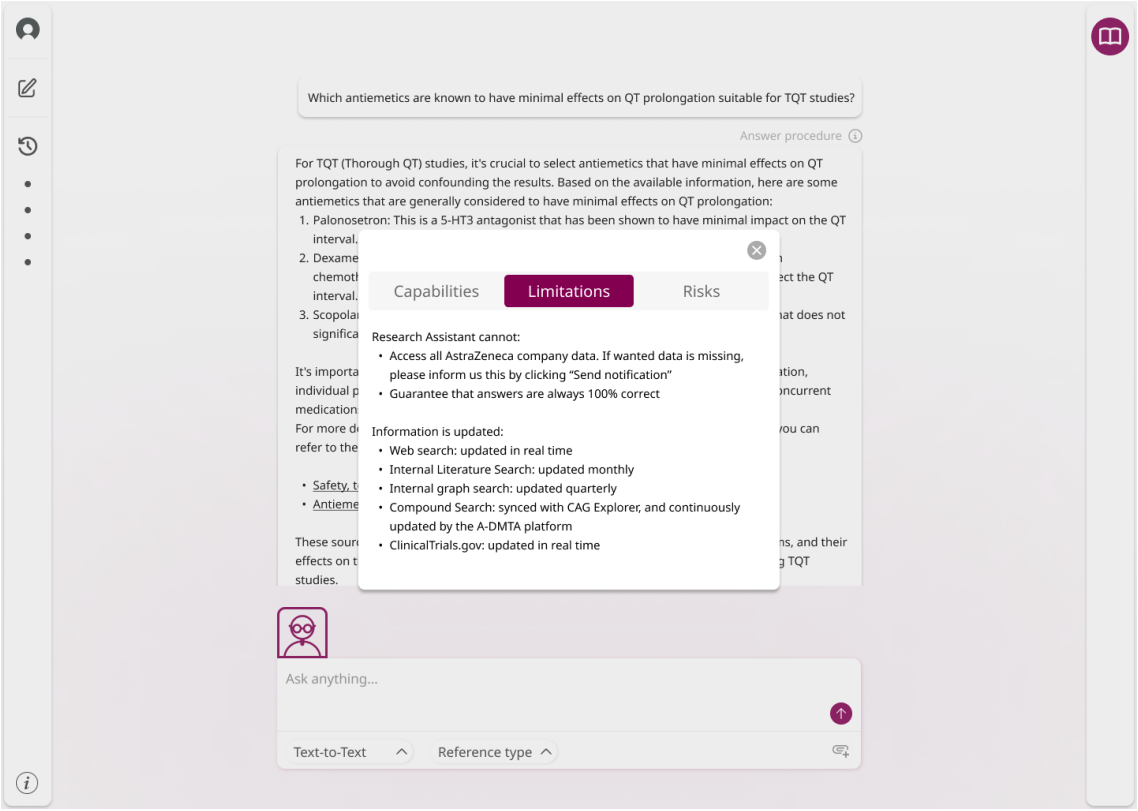


Figure 7.22: Information pop-up

8

Discussion

This chapter discusses the project's theory, results, process, ethical considerations, and future work.

8.1 Theory

Computers are social actors, CASA, and anthropomorphism are two theories that was initially discussed that could potentially have an impact on our work. These are theories that concern how people interact with objects, (see section 3.5). The theories was hypothesized that they could have an impact in how the communication, interaction, and dialogue with the user were phrased. From the interview results, it was discovered that two participants had opposing views in how to interact with Research Assistant (see table 7.6) whereas the rest of the interviewees did not mention the communication at all. This variation in perspectives highlights the complexity of designing a conversational agent that addresses different user preferences. There was discussions concerning whether to integrate personalization in the application, that users would be able to choose the type of tone Research Assistant should communicate in, to be more human-like or to be more robot-like depending on preference. This idea was, however, abandoned due to modes and references were seen as more important focus points, and personalization of tone was abandoned. Furthermore, the decision was also influenced by technical limitations in what possibilities there was to prototype and test different conversations with the participants. However, it is worth to mention that questions concerning the tone of Research Assistant were lacking in both the survey and the interviews, but that two participants still talked about this could indicate that this is an area that has been overlooked.

The Technology Acceptance Model (TAM) (see section 3.6.1), and The Task Technology Fit theory, TTF (see section 3.6.2), are two theories that concerns the prediction of user acceptance in regards to technology. These theories has been taken into consideration when designing the user tests. The Usability Metric for User Experience (UMUX) questionnaire aligns with TAM since it is covering two critical aspects of the model, ease of use and usefulness. In the A/B testing, the user had to conduct the same task in different interfaces and afterwards fill out the UMUX questionnaire. In this way, the fit between the technology (interface) and the user tasks are evaluated, since the interface is assessed after each task with

the UMUX. The theories were useful for providing a framework of aspects, such as evaluation of ease of use in order to measure user satisfaction.

Cognitive load theory (see section 3.3) was a useful theory when different design solutions were ideated and visualized. It was a useful framework when discussing the design solutions, in order to be able to determine which solution would reduce cognitive load and hypothesized to be better.

8.2 Process

The process of the thesis will be discussed below, in terms of the utility and results of the methods used.

8.2.1 Survey

The survey was accessible on internal channels to 9 000 people, where 54 responses were recorded. This is a low response rate of 0.6% participation. However, there was no way to confirm whether the survey was actually viewed by all 9,000 intended recipients. It is possible that the survey did not reach the entire group, which could explain the unusually low response rate. With a higher response rate, the results of the survey could have been more generalizable.

Some of the questions in the survey were not mandatory to answer, which was decided to avoid forcing the participants to answer something that was not considered relevant to them. However, it led to some questions not being answered by as many participants as wanted. If the survey was to be redone, all questions could be made mandatory to ensure that more answers and insights could be collected. It would also make it easier to compare the quantitative questions results from the survey with the later evaluations of other used methods. Making all questions mandatory could however force the participant to answer a question they do not understand, or choose an alternative that they do not agree with, which would make the results inaccurate.

8.2.2 Interviews

The interviews were a fruitful source to gather rich data. The survey served as an initial overview, but to get beyond the surface, a deeper and more thorough investigation was needed. All the participants that were interviewed were very positive regarding Research Assistant. Therefore, it was necessary to prompt the participants a bit to get beyond the initial user satisfaction and to find areas of improvement. Thus, the method of interviewing was fruitful in the exploration phase since the ability to further prompt the participant to explain their answers was available.

8.2.3 User Testing

In contrast with the interviews, the ability to prompt the user further was something that vanished during the user testing due to time limits. For the users to conduct all of the six tasks, fill out a questionnaire after each task and respond on two or three questions, was all the time that was available. Therefore, the qualitative insights from the user testing were not as rich compared with the in-depth interviews, but did provide an initial ground to explore further.

8.2.4 Usability Metrics for User Experience

The utility of Usability Metrics for User Experience (UMUX) was good in terms of providing a quantitative way to asses the usability. It was noted in the interviews that all participants were very positive. UMUX provided a numeric value for their satisfaction, which offers a clearer comparison than words alone, as individual expressions of satisfaction can vary considerably. UMUX also provided a benchmark against industry standards of their ratings, that validated their satisfaction. A disadvantage of using UMUX was that it was a bit limited, including that the measurements did not reveal any new insights or areas of improvement.

8.2.5 The result of the user tests

The user tests and use of UMUX confirmed that the participants consistently rate very positively, which consequently affects the result. The mean for the current workflow, the interface that is in use today, scored 87, which is seen as *excellent* according to Bangorts framework, see table 7.7.

The high scores may hide minor issues that users have not articulated during testing, or needs that users did not explicitly ask for but still find valuable. This was a phenomena that was observed during the interviews, where participants stated their overall satisfaction over RA. However, when examples were shown that displayed other design solutions than the current one, all participants found the new design solution to be better than the current one. The high scores may also indicate complacency, as one user said during one of the interviews: ”Certain things which might be improved, but, you know, it’s already a luxury”(Participant 1).

Although all participants were in general very positive, participant 2 in user test stood out. Participant 2 scored 100, also known as *best imaginable* on three versions, indicating a very high satisfaction.

Participant 3 scored a score of 29 in modes A, which is far below *OK*. The reason for the low score seems to be due to lack of understanding how it should be used. ”I don’t know how I would use this to Be honest. Now. I got confused about these options. I didn’t like them much” (Participant 3).

Reference B is the one with the smallest variation in values, whereas Modes A is the one with the biggest variation of values seen in figure 7.4. Reference B

was the version with the highest mean, indicating users were very pleased with the solution. In contrast, some users were very dissatisfied with Modes A (Participant 3) whereas Participant 5 scored 100. The difference might stem from personal preferences when assessing the utility of altering the response and the need to understand exactly how the selection would impact the response.

Reference B is the version with the highest mean, closely followed by Reference A. The users makes use of references whenever the user uses the application, reviewing references and assessing whether the information and source is credible. The design solution for references made use of the space on the side, displaying references to the side, allowing the participant to review the references simultaneously as reviewing the response from RA. This was probably the main cause for the high satisfaction, since reviewing references on the side are theorized to reduce cognitive load, off loading short term memory and enable more easily contextual recall. When conducting the interviews, some other alternatives was also tested, to get a direction of what we should focus further on. The results from those interviews when presenting references to the side was ”Brilliant. I love it. That’s exactly, yeah, that solves the problem. That just solves the problem right there. That’s perfect” (Participant 3). The previous solution had references displayed after the full response, forcing the user to go up and down multiple times to review the references.

The main difference between Reference A and B is that Reference B makes more use of information hierarchy, guiding the user where to look by using different font and larger text for the journal and year. This might also be the reason for why alternative B scored higher on usability.

The user group
The user group of scientists that are experts within their fields, could primarily be observed in their workflow. The users’ daily tasks require them to be very thorough, which was observed during user testing, when participants quickly noticed any scientific factual errors. In the interviews, but particularly in the user testing, it became obvious that the user group of scientist needed clear descriptions for what each function entailed to be able to feel that they trusted and wanted to utilize the function. This might stem from that their work requires them to be very thorough, which influences other areas they are in touch with.

8.2.6 A/B testing

A/B testing was a useful method as the focus was to explore wide, and this approach enabled us to explore different concepts and receive insights regarding which overarching features and design solutions that resonates with the target audience. However, the most prominent benefit with A/B testing was that the different versions gave the participants something to reflect upon, when they could compare pros and cons between the different versions.

8.2.7 Net Promoter Score

Net Promoter Score as a measurement was a good method to assess satisfaction of RA for a wider audience. It also shows that there are users that are not as satisfied as the participants we interviewed and conducted user tests with. In the user testing, 80% of the participants were promoters and 20% were passives, and no detractors. This can be contrasted with the survey, where 61% were promoters, 31% were passives and 8 % were detractors.

8.2.8 Factors to consider influencing the result

The users signed up voluntarily to participate in the different measurements. When conducting the interviews, it was quickly realized that all participants were very positive towards RA and the same phenomena could be observed during the user testing. At the same time, the survey revealed that not all users were positive. This could mean that participation in the interviews and user tests was somewhat skewed and did not reflect all users. Social desirability bias see section 3.7.4) could potentially account for the phenomenon observed. However, the fact that the quantitative result of the current workflow in place in RA today scored a higher mean than the new workflow, designed by the researchers, suggests that social desirability bias have not influenced the outcome.

RA is created by an internal group at the company. If the same party would conduct this type of evaluation, it could be that evaluation apprehension, see section 3.7.5) would refrain participants from stating their true opinions (if negative) in fear of being perceived in a negative way. However, when an independent party like the research setting for this project is evaluating the application, both evaluation apprehension and social desirability bias should be avoided. The researchers for this project are not responsible for the development of the application, and therefore, should be a safe space for the participants to state their opinions.

Rather, the reason for the very positive users may be that the RA is a new tool that has only been around for one year. Before RA, the users had to do the information search more or less manually with the help of search engines like Google. In comparison, utilizing a conversational agent where the search is tailored exactly after your own words, definitely came with improvements. This may also be one of the reasons why it was necessary to prompt the user a bit to reach areas of improvement, to reach beyond their overarching opinion of satisfaction.

8.3 Limitations

That all the participants in the interview and user testing were very positive is something that was taken into consideration that affects the result. It is hard to argue that the result can be generalized for all users of the application, when the

survey revealed that there are some participants that are not as satisfied as the participants that participated in the interviews and the user testing.

There were efforts to reach out to more hesitant or negative participants as well, but without success. Reasons for this can be discussed, it may be everything from that is feels easier to provide good feedback than bad or that users who like RA feels more engaged with the application and are more inclined to participate.

8.4 Ethical considerations

Ethical considerations of users
The methods of survey, interviews, and user testing led to a thorough ethical consideration of participants' privacy concerns. All participants received prior information to ensure they were well informed and prepared for their participation. All users were informed that their answers would be anonymized, about who the data would be shared with (which was the project team), and that anonymized data and opinions could be included in our thesis work. All participants in the interviews and user tests consented to the use of their feedback and agreed to the recording of the sessions for evaluation purposes.

Ethical considerations of improving a conversational agent
Trust was mentioned throughout the project, as it was an important factor influencing users willingness to use the application. During this project, there have been efforts in terms of design in how to increase the level of trust with users of Research Assistant. One could reflect about the researchers responsibility in increasing the level of trust. The science field require great rigor and a thorough approach, as well as a critical mindset to critically reflect upon information to provide accuracy. One might fear that if trust levels increase too much, it could lead to danger by over-relying on a tool powered by artificial intelligence, which could possibly be wrong. In that way, there is a responsibility in developing the application to maintain the balance between trust and over-reliance when utilizing the application.

8.5 Future work

For future work, it would be beneficial to include a larger amount participants to the research to be able to generalize and validate the result. It would be particularly useful to include more opinions from non-users, to understand their perspective for not using the application. It would also be useful to be able to reach out to participants with more negative opinions than the ones included in the interviews and user tests, to get a different perspective. Furthermore, it would also be beneficial to test Research Assistant several years forward, to see whether the user satisfaction stays at the same high levels or decreases.

9

Conclusion

The thesis aimed to investigate the design of conversational agents and explore how they are used. The purpose was to identify areas of improvements on an existing conversational agent, Research Assistant, as well as to detect which factors that are beneficial for improvements of conversational agents, for an increased value of the product. The project had the goal to address a gap in previous studies, by targeting a specific user group, which were scientists using a conversational agent for research purposes, internally and in a professional context.

The research question was:

1. *What are important factors for designing conversational agents that influence user satisfaction in the context of internal field-specific experts, and how can these factors be realized?*

The answer was found through a thorough investigation of the usage of the application and its users, with several methods used to find insights and evaluate solutions. Each chosen method was picked and adapted to the results of the previous method used, as well as its evaluation. The survey created a broad overview of the usage of the application, where critical aspects to investigate further were incorporated in the interview. The in-depth interviews were evaluated with a thematic analysis, resulting in a list of needs and explanations, as well as insight cards, that were used to ideate and design the prototype. The prototype was then evaluated with user testing, creating both quantitative and qualitative insights. The insights from all methods were taken into consideration, to result in identified important factors. The answer to the research question is thereby seen in table 9.1:

No.	Important factor
1	**Research Foundations**
1.1	Consider the experts' mental model
1.2	Offer visual aids to elicit participant feedback during design research
2	**Trust**
2.1	Offer references for all information
2.2	Clarify the source of the publication
2.3	Facilitate how to distinguish between internal and external data
3	**Usability**
3.1	Set visual hierarchy between elements
3.2	Incorporate visual response when possible, to aid understanding and learning
3.3	Prioritize ease of use
3.4	Ensure a clean interface
4	**Time Efficiency**
4.1	Ensure the application accelerates the users' workflow
4.2	Integrate only the most essential functions
5	**Transparency**
5.1	Provide explanations to how selections affect the response
5.2	Use clear communication and explain the limitations of the application for enhanced understanding
6	**Engagement**
6.1	Offer follow-up questions
6.2	Encourage interaction with the application

Table 9.1: The identified important factors

These are the important factors to consider when designing conversational agents, that influence user satisfaction, in the context of internal field-specific experts. The factors reflect the insights from the used methods, and can be seen visually realized in the design of the prototype, which was a design influenced by theory, interview results, user tests results, as well as the important factors.

Bibliography

[1] K. L. Ehsani, E. R. Rhythm, M. H. K. Mehedi, and A. A. Rasel, "A comparative analysis of customer service chatbots: Efficiency, usability and application," 2023. DOI: 10.1109/CATS58046.2023.10424303. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85186141213&doi=10.1109%2fCATS58046.2023.10424303&partnerID=40&md5=4a052209d493d53ed12f2a75676f96c9.

[2] C. Heitzinger and S. Woltran, "A short introduction to artificial intelligence: Methods, success stories, and current limitations," in *Introduction to Digital Humanism: A Textbook*, H. Werthner, C. Ghezzi, J. Kramer, *et al.*, Eds. Cham: Springer Nature Switzerland, 2024, pp. 135–149, ISBN: 978-3-031-45304-5. DOI: 10.1007/978-3-031-45304-5_9. [Online]. Available: https://doi.org/10.1007/978-3-031-45304-5_9.

[3] S. Kusal, S. Patil, J. Choudrie, K. Kotecha, S. Mishra, and A. Abraham, "Ai-based conversational agents: A scoping review from technologies to future directions," *IEEE Access*, vol. 10, pp. 92 337–92 356, 2022. DOI: 10.1109/ACCESS.2022.3201144. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85137594701&doi=10.1109%2fACCESS.2022.3201144&partnerID=40&md5=55329c4e112da4c7031bb467bc35e7b7.

[4] S. Marsland, *Machine Learning: An Algorithmic Perspective, Second Edition* (Chapman & Hall/CRC Machine Learning & Pattern Recognition). CRC Press, 2014, ISBN: 9781466583337. [Online]. Available: https://books.google.se/books?id=6GvSBQAAQBAJ.

[5] A. Dubey and A. Rasool, "Recent advances and applications of deep learning technique," *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 13, 2022. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85134404695&partnerID=40&md5=97a9aabd4233c153a12069e21e9ec154.

[6] N. Karanikolas, E. Manga, N. Samaridi, E. Tousidou, and M. Vassilakopoulos, "Large language models versus natural language understanding and generation," in *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics*, ser. PCI '23, Lamia, Greece: Association for Computing Machinery, 2024, pp. 278–290, ISBN: 9798400716263. DOI: 10.1145/3635059.3635104. [Online]. Available: https://doi.org/10.1145/3635059.3635104.

[7] B. Mallikarjuna and P. Chittemsetty, "Generative artificial intelligence: Fundamentals and evolution," in *Generative AI: Current Trends and Applications*,

K. Raza, N. Ahmad, and D. Singh, Eds. Singapore: Springer Nature Singapore, 2024, pp. 3–17, ISBN: 978-981-97-8460-8. DOI: 10.1007/978-981-97-8460-8_1. [Online]. Available: https://doi.org/10.1007/978-981-97-8460-8_1.

[8] A. Singh, A. Ehtesham, S. Kumar, and T. T. Khoei, *Agentic retrieval-augmented generation: A survey on agentic rag*, 2025. arXiv: 2501.09136 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2501.09136.

[9] A. K. Shahade and P. V. Deshmukh, "Enhancing natural language processing: A comprehensive review of retrieval augmented generation," in *2024 4th International Conference on Sustainable Expert Systems (ICSES)*, 2024, pp. 609–611. DOI: 10.1109/ICSES63445.2024.10763224.

[10] E. Liddy, "Natural language processing," *Encyclopedia of Library and Information Science*, 2001.

[11] S. Roukos, "Natural language understanding," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. A. Huang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 617–626, ISBN: 978-3-540-49127-9. DOI: 10.1007/978-3-540-49127-9_31. [Online]. Available: https://doi.org/10.1007/978-3-540-49127-9_31.

[12] Y. Liu, S. Zhou, Y. Ma, and X. Luo, "A review of deep learning-based natural language generation research," 2024, pp. 331–335. DOI: 10.1109/ICNLP60986.2024.10692905. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85207653901&doi=10.1109%2fICNLP60986.2024.10692905&partnerID=40&md5=6b89254e39631cb6119d201f3e1b0082.

[13] D. B. Acharya, K. Kuppan, and B. Divya, "Agentic ai: Autonomous intelligence for complex goals - a comprehensive survey," *IEEE Access*, vol. 13, pp. 18 912–18 936, 2025, Cited by: 0; All Open Access, Gold Open Access. DOI: 10.1109/ACCESS.2025.3532853. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85216361734&doi=10.1109%2fACCESS.2025.3532853&partnerID=40&md5=2f9f2d188d93aa35955a93254893da0f.

[14] OpenAI, *Introducing chatgpt*. [Online]. Available: https://openai.com/index/chatgpt/ (visited on 02/03/2025).

[15] T. Kaufmann, S. Ball, J. Beck, E. Hüllermeier, and F. Kreuter, "On the challenges and practices of reinforcement learning from real human feedback," *Communications in Computer and Information Science*, vol. 2134 CCIS, pp. 276–294, 2025. DOI: 10.1007/978-3-031-74627-7_21. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85215603670&doi=10.1007%2f978-3-031-74627-7_21&partnerID=40&md5=ecc3d6384a34774ecd1f696535560d65.

[16] Anthropic, *Meet claude*. [Online]. Available: https://www.anthropic.com/claude (visited on 02/03/2025).

[17] Microsoft, *Microsoft 365 copilot overview*. [Online]. Available: https://learn.microsoft.com/en-us/copilot/microsoft-365/microsoft-365-copilot-overview (visited on 02/03/2025).

[18] IBM, *What is a chatbot?* [Online]. Available: https://www.ibm.com/think/topics/chatbots (visited on 01/30/2025).

[19] G. Laban and T. Araujo, "Working together with conversational agents: The relationship of perceived cooperation with service performance evaluations," in *Chatbot Research and Design*, A. Følstad, T. Araujo, S. Papadopoulos, *et al.*, Eds., Cham: Springer International Publishing, 2020, pp. 215–228, ISBN: 978-3-030-39540-7.

[20] S. Schöbel, A. Schmitt, D. Benner, *et al.*, "Charting the evolution and future of conversational agents: A research agenda along five waves and new frontiers," *Information Systems Frontiers*, vol. 26, pp. 729–754, 2024. DOI: 10.1007/s10796-023-10375-9. [Online]. Available: https://doi.org/10.1007/s10796-023-10375-9.

[21] J. Weizenbaum, "Elizaa computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, no. 1, pp. 36–45, Jan. 1966, ISSN: 0001-0782. DOI: 10.1145/365153.365168. [Online]. Available: https://doi.org/10.1145/365153.365168.

[22] L. Switzky, "Eliza effects: Pygmalion and the early development of artificial intelligence," *Shaw*, vol. 40, no. 1, pp. 50–68, Jun. 2020, ISSN: 0741-5842. DOI: 10.5325/shaw.40.1.0050. eprint: https://scholarlypublishingcollective.org/psup/shaw/article-pdf/40/1/50/1221119/shaw_40_1_50.pdf. [Online]. Available: https://doi.org/10.5325/shaw.40.1.0050.

[23] S. Kelly, S.-A. Kaye, and O. Oviedo-Trespalacios, "What factors contribute to the acceptance of artificial intelligence? a systematic review," *Telematics and Informatics*, vol. 77, p. 101 925, 2023.

[24] L. Gkinko and A. Elbanna, "Designing trust: The formation of employees trust in conversational ai in the digital workplace," *Journal of Business Research*, vol. 158, p. 113 707, 2023, ISSN: 0148-2963. DOI: https://doi.org/10.1016/j.jbusres.2023.113707. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0148296323000656.

[25] M. Jain, P. Kumar, R. Kota, and S. N. Patel, "Evaluating and informing the design of chatbots," in *Proceedings of the 2018 Designing Interactive Systems Conference*, ser. DIS '18, Hong Kong, China: Association for Computing Machinery, 2018, pp. 895–906, ISBN: 9781450351980. DOI: 10.1145/3196709.3196735. [Online]. Available: https://doi.org/10.1145/3196709.3196735.

[26] J. W. Treem, S. L. Dailey, C. S. Pierce, and P. M. Leonardi, "Bringing technological frames to work: How previous experience with social media shapes the technology's meaning in an organization," *Journal of Communication*, vol. 65, no. 2, pp. 396–422, Mar. 2015, ISSN: 0021-9916. DOI: 10.1111/jcom.12149. eprint: https://academic.oup.com/joc/article-pdf/65/2/396/22321044/jjnlcom0396.pdf. [Online]. Available: https://doi.org/10.1111/jcom.12149.

[27] X. Yang and M. Aurisicchio, "Designing conversational agents: A self-determination theory approach," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI '21, Yokohama, Japan: Association for Computing Machinery, 2021, ISBN: 9781450380966. DOI: 10.1145/3411764.3445445. [Online]. Available: https://doi.org/10.1145/3411764.3445445.

[28] C. S. Kopplin, "Chatbots in the workplace: A technology acceptance study applying uses and gratifications in coworking spaces," *Journal of Organizational Computing and Electronic Commerce*, vol. 32, no. 3-4, pp. 232–257, 2022. DOI: 10.1080/10919392.2023.2215666. eprint: https://doi.org/10.1080/10919392.2023.2215666. [Online]. Available: https://doi.org/10.1080/10919392.2023.2215666.

[29] IBM, *Nlp vs. nlu vs. nlg: The differences between three natural language processing concepts*. [Online]. Available: https://www.ibm.com/think/topics/nlp-vs-nlu-vs-nlg (visited on 01/30/2025).

[30] OpenAI, *Introducing chatgpt enterprise*. [Online]. Available: https://openai.com/index/introducing-chatgpt-enterprise/ (visited on 01/15/2025).

[31] IBM, *What are ai agents?* [Online]. Available: https://www.ibm.com/think/topics/ai-agents (visited on 01/30/2025).

[32] IBM, *What is conversational ai?* [Online]. Available: https://www.ibm.com/think/topics/conversational-ai?mhsrc=ibmsearch_a&mhq=conversational%20ai (visited on 01/30/2025).

[33] A. L. Kotian, R. Nandipi, U. M, U. R. S, VARSHAUK, and V. G. T, "A systematic review on human and computer interaction," in *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, 2024, pp. 1214–1218. DOI: 10.1109/IDCIoT59759.2024.10467622.

[34] J. Zimmerman and J. Forlizzi, "Research through design in hci," in *Ways of Knowing in HCI*, J. S. Olson and W. A. Kellogg, Eds. New York, NY: Springer New York, 2014, pp. 167–189, ISBN: 978-1-4939-0378-8. DOI: 10.1007/978-1-4939-0378-8_8. [Online]. Available: https://doi.org/10.1007/978-1-4939-0378-8_8.

[35] W. Gaver, "What should we expect from research through design?" *Conference on Human Factors in Computing Systems - Proceedings*, May 2012. DOI: 10.1145/2207676.2208538.

[36] J. Sweller, J. J. G. van Merrienboer, and F. G. W. C. Paas, "Cognitive architecture and instructional design," *Educational Psychology Review*, vol. 10, no. 3, pp. 251–296, 1998, ISSN: 1040726X, 1573336X. [Online]. Available: http://www.jstor.org/stable/23359412 (visited on 05/20/2025).

[37] A. D. Baddeley and G. Hitch, "Working memory," *Psychology of Learning and Motivation - Advances in Research and Theory*, vol. 8, no. C, pp. 47–89, 1974, Cited by: 8947. DOI: 10.1016/S0079-7421(08)60452-1. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-77956965422&doi=10.1016%2fS0079-7421%2808%2960452-1&partnerID=40&md5=6c7ad137e7f919517ddb7b738506c434.

[38] G. Vallar, *Short-term memory*. 2017, pp. 367–381, Cited by: 8. DOI: 10.1016/B978-0-12-809324-5.03170-9. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85079286202&doi=10.1016%2fB978-0-12-809324-5.03170-9&partnerID=40&md5=52d85e49bad225a36b6433e4b8ac10c1.

[39] C. Nass, J. Steuer, and E. R. Tauber, "Computers are social actors," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1994, pp. 72–78.

[40] K. E. Arleen Salles and M. Farisco, "Anthropomorphism in ai," *AJOB Neuroscience*, vol. 11, no. 2, pp. 88–95, 2020, PMID: 32228388. DOI: 10.1080/21507740.2020.1740350. eprint: https://doi.org/10.1080/21507740.2020.1740350. [Online]. Available: https://doi.org/10.1080/21507740.2020.1740350.

[41] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989, ISSN: 02767783, 21629730. [Online]. Available: http://www.jstor.org/stable/249008 (visited on 01/21/2025).

[42] S. Y. Yousafzai, G. R. Foxall, and J. G. Pallister, "Technology acceptance: A meta-analysis of the tam: Part 1," *Journal of modelling in management*, vol. 2, no. 3, pp. 251–280, 2007.

[43] D. L. Goodhue and R. L. Thompson, "Task-technology fit and individual performance," *MIS Quarterly*, vol. 19, no. 2, pp. 213–236, 1995, ISSN: 02767783, 21629730. [Online]. Available: http://www.jstor.org/stable/249689 (visited on 01/21/2025).

[44] International Organization for Standardization, *Iso 9241-11:2018(en) - ergonomics of human-system interaction – part 11: Usability: Definitions and concepts*, Accessed: 2025-02-04, 2018. [Online]. Available: https://www.iso.org/standard/63500.html.

[45] D. A. Norman, *The Design of Everyday Things*. Basic Books, 2013.

[46] M. Bordegoni, M. Carulli, and E. Spadoni, "User experience and user experience design," in *Prototyping User eXperience in eXtended Reality*. Cham: Springer Nature Switzerland, 2023, pp. 11–28, ISBN: 978-3-031-39683-0. DOI: 10.1007/978-3-031-39683-0_2. [Online]. Available: https://doi.org/10.1007/978-3-031-39683-0_2.

[47] A. Pitale and A. Bhumgara, "Human computer interaction strategies-designing the user interface," in *Proceedings of the 2nd International Conference on Smart Systems and Inventive Technology, ICSSIT 2019*, 2019, pp. 752–758. DOI: 10.1109/ICSSIT46314.2019.8987819. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85080070953&doi=10.1109%2fICSSIT46314.2019.8987819&partnerID=40&md5=c2c7467dc3984e1871e8df22d466f3d0.

[48] A. Cooper, R. Reimann, D. Cronin, and C. Noessel, *About Face: The Essentials of Interaction Design*. Hoboken, NJ: John Wiley & Sons, Incorporated, 2014.

[49] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Proceedings 1996 IEEE Symposium on Visual Languages*, 1996, pp. 336–343. DOI: 10.1109/VL.1996.545307.

[50] M. Wahid, A. Waris, S. O. Gilani, and R. Subramanian, "The effect of eye movements in response to different types of scenes using a graph-based visual saliency algorithm," *Applied Sciences (Switzerland)*, vol. 9, no. 24, 2019, Cited by: 5; All Open Access, Gold Open Access. DOI: 10.3390/app9245378. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-

s2.0-85077337492&doi=10.3390%2fapp9245378&partnerID=40&md5=2bc7d172c85e5e27590fc51b9f09cc3a.

[51] O. Y. Zhu, D. Greene, and S. Dolnicar, "Should the risk of social desirability bias in survey studies be assessed at the level of each pro-environmental behaviour?" *Tourism Management*, vol. 104, p. 104 933, 2024, ISSN: 0261-5177. DOI: https://doi.org/10.1016/j.tourman.2024.104933. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0261517724000529.

[52] S. Jahedizadeh, B. Ghonsooly, and A. H. Fatemi, "Student evaluation apprehension: An interdisciplinary review of determinants and ramifications," *Polish Psychological Bulletin*, vol. 50, no. 3, pp. 226–236, 2019, Cited by: 2; All Open Access, Gold Open Access. DOI: 10.24425/ppb.2019.130695. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85084637993&doi=10.24425%2fppb.2019.130695&partnerID=40&md5=228b8a74d7f5fbcf5c038022fd780fb1.

[53] D. Leslie, "Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of ai systems in the public sector," en, 2019. DOI: 10.5281/ZENODO.3240529. [Online]. Available: https://zenodo.org/record/3240529.

[54] H. Snyder, "Literature review as a research methodology: An overview and guidelines," *Journal of Business Research*, vol. 104, pp. 333–339, 2019, ISSN: 0148-2963. DOI: https://doi.org/10.1016/j.jbusres.2019.07.039. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0148296319304564.

[55] H. Sharp, Y. Rogers, and J. Preece, *Interaction Design: Beyond Human-Computer Interaction*, 2nd. Chichester, England: John Wiley & Sons, 2007.

[56] B. Hannington and B. Martin, *Universal Methods of Design, Expanded and Revised*. Minneapolis: Quarto Publishing Group USA, 2019, Available from: ProQuest Ebook Central. [4 February 2025].

[57] M. A. Laitinen, "Net promoter score as indicator of library customers' perception," *Journal of Library Administration*, vol. 58, no. 4, pp. 394–406, 2018. DOI: 10.1080/01930826.2018.1448655. eprint: https://doi.org/10.1080/01930826.2018.1448655. [Online]. Available: https://doi.org/10.1080/01930826.2018.1448655.

[58] R. Janghorban, R. Latifnejad Roudsari, and A. Taghipour, "Pilot study in qualitative research: The roles and values," *HAYAT*, vol. 19, no. 4, pp. 1–5, 2014, Cited by: 19. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84896448646&partnerID=40&md5=1a1dcabb7e42268fb32e98dd6768a65e.

[59] R. I. Sutton and A. Hargadon, "Brainstorming groups in context: Effectiveness in a product design firm," *Administrative Science Quarterly*, vol. 41, no. 4, pp. 685–718, 1996, Cited by: 750. DOI: 10.2307/2393872. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0030354997&doi=10.2307%2f2393872&partnerID=40&md5=0ec818b35eb0ceb60c82eef2d0cd38ed.

[60] Google, *Crazy 8's*. [Online]. Available: https://designsprintkit.withgoogle.com/methodology/phase3-sketch/crazy-8s (visited on 04/15/2025).

[61] S. Gibbons, *Dot voting: A simple decision-making and prioritizing technique in ux*. [Online]. Available: https://www.nngroup.com/articles/dot-voting/ (visited on 04/15/2025).

[62] S. Garner and D. McDonagh-Philp, "Problem interpretation and resolution via visual stimuli: The use of mood boards in design education," *Journal of Art & Design Education*, vol. 20, no. 1, pp. 57–64, 2001.

[63] B. Tversky and M. Suwa, *Thinking with Sketches*. 2009, Cited by: 116. DOI: 10.1093/acprof:oso/9780195381634.003.0004. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84920928687&doi=10.1093%2facprof%3aoso%2f9780195381634.003.0004&partnerID=40&md5=43ec7f0218a3d283bc237e51847d200f.

[64] M. J. Hamm, *Wireframing essentials*. Packt Publishing Ltd, 2014.

[65] R. B. Johnson and A. J. Onwuegbuzie, "Mixed methods research: A research paradigm whose time has come," *Educational Researcher*, vol. 33, no. 7, pp. 14–26, 2004, Cited by: 5025; All Open Access, Green Open Access. DOI: 10.3102/0013189X033007014. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84993820511&doi=10.3102%2f0013189X033007014&partnerID=40&md5=d692cd96ab991f4125581a4a67f27adb.

[66] F. Heimerl, S. Lohmann, S. Lange, and T. Ertl, "Word cloud explorer: Text analytics based on word clouds," in *2014 47th Hawaii International Conference on System Sciences*, 2014, pp. 1833–1842. DOI: 10.1109/HICSS.2014.231.

[67] V. Braun and V. Clarke, "Thematic analysis," in *Encyclopedia of quality of life and well-being research*, Springer, 2024, pp. 7187–7193.

[68] S. Riihiaho, "Usability testing," in *The Wiley Handbook of Human Computer Interaction*. John Wiley & Sons, Ltd, 2018, ch. 14, pp. 255–275, ISBN: 9781118976005. DOI: https://doi.org/10.1002/9781118976005.ch14. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118976005.ch14. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118976005.ch14.

[69] G. Charness, U. Gneezy, and M. A. Kuhn, "Experimental methods: Between-subject and within-subject design," *Journal of Economic Behavior and Organization*, vol. 81, no. 1, pp. 1–8, 2012, Cited by: 954. DOI: 10.1016/j.jebo.2011.08.009. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-80054064387&doi=10.1016%2fj.jebo.2011.08.009&partnerID=40&md5=779f02ced55ef76c6fbf0c0360c8c3c4.

[70] K. A. Ericsson and H. A. Simon, "Verbal reports as data," *Psychological Review*, vol. 87, no. 3, pp. 215–251, 1980, Cited by: 3015. DOI: 10.1037/0033-295X.87.3.215. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-58149405625&doi=10.1037%2f0033-295X.87.3.215&partnerID=40&md5=9d4fe373185813e51dbdd4a953fc86a3.

[71] K. Finstad, "The usability metric for user experience," *Interacting with Computers*, vol. 22, no. 5, pp. 323–327, 2010, Cited by: 442. DOI: 10.1016/j.

intcom.2010.04.004. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-77955429964&doi=10.1016%2fj.intcom.2010.04.004&partnerID=40&md5=9a09eea86b5cf3fde0ed2c271de59e11.

[72] J. Mumu, B. Tanujaya, R. Charitas, and I. Prahmana, "Likert scale in social sciences research: Problems and difficulties," *FWU Journal of Social Sciences*, vol. 16, no. 4, pp. 89–101, 2022, Cited by: 50; All Open Access, Bronze Open Access. DOI: 10.51709/19951272/Winter2022/7. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85145572957&doi=10.51709%2f19951272%2fWinter2022%2f7&partnerID=40&md5=70c854143f0b610d6a9142e83b753183.

[73] A. Bangor, P. T. Kortum, and J. T. Miller, "An empirical evaluation of the system usability scale," *International Journal of HumanComputer Interaction*, vol. 24, no. 6, pp. 574–594, 2008. DOI: 10.1080/10447310802205776. eprint: https://doi.org/10.1080/10447310802205776. [Online]. Available: https://doi.org/10.1080/10447310802205776.

[74] World Wide Web Consortium (W3C), *Understanding WCAG 2.1 - contrast minimum*, Accessed: 2025-05-27, 2025. [Online]. Available: https://www.w3.org/WAI/WCAG21/Understanding/contrast-minimum.html.

[75] S. A. Eide, A.-M. Poljac, and F. E. Sandnes, "Image search versus text search revisited: A simple experiment using a kahoot quiz," in *International Conference on Human-Computer Interaction*, Springer, 2021, pp. 186–196.

A

Appendix: Survey

INTRODUCTION
1. Gender
Woman
Man
Non-binary
Prefer not to say
Other [text]

2. Age
18-29
30-39
40-49
50-59
60+

3. Which department do you belong to? (e.g. IT)

4. What is your role?

Workflow

5. Do you use [RA]?
Yes *Go to next question*
No *Go to section Not using [RA]*

6. How often do you use [RA]?
Only tried once
1-5 times a month
1-5 times a week
Almost every day

7. For what tasks in your work do you use Research Assistant?

8. Rate how well you were able to use the Research Assistant on

your first attempt.
Not at all 1-6 Very well

9. Do you use the chat history? (e.g. go back to previous search queries, what you have searched before)
Yes
No
I dont know

10. Do you use the different modes (Factual/Balanced/Creative)? (e.g. switching between the different modes)
Yes
No I dont need them
No I dont understand the modes
I dont understand the question

USER EXPERIENCE

11. What are your main benefits of using Research Assistant?

12. What are your main pain points of using Research Assistant?

13. Would you like to add any extra functions or personalization in Research Assistant?
Yes *Go to next question*
No, its not needed *Skip next question*

14. What could the extra functions be? (You can add as many as you would like)

INTERFACE

15. Rate the interface of [RA]
I find [RA] easy to use: Strongly disagree 1-6 Strongly agree
I am satisfied with my experience using Research Assistant: Strongly disagree 1-6 Strongly agree

IF NOT USING RA

16. What is the reason for not using [RA]?
I don't have a need for this tool

I don't see the value
I find the replies incorrect or insufficient
Other [text]

17. When in your work could you include Research Assistant?

FINAL NOTES

18. Do you use any of the following [department] tools? Check all boxes that apply. If you do not use any of these tools, please proceed to next question.
[Five application alternatives]

19. Do you use any other AI chatbots at work? Check all boxes that apply. If you do not use any of these tools, please proceed to next question.
[Two application alternatives + Other]

20. How likely are you to recommend Research Assistant to a colleague? 1-10

21. Is there anything you would like to let us know?

22. Would you like to participate in a follow-up interview and take the opportunity to influence the future development of Research Assistant?
Yes
No

B

Appendix: Email template interviews

Subject: Invitation to 45–60 Minute Master Thesis Interview / Invitation to Research Assistant Interview - Master Thesis Students

Hi [Name],

Thank you for responding to our master thesis survey about Research Assistant. We will now be conducting interviews to gain a more in-depth understanding of the usage of the application and its users and are therefore reaching out to you.

Interview details:
When: Please respond to this email with suggestions of suitable times for you during 3rd–7th of March.
Duration: 45–60 minutes
Format: Teams meeting, conducted by one of us thesis students, either Sara Börjesson or Karin Örn Andersson.
Recording: We kindly ask for your permission to record the session for analysis purposes. The recording will be kept within the project group and used solely for research purposes.
Screen sharing: We will also ask you to share your screen, to allow us to observe your workflow with the application.

Before the interview:
Please prepare by opening Research Assistant and reflect on when during your work you use the application.

Please let us know if you are totally occupied during 3rd–7th of March with another suggested date, and we will do our best to accommodate your schedule. We are very thankful for your contribution to our Master thesis and to the improvement of the Research Assistant.

If you have any questions or concerns, feel free to reach out to us directly. We are looking forward to the interview and to learn from your experiences.

C

Appendix: Interview questions

INTRODUCTION
Hi and welcome [participants name]! How are you today?

My name is [name] and I am a thesis worker here at AstraZeneca. Together with [Sara/Karin], I am doing my master thesis at Chalmers, at the master Interaction Design. The purpose of our thesis is to explore and improve Research Assistant, your answers will therefore contribute to the further development of the application and the insights will be included in our thesis report. However, all answers will be anonymized.

Today we want to understand your workflow, needs and potential pain points. Do you have any questions before we begin?

Great! If it is okay, I would like to record this session so my colleagues and I can review our discussions.

ROLE & WORK Please describe your role and work with a few sentences/with a minute or two.

How long have you been in Pharma?

WORKFLOW
Describe your process when you use the Research Assistant. Please keep in mind that I do not have a scientific background.

What happens next after you have received an answer from Research Assistant?

What steps would you take if the Research Assistant doesn't display information you believe is available?

Has your way of working changed since starting using RA?

PROS & CONS
Describe your main benefits/pain points of using Research Assistant

TRUST
Do you trust the answer from Research Assistant?
If NO, what would make you trust the RA system's responses?
IF YES, What makes you trust the Research Assistant's responses?

Do you verify the information?
Why do you verify it?
When do you verify?
Do you always do that or is it in certain circumstances?

PROMPTING
Describe your thought process when writing a prompt

SECTION DESIGN SUGGESTIONS

MODES
Show the mode slide with only the word and ask the questions
Description: Now we will talk about modes.
Do you use the different modes? (If users asks: what are the modes? - Show image of Research Assistant with the modes)
If yes, What does the modes mean for you?
If yes, When do you use the different modes?
Can they be improved and how?
If no, why not?
What modes would you like to have?

Showing modes visual
Description: Here you can see two new suggestions of how to visualize the modes, one with buttons below and information-symbol to see the descriptions, and one with a drop-down-menu where you can see the descriptions right away.

Questions modes when showing design suggestions:
What do you think when you see this design?
In terms of improvement, would you want to see this in a different way? Would you want something totally different?

REFERENCES
Now we will move on to references.
Show the reference slide with only the word and ask the questions
Now we will talk about references in Research Assistant.
Whats your opinion on the way the references are presented?
Can you elaborate, what could be improved?
What type of information would you like to see?

Showing references visual

Here you can see a new suggestion of how references are shown. When clicking on a number to the left, the reference is shown to the right. You can also see if the reference is from internal or external data.

Questions when showing design suggestion:
What do you think when you see this design?
In terms of improvement, would you want to see this in a different way? Would you want something totally different?

Showing functions visual
Now we will move on to functions in research assistant.
Description: Here you can see a new suggestion of settings and functions in the interface. You can choose level of depth, reference type and length of answer the research assistant would give / how it should answer.

Questions when showing design suggestion:
What do you think of these settings?
Are there any other settings or functions that would be important to you or that you would like to see or add?

END NOTES Is there anything else you would like to add?

Thank the participant
Thank you so much for your time! We are so grateful for your participation and this will help us in the development of Research Assistant and for our master thesis.

D

Appendix: Interview questions for Non-user of RA

INTRODUCTION
Hi and welcome [participant's name]! How are you today?

My name is [name] and I am a thesis worker here at AstraZeneca. Together with [Sara/Karin], I am doing my master thesis at Chalmers, in the master Interaction Design. The purpose of our thesis is to explore and improve Research Assistant; your answers will therefore contribute to the further development of the application and the insights will be included in our thesis report. However, all answers will be anonymized.

Today we want to understand your workflow, needs, and potential pain points. Do you have any questions before we begin?

Great! If it is okay, I would like to record this session so my colleagues and I can review our discussions.

ROLE & WORK
Please describe your role and work with a few sentences. Keep in mind that I do not have a scientific background. How long have you been in Pharma?

WORKFLOW
Describe a general work process when you are doing a literature search. How do you find information internally at AZ? How do you find information externally?

VERIFICATION
Do you verify the information? Why do you verify it? When do you verify? Do you always do that or is it in certain circumstances?

AI TOOLS
What is your opinion on AI tools? Do you use any AI tools in your work? What steps would you take if you don't find information you believe is available? Has your way of working changed since you started using AI tools?

TRUST
Do you trust the answers from the AI tool? If NO, what would make you trust the

responses? IF YES, What makes you trust the responses?

PROS AND CONS
Describe your main pain points/benefits when using AZ ChatGPT?

NEW TOOLS
How do you find new tools at AZ? How do you want to find new tools? Where? How open-minded would you say you are to try new technology or tools?

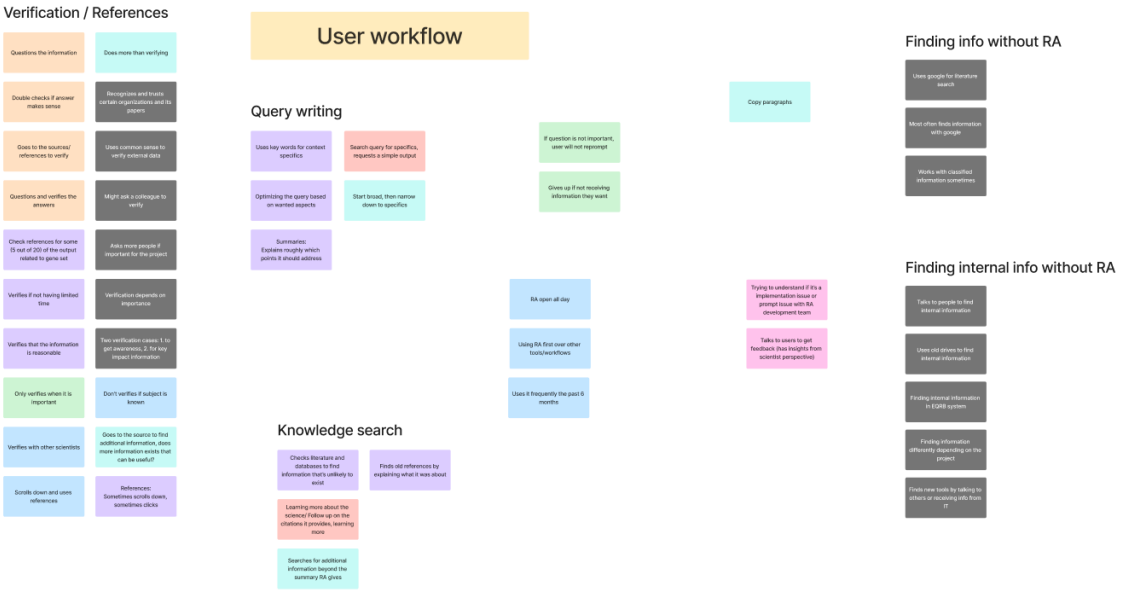
END NOTES
Is there anything else you would like to add?

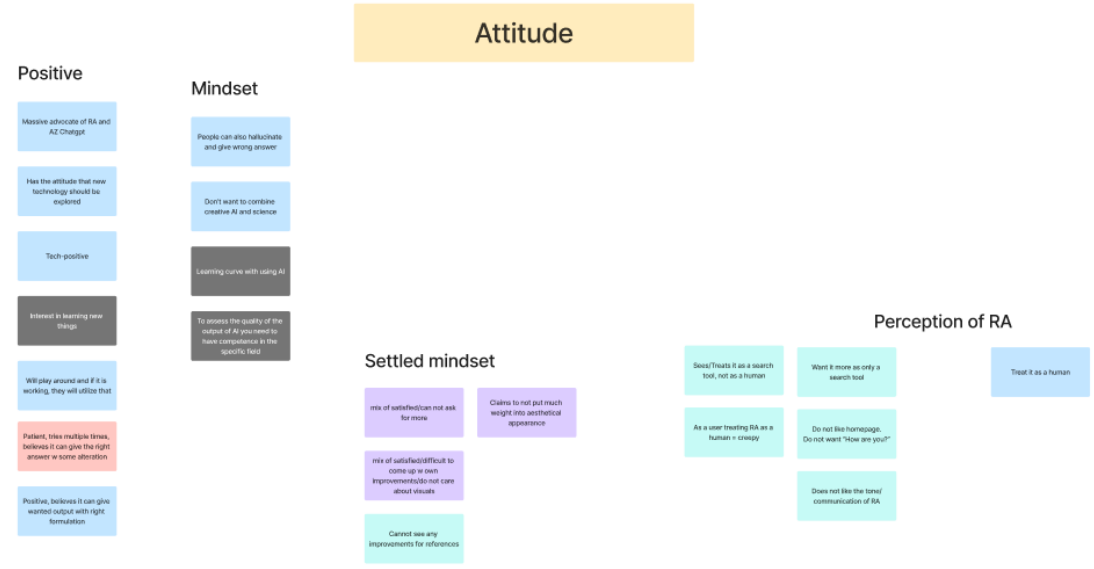
Thank the participant
Thank you so much for your time! We are so grateful for your participation, and this will help us in the development of Research Assistant and for our master thesis.

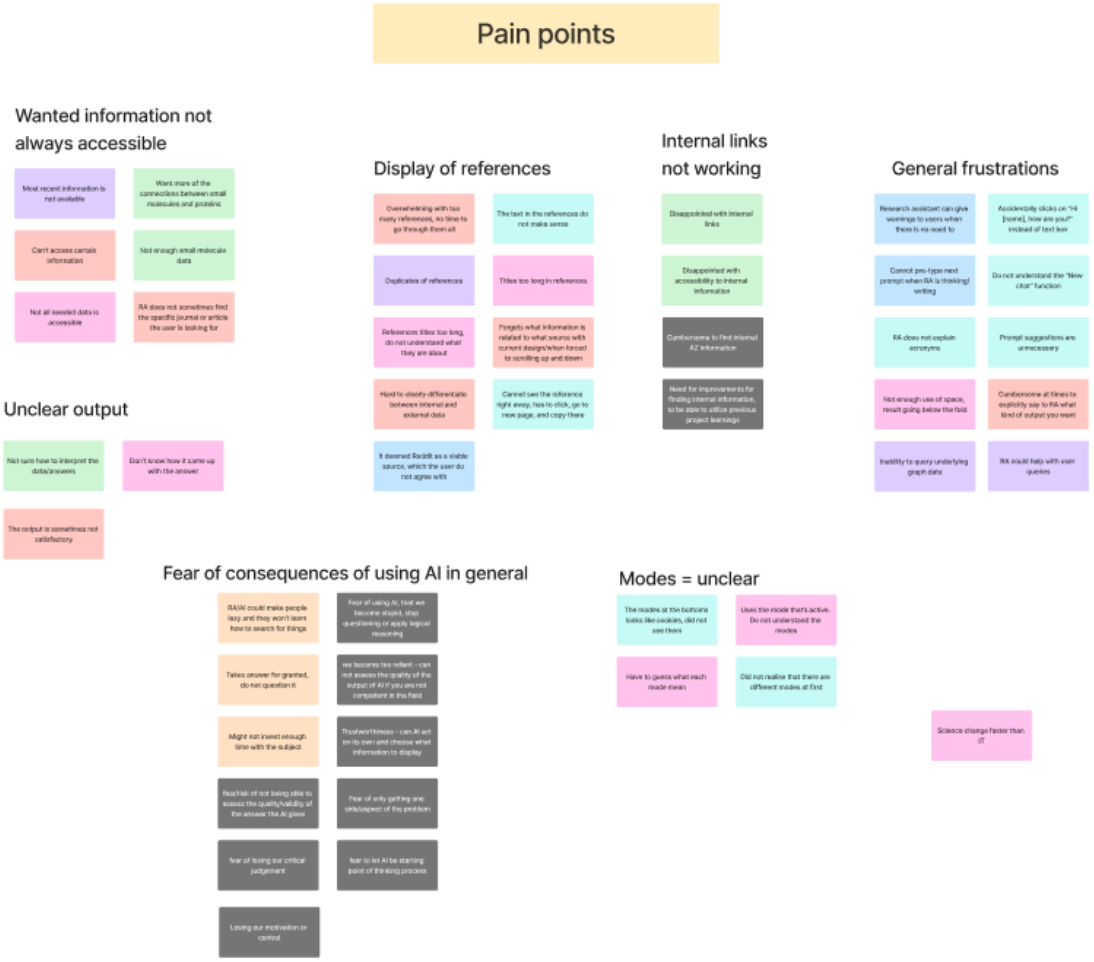
E

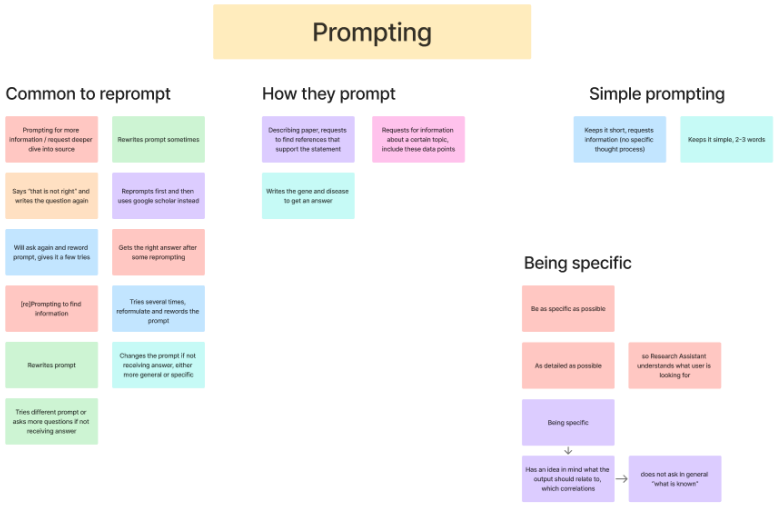
Appendix: Interview Analysis

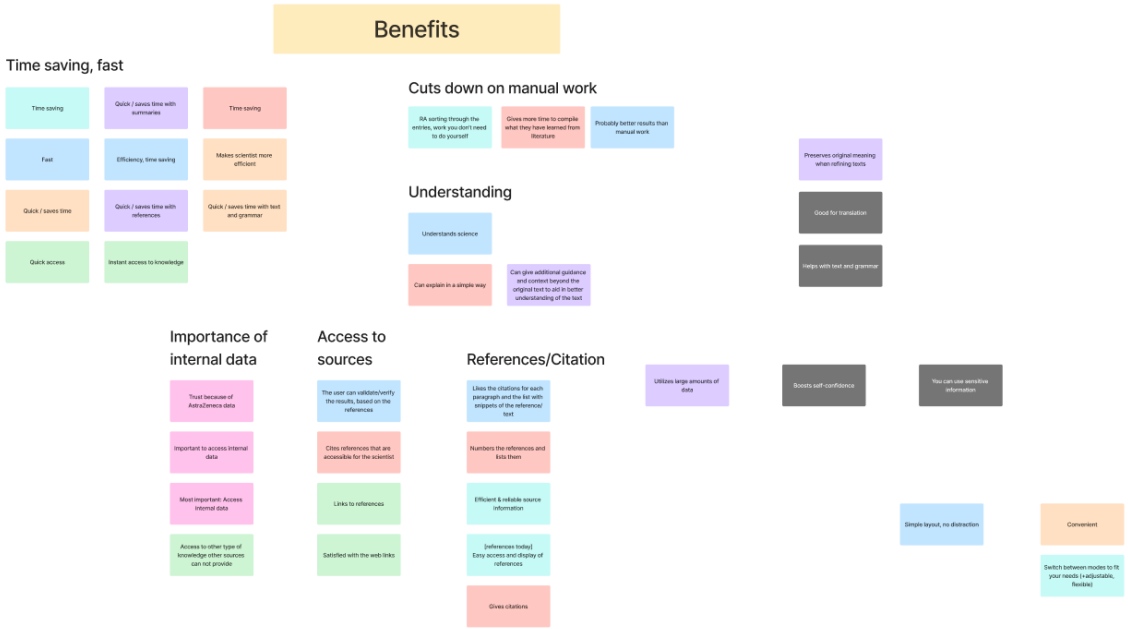

Use cases
Understanding
Investigate new angle to a subject
Learn new subjects
Can get summaries
Aid in understanding how the compound is targeting
Aid in understanding the project, how things are related/interconnected
Can read documents and present it for the user
Background information for projects
Understanding
Helps with presentations and graphic suggestions
Background for projects / Aid in understanding the project, how things are related/interconnected
Aid for understanding the project, how things are related/interconnected
Explore and understand targets
Understand history of projects
Summaries
Summary (of papers)
Create summary tables
Literature summary, get familiar with topic
Literature survey/search
Inspiration tool for literature
To read large articles
Summarization of subject, key points
Quick summary of own report/statement
Get quick answers to questions
Linking compounds with something more specific, like a disease
Hypothesis generation
Improve SOPs (standard operation procedures)
Translation
Find information
Retrieve/find old references/ information
Information searching - has this been processed/worked on before?
Checking if information is available
Data verification
Writing aid
Writing aid
Email creation with desired/ suitable tone
Write documents for authorities
For annual performance reviews
Create presentations
Preparing presentations & introductions
Internal number searching, multiple at a time
Goes through millions of compounds, iterative process
Checks correlations for all the genes in the gene set
Data integration

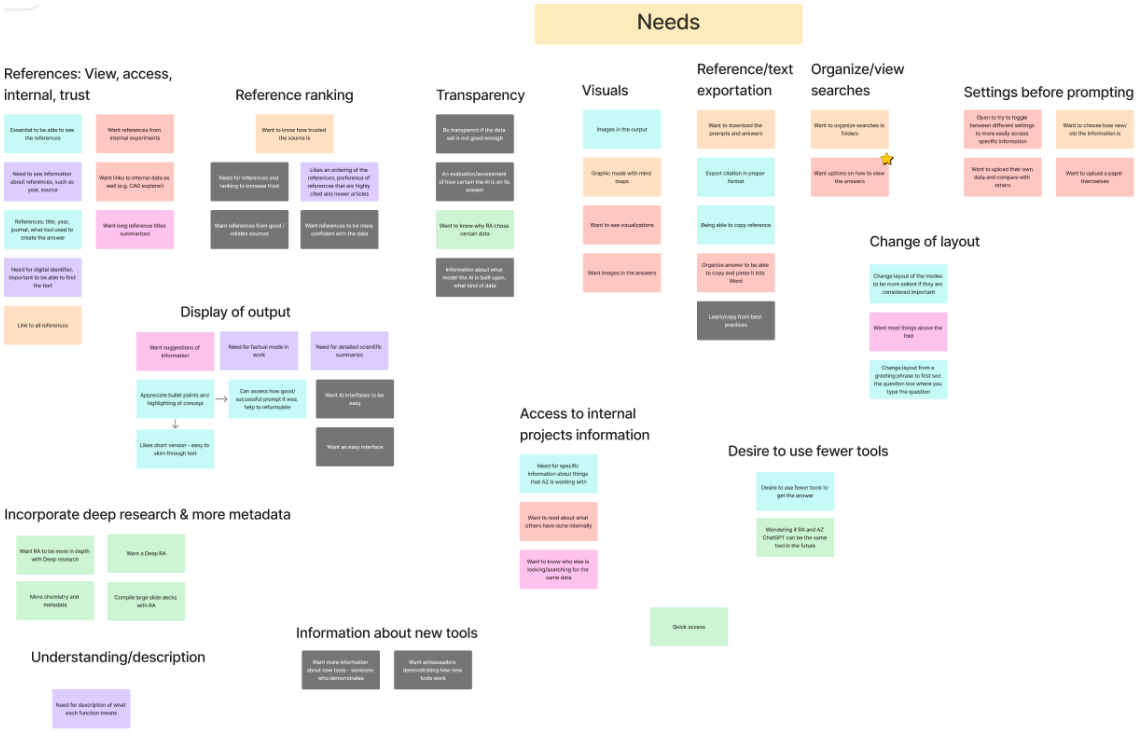

Verification / References
Questions the information
Does more than verifying
Double checks if answer makes sense
Recognizes and trusts certain organizations and its papers
Goes to the sources/ references to verify
Uses common sense to verify external data
Questions and verifies the answers
Might ask a colleague to verify
Check references for some (5 out of 20) of the output related to gene set
Asks more people if important for the project
Verifies if not having limited time
Verification depends on importance
Verifies that the information is reasonable
Two verification cases: 1. to get awareness, 2. for key impact information
Only verifies when it is important
Don't verifies if subject is known
Verifies with other scientists
Goes to the source to find additional information, does more information exists that can be useful?
Scrolls down and uses references
References: Sometimes scrolls down, sometimes clicks
User workflow
Query writing
Uses key words for context specifics
Search query for specifics, requests a simple output
Optimizing the query based on wanted aspects
Start broad, then narrow down to specifics
Summaries: Explains roughly which points it should address
If question is not important, user will not reprompt
Gives up if not receiving information they want
Copy paragraphs
Knowledge search
Checks literature and databases to find information that's unlikely to exist
Finds old references by explaining what it was about
Learning more about the science/ Follow up on the citations it provides, learning more
Searches for additional information beyond the summary RA gives
RA open all day
Using RA first over other tools/workflows
Uses it frequently the past 6 months
Trying to understand if it's a implementation issue or prompt issue with RA development team
Talks to users to get feedback (has insights from scientist perspective)
Finding info without RA
Uses google for literature search
Most often finds information with google
Works with classified information sometimes
Finding internal info without RA
Talks to people to find internal information
Uses old drives to find internal information
Finding internal information in EQRB system
Finding information differently depending on the project
Finds new tools by talking to others or receiving info from IT

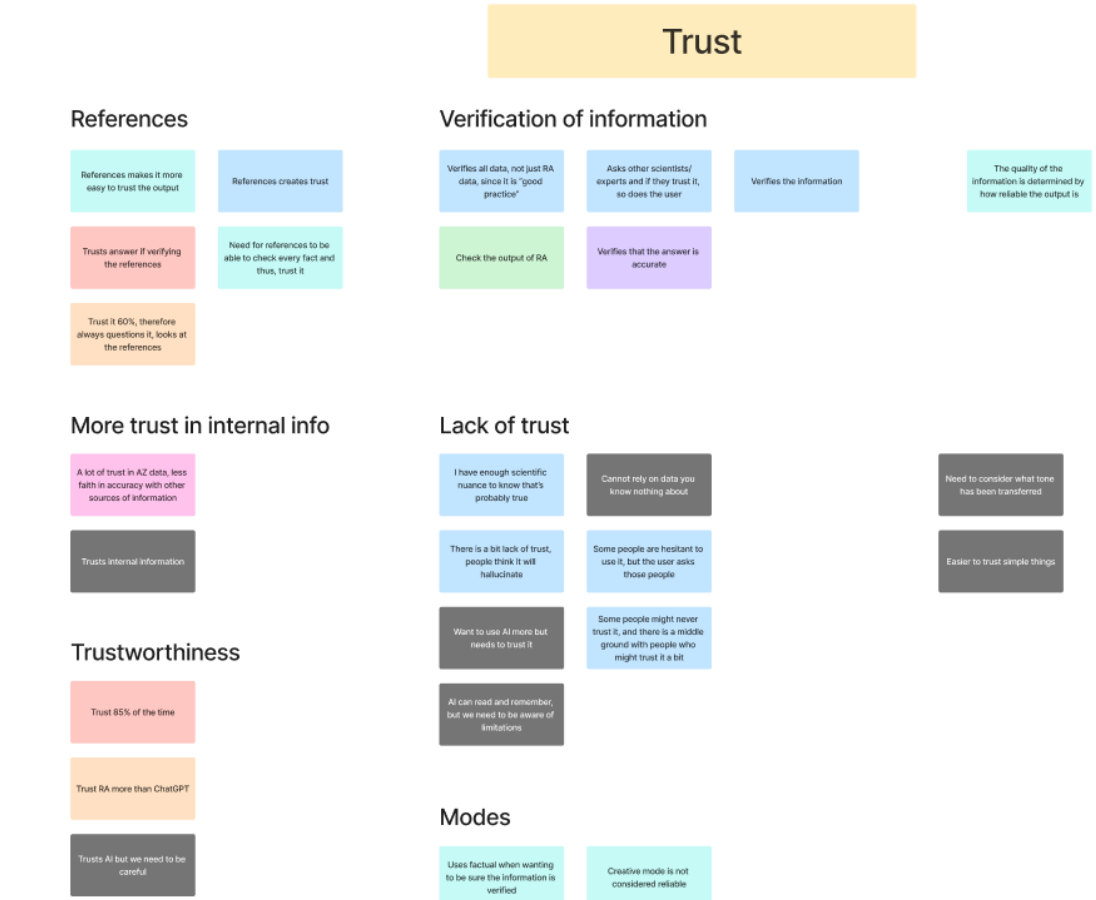

Attitude
Positive
Massive advocate of RA and AZ Chatgpt
Has the attitude that new technology should be explored
Tech-positive
Interest in learning new things
Will play around and if it is working, they will utilize that
Patient, tries multiple times, believes it can give the right answer w some alteration
Positive, believes it can give wanted output with right formulation
Mindset
People can also hallucinate and give wrong answer
Don't want to combine creative AI and science
Learning curve with using AI
To assess the quality of the output of AI you need to have competence in the specific field
Settled mindset
mix of satisfied/can not ask for more
Claims to not put much weight into aesthetical appearance
mix of satisfied/difficult to come up w own improvements/do not care about visuals
Cannot see any improvements for references
Perception of RA
Sees/Treats it as a search tool, not as a human
Want it more as only a search tool
Treat it as a human
As a user treating RA as a human = creepy
Do not like homepage. Do not want "How are you?"
Does not like the tone/ communication of RA

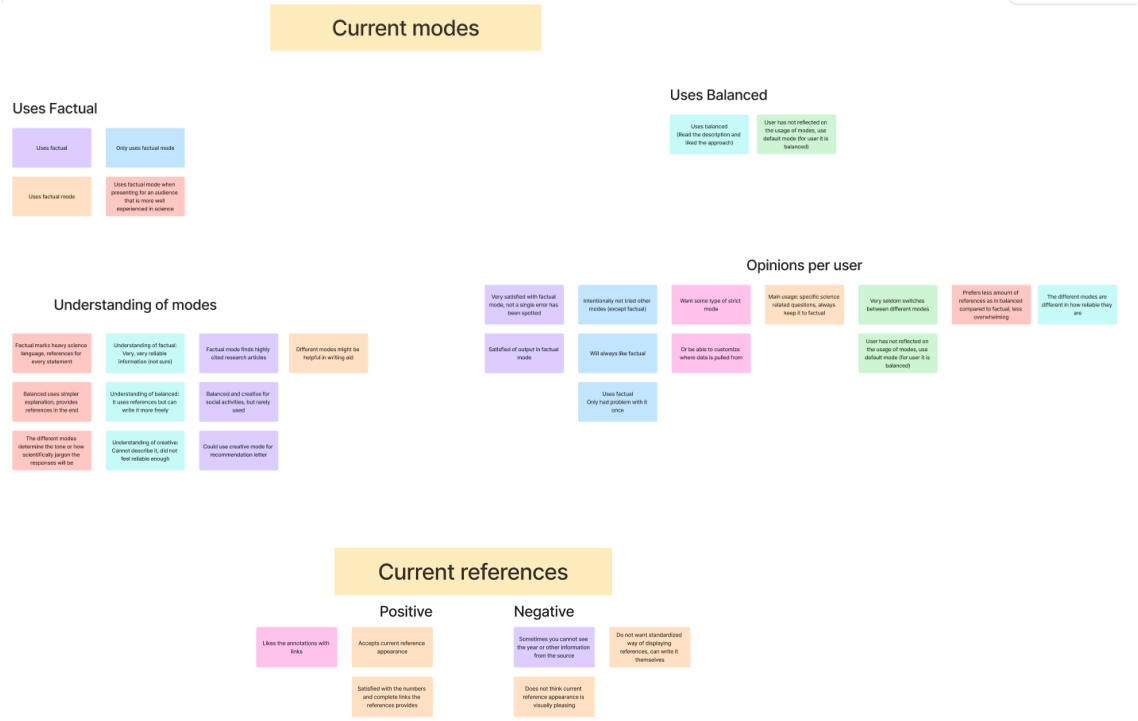

Pain points
Wanted information not always accessible
Most recent information is not available
Want more of the connections between small molecules and proteins
Can't access certain information
Not enough small molecule data
Not all needed data is accessible
RA does not sometimes find the specific journal or article the user is looking for
Unclear output
Not sure how to interpret the data/answers
Don't know how it came up with the answer
The output is sometimes not satisfactory
Display of references
Overwhelming with too many references, no time to go through them all
The text in the references do not make sense
Duplicates of references
Titles too long in references
References titles too long, do not understand what they are about
Forgets what information is related to what source with current design/when forced to scrolling up and down
Hard to closely differentiate between internal and external data
Cannot see the reference right away, has to click, go to new page, and copy there
It deemed Reddit as a viable source, which the user do not agree with
Internal links not working
Disappointed with internal links
Disappointed with accessibility to internal information
Cumbersome to find internal AZ information
Need for improvements for finding internal information, to be able to utilize previous project learnings
General frustrations
Research assistant can give warnings to users when there is no need to
Accidentally clicks on "Hi [name], how are you?" instead of text box
Cannot pre-type next prompt when RA is thinking/ writing
Do not understand the "New chat" function
RA does not explain acronyms
Prompt suggestions are unnecessary
Not enough use of space, result going below the fold
Cumbersome at times to explicitly say to RA what kind of output you want
Inability to query underlying graph data
RA could help with user queries
Fear of consequences of using AI in general
RAIA could make people lazy and they won't learn how to search for things
Fear of using AI, that we become stupid, stop questioning or apply logical reasoning
Takes answer for granted, do not question it
we become too reliant - can not assess the quality of the output of AI if you are not competent in the field
Might not invest enough time with the subject
Trustworthiness - can AI act on its own and choose what information to display
Risk/risk of not being able to assess the quality/validity of the answer the AI gives
Fear of only getting one side/aspect of the problem
fear of losing our critical judgement
fear to let AI be starting point of thinking process
Losing our motivation or control
Modes = unclear
The modes at the bottoms looks like cookies, did not see them
Uses the mode that's active. Do not understand the modes
Have to guess what each mode mean
Did not realise that there are different modes at first
Science change faster than IT


Prompting
Common to reprompt
Prompting for more information / request deeper dive into source
Rewrites prompt sometimes
Says "that is not right" and writes the question again
Reprompts first and then uses google scholar instead
Will ask again and reword prompt, gives it a few tries
Gets the right answer after some reprompting
[re]Prompting to find information
Tries several times, reformulate and rewords the prompt
Rewrites prompt
Changes the prompt if not receiving answer, either more general or specific
Tries different prompt or asks more questions if not receiving answer
How they prompt
Describing paper, requests to find references that support the statement
Requests for information about a certain topic, include these data points
Writes the gene and disease to get an answer
Simple prompting
Keeps it short, requests information (no specific thought process)
Keeps it simple, 2-3 words
Being specific
Be as specific as possible
As detailed as possible
so Research Assistant understands what user is looking for
Being specific
Has an idea in mind what the output should relate to, which correlations
does not ask in general "what is known"

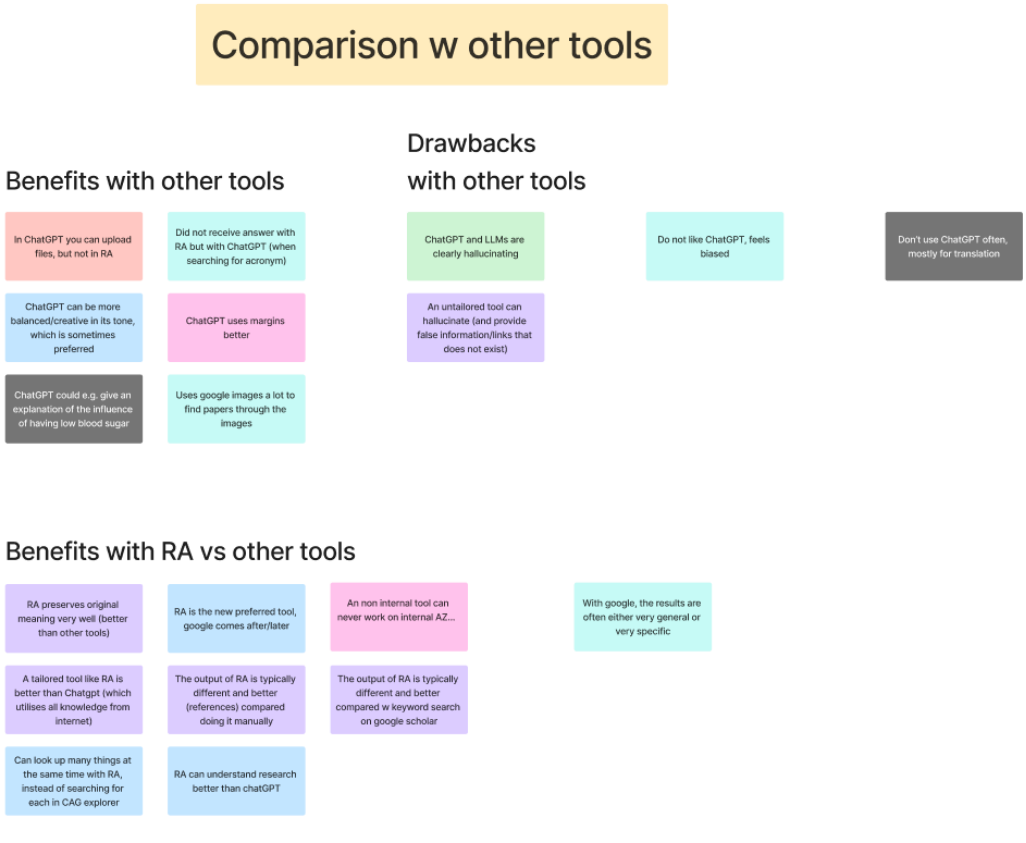

Benefits
Time saving, fast
Time saving
Quick / saves time with summaries
Time saving
Fast
Efficiency, time saving
Makes scientist more efficient
Quick / saves time
Quick / saves time with references
Quick / saves time with text and grammar
Quick access
Instant access to knowledge
Cuts down on manual work
RA sorting through the entries, work you don't need to do yourself
Gives more time to compile what they have learned from literature
Probably better results than manual work
Understanding
Understands science
Can explain in a simple way
Can give additional guidance and context beyond the original text to aid in better understanding of the text
Preserves original meaning when refining texts
Good for translation
Helps with text and grammar
Importance of internal data
Trust because of AstraZeneca data
Important to access internal data
Most important: Access internal data
Access to other type of knowledge other sources can not provide
Access to sources
The user can validate/verify the results, based on the references
Cites references that are accessible for the scientist
Links to references
Satisfied with the web links
References/Citation
Likes the citations for each paragraph and the list with snippets of the references' text
Numbers the references and lists them
Efficient & reliable source information
[references today] Easy access and display of references
Gives citations
Utilizes large amounts of data
Boosts self-confidence
You can use sensitive information
Simple layout, no distraction
Convenient
Switch between modes to fit your needs (=adjustable, flexible)

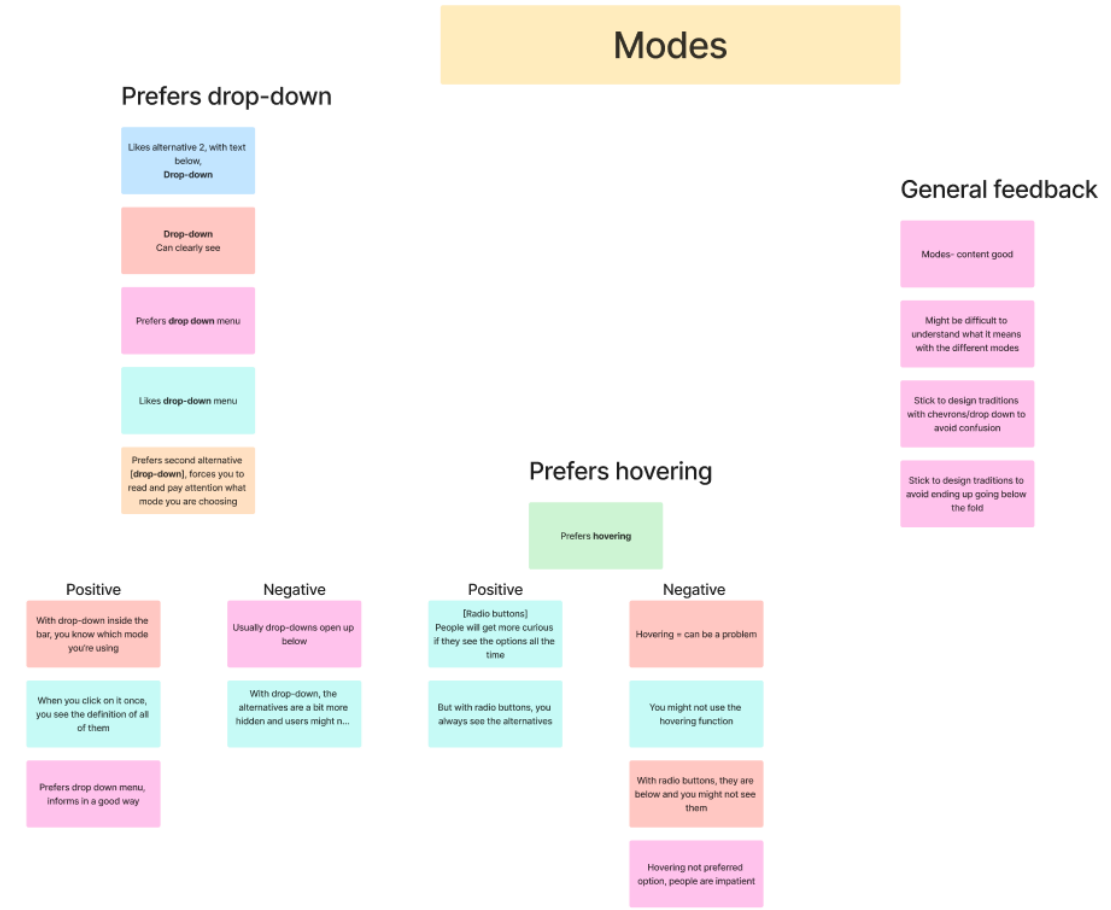

Needs
References: View, access, internal, trust
Reference ranking
Transparency
Visuals
Reference/text exportation
Organize/view searches
Settings before prompting
Change of layout
Display of output
Access to internal projects information
Desire to use fewer tools
Incorporate deep research & more metadata
Information about new tools
Understanding/description

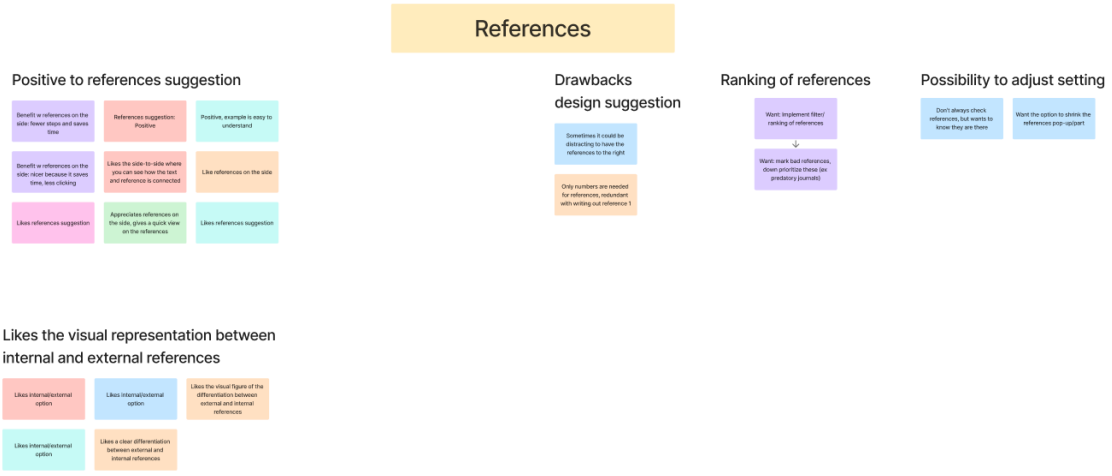

Trust
References
References makes it more easy to trust the output
References creates trust
Trusts answer if verifying the references
Need for references to be able to check every fact and thus, trust it
Trust it 60%, therefore always questions it, looks at the references
Verification of information
Verifies all data, not just RA data, since it is "good practice"
Asks other scientists/ experts and if they trust it, so does the user
Verifies the information
The quality of the information is determined by how reliable the output is
Check the output of RA
Verifies that the answer is accurate
More trust in internal info
A lot of trust in AZ data, less faith in accuracy with other sources of information
Trusts internal information
Lack of trust
I have enough scientific nuance to know that's probably true
Cannot rely on data you know nothing about
Need to consider what tone has been transferred
There is a bit lack of trust, people think it will hallucinate
Some people are hesitant to use it, but the user asks those people
Easier to trust simple things
Want to use AI more but needs to trust it
Some people might never trust it, and there is a middle ground with people who might trust it a bit
AI can read and remember, but we need to be aware of limitations
Trustworthiness
Trust 85% of the time
Trust RA more than ChatGPT
Trusts AI but we need to be careful
Consistent correct output without error
Modes
Uses factual when wanting to be sure the information is verified
Creative mode is not considered reliable
Creative mode is not used, feels to general
[Creative] When giving too much freedom/flexibility to the model, feels less reliable/trustworthy

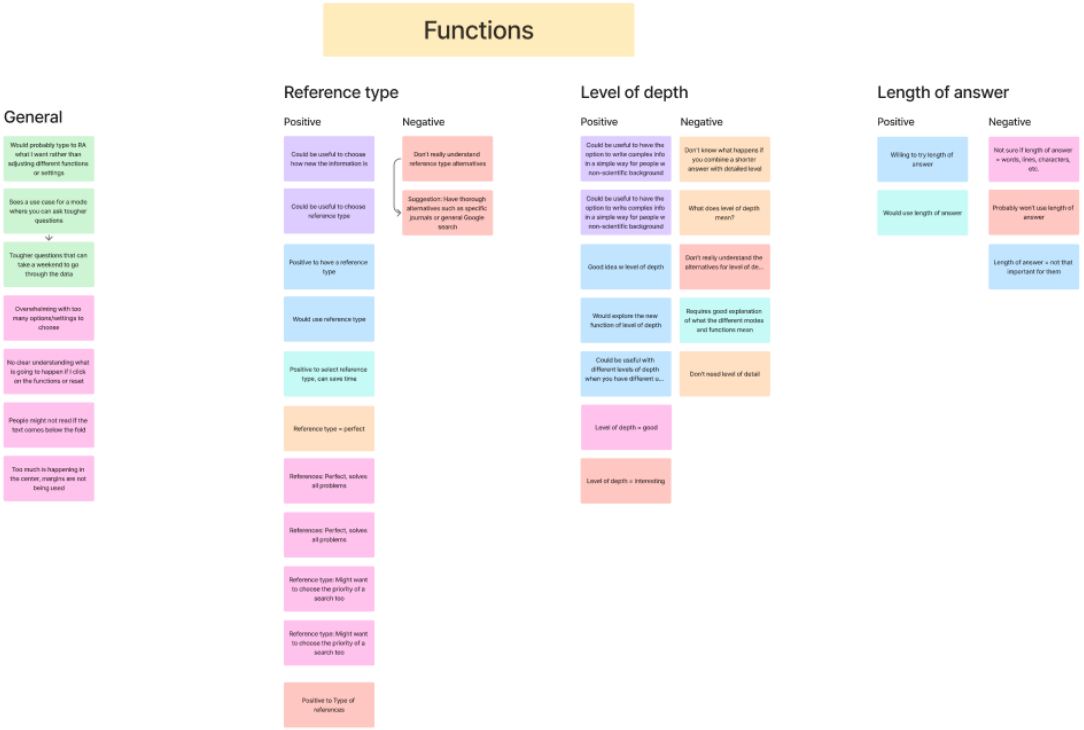

Current modes
Uses Factual
Uses factual
Only uses factual mode
Uses factual mode
Uses factual mode when presenting for an audience that is more well experienced in science
Uses Balanced
Uses balanced (liked the description and liked the approach)
User has not reflected on the usage of modes, use default mode (for user it is balanced)
Understanding of modes
Factual marks heavy science language, references for every statement
Understanding of factual: Very, very reliable information (not sure)
Factual mode finds highly cited research articles
Different modes might be helpful in writing aid
Balanced uses simpler explanation, provides references in the end
Understanding of balanced: It uses references but can write it more freely
Balanced and creative for social activities, but rarely used
The different modes determine the tone or how scientifically jargon the responses will be
Understanding of creative: Cannot describe it, did not feel reliable enough
Could use creative mode for recommendation letter
Opinions per user
Very satisfied with factual mode, not a single error has been spotted
Intentionally not tried other modes (except factual)
Want some type of strict mode
Main usage: specific science related questions, always keep it to factual
Very seldom switches between different modes
Prefers less amount of references as in balanced compared to factual, less overwhelming
The different modes are different in how reliable they are
Satisfied of output in factual mode
Will always like factual
Or be able to customize where data is pulled from
User has not reflected on the usage of modes, use default mode (for user it is balanced)
Uses factual
Only had problem with it once

Current references
Positive
Likes the annotations with links
Accepts current reference appearance
Satisfied with the numbers and complete links the references provides
Negative
Sometimes you cannot see the year or other information from the source
Do not want standardized way of displaying references, can write it themselves
Does not think current reference appearance is visually pleasing


Comparison w other tools
Benefits with other tools
In ChatGPT you can upload files, but not in RA
Did not receive answer with RA but with ChatGPT (when searching for acronym)
ChatGPT can be more balanced/creative in its tone, which is sometimes preferred
ChatGPT uses margins better
ChatGPT could e.g. give an explanation of the influence of having low blood sugar
Uses google images a lot to find papers through the images
Drawbacks with other tools
ChatGPT and LLMs are clearly hallucinating
Do not like ChatGPT, feels biased
Don't use ChatGPT often, mostly for translation
An untailored tool can hallucinate (and provide false information/links that does not exist)
Benefits with RA vs other tools
RA preserves original meaning very well (better than other tools)
RA is the new preferred tool, google comes after/later
An non internal tool can never work on internal AZ...
With google, the results are often either very general or very specific
A tailored tool like RA is better than Chatgpt (which utilises all knowledge from internet)
The output of RA is typically different and better (references) compared doing it manually
The output of RA is typically different and better compared w keyword search on google scholar
Can look up many things at the same time with RA, instead of searching for each in CAG explorer
RA can understand research better than chatGPT


Modes
Prefers drop-down
Likes alternative 2, with text below, Drop-down
Drop-down Can clearly see
Prefers drop down menu
Likes drop-down menu
Prefers second alternative [drop-down], forces you to read and pay attention what mode you are choosing
General feedback
Modes- content good
Might be difficult to understand what it means with the different modes
Stick to design traditions with chevrons/drop down to avoid confusion
Stick to design traditions to avoid ending up going below the fold
Prefers hovering
Prefers hovering
Positive
With drop-down inside the bar, you know which mode you're using
When you click on it once, you see the definition of all of them
Prefers drop down menu, informs in a good way
Negative
Usually drop-downs open up below
With drop-down, the alternatives are a bit more hidden and users might n...
Positive
[Radio buttons] People will get more curious if they see the options all the time
But with radio buttons, you always see the alternatives
Negative
Hovering = can be a problem
You might not use the hovering function
With radio buttons, they are below and you might not see them
Hovering not preferred option, people are impatient


References
Positive to references suggestion
Drawbacks design suggestion
Ranking of references
Possibility to adjust setting
Likes the visual representation between internal and external references


Functions
General
Would probably type to RA what I want rather than adjusting different functions or settings
Sees a use case for a mode where you can ask tougher questions
Tougher questions that can take a weekend to go through the data
Overwhelming with too many options/settings to choose
No clear understanding what is going to happen if I click on the functions or reset
People might not read if the text comes below the fold
Too much is happening in the center, margins are not being used
Reference type
Positive
Negative
Could be useful to choose how new the information is
Could be useful to choose reference type
Positive to have a reference type
Would use reference type
Positive to select reference type, can save time
Reference type = perfect
References: Perfect, solves all problems
References: Perfect, solves all problems
Reference type: Might want to choose the priority of a search too
Reference type: Might want to choose the priority of a search too
Positive to Type of references
Don't really understand reference type alternatives
Suggestion: Have thorough alternatives such as specific journals or general Google search
Level of depth
Positive
Negative
Could be useful to have the option to write complex info in a simple way for people w non-scientific background
Could be useful to have the option to write complex info in a simple way for people w non-scientific background
Good idea w level of depth
Would explore the new function of level of depth
Could be useful with different levels of depth when you have different u...
Level of depth = good
Level of depth = interesting
Don't know what happens if you combine a shorter answer with detailed level
What does level of depth mean?
Don't really understand the alternatives for level of de...
Requires good explanation of what the different modes and functions mean
Don't need level of detail
Length of answer
Positive
Negative
Willing to try length of answer
Would use length of answer
Not sure if length of answer = words, lines, characters, etc.
Probably won't use length of answer
Length of answer = not that important for them

F

Appendix: List of Needs and Why's

What does the user need?	Why?
Tool for gaining more knowledge	To understand the project better
	To get background information
	Find new angles for projects
	How things are interconnected
	Explore targets
	Get additional guidance beyond the summaries
To get summaries	Save time
	Get inspiration for literature
	Get familiar with topics
	Not have to read large documents
Writing aid	

References / Verification	See if information is reasonable
	See if information makes sense
	To find additional information
	To find previous references by explaining what it was about
	Too many references can be overwhelming
	Reference titles are sometimes too long
	Want to distinguish internal and external data
	Want reference right away, not having to click and copy in new d
	To see the reference without scrolling down
	To see links to internal information as well
	Links to internal data
	Title, year, journal
	Digital identifier to find the text/info
	Ranking of references
	Highly cited references to increase trust
	Newer articles to increase trust
	Reliable sources to increase trust
	User can validate the result with the references
	The annotations with links ease work
	Year and other information is needed to use the references
	References makes you trust the output
	References and verification creates trust
	Verification with references creates trust
	With references side-by-side, less steps are needed
	With references side-by-side, more time is saved
	References side-by-side gives the user a quick overlook
	Internal/external option for references clarifies them
	Internal/external option for references is positive
	Internal/external option ease the work with the references
	Ranking of references would be beneficial
	Marking of bad references would ease the work long-term
	Hide/show references would minimize risk of feeling overwhelmed
	Type of references would ease with finding wanted information
	Type of references would be used
	Type of references: might want to choose the priority of the search
	Type of references: could be useful to choose specific journals
Up to date information	Want connections between small molecules and proteins
	Not all needed data is accessible currently

Clear communication	The box with "Hi [name], how are you?" looks like a text box for prompting
	"New chat" function is unclear
	Prompt suggestions are not necessary
	To be able to type next prompt while RA is giving an answer
	Unclear how RA came up with the answer
	Unclear how to interpret the data/answers
	To understand why the AI chose certain data
	Evaluation of how certain the AI is to increase trust
	See prompt box first instead of "Hi [name]" to type question faster
	Modes more salient so user can use them easier
	Need for description of what each function means
	Hard to understand what happens if you click certain buttons/ settings
	Useful to choose how new the information in the output is
Visuals	Visuals to understand better
	Graphic mode to see mindmaps
Exportation	Exportation/copy of references to cite easily in other tools
	Copy & paste answer/text to add to Word

Organization	To organize answers in folders and find them easier
	Option on how to view the answers to adapt to users needs
Prompt settings	Upload own data/files to be able to compare
	To choose how old the data is to adapt to users needs
Save time - fast tool	Makes scientist more efficient
	Instant access to knowledge
	Cuts down work load, don't need to go through all entries themselves
	Gives more time to compile what they have learned instead of reading
	Better results than manual work
	Saves time / ease with grammar and text
Modes	Modes determine the tone or scientific jargon of responses
	Modes can be helpful for writing aid
	Want some type of strict mode
	Different modes determine how reliable the output is
	Factual mode to be scientific
	Factual to get reliable information
	Factual is used when wanting the information to be verified
	Factual to get heavy science language
	Factual to get references for every statement
	Factual to get highly cited articles
	Factual to create specific science related questions
	Balanced because too many references can be overwhelming
	Balanced because the description seemed like a good approach
	Balanced provides simpler explanations
	Balanced or creative can be used for social activities
	Creative can be used for recommendation letters
	Creative feels to general
	Creative means too much freedom, which feels less reliable
	Drop-down for modes is positive
	With a drop-down you can clearly see
	A drop-down forces the user to read the description and understand them better
	With a drop-down it is very clear what modes is in use
	It helps to see the definitions of the modes easy
	With a drop-down, the alternatives are a bit hidden
	With radio buttons, the alternatives are always visible
	Hovering can be a problem
	Hovering might not work since people are impatient
	Radio buttons makes the user more curious since they see the alternatives
	The modes descriptions might be hard to understand

Minimalistic layout	Too many settings can be overwhelming
	With too much text/information, user might not read it
Level of depth	Level of depth would be used
	Level of depth is interesting
	Level of depth could be used for different use cases
	Unclear what happens when combining a shorter answer with detailed level
	Level of depth is a bit unclear
	They require a good explanation of what the settings mean

G

Appendix: Insight cards

Insight theme References 1

The user needs references

- The references enables the participant to verify that the information provided is accurate.
- Usage of references are required in their work as scientists as it is essential for them and others to be able to verify the validity of the information

"I really need reliable sources and to be able to check every fact."
Participant 4

"So I will go to the source and make sure that whatever is coming out of the research assistant is making any sense."
Participant 6

"It's good practice regardless of what you're using to verify. I never trust a single data point. I think it's just good scientific methodology. Never trust anything until you verified it either with a different source, Different person or actually done it yourself in the lab and go and check."
Participant 3

Recommendations:

Display of references in a structured manor

Insight theme References 2

The user needs to distinguish references efficiently

- Gives a the user a quick overview of the references presented, where some references are more reliable than others
- Internal references are more trustworthy
- Easier to distinguish where data is from
- Makes it easier to trust the output

"So if I know it's looking at AstraZeneca data, I have a lot of trust in what it's saying as it moves beyond that, just as a user of tools on the internet, then I have less and less faith that the information is accurate, and I think that's just the nature of things"
Participant 2

"I think showing internal/external is a really good idea, showing the differentiation. And it's nice to know what kind of percentage is internal and external."
Participant 3

"[Reference type] I've got all these, but which one is becoming the priority?
Participant 2

Recommendations:

Visualise if the reference is from internal or external source

Let the user choose where to find the data

Provide reference title, year, author, and link

Prioritize and/or view distribution of reference type

Insight theme References 3

The user needs to be able to reference the source in other tools

- They use referencing to justify and strengthen their choices
- It would save time if they get the references in the correct format when adding in another document

"It would be great if not only copy. But we had the option to export it in a In the citation mode, so you know without the link but in a proper citation format. So you know, with the, like, no matter which citation format because there are many standards, but at least one."
Participant 4

"So yeah, but the possibility to just click and then copy the reference, that would be awesome."
Participant 4

Recommendations:

Provide an exportation button for references, to let them copy & paste it into other documents

Insight theme Modes 1

The user needs guidance in how to interpret the modes

- There is a discrepancy in user understanding regarding what the different modes mean
- Many users did not realize or utilize that there are different modes available

"I think I reflect now. I mean, I, I probably just use the default mode and I haven't really reflected on. Actually."
Participant 5

"I've intentionally not tried the other modes. Because I want to keep it to facts. And I worry that if I went balanced and especially if I went creative, that's when it would start giving you more, not hallucinating, but I want factual science. I don't want creative AI input into science at the moment."
Participant 3

Recommendations:

Provide a clear description of what each mode mean

Let the description of modes be easy to find and see

Insight theme Modes 2

The user needs guidance in how to use the modes

- There is a discrepancy in user understanding regarding the different modes mean
- Many users did not realize or utilize that there were different modes available

"I think I reflect now. I mean, I, I probably just use the default mode and I haven't really reflected on. Actually."
Participant 5

"I've intentionally not tried the other modes. Because I want to keep it to facts. And I worry that if I went balanced and especially if I went creative, that's when it would start giving you more, not hallucinating, but I want factual science. I don't want creative AI input into science at the moment."
Participant 3

Recommendations:

Make the user curious to try the different modes

Be clear to the user what mode they are using

Insight theme Transparency 1

The user needs information about how it came up with its answer and guidance in how to interpret it

- Allows the user to get a better understanding of how to interpret its answer
- Can make the user understand what more it can do for them

"How did it come up with the answer it did, where did this data come from, how do I know? How do I get a sense of how deep can I go, you know, how do I get follow-up information"
Participant 2

"Maybe a bit more understanding of you know, how to... When we ask the questions, how to interpret that and sort of make the link. That's one area I see that definitely could be improved."
Participant 5

"Now it doesn't have access to all the data it needs. And it's kind of hard to tell. You know Is it AstraZeneca data that's missing because it goes through a couple of layers of agents, it's like, OK, well, we didn't have an AstraZeneca. Should we check this? And then it's, it's, I think the more broad it gets, the less reliable the results may be."
Participant 2

Recommendations:

Give a statement of its information retrieval procedure and how it reach its conclusion

Provide examples of follow up questions or examples of what it can do

Insight theme Transparency 2

The user needs to understand what limitations there are with the application

- They are frustrated when not receiving wanted information even when it is not available
- They need to know the limitations to understand the possibilities of the application

"So, why couldn't we link to, to that source when we, we have it, so. I think that's the only thing I reflected on when it comes to references that the web I think we are fine, but internal to. Databases and sources doesn't seem to be as, as, as nice."
Participant 4

"There have been times where I've tried to ask it to look at a specific journal or specific article, but it wasn't able to do that."
Participant 8

Recommendations:

Communicate which limitations there are in a structured way and in one place

Communicate the possibilities

Insight theme Communication 1

The user needs clear communication of the interface

- The user want to save time with the application
  - With clearer communication, they can use the application more efficiently
- It is confusing what all settings and buttons do

"Sometimes it can be overwhelming when they're giving you over 50 resources, and I want to think of how can this be maybe like a little shorter cause I don't need that many references, maybe just a handful that I can easily go through, cause no one really wants to go through 50 references."
Participant 8

"This is why the first time I didn't realize that it was something that I had to pay attention because I thought that it was something about the cookies"
Participant 4

Recommendations:

Keep coherency with design of buttons and information pop-ups

Not provide too much information at the same time

Insight theme Communication 2

The user needs a visual way to interpret the information

"I don't know if it would be possible to have like a graphic mode, so to get your answer in some sort of mind map kind of thing, that would be really cool."
Participant 6

- Several users have indicated that they are visual learners, that it would simplify the process of understanding and learning new concepts
  - Can save time and energy from the user

"Another thing that I've noticed in Research Assistant that I would love to see is more visualization."
Participant 8

"I would love instead of only getting text. Also to get images of that, for example, that metabolic pathway... I'm a visual learner."
Participant 4

Recommendations:

Provide visuals of the answers from the application

Insight theme Communication 3

The user needs clear communication of the information

"That data is really important to keep around. Because if I'm using something like RA, I want to know what's already been done?"
Participant 2

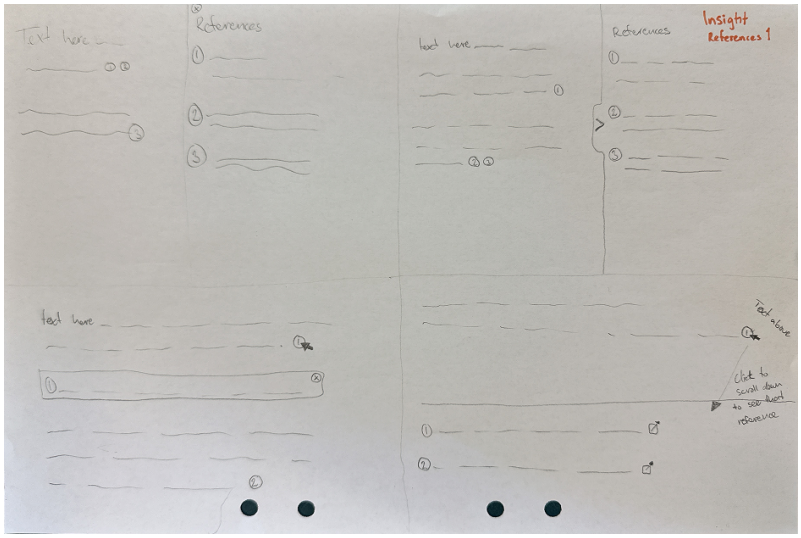
- The user want to save time with the application
  - With clearer communication, they can use the application more efficiently

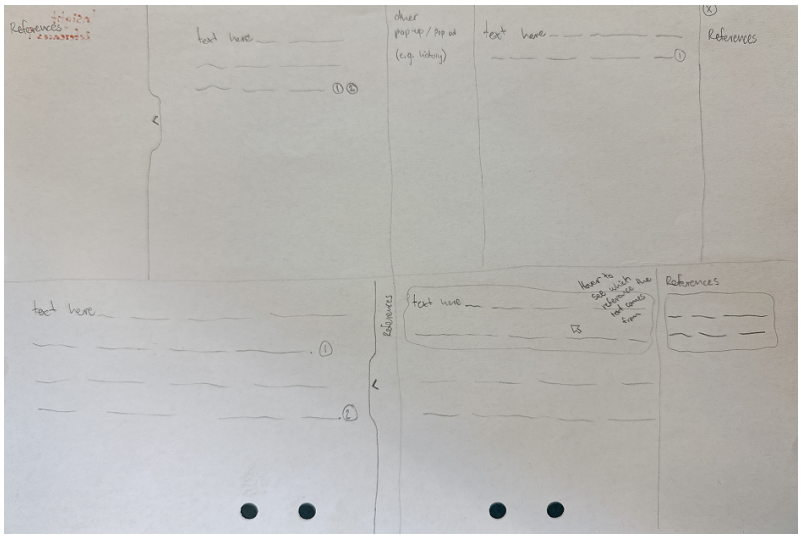
Recommendations:

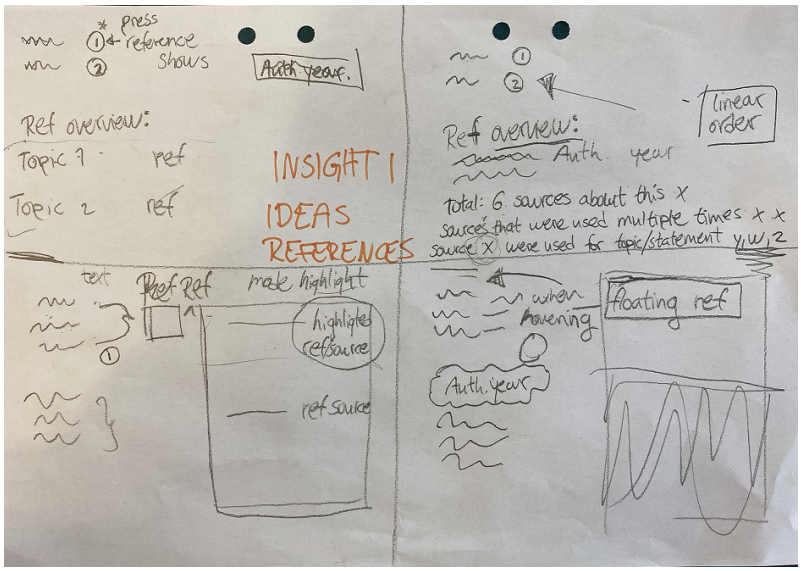
Provide data in a clear and efficient way

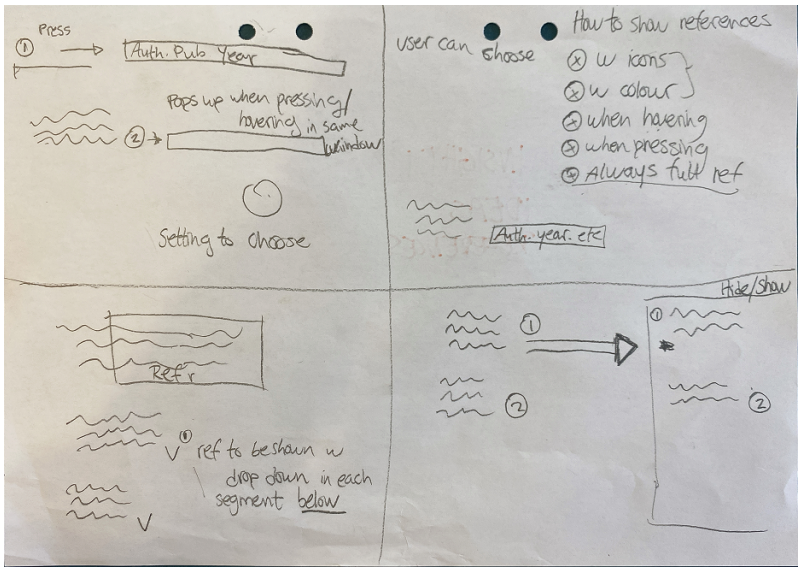
H

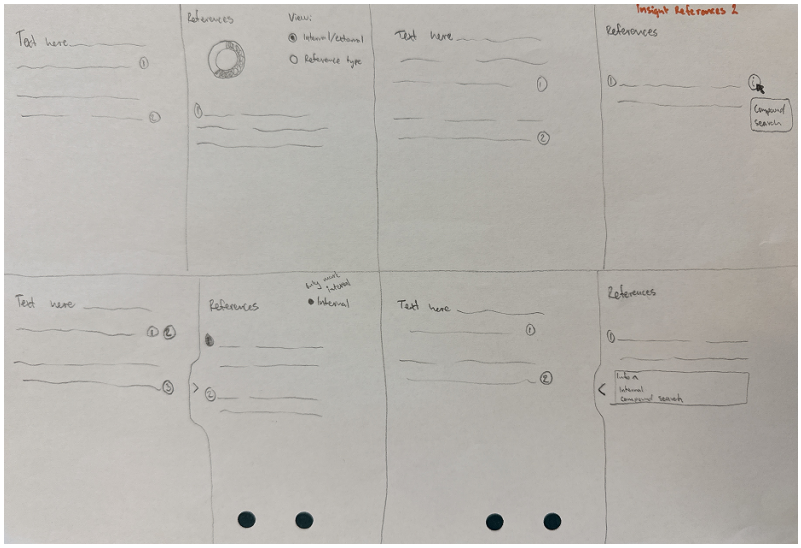
Appendix: Sketches

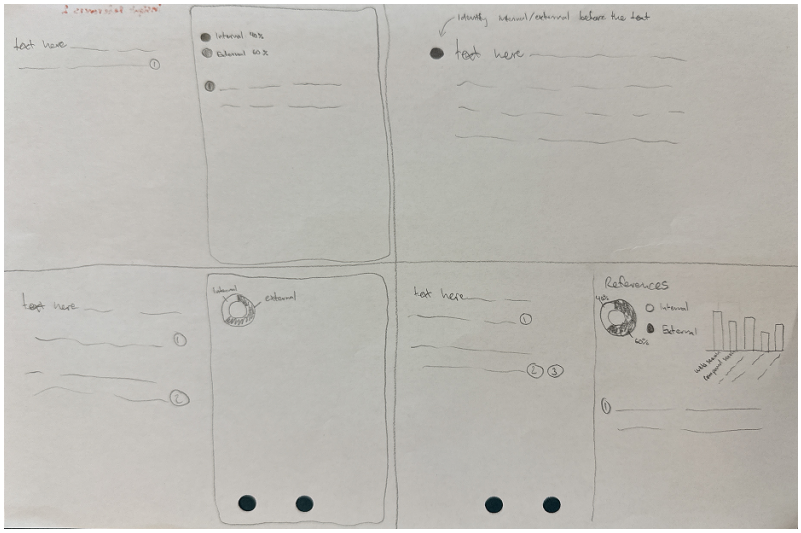


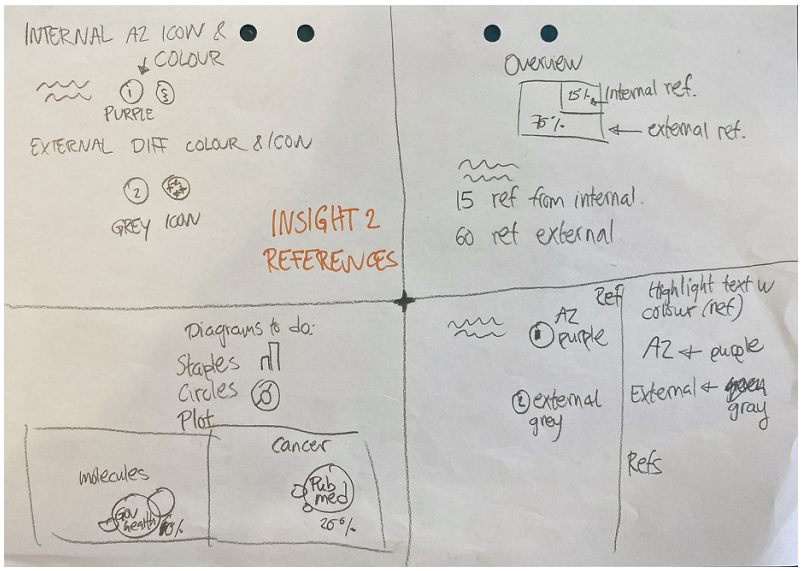

References
text here
other pop up / pop out (e.g. history)
text here
References
text here
References
text here
References

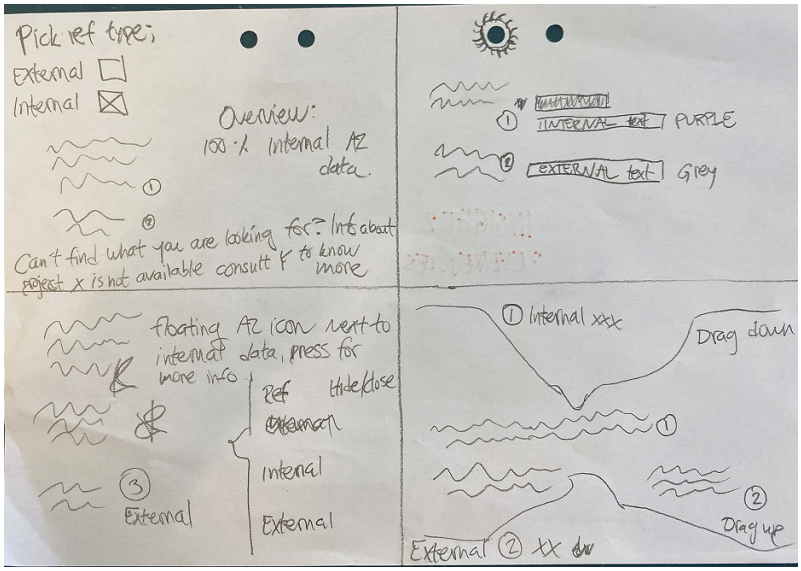

Press
reference
shows
Auth. year.
Ref overview:
Topic 1
ref
Topic 2
ref
INSIGHT 1
IDEAS
REFERENCES
text
Ref
make highlight
highlights ref source
ref source
linear order
Ref overview:
Auth. year
total: 6 sources about this x
sources that were used multiple times x x
source x were used for topic/statement y, w, z
when hovering
floating ref
Auth. year

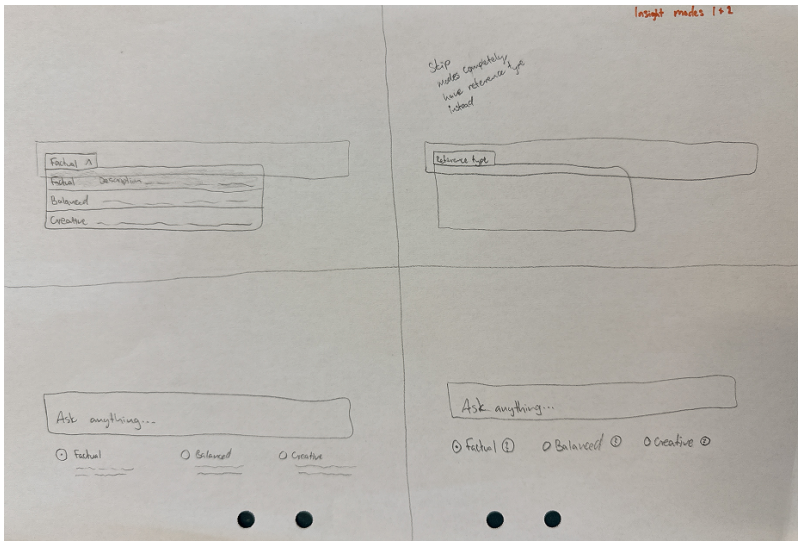

Press
Auth. Pub Year
pops up when pressing/ hovering in same window
Setting to choose
user can choose
How to show references
w icons
w colour
when hovering
when pressing
Always full ref
Auth. year. etc
Hide/show
Ref
ref to be shown w drop down in each segment below

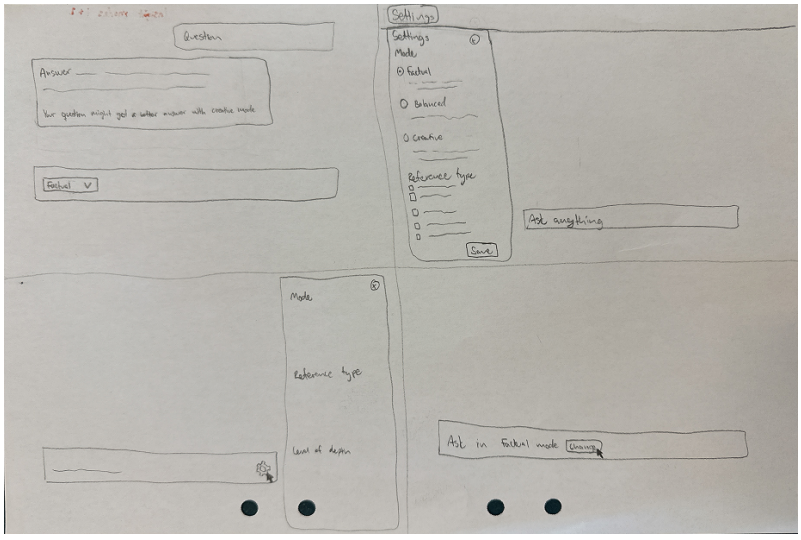

Text here
References
View:
Internal/external
Reference type
Insight References 2
Compound search
Only count internal
Internal
Take a internal compound search

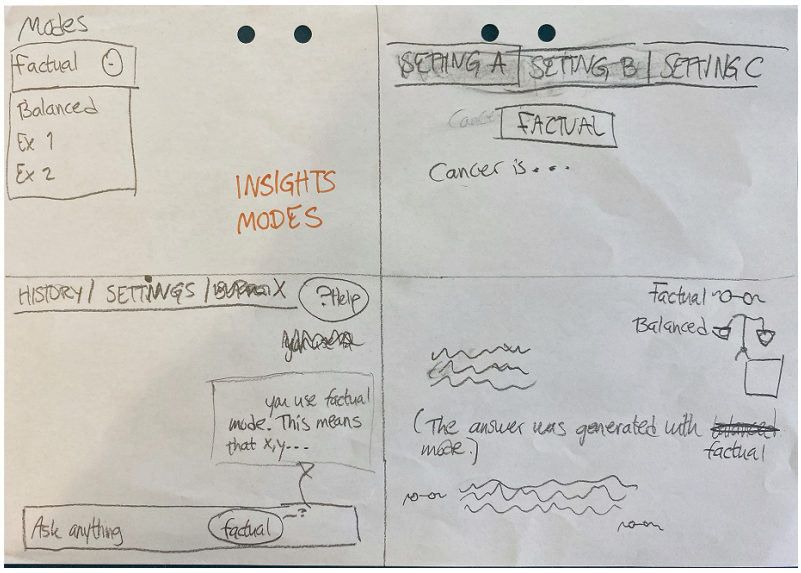

Internal 40%
External 60%
Identify internal/external before the text
text here
References
Internal
External

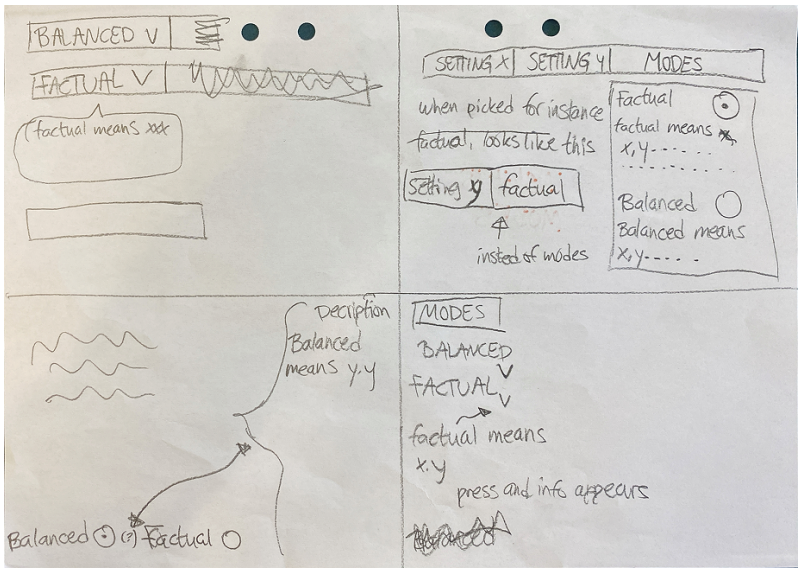

INTERNAL A2 ICON & COLOUR
PURPLE
EXTERNAL DIFF COLOUR & ICON
GREY ICON
INSIGHT 2 REFERENCES
Overview
15% Internal ref.
75% external ref.
15 ref from internal.
60 ref external
Diagrams to do:
Staples
Circles
Plot
Cancer
Molecules
Pub med
25%
Ref
A2 purple
external grey
highlight text w colour (ref)
A2 ← purple
External ← gray
Refs

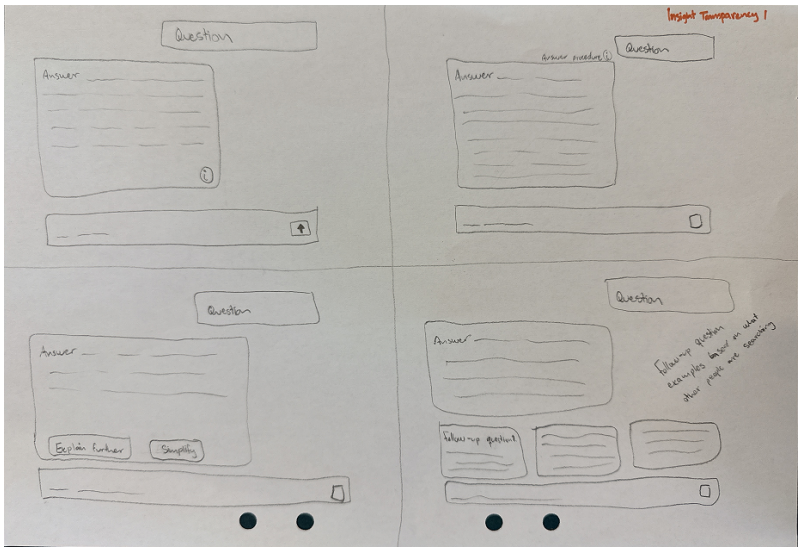

Pick ref type:
External
Internal
Overview:
100 % internal AZ data.
Can't find what you are looking for? Info about project X is not available consult F to know more
floating AZ icon next to internal data, press for more info
Ref
Hide/close
Internal
External
External
Internal xxx
Drag down
Drag up
External

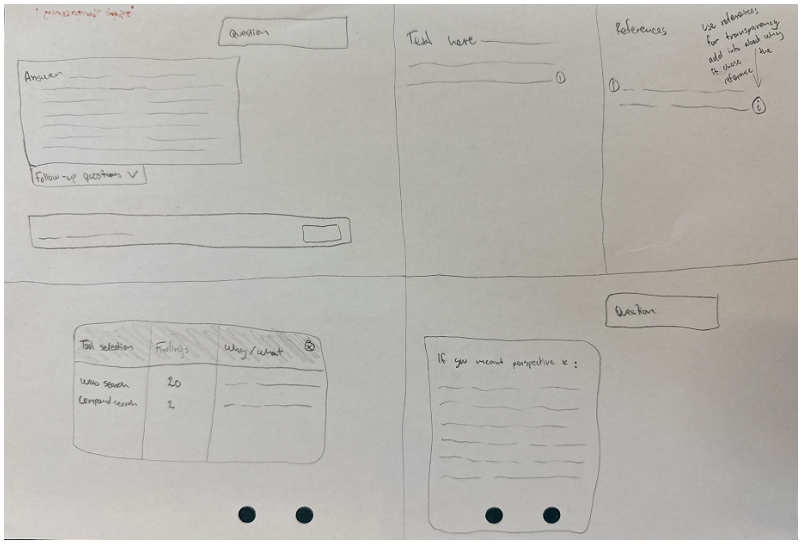

Insight modes 1+2
Factual
Balanced
Creative
Reference type
Ask anything...
Factual
Balanced
Creative
Ask anything...
Factual
Balanced
Creative

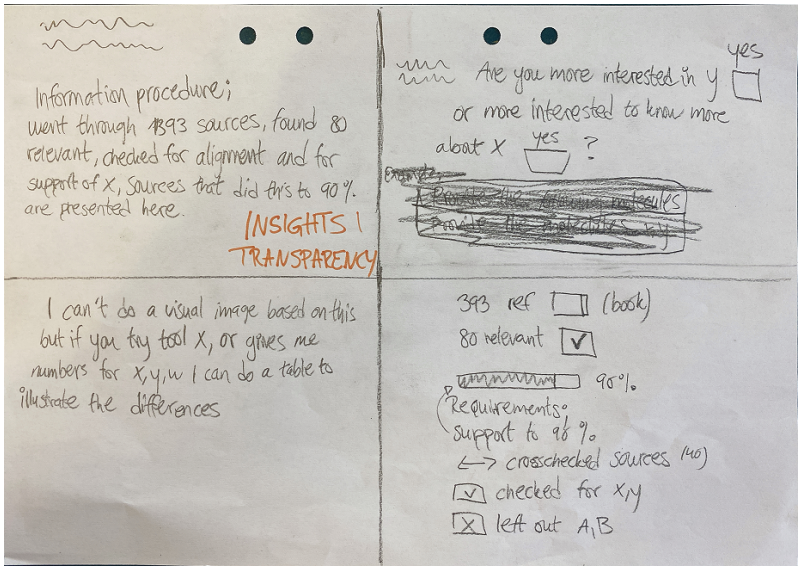

Question
Answer
Factual
Settings
Settings
Mode
Factual
Balanced
Creative
Reference type
Save
Ask anything
Mode
Reference type
Ask in Factual mode

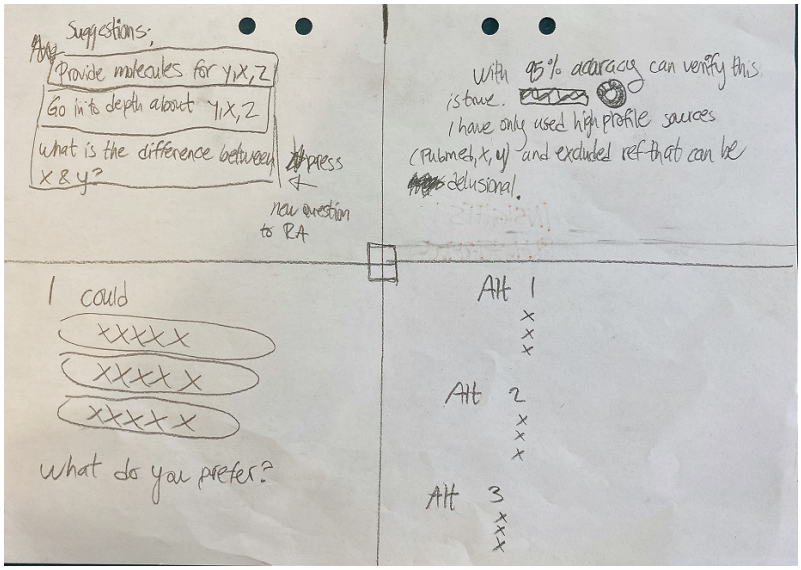

Modes
Factual
Balanced
Ex 1
Ex 2
INSIGHTS
MODES
SETTING A
SETTING B
SETTING C
FACTUAL
Cancer is...
HISTORY
SETTINGS
?Help
you use factual mode. This means that x,y---
Ask anything
factual
Factual
Balanced
(The answer was generated with factual mode.)

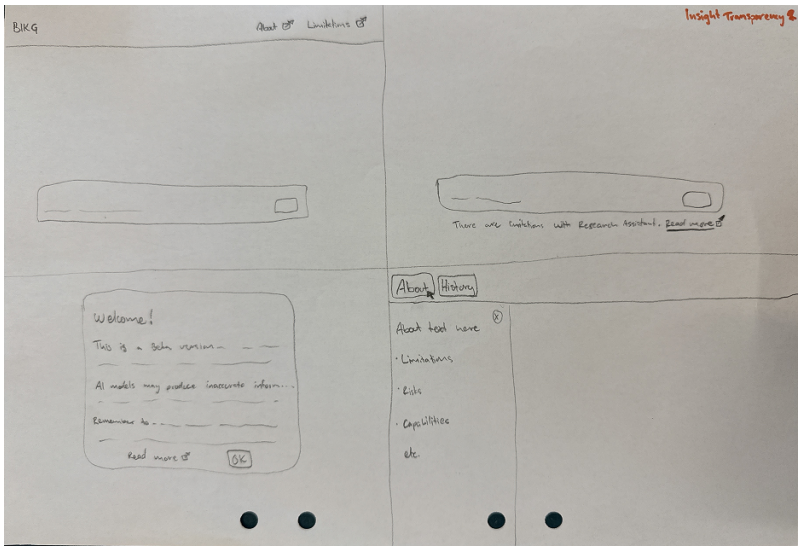

BALANCED
FACTUAL
factual means xxx
SETTING X
SETTING Y
MODES
when picked for instance factual, looks like this
Setting y
factual
instead of modes
Factual
factual means x,y
Balanced
Balanced means x,y
Description
Balanced means y,y
Balanced
Factual
MODES
BALANCED
FACTUAL
factual means x,y
press and info appears

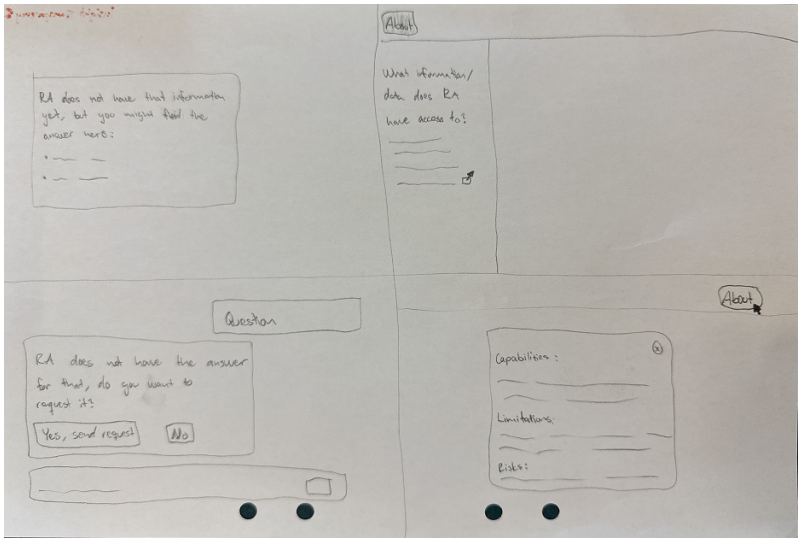

Question
Answer
Insight Transparency !
Explain further
Simplify

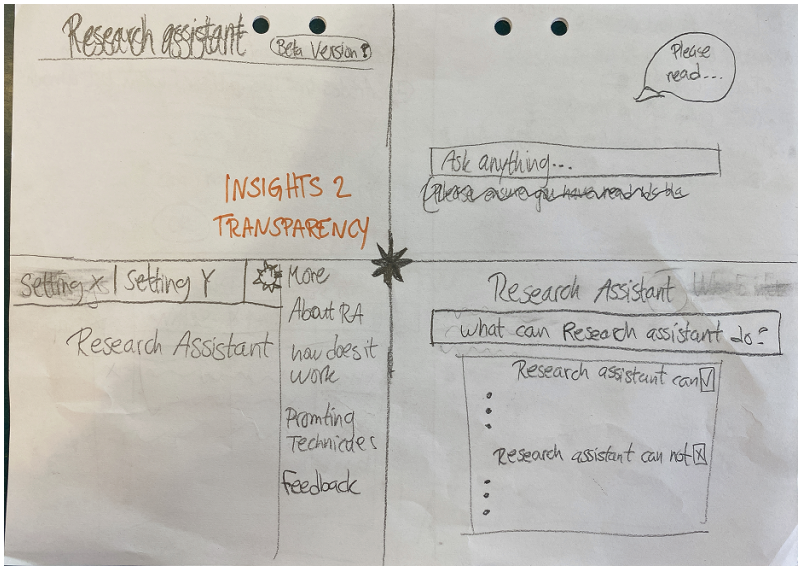

Question
Answer
Follow-up questions
Text here
References
Use references for transparency add info about why it uses reference
Tool selection
Findings
Why / what
Question
If you meant perspective x :

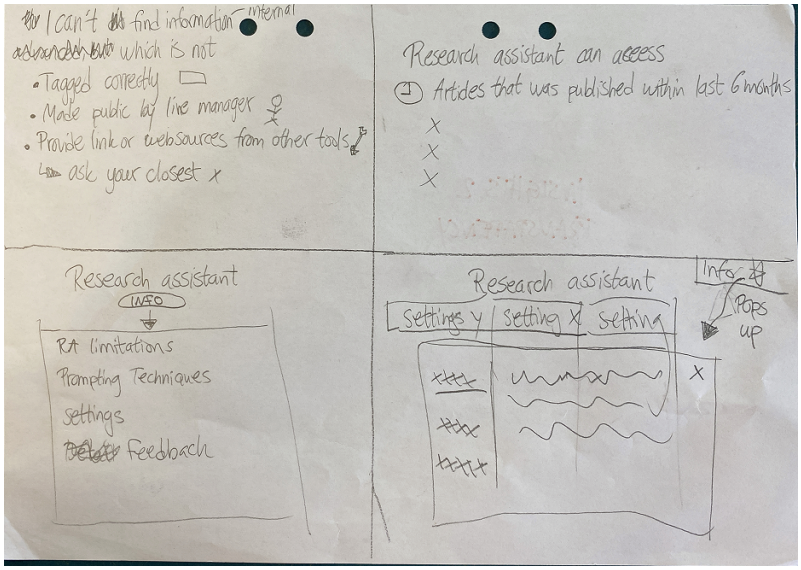

Information procedure;
went through 1893 sources, found 80 relevant, checked for alignment and for support of X, Sources that did this to 90 %. are presented here.
INSIGHTS / TRANSPARENCY
Are you more interested in y
or more interested to know more about X
yes
?
yes
I can't do a visual image based on this but if you try tool X, or gives me numbers for X,y,w I can do a table to illustrate the differences
393 ref (book)
80 relevant
90%
Requirements:
support to 98 %
crosschecked sources (40)
checked for X,y
left out A,B

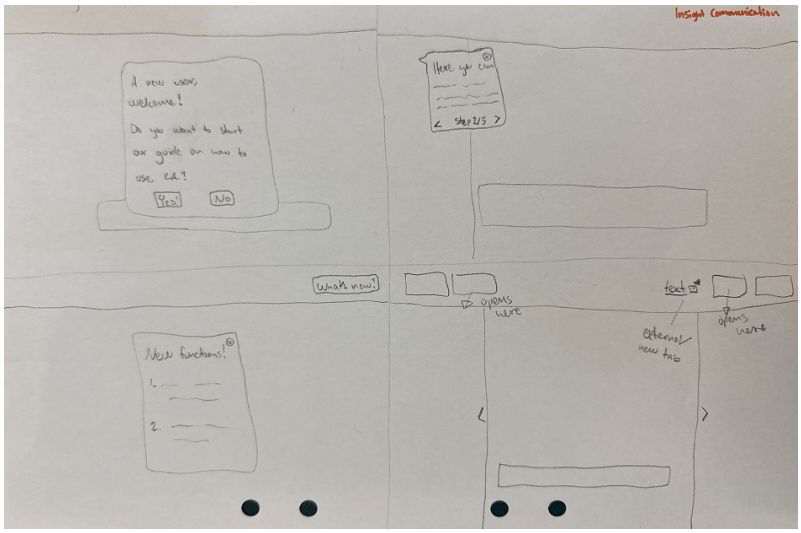

Suggestions;
Provide molecules for y,x,z
Go in to depth about y,x,z
what is the difference between x & y?
press
new question to RA
With 95% accuracy can verify this is true.
I have only used high profile sources (Pubmed, x, y) and excluded ref that can be delusional.
I could
what do you prefer?
Alt 1
Alt 2
Alt 3

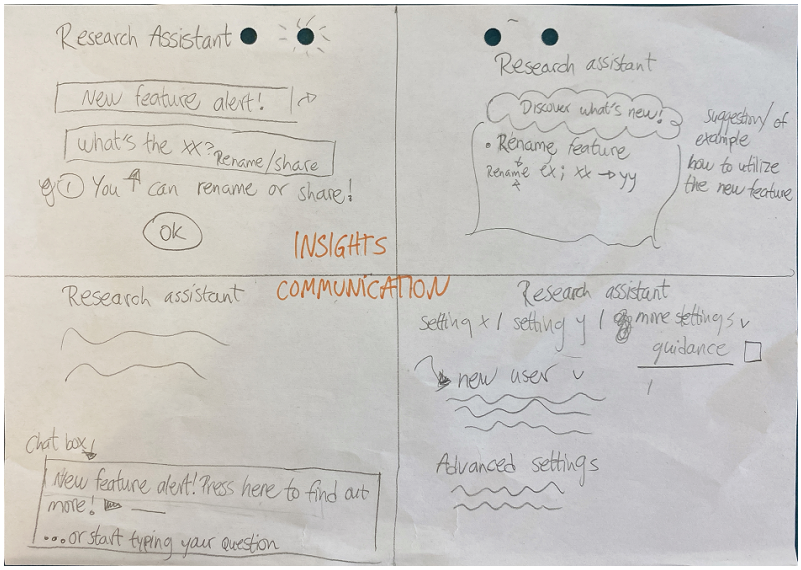

BJKG
About
Limitations
Insight Transparency 2
There are limitations with Research Assistant. Read more
Welcome!
This is a Beta version
AI models may produce inaccurate inform...
Remember to
Read more
OK
About
History
About text here
· Limitations
· Risks
· Capabilities
etc.

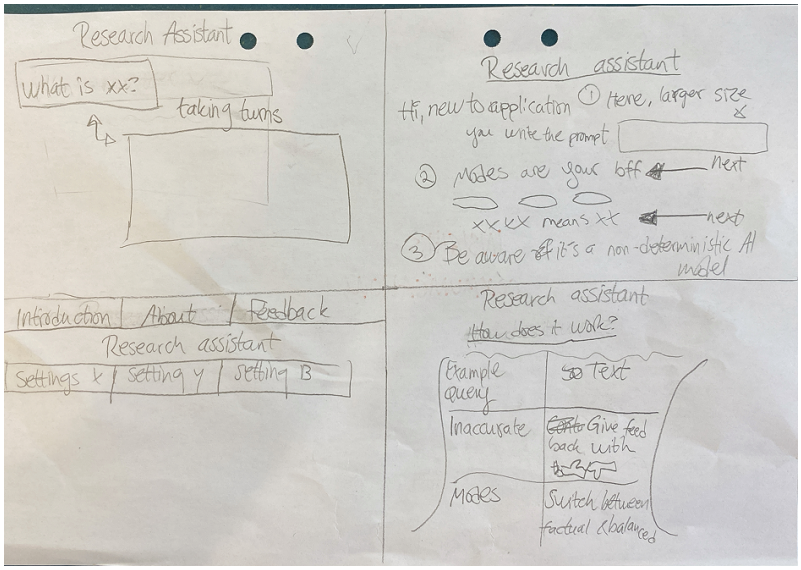

About
RA does not have that information yet, but you might find the answer here:
What information/ data does RA have access to?
Question
RA does not have the answer for that, do you want to request it?
Yes, send request
No
About
Capabilities:
Limitations:
Risks:

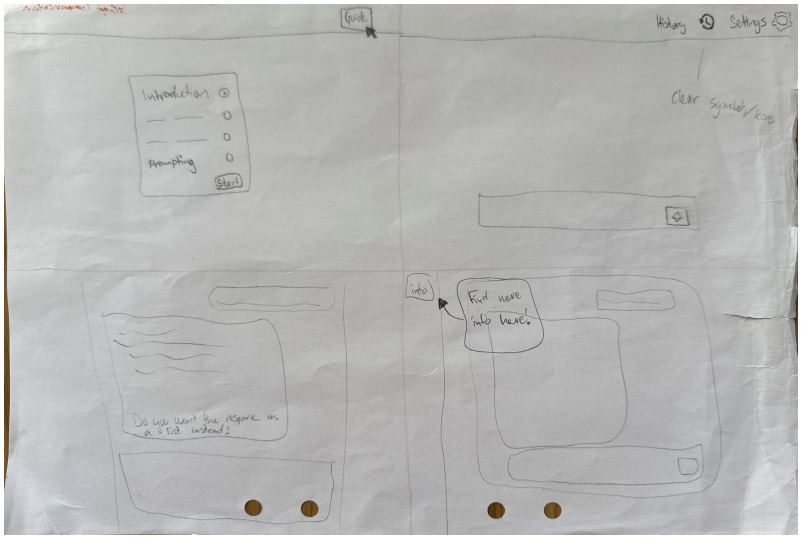

Research assistant
Beta Version
Please read...
INSIGHTS 2
TRANSPARENCY
Ask anything...
Setting X
Setting Y
More
About RA
how does it work
Prompting Techniques
Feedback
Research Assistant
Research Assistant
What can Research assistant do?
Research assistant can
Research assistant can not


I can't find information which is not
Tagged correctly
Make public by line manager
Provide link or websources from other tools
ask your closest x
Research assistant can access
Artides that was published within last 6 months
Research assistant
INFO
RA limitations
Prompting Techniques
settings
Feedback
Research assistant
Setting y
Setting x
Setting
Info
Pops up


Insight Communication
A new user, welcome!
Do you want to start or guide on how to use RA?
Yes!
No
What's new!
New functions!
1.
2.
text
opens here
external new tab
opens here


Research Assistant
New feature alert!
What's the xx? Rename/share
You can rename or share!
Ok
INSIGHTS
COMMUNICATION
Research assistant
Chat box
New feature alert! Press here to find out more!
...or start typing your question
Research assistant
Discover what's new!
Rename feature
Rename ex; xx → yy
Suggestion of examples how to utilize the new feature
Research assistant
setting x
setting y
more settings
guidance
new user
Advanced settings

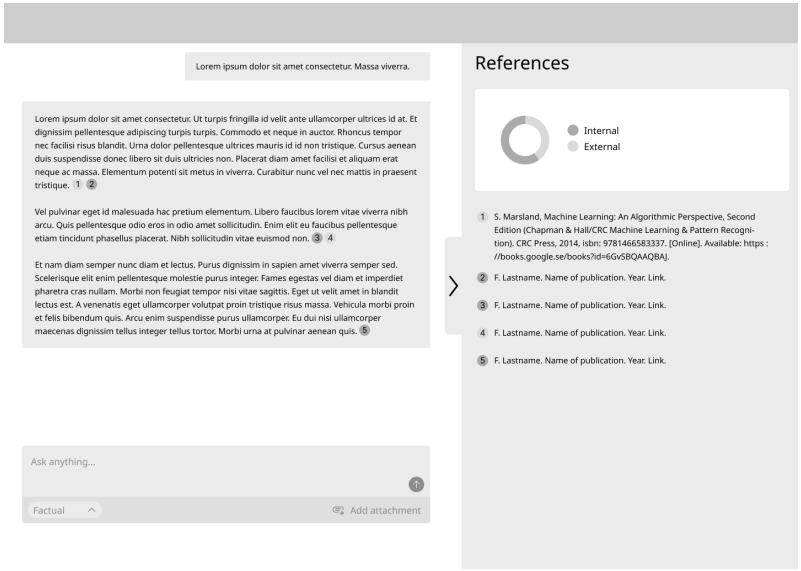

Research Assistant
what is xx?
taking turns
Introduction
About
Feedback
Research assistant
Settings x
Setting y
setting B
Research assistant
Hi, new to application
1 Here, larger size
you write the prompt
2 Modes are your bff
Next
xx xx means xx
next
3 Be aware of it's a non-deterministic AI model
Research assistant
How does it work?
Example query
Text
Inaccurate
Give feed back with
Modes
Switch between factual &balanced

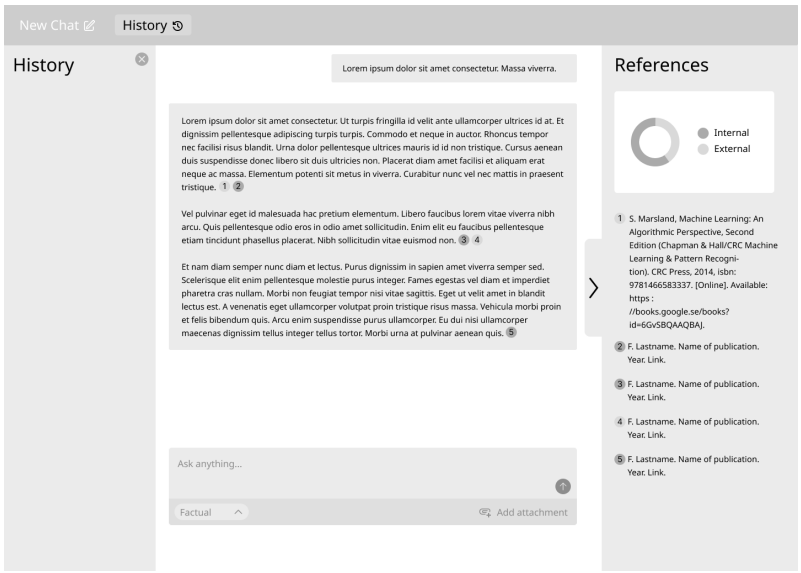

Quick
Introduction
Prompting
Start
History
Settings
Clear
Info
First name into here!
Do you want the response as a list instead?

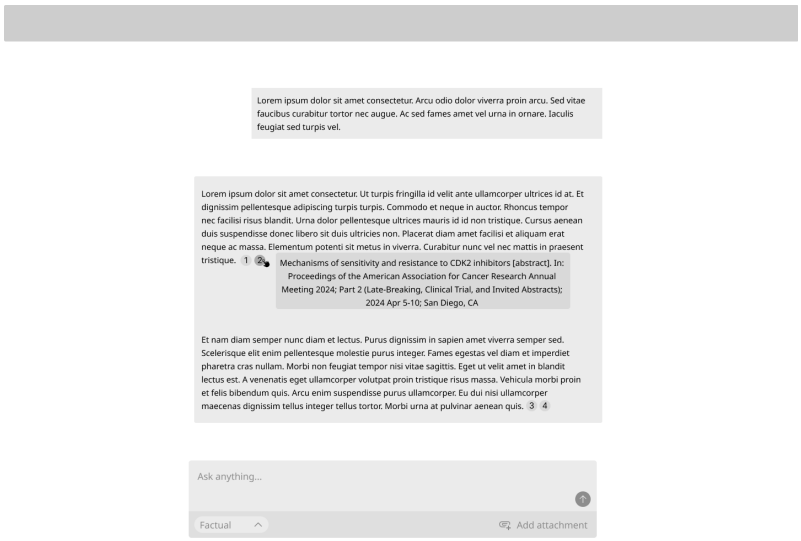
I

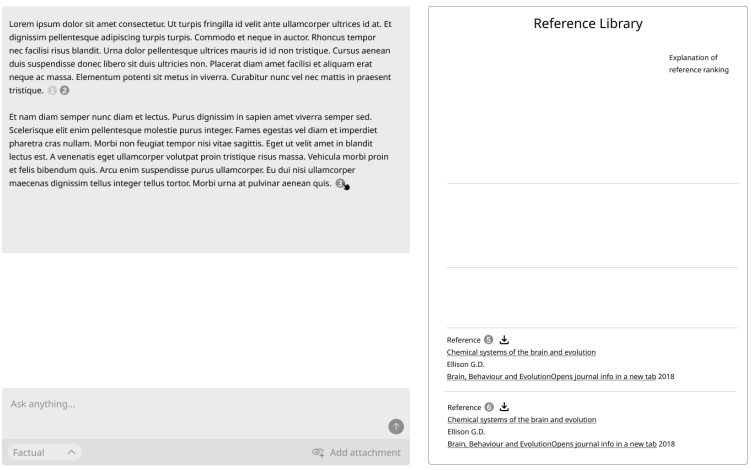
Appendix: Wireframes

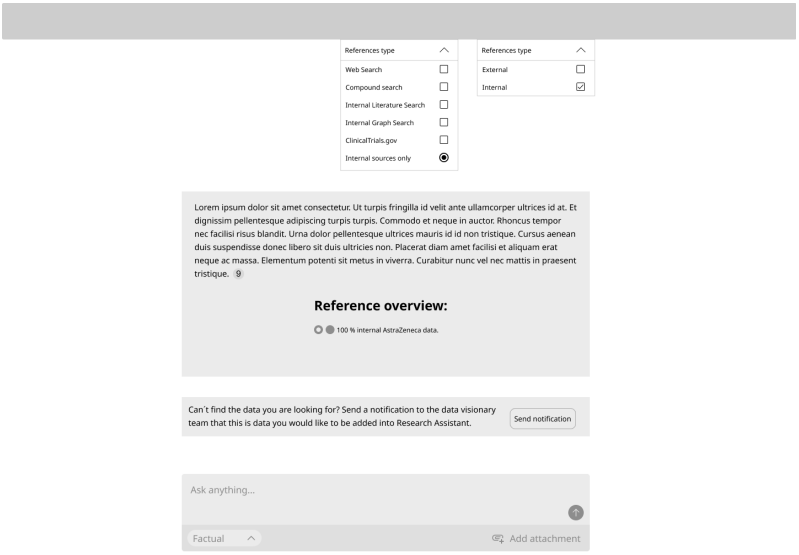
References 1 + 2



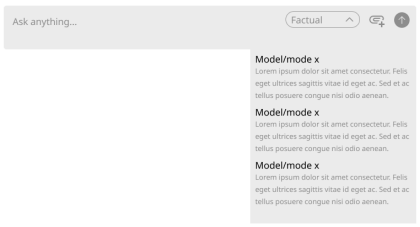


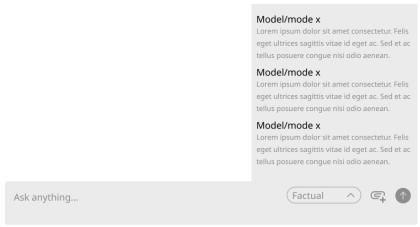

Lorem ipsum dolor sit amet consectetur. Arcu odio dolor viverra proin arcu. Sed vitae faucibus curabitur tortor nec augue. Ac sed fames amet vel urna in ornare. Iaculis feugiat sed turpis vel.
Lorem ipsum dolor sit amet consectetur. Ut turpis fringilla id velit ante ullamcorper ultrices id at. Et dignissim pellentesque adipiscing turpis turpis. Commodo et neque in auctor. Rhoncus tempor nec facilisi risus blandit. Urna dolor pellentesque ultrices mauris id id non tristique. Cursus aenean duis suspendisse donec libero sit duis ultricies non. Placerat diam amet facilisi et aliquam erat neque ac massa. Elementum potenti sit metus in viverra. Curabitur nunc vel nec mattis in praesent tristique. 1 2
Mechanisms of sensitivity and resistance to CDK2 inhibitors [abstract]. In: Proceedings of the American Association for Cancer Research Annual Meeting 2024; Part 2 (Late-Breaking, Clinical Trial, and Invited Abstracts); 2024 Apr 5-10; San Diego, CA
Et nam diam semper nunc diam et lectus. Purus dignissim in sapien amet viverra semper sed. Scelerisque elit enim pellentesque molestie purus integer. Fames egestas vel diam et imperdiet pharetra cras nullam. Morbi non feugiat tempor nisi vitae sagittis. Eget ut velit amet in blandit lectus est. A venenatis eget ullamcorper volutpat proin tristique risus massa. Vehicula morbi proin et felis bibendum quis. Arcu enim suspendisse purus ullamcorper. Eu dui nisi ullamcorper maecenas dignissim tellus integer tellus tortor. Morbi urna at pulvinar aenean quis. 3 4
Ask anything...
Factual
Add attachment


Lorem ipsum dolor sit amet consectetur. Ut turpis fringilla id velit ante ullamcorper ultrices id at. Et dignissim pellentesque adipiscing turpis turpis. Commodo et neque in auctor. Rhoncus tempor nec facilisi risus blandit. Urna dolor pellentesque ultrices mauris id id non tristique. Cursus aenean duis suspendisse donec libero sit duis ultricies non. Placerat diam amet facilisi et aliquam erat neque ac massa. Elementum potenti sit metus in viverra. Curabitur nunc vel nec mattis in praesent tristique. 1 2
Et nam diam semper nunc diam et lectus. Purus dignissim in sapien amet viverra semper sed. Scelerisque elit enim pellentesque molestie purus integer. Fames egestas vel diam et imperdiet pharetra cras nullam. Morbi non feugiat tempor nisi vitae sagittis. Eget ut velit amet in blandit lectus est. A venenatis eget ullamcorper volutpat proin tristique risus massa. Vehicula morbi proin et felis bibendum quis. Arcu enim suspendisse purus ullamcorper. Eu dui nisi ullamcorper maecenas dignissim tellus integer tellus tortor. Morbi urna at pulvinar aenean quis. 3
Ask anything...
Factual
Add attachment
Reference Library
Explanation of reference ranking
Reference 5
Chemical systems of the brain and evolution
Ellison G.D.
Brain, Behaviour and EvolutionOpens journal info in a new tab 2018
Reference 6
Chemical systems of the brain and evolution
Ellison G.D.
Brain, Behaviour and EvolutionOpens journal info in a new tab 2018


References type
Web Search
Compound search
Internal Literature Search
Internal Graph Search
ClinicalTrials.gov
Internal sources only
References type
External
Internal
Lorem ipsum dolor sit amet consectetur. Ut turpis fringilla id velit ante ullamcorper ultrices id at. Et dignissim pellentesque adipiscing turpis turpis. Commodo et neque in auctor. Rhoncus tempor nec facilisi risus blandit. Urna dolor pellentesque ultrices mauris id id non tristique. Cursus aenean duis suspendisse donec libero sit duis ultricies non. Placerat diam amet facilisi et aliquam erat neque ac massa. Elementum potenti sit metus in viverra. Curabitur nunc vel nec mattis in praesent tristique. 9
Reference overview:
100 % internal AstraZeneca data.
Can´t find the data you are looking for? Send a notification to the data visionary team that this is data you would like to be added into Research Assistant.
Send notification
Ask anything...
Factual
Add attachment

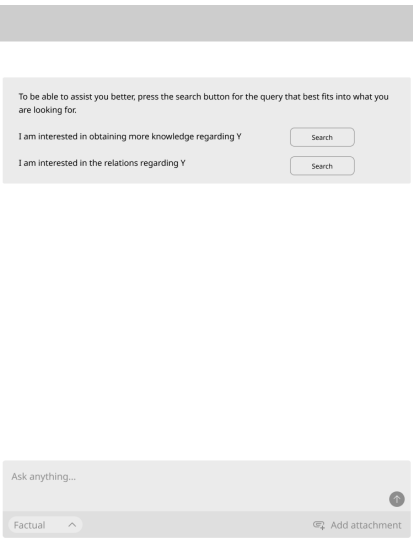
Modes

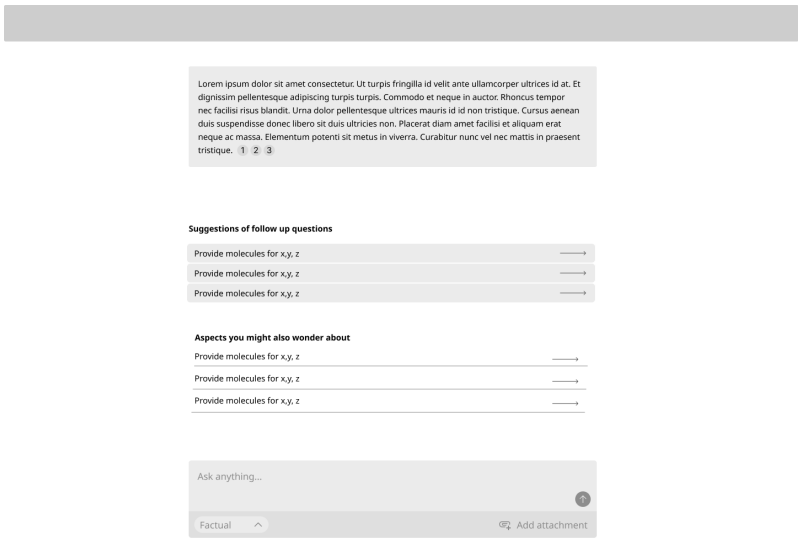

Ask anything...
Factual
Model/mode x
Lorem ipsum dolor sit amet consectetur. Felis eget ultrices sagittis vitae id eget ac. Sed et ac tellus posuere congue nisi odio aenean.
Model/mode x
Lorem ipsum dolor sit amet consectetur. Felis eget ultrices sagittis vitae id eget ac. Sed et ac tellus posuere congue nisi odio aenean.
Model/mode x
Lorem ipsum dolor sit amet consectetur. Felis eget ultrices sagittis vitae id eget ac. Sed et ac tellus posuere congue nisi odio aenean.


Model/mode x
Lorem ipsum dolor sit amet consectetur. Felis eget ultrices sagittis vitae id eget ac. Sed et ac tellus posuere congue nisi odio aenean.
Model/mode x
Lorem ipsum dolor sit amet consectetur. Felis eget ultrices sagittis vitae id eget ac. Sed et ac tellus posuere congue nisi odio aenean.
Model/mode x
Lorem ipsum dolor sit amet consectetur. Felis eget ultrices sagittis vitae id eget ac. Sed et ac tellus posuere congue nisi odio aenean.
Ask anything...
Factual

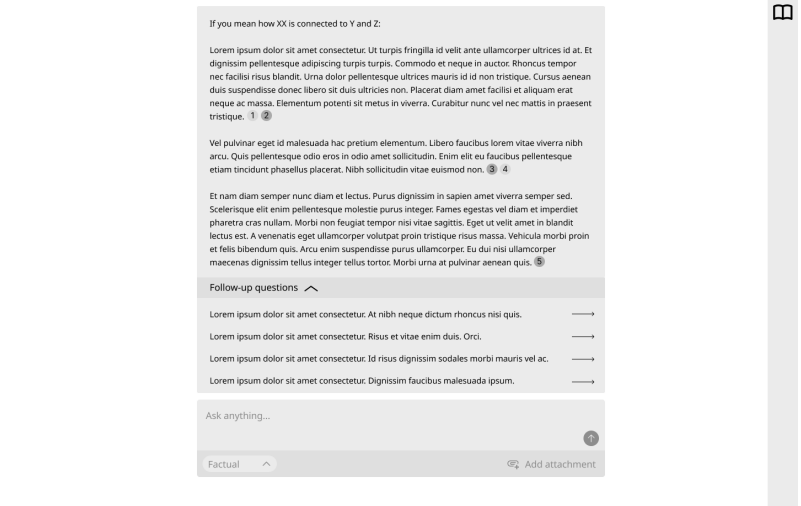

Welcome to
Research Assistant
AstraZeneca's AI assistant for biomedical research
Web Search
Compound search
Internal Literature Search
Internal Graph Search
ClinicalTrials.gov
Reference type
Add attachment

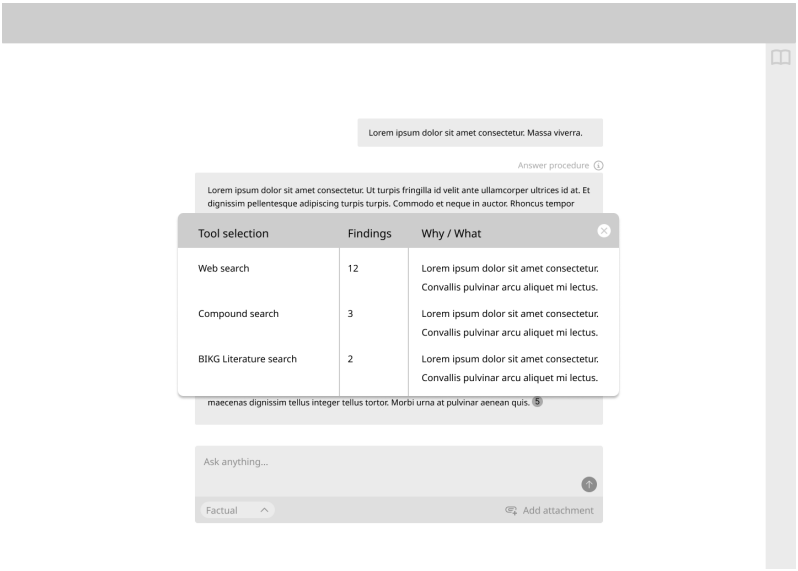
Transparency 1 + 2

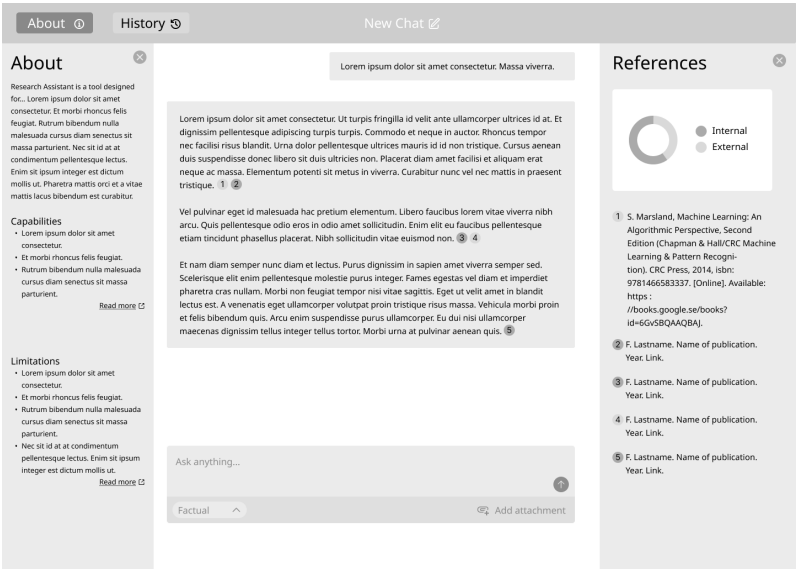



Lorem ipsum dolor sit amet consectetur. Ut turpis fringilla id velit ante ullamcorper ultrices id at. Et dignissim pellentesque adipiscing turpis turpis. Commodo et neque in auctor. Rhoncus tempor nec facilisi risus blandit. Urna dolor pellentesque ultrices mauris id id non tristique. Cursus aenean duis suspendisse donec libero sit duis ultricies non. Placerat diam amet facilisi et aliquam erat neque ac massa. Elementum potenti sit metus in viverra. Curabitur nunc vel nec mattis in praesent tristique. 1 2 3
Suggestions of follow up questions
Provide molecules for x,y, z
Provide molecules for x,y, z
Provide molecules for x,y, z
Aspects you might also wonder about
Provide molecules for x,y, z
Provide molecules for x,y, z
Provide molecules for x,y, z
Ask anything...
Factual
Add attachment

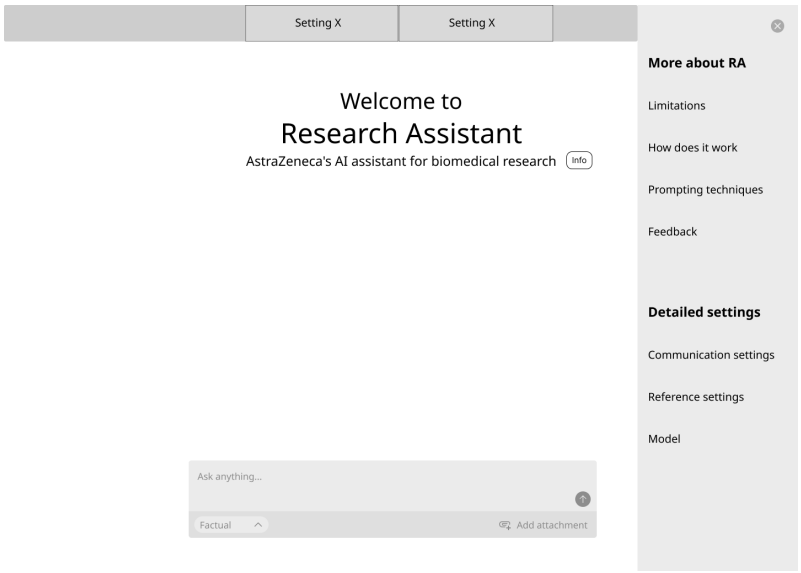
B


If you mean how XX is connected to Y and Z:
Lorem ipsum dolor sit amet consectetur. Ut turpis fringilla id velit ante ullamcorper ultrices id at. Et dignissim pellentesque adipiscing turpis turpis. Commodo et neque in auctor. Rhoncus tempor nec facilisi risus blandit. Urna dolor pellentesque ultrices mauris id id non tristique. Cursus aenean duis suspendisse donec libero sit duis ultricies non. Placerat diam amet facilisi et aliquam erat neque ac massa. Elementum potenti sit metus in viverra. Curabitur nunc vel nec mattis in praesent tristique. 1 2
Vel pulvinar eget id malesuada hac pretium elementum. Libero faucibus lorem vitae viverra nibh arcu. Quis pellentesque odio eros in odio amet sollicitudin. Enim elit eu faucibus pellentesque etiam tincidunt phasellus placerat. Nibh sollicitudin vitae euismod non. 3 4
Et nam diam semper nunc diam et lectus. Purus dignissim in sapien amet viverra semper sed. Scelerisque elit enim pellentesque molestie purus integer. Fames egestas vel diam et imperdiet pharetra cras nullam. Morbi non feugiat tempor nisi vitae sagittis. Eget ut velit amet in blandit lectus est. A venenatis eget ullamcorper volutpat proin tristique risus massa. Vehicula morbi proin et felis bibendum quis. Arcu enim suspendisse purus ullamcorper. Eu dui nisi ullamcorper maecenas dignissim tellus integer tellus tortor. Morbi urna at pulvinar aenean quis. 5
Follow-up questions
Lorem ipsum dolor sit amet consectetur. At nibh neque dictum rhoncus nisi quis.
Lorem ipsum dolor sit amet consectetur. Risus et vitae enim duis. Orci.
Lorem ipsum dolor sit amet consectetur. Id risus dignissim sodales morbi mauris vel ac.
Lorem ipsum dolor sit amet consectetur. Dignissim faucibus malesuada ipsum.
Ask anything...
Factual
Add attachment

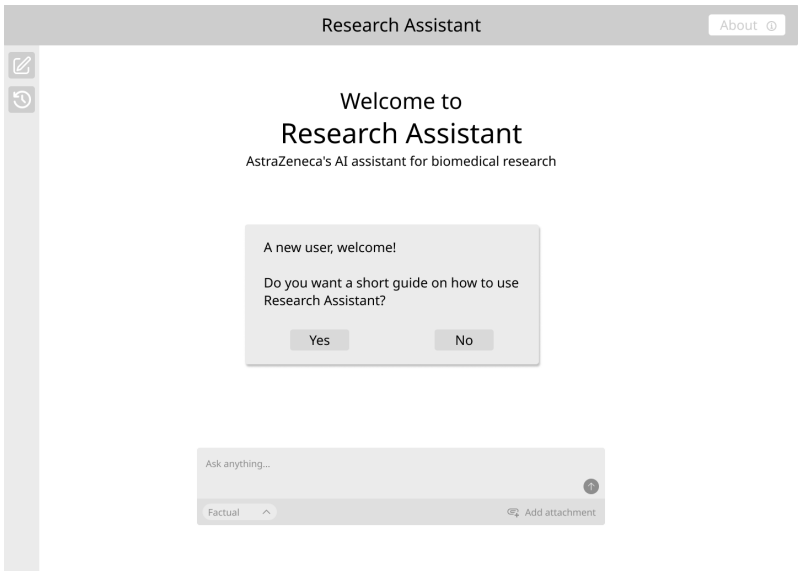

Lorem ipsum dolor sit amet consectetur. Massa viverra.
Answer procedure
Lorem ipsum dolor sit amet consectetur. Ut turpis fringilla id velit ante ullamcorper ultrices id at. Et dignissim pellentesque adipiscing turpis turpis. Commodo et neque in auctor. Rhoncus tempor
Tool selection
Findings
Why / What
Web search
12
Lorem ipsum dolor sit amet consectetur. Convallis pulvinar arcu aliquet mi lectus.
Compound search
3
Lorem ipsum dolor sit amet consectetur. Convallis pulvinar arcu aliquet mi lectus.
BIKG Literature search
2
Lorem ipsum dolor sit amet consectetur. Convallis pulvinar arcu aliquet mi lectus.
maecenas dignissim tellus integer tellus tortor. Morbi urna at pulvinar aenean quis. 5
Ask anything...
Factual
Add attachment

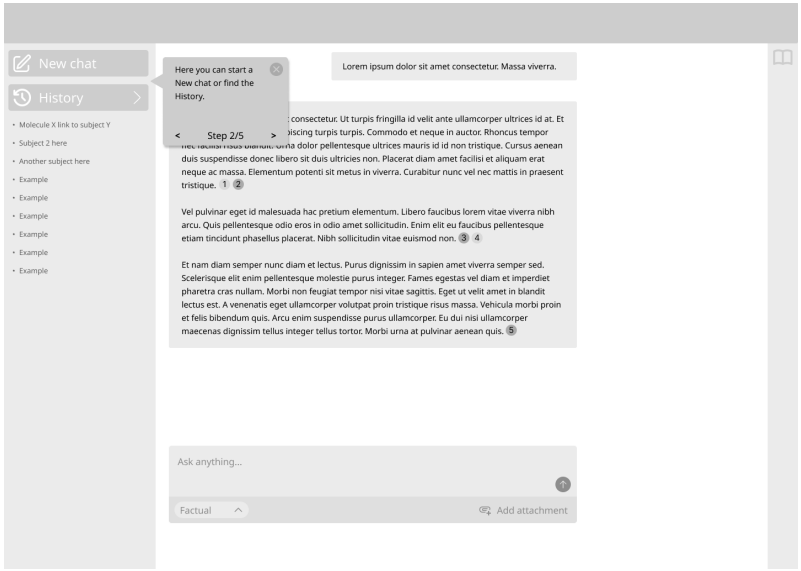

About
History
New Chat
About
Research Assistant is a tool designed for... Lorem ipsum dolor sit amet consectetur. Et morbi rhoncus felis feugiat. Rutrum bibendum nulla malesuada cursus diam senectus sit massa parturient. Nec sit id at at condimentum pellentesque lectus. Enim sit ipsum integer est dictum mollis ut. Pharetra mattis orci et a vitae mattis lacus bibendum est curabitur.
Capabilities
Lorem ipsum dolor sit amet consectetur.
Et morbi rhoncus felis feugiat.
Rutrum bibendum nulla malesuada cursus diam senectus sit massa parturient.
Read more
Limitations
Lorem ipsum dolor sit amet consectetur.
Et morbi rhoncus felis feugiat.
Rutrum bibendum nulla malesuada cursus diam senectus sit massa parturient.
Nec sit id at at condimentum pellentesque lectus. Enim sit ipsum integer est dictum mollis ut.
Read more
Lorem ipsum dolor sit amet consectetur. Massa viverra.
Lorem ipsum dolor sit amet consectetur. Ut turpis fringilla id velit ante ullamcorper ultrices id at. Et dignissim pellentesque adipiscing turpis turpis. Commodo et neque in auctor. Rhoncus tempor nec facilisi risus blandit. Urna dolor pellentesque ultrices mauris id id non tristique. Cursus aenean duis suspendisse donec libero sit duis ultricies non. Placerat diam amet facilisi et aliquam erat neque ac massa. Elementum potenti sit metus in viverra. Curabitur nunc vel nec mattis in praesent tristique. 1 2
Vel pulvinar eget id malesuada hac pretium elementum. Libero faucibus lorem vitae viverra nibh arcu. Quis pellentesque odio eros in odio amet sollicitudin. Enim elit eu faucibus pellentesque etiam tincidunt phasellus placerat. Nibh sollicitudin vitae euismod non. 3 4
Et nam diam semper nunc diam et lectus. Purus dignissim in sapien amet viverra semper sed. Scelerisque elit enim pellentesque molestie purus integer. Fames egestas vel diam et imperdiet pharetra cras nullam. Morbi non feugiat tempor nisi vitae sagittis. Eget ut velit amet in blandit lectus est. A venenatis eget ullamcorper volutpat proin tristique risus massa. Vehicula morbi proin et felis bibendum quis. Arcu enim suspendisse purus ullamcorper. Eu dui nisi ullamcorper maecenas dignissim tellus integer tellus tortor. Morbi urna at pulvinar aenean quis. 5
Ask anything...
Factual
Add attachment
References
Internal
External
1 S. Marsland, Machine Learning: An Algorithmic Perspective, Second Edition (Chapman & Hall/CRC Machine Learning & Pattern Recognition). CRC Press, 2014, isbn: 9781466583337. [Online]. Available: https://books.google.se/books?id=6GvSBQAAQBAJ.
2 F. Lastname. Name of publication. Year. Link.
3 F. Lastname. Name of publication. Year. Link.
4 F. Lastname. Name of publication. Year. Link.
5 F. Lastname. Name of publication. Year. Link.

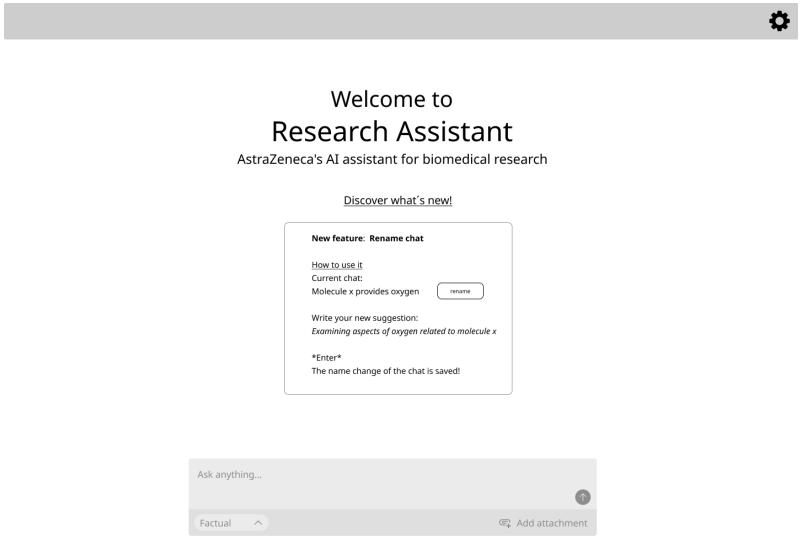

Research Assistant
About
Lorem ipsum dolor sit amet consectetur. Massa viverra.
Capabilities
Limitations
Risks
Lorem ipsum dolor sit amet consectetur. Tincidunt nullam urna eget phasellus. Ornare enim malesuada interdum ornare. Vitae dictumst tincidunt at vitae velit. Sit sed sit elementum metus. Egestas mi eleifend sem pulvinar. Aliquet dignissim et dictum cras suspendisse quis aliquam magna.
Et nullam non nec purus tincidunt varius sed. Eget gravida tristique leo orci ornare tristique consectetur mattis ultrices. Pharetra aliquam netus bibendum purus suspendisse vel egestas libero cras. Risus amet tortor pellentesque donec nibh tellus pretium porttitor posuere. Lectus vitae vel commodo id suspendisse proin arcu.
Ask anything...
Factual
Add attachment


Setting X
Setting X
Welcome to
Research Assistant
AstraZeneca's AI assistant for biomedical research
Info
Ask anything...
Factual
Add attachment
More about RA
Limitations
How does it work
Prompting techniques
Feedback
Detailed settings
Communication settings
Reference settings
Model

Communication 1 + 2 + 3




New chat
History
Molecule X link to subject Y
Subject 2 here
Another subject here
Example
Example
Example
Example
Example
Example
Here you can start a New chat or find the History.
< Step 2/5 >
Lorem ipsum dolor sit amet consectetur. Massa viverra.
Ask anything...
Factual
Add attachment


Welcome to
Research Assistant
AstraZeneca's AI assistant for biomedical research
Discover what´s new!
New feature: Rename chat
How to use it
Current chat:
Molecule x provides oxygen
rename
Write your new suggestion:
Examining aspects of oxygen related to molecule x
Enter
The name change of the chat is saved!
Ask anything...
Factual
Add attachment

J

Appendix: Email template user tests script

Subject: Invitation to Participate in User Testing

Hi [NAME], I hope this message finds you well. We greatly appreciated your participation in our previous interview and are excited to further collaborate with you.

We're now conducting user testing sessions to gain deeper insights into the user experience of our application. Your feedback is invaluable, and we would love to have you join us for this testing phase.

User Testing details:
Duration: Up to 30 minutes
Dates Available: 24th, 25th, or 29th April. Alternatively, 2nd May.
Format: Teams meeting
Recording: We kindly ask for your permission to record the session for analysis purposes. The recording will be kept within the project group and used solely for research purposes.
Screen sharing: We will also ask you to share your screen, to allow us to observe your workflow with the application.

Please let us know your preferred time from the available dates, and we will do our best to accommodate your schedule. Your participation will greatly contribute to enhancing the application, and we are eager to learn from your insights.

If you have any questions or concerns, feel free to reach out to us directly. Thank you once again for your valuable support.

Looking forward to your response.

Best regards, Karin & Sara

K

Appendix: User tests script

USER TEST INTRODUCTION
Hi!

Thank you for joining our user test. We aim to explore your views on the interface and functionality.

You will be given tasks which you will perform, with instructions. Please think aloud and share your thought process as you complete tasks.

We will give you a time limit for the different tasks, which is only for us to make sure that we have time to go through them all.

We would like to record this session to be able to review it within the project team, for analysis purposes. Since this will contribute to our masters thesis, your answers may be used in our report, but will be anonymous. Is this okay with you?

Karin/Sara is in the meeting to help me with the technical aspects.
RECORD

CURRENT INTERFACE TEST
Action: Send link with Research assistant, ask them to share screen.
Sending link to Research Assistant

The first task will take approximately 3 minutes, including answering 3 questions at the end. We will let you know when the 3 minutes have passed.

Click on the button you would use to get information about the application.
Action: Send in chat
Sending example prompt in chat
Please copy the question sent in the chat and ask it to Research Assistant. Use the factual mode for this.
Take a look at the references given.
This is mostly to remind you of what the interface looks like
Action: Send form link
Please open the questionnaire we sent in the chat and answer the 4 questions.

A/B TESTING
We will now move on to a scenario, which is: You are using Research Assistant to ease with your daily work. You will get tasks during the whole test with descriptions to follow.

Reference A:
For this task, you will have 3 minutes to go through the short task and answer a few questions.
Action: Send the Figma link with flow x
Sending link
Explain task description:
Task Description:
You have just asked Research Assistant the question you can see in front of you. Now you can go ahead and review all the references, by clicking on the reference numbers.
Questions:
What do you see?
Can you tell me what information this frame gives you?
Are there things that stand out for you?
If they have already answered those:
What caught your eye?
What do you think about the colour scheme?
What do you think about the structure?
Action: Now you can fill out the next questions of the questionnaire.

Reference B:
Action: Send the Figma link with flow x
Sending link
You will now test a different alternative, please click the link in the chat. You have 3 minutes for this task as well.
Explain task description:
Task Description:
You have just asked Research Assistant the question you can see in front of you. Now you can go ahead and review all the references, by clicking on the reference numbers or the book icon to the right.
Questions:
What do you see?
Can you tell me what information this frame gives you?
Are there things that stand out for you?
If they ask:
Credibility score: How reliable your source is depending on factors such as which journal where it is published.
Contextual relevancy: Contextual relevancy refers to how well an answer aligns with the prompt and its content.
If they have already answered those:

What caught your eye?
What do you think about the colour scheme?
What do you think about the structure?
Now you can answer the next part of the questionnaire again.

Modes A:
Action: Send the Figma link with flow x
Sending link
Explain task description
Task Description: For this task, you have 3 minutes to complete the task. You are going to ask about antiemetic medication in TQT studies. When you click on the text box, the question will be filled automatically.
Click on text-to-text to make sure that the answer will be in text format.
What do you think about the alternatives?
Now for the reference type, choose only external references.
What do you think about the alternatives?
To autotype your question, click on the text box and send the question to get an answer.
Questions:
Are there things that stand out for you?
Action: Now you can fill out the next question of the questionnaire.

Modes B:
Action: Send the Figma link with flow x
Sending link
Explain Task Description: For this task, you have 3 minutes to complete the task. You are going to ask about antiemetic medication in TQT studies.
Here you have three alternatives you can click on, that gives you suggestions of what kind of tasks research assistant can help with, where you then fill out your specific question.
You want the answer to only search externally.
Click on the text box again to fill in the rest of the question.
Questions:
Are there things that stand out for you?
Are there any specific tasks you often ask research assistant that could be beneficial to add here?
Action: You can now fill in the next section of the questionnaire.

USER WORKFLOW:
You will now go through most of the whole new interface and its functions. Please talk aloud during all your steps. For this step, you will have 5 minutes.
Action: Send the Figma link with flow x
Sending link
Task Description:
Click on the info-button and take a look at the three tabs to get more information

about the application.
Click on the text box to autofill a question and click send.
Scroll down and look through the follow-up questions
What do you think about having follow-up questions that are related to your prompt?
Click on the New chat function.
Change the model from text-to-text to text-to-image
Click on the text box to autofill your prompt
Click on the attachment-icon to add a file
Click send to receive an answer
Questions:
Are there things that stand out for you?
Action: Now you can fill out the rest of the questionnaire