# Assistance Required: A Qualitative Study of Researcher Needs for AI Research Assistants

by

Marc Otten

In partial fulfillment of the requirements for the degree of
Master of Science
at the Delft University of Technology,
to be defended publicly on June 17th, 2025

| | |
|---|---|
| Faculty: | Electrical Engineering, Mathematics and Computer Science |
| Programme: | Master Computer Science |
| Track: | Software Technology |
| Research Group: | Web Information Systems |
| Graduation Committee: | Jie Yang, Supervisor |
| | Pradeep Murukannaiah, Supervisor |
| | Luciano Cavalcante Siebert |

An electronic version of this thesis is available at `http://repository.tudelft.nl`

TU Delft

# Assistance Required:
# A Qualitative Study of Researcher Needs for AI Research Assistants

**Marc Otten**
Delft University of Technology
Delft, Netherlands

## Abstract

The use of research assistants has increased significantly, providing support and automation for researchers. However, there is limited research on researchers using research assistants and what assistance researchers require for each research stage. We interview researchers to gather insights into their opinions and usage, and afterwards develop a prototype. Researchers highlight that the most difficult part of research is the experiment design, which is reflected in the lack of literature. Participants use research assistants for reading literature and writing, but request more support. The research assistant must be factual, transparent, and correct, and including a conversation allows for feedback and discussion. We evaluate the prototype for the experiment design phase, highlighting the effectiveness of the component architecture by generating correct experiments.

## 1 Introduction

Research follows the scientific method by having an idea, performing an experiment, and finishing with writing the results. Recently, Large Language Models (LLM) have started to assist or even automate parts of the process, resulting in papers similar to an early-stage researcher (Lu et al., 2024). However, the literature focuses on automating research instead of assisting. This is not a good direction due to hallucinations and ethical concerns. Involving users increases performance, and it is important to consider them when designing research assistants (Schmidgall et al., 2025). Additionally, the literature focuses on ideation and writing, often leaving out the experiment design.

Therefore, this explorative study looks into Artificial Intelligence (AI) research assistants, focusing on users' desires and opinions. We focus on three research questions and discuss the methods below.

- **RQ1:** What requirements do users desire for a research assistant?

For this question, we perform a user study to gather insights and identify concrete expectations of a research assistant. Additionally, we discuss the research process, allowing for better assistants.

- **RQ2:** How can a research assistant realize these requirements through functionalities?

In addition to a user study, we perform a study on existing research assistants to compare the previous insights. We highlight the functionalities needed to realize the requirements.

- **RQ3:** What are effective design choices for a research assistant in the experiment design phase using a conversation?

We design a prototype using the previous insights to find out how well we can realize the requirements and functionalities to assist with experiment design. We highlight effective design choices and limitations, focusing on the conversational approach. A graphical overview of the research question and methods can be seen in Figure 1

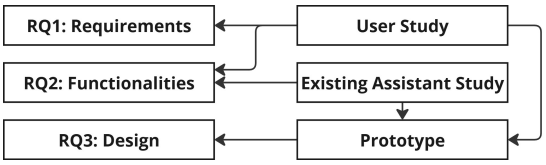


Figure 1: Research Questions & Methods Structure

The interviews show that participants have the most trouble with the experiment design phase, but enjoy it nonetheless. Participants use research assistants for gathering and summarizing literature or writing, but not for other research tasks. This is due to limited supportive capabilities and difficulties with writing prompts. Participants note they would use research assistants more if the performance were better. An ideal research assistant must be factual and efficient and support the entire process, allowing for personalized discussion and feedback.

The component design and conversations are effective and mimic participants' approach to experiment design, making it intuitive. The prototype is able to generate fitting experiment designs with the cooperation of a user. Summarizing the experiment design sections of related literature provides the best fitting information, according to our

experiment. However, future research should perform more interviews to validate our insights and perform a more rigorous evaluation of a research assistant for experiment design.

## 2 Related Work

### 2.1 Models & Techniques

Using LLMs, such as GPT and Llama, different tasks can be sped up and automated (Cambon et al., 2023; OpenAI et al., 2024; Grattafiori et al., 2024; Vaswani et al., 2023). Researchers are looking for new applications for research and techniques to improve performance and reliability. Retrieval Augmented Generation (RAG) enhances the prompt with relevant knowledge for the question and reduces hallucinations (Lewis et al., 2021). RAG relies on retrieving correct and fitting data, and there are multiple effective approaches for this difficult task (Gao et al., 2024; Chen et al., 2023; Peng et al., 2023). We will provide literature to the prototype for relevant information, like RAG, but we do not retrieve it automatically. For the question answering task, prompt techniques such as Chain of Thought and ReAct are effective, showing reasoning steps to derive an answer (Wei et al., 2022; Yao et al., 2023). However, the research process is iterative, and we therefore do not use these approaches, but we will include automated feedback since this reduces mistakes (Baek et al., 2025).

### 2.2 Research Process

There are many options to automate or assist with the research process. Research starts with an idea, and ideas generated by AI are novel and exciting, but lack feasibility and uniqueness (Si et al., 2024). The imagination of researchers is not negatively impacted (Ashkinaze et al., 2024). Using LLMs to assist with ideation makes ideas more semantically similar on a group level by not making diverse suggestions (Anderson et al., 2024). Models that provide feedback for the writing phase lower writing diversity (Padmakumar and He, 2024). Models are able to come up with robust hypotheses (Zhou et al., 2024). Iteratively refining hypotheses allows for high-quality novel hypotheses, even for open domain questions (Yang et al., 2024; Qiu et al., 2024). LLMs can assist during the experiment design phase and identify and train fitting models for Machine Learning (ML) research (Zhang et al., 2023). By providing research goals and data, it is possible to automate the research process and produce findings autonomously (Ifargan et al., 2024; Guo et al., 2024). It is also possible to perform simple real-world chemical experiments (Boiko et al., 2023). Smaller experiments are more feasible to be generated and executed, especially for ML (Huang et al., 2024). However, not all experiments can be automatically performed, so generating ideas, hypotheses, and experiment designs is a good alternative (Baek et al., 2025). The experiment design task is challenging, therefore, we develop and evaluate a new possible design to provide new insights.

It is possible to merge different systems, increasing their capabilities and potential (Li et al., 2024). It is possible to generate full-length papers with figures from an initial seed idea that is iteratively improved upon at the level of an early-stage researcher (Lu et al., 2024). New designs of research assistants use multiple agents with a distinct role instead of one (Ghafarollahi and Buehler, 2024a,b). As well as involving humans in the process, instead of automating, resulting in higher-quality results using conversations for feedback and discussion (Gottweis et al., 2025; Schmidgall et al., 2025). We adopt a multi-agent structure and human involvement in the prototype to verify their effectiveness.

### 2.3 Existing Tools

Gathering knowledge is a crucial part of the research process and is mainly done by reading literature. Tools exist that highlight important sentences and references, explain symbols and formulas, and summarize literature (Lo et al., 2023). Other tools gather literature and generate an article with it to answer your questions (Shao et al., 2024). With sufficient information or clarification, these systems can retrieve relevant papers (Lála et al., 2023). But sometimes not all information is present, and a user might not know what to search for. Using interactive intent modeling, exploratory search becomes significantly easier and adapts every search (Ruotsalo et al., 2018). Some tools can answer questions, summarize texts, and recommend new literature all in one, thereby improving usability (Zheng et al., 2024). LLMs can provide reviews and feedback, but their focus is not aligned with human reviewers, and LLMs lack performance on complex issues (Liang et al., 2023). These tools are effective, but we will interview researchers directly to identify other important tasks to assist with.

Most research assistant tools do not have public research associated with them due to the financial incentives not to let the competition catch up and

attract users to their service. The limited research evaluating existing tools shows vast differences in databases and algorithms used, as outputs differ drastically for the same task (Danler et al., 2024). Tools for writing are widely used for paraphrasing, grammar, references, citations, and proofreading (Guhan et al., 2023). Most online tools allow for conversations with the user, but this aspect is not well-researched. Conversations are often mentioned as a side note and always increase performance and usability (Ifargan et al., 2024; Zheng et al., 2024; Schmidgall et al., 2025; Ruotsalo et al., 2018; Gottweis et al., 2025). Therefore, this research focuses on the conversational aspect.

### 2.4 Benchmarks

LLMs are often closed source, but there is a rise in open source models. However, there can be disparities between long-term reasoning and instruction following abilities in some tasks (Liu et al., 2023). Multiple benchmarks exist to evaluate LLMs, testing their capabilities to conduct ML research tasks (Huang et al., 2024; Chen et al., 2024). Other benchmarks focus on general capabilities, such as organizing knowledge or interpreting figures (Kang and Xiong, 2025; Roberts et al., 2024). Benchmarks for data-driven research highlight the need for further advancements (Majumder et al., 2024; Jing et al., 2025). However, benchmarks for experiment design do not exist, so we derive guidelines to evaluate this task.

## 3 User Study

We perform a user study to answer **RQ1** and **RQ2**, providing insights and identifying requirements and functionalities of research assistants. After analysing the interviews, we identified four themes. To answer **RQ1**, we use the first three themes, and for **RQ2**, we use the last theme combined with our study on existing research assistants.

### 3.1 Methodology

Interviews, contrary to surveys, allow us to get to the essence by asking follow-up questions. We do not know what participants will respond with, and this format allows us to explore more in-depth. Additionally, we only need to mention examples if the participant needs them, reducing bias in their answers. A disadvantage of interviews is that they are time-consuming, resulting in fewer possible interviews. Additionally, the analysis takes longer due to the qualitative nature of the insights and opinions.

Table 1: Interview Structure

| Section | Topic |
|---|---|
| Administrative | Consent, information, and interview structure |
| Background | Research experience and AI understanding |
| Work Practice | Research process and enjoyment |
| Tool Evaluation | Hands-on experience and experiment design |
| Current Usage | What, why, and how participants use research assistants |
| Needs | Desired (future) functionalities and needs |
| Closing | Additional remarks and confirm consent |

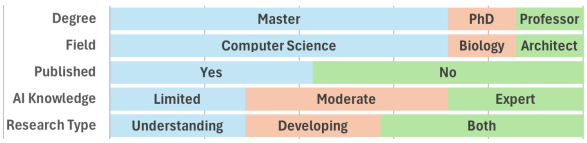


Figure 2: Interview participant distribution

The goal is to make broad questions, forcing the participants to think and provide detailed answers. Table 1 shows the interview structure.

In addition to the interviews, we study existing research assistants that are available online. This is necessary to compare what is available to researchers with the literature. Not all tools have associated literature, meaning the performance and capabilities of online tools might not reflect the literature. This allows us to perform the interviews with a comprehensive understanding of the domain and validate the insights from the interviews.

### 3.2 Experiment Setup

The Delft University of Technology Human Research Ethics Committee approved the interviews and data management. We gather participants directly through our social networks.

There are seven participants in this study, and the distribution and details are shown in Figure 2. The diversity of the participants allows us to gather diverse insights and opinions. For a better experience, the interviews are not transcribed during the interview but recorded and transcribed afterwards.

The research assistant used for the tool evaluation section is NotebookLM [1], and its three main goals are: effective learning, organizing thoughts, and inspiring new ideas. The user provides literature that the assistant uses to update its knowledge, allowing it to answer questions, provide feedback,

[1] https://notebooklm.google/

and have discussions. This is done via a conversation, and the tool can assist with generating hypotheses and designing experiments. We select this tool for the interview due to its capabilities to do more than find and discuss literature.

The next step is to analyze the interviews via reflexive thematic analysis (Braun and Clarke, 2021). Reflexive means involving the researcher's experiences and knowledge to reflect on the potential insights. Thematic analysis is the standard approach for qualitative analysis and aims to find themes in the data. The interview recordings are transcribed, and the interesting statements in the transcription are highlighted and coded. We perform two rounds of coding to ensure consistency and allow the codes to evolve gradually. We end up with 28 codes, which are grouped, and themes are derived from these groups. This is an iterative process, the themes are further refined to provide fitting insights for the research. Lastly, each theme gets an abstract explaining the central principles and contributions. The transcriptions are not publicly available, but the insights and statistics are discussed in this study.

To perform our study on existing research assistants, we find and select publicly available ones that are actively used and maintained. Additionally, they have to be free to use and assist with the research process, not automate it. There must also be diversity in the tasks they support. We identify research assistants by looking online and asking our network for research assistants they use.

### 3.3 RQ1: Research Assistant requirements

#### 3.3.1 Participants' Research Process

The first theme about the research process provides valuable insights and highlights which parts are difficult or enjoyable. The phases that research consists of, according to the interview participants, are as follows: ideation, experiment design, result & conclusion analysis, and writing, with ideation being broad and containing the hypothesis phase.

> Participants' research process is the same as the traditional research approach.

Noteworthy additional phases, mentioned by some participants, consist of familiarization, background knowledge, discussion, and research structure. However, it can be argued that these are not separate phases but subtasks of the traditional phases. Familiarization focuses on getting familiar with a topic by reading general literature to form an idea. Not all participants explicitly mention this because some already have an idea for their research. Background knowledge consists of reading specific literature after forming an idea. The discussion phase discusses the research with peers and supervisors to identify interesting aspects and check the reasoning. The research structure phase focuses on clarifying the goal, the planning, and the expected results, and is specifically mentioned by master's students because they are evaluated on their process. Participants highlight that the research process, especially the experiment design phase, is iterative and involves returning to previous phases to update them.

The opinions on the research process are:

- **Reading Literature -** Straightforward, but generally not enjoyable, although it differs per paper.
- **Ideation -** Designing a novel and feasible idea is difficult.
- **Experiment Design -** Most enjoyable because participants can apply their knowledge and gather insights or results. But it comes at the expense of being the most difficult and time-consuming phase. However, repetitive tasks such as reading transcripts or running long experiments are not enjoyable.
- **Result & Conclusion -** Divided opinions, some participants have conclusive outputs and thus find it easier, whereas others have qualitative outputs requiring more effort to extract conclusions.
- **Writing -** Mostly neutral or slightly negative opinions. But participants do note it is interesting to see everything coming together.

> Overall, the later parts of the research are more enjoyable because participants can focus on their own research.

#### 3.3.2 User & AI Expectations

The second theme, user & AI expectations, highlights what users expect and what is expected of them, as well as the interaction between the user and the research assistant.

> Participants mention neither they nor the assistant can be perfect, and they can both make mistakes.

This means both parties should be critical and think logically for themselves. Mistakes should be

pointed out and corrected, allowing for a better process continuation. Additionally, either party should indicate if information is unknown, and the needed information should be gathered or exchanged, as participants want to know the certainty the assistant has that it is correct. Questions from the user or the assistant highlight possible misunderstandings or incomplete information and should be answered. For efficient cooperation, all text must be high quality and grammatically correct.

Participants want to always be in control and have the final say in disagreements, ensuring it is always possible to move forward. However, this does not mean the assistant should always assume the user is correct. Additionally, the user must be critical of the assistant's performance and decide if something is possible for the assistant.

Participants highlight that NotebookLM is useful and were unaware that research assistants had such capabilities. Participants note that writing a fitting and correct prompt is difficult. They consider using research assistants more often, especially if performance increases. It is also important that a tool assists them in research and not automates it, as they do not want to be replaced. Researchers should disclose if they use AI and why it is applicable. Lastly, some tasks need to be performed manually, especially when sensitive data is involved.

#### 3.3.3 Assistant Requirements

Now we discuss the third theme regarding the requirements that participants expect a research assistant to have. Requirements in this context are abstract concepts that the tool needs to realize. The main requirements mentioned by participants are about the output, but some are specific characteristics of the assistant or process itself.

> The most important requirement is that the output must be factually correct

This is because LLMs sometimes produce false statements that users possibly believe. Additionally, the output should be complete and contain all relevant information and logical reasoning. When the user provides new literature, the assistant should be able to learn, understand, and link it to other literature. Furthermore, it should be reliable, meaning the output and reasoning are consistent when the same prompt is used. The assistant should be trustworthy by consistently producing good answers.

> Transparency is another important requirement, requiring the assistant to show the source of the information.

Participants who are more knowledgeable about AI are concerned about hallucinations and require the research assistant to indicate if it does not know the answer. The assistant must be efficient, meaning it should not waste effort, have reasonable response times, and be a good option for a specific purpose. The assistant must be adaptable by being able to change between different tasks, user demands, and scopes. It should also be unbiased and not have strong opinions, but still be critical and detect if the user is wrong.

### 3.4 RQ2: Research Assistant Functionalities

#### 3.4.1 Existing Research Assistants Study

To find the required functionalities of a research assistant, we first discuss our study of existing online tools. A more detailed overview of the specific research assistants evaluated and our findings can be seen in Appendix B. There is a large disparity in performance, with free tools often having associated literature about their designs, such as ORKG Ask [2] and STORM [3], but performing worse than paid tools with no associated literature.

> The main existing functionalities are retrieving, summarizing, and discussing literature to answer questions.

This assists with ideation, but if a chat is available, then formulating a hypothesis is possible as well. Some tools focus purely on writing by generating text or finding references. However, a minority allows for a sophisticated discussion, such as NotebookLM, and can assist with designing experiments. The main limitation is the need to find and upload literature manually. No tool can currently assist with the entire research process, forcing users to switch between them for complete support. The biggest LLMs, such as ChatGPT, can provide some assistance due to their inherent reasoning capabilities, but are not specifically designed for this.

#### 3.4.2 Assistant Needs

The last theme showcases the required and desired needs and concrete functionalities of research assistants. All participants indicate that a chat function-

[2] https://ask.orkg.org/

[3] https://storm.genie.stanford.edu/

ality is useful. A chat mimics the iterative nature of research and allows users to elaborate, correct, or follow up on the previous response.

> A chat mimics the iterative nature of research and allows users to elaborate, correct, or follow up on the previous response.

But finding information in a conversation can be tedious. A functionality that provides a summary or highlights would alleviate this issue. Being able to steer the assistant is a necessity, and by selecting sources, filtering aspects, and specifying the scope, users get personalized assistance. The ability to upload documents is important, especially for ones not publicly available. Some participants want the ability to specify output length, since the outputs are often too long. This makes a tool more user-friendly, but can lead to longer conversations.

> A crucial functionality that increases trust and transparency, which participants find very important, is listing sources.

Participants want the ability to take notes separately from the conversation to remember their insights and use them for inspiration later. The research assistant should support the entire research process because switching between assistants to get adequate support is a reason for participants not to use them. Participants doing computer science research want assistance with coding, and other research fields likely desire other specific functionalities. Some participants struggle with writing and request assistance with grammar. Additionally, it should suggest a logical storyline, be able to identify what is relevant, and cite literature.

Participants who recently started new research focused on literature functionalities, both retrieving and absorbing the information. The main approach for absorbing information is summarizing the literature, making it easier and faster to digest. An alternative possibility is highlighting important sentences, allowing users to read the surrounding text that contains details or clarifications. Additionally, the capability of generating podcasts would also be useful. Participants want the assistant to be able to generate and understand images, which simultaneously allows for more knowledge extraction.

> Participants find reviewing the most important functionality for a research assistant, providing criticism and feedback on the literature or user.

Participants want other perspectives and confirmation on the interesting aspects of their research. The assistant must be able to hold a discussion and understand other viewpoints, as well as highlight possible new steps or aspects for the research, making it more concrete. Supervisors or peers normally provide feedback, but a research assistant can do this more often and earlier. Lastly, another key capability is being proactive by making suggestions and pointing out mistakes of the user. However, there is a fine line between being proactive and automating the research process.

## 4 Prototype

We design the prototype for the experiment design phase. We select this phase due to the limited representation in research, and our user study highlighting this phase to be the most difficult. The goal of the prototype is to highlight how well we can realize the requirements and functionalities found in the user study. We evaluate the effectiveness of assisting with the experiment design phase and gather insights into the usability and conversational aspect. Using this evaluation and insights, we answer **RQ3** and discuss the design choices. The code and evaluation outputs are publicly available[4].

### 4.1 Methodology

The design is based on the insights from the user study, implementing a conversation. The conversation allows the prototype to provide feedback, ask questions to the user, and mimic the iterative nature by performing the process in steps, as these are the main requirements according to the participants. The design can be seen in Figure 3 and is also based on multi-agent systems seen in the literature (Gottweis et al., 2025; Ghafarollahi and Buehler, 2024a,b; Schmidgall et al., 2025).

It consists of multiple components in a set order, each with its distinct task. The structure of the design mimics the research process from the user study results. We constrain the prototype design and experiments to the computer science field to allow us to evaluate the outputs. The model and parameter details are highlighted in Appendix A.

The first component of the prototype, summarization, focuses on preparing the literature for later use. This step reduces the size of the literature, allowing it to fit in the context window of the model.

[4] https://github.com/delftcrowd/RA_Experiment_Design/tree/main

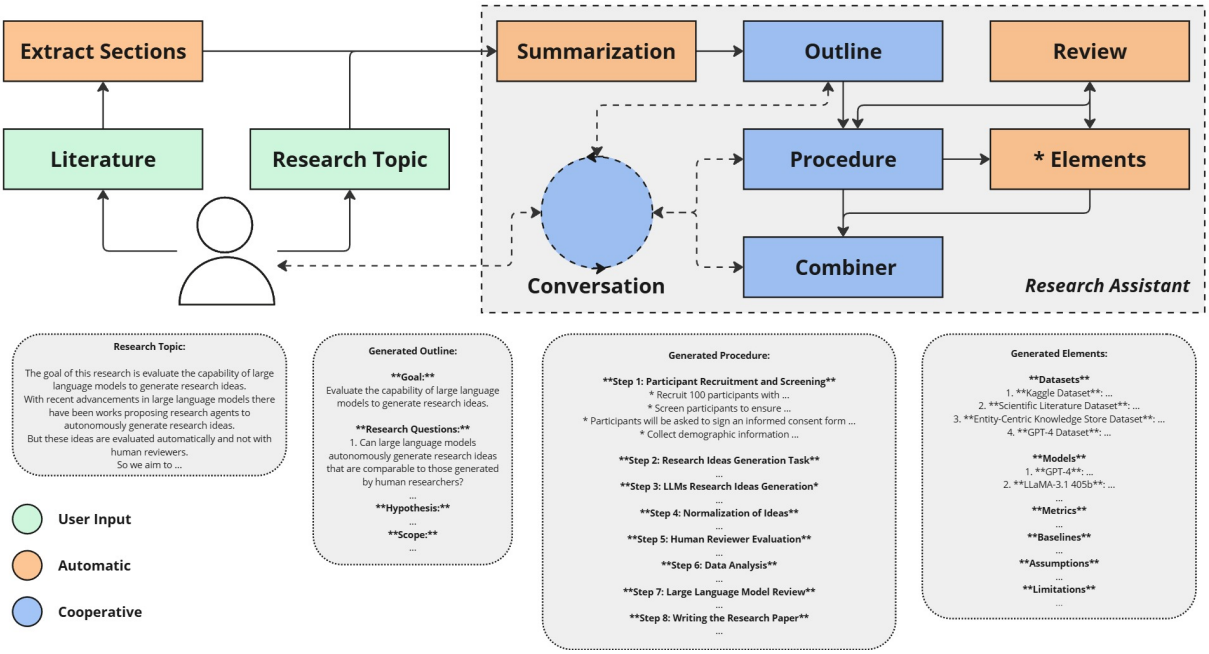


Figure 3: Prototype Design & Workflow With Examples

The outline component asks the user questions to help finalize the details for the research, which consists of the goal, research questions, hypotheses, and scope. Then, the procedure component designs a general stepwise experiment procedure with the user. This ensures the procedure is well structured and logical because energy is not wasted on the details. The elements for the prototype are: datasets, models, metrics, baselines, assumptions, and limitations. Other research fields will need to add or remove specific categories of elements. Each element is a separate component since identifying all elements in a single component is too difficult and results in lower-quality elements. The review component reviews the initial procedure and elements, and the respective components revise them. Due to the nature of the conversation in the procedure component, the revision is only done initially to give the user more control. The last component combines the elements and procedure into a coherent and detailed experiment design. The user converses with the combiner component as well, changing the details, elements, or even steps if needed.

The system prompts of each component are general and adhere to a specific structure and can be found in Appendix C. Strict and precise system prompts can result in the model repeating itself indefinitely. The prompt starts with a role description followed by the task and goal. Then, an explanation of the resources, which consist of the summaries and outputs from previous components. It ends with the restrictions and details on the output.

## 4.2 Evaluation

We perform two distinct experiments on the prototype and evaluate them using experiments from two papers. We replicate the experiment designs from the research of (Si et al., 2024) about LLM *ideation* and (van Dam et al., 2023) about *enriching* code completions.

The first experiment evaluates which sections of a paper are the most useful to summarize. There are many possible combinations of sections to test, so we performed a preliminary analysis and analyzed the summaries. This reveals that there needs to be sufficient information and limited noise to summarize the paper effectively. The top three detailed and complete summaries are from design-focused, result-focused, and the entire paper, so we use these for the prototype evaluation. All options contain the abstract and introduction. The result-focused sections also contain the discussion and conclusion. We use the best combination of sections from the first experiment in the second experiment.

The second experiment focuses on the components of the prototype and their effectiveness. It consists of four alternate structures of the prototype: no summarization, no outline, no review, and merging the procedure, element, and combiner components.

We construct guidelines to ensure evaluation is consistent and mimics real usage of the system.

- The user must not respond with multiple actions for the system to perform.
- The actions must be precise and not vague.

Table 2: Main observed differences from the section experiment

| Sections | Wording | Procedure | Elements |
|---|---|---|---|
| Design | Descriptive | - Focus on data & data analysis<br>- A lot of details | **Ideation** - Correct models, metrics, assumptions, & limitations<br>**Enriching** - Correct models, metrics, baselines, assumptions, & limitations |
| Result | Concrete | - Focus on gathering more results<br>- Some details | **Ideation** - Correct models, assumptions, & limitations. Metrics contain no explanation on how to measure<br>**Enriching** - Correct baselines, assumptions, & limitations. Limited quality models & metrics. |
| All | General & broad | - Focus on general research practices<br>- Limited details | **Ideation** - Correct models & limitations. Assumptions are hallucinated statements and metrics are hard to quantify<br>**Enriching** - Correct models, metrics, assumptions, limitations. Incorrect baselines. |

- When answering a question from the assistant, the input must not contain additional actions.
- The user must handle the model response from top to bottom and use correct grammar.
- The user should only ask for feedback or ideas from the system at the end of a task.

We analyze the conversation and intermediary outputs of each component to observe differences in the process and interpret their effects. The guideline to evaluate the experiment is constructed by analysing three different experiment design guidelines from: ACL Responsible NLP Research [5], AAAI Reproducibility [6] Checklist, and ACL Review Guidelines 3.4 [7]. These guidelines focus on writing and not the design itself, but are still a fitting source of inspiration. The experiment design must contain all relevant parameters, variables, design choices, motivation, and references. The evaluation procedure and metrics must be applicable and allow for comparison with existing research. All limitations and potential risks should be listed, and their effects should be explained. These elements allow us to check the experiment design.

### 4.3 RQ3: Experiment Design Assistant

#### 4.3.1 General Insights

First, we list some general insights observed when evaluating the prototype. Extracting the text from a PDF includes footnotes and figure subtexts. Removing this noise and normalizing newlines could improve performance, but it is not trivial to do automatically. Parts of the user topic can bleed into the summary, which is not ideal, but results in a more detailed summary. The model might not perform all requests by the user if there are too many in one prompt. The components are instructed not to make changes that the user did not specify, but this can still happen. A larger model with more inherent reasoning capabilities will possibly better adhere to the system prompts, but could increase runtime and thus reduce user experience. The combiner component does not always add all elements to the procedure or adds the elements to multiple or wrong spots, even when there is a logical place for them. This is likely due to the large number of elements and varying applicability. A conversation in the element component would allow the user to filter the best elements and thus decrease the number of elements, making it easier to incorporate them. This approach would also allow the user to specify the types of elements that are applicable for their research. The final experiment design contains most aspects of our evaluation guidelines. However, the risks and motivations are not explicitly mentioned, but present in the conversation. Adding references is not feasible and results in hallucinations.

> The conversation allows for discussing and iterating on the users' vision and is effective.

#### 4.3.2 Section Experiment

We list the main observations from the experiment using different sections for summarization in Table 2. We will interpret and discuss these observations, for example outputs see Appendix D.

The main difference is in the wording of the outputs, mimicking the tone of the summaries. Using all sections also introduces some noise. The differences in the outlines are only small. However, the procedures differ significantly, but all end up with the same steps, except for one step. Each element group, regardless of the summary used, contains many elements with substantial differences in qual-

[5] https://aclrollingreview.org/responsibleNLPresearch/

[6] https://aaai.org/conference/aaai/aaai-23/reproducibility-checklist/

[7] https://aclrollingreview.org/reviewerguidelines#paper-issues

Table 3: Main observed differences from the component experiment

| Experiment | Effect | Observations |
|---|---|---|
| No Summarization | This means there is no literature available for additional information | - The user has to specify and confirm the hypothesis<br>- Questions are more focused on the research topic<br>- Simpler wording and tone<br>- Fewer details and more ambiguous wording.<br>- It has difficulties identifying elements. |
| No Outline | The research topic given directly to the procedure component, resulting in less clarity on the research | - Additional procedure step to clarify the topic.<br>- Other steps are the same, but shorter and less detailed.<br>- Elements are less fitting. |
| No Review | The automatic reviews for the procedure and elements are removed, giving the components no room for iteration | - Significantly less information in procedure steps.<br>- More elements, but less fitting.<br>- Combiner struggles to incorporate the larger number of elements. |
| Merged Components | The procedure, elements, and combiner component are merged into one new component | - Result is comparable to a procedure, meaning it misses most elements.<br>- Shorter and less detailed output<br>- Can include additional non-important sections, such as timeline and future directions. |

ity. The datasets and baselines for *ideation* and the datasets for *enriching* are completely different, likely due to these elements being less applicable for the respective research. The result sections produce the worst metrics, possibly because metrics are explained in a methodology section, and this information is thus not present. There are no significant differences in the combiner component, except when using all sections, the elements are added to new steps instead of being incorporated.

Overall, the sections mainly affect the wording, details in the procedures, and quality of the elements. Design sections result in the best experiment designs because they contain the most information with the least noise.

#### 4.3.3 Component Experiment

We list the observations of this experiment in Table 3, for example outputs see Appendix E

Removing the summarization significantly impacts the prototype. As seen in the observations, the literature provides important information and changes the overall wording and tone. However, removing the literature reduces noise and results in more general-style questions, which is favorable for the outline. The procedure is shorter and lacks details, but it is still logical and fitting. By missing details, the user possibly puts less thought into the research and can miss important parts. The fewer details, as well as the ambiguous wording, mean the model now makes more mistakes, but the user can remedy this via the conversation. The element component struggles with identifying elements. It now often selects elements from the procedure itself instead of new ones. The models do not adhere to the types needed for the research, and the metrics are either vague or fabricated. For *enriching* the assumptions now contain critiques of the choices made, but for *ideation* the assumptions are only critiques. However, the limitations are unaffected. The combiner component is unchanged.

When excluding the outline component, we lose details, and elements are less fitting, highlighting how valuable clarifying the research is.

The reviews highlight areas that can use more information in the procedures, and indicate where elements need to be added or when elements are too similar or redundant. Adding the opportunity to iterate thus increases performance.

Lastly, merging the procedure, element, and combiner into one component highlights how effective splitting up a large task is.

The prototype highlights that we can assist with experiment design, and simplifying or removing components results in worse performance. Literature provides valuable information, but can introduce noise. Clarifying the outline makes it more detailed and fitting. Reviews allow for automatic iteration. Splitting up a large task is effective and intuitive.

However, as the experiments demonstrate, minor adjustments will likely improve performance and user experience. These adjustments consist of excluding the literature in the outline component and thus providing a more general approach and focus, as well as including the user in element selection to reduce the number of elements.

## 5 Discussion

### 5.1 Threats to Validity

The internal threats to validity consist of interpreting answers in the user study and the prototype implementation. One researcher performed the interviews and analysis, increasing subjectivity. However, there was communication with the supervisors to provide feedback and insights. Multiple rounds of coding were conducted to improve the rigor of the analysis. The model parameters or system prompts might not be optimal, but we conducted multiple tests to select and tune them. The implementation and prompts in the experiment were double-checked, but there can always be errors or mistakes. The papers used for designing and evaluating the prototype were different, but limited and possibly not representative.

The external threats to validity focus on the generalizability and qualitative nature of the results. The study is formative and had limited time to be more rigorous. The user study had seven participants, which might not reflect the wider research audience. The prototype is limited to computer science, and the differences between fields are not accounted for. The evaluation of the user study and prototype is qualitative and can thus be subjective. The model used, Llama 3.1 8B, is small compared to the models used in literature. Larger models likely have different performance and possibly require different components for the design.

### 5.2 Implications for Future Work

We showcase insights into the requirements and functionalities according to our user study. But, since the study is formative, we make some recommendations and topics for future work. The first option is replicating the user study, making it more concrete by having more participants from varying fields and levels. The interview can be adapted to have a more specific focus instead of the general insights we produce. Using multiple interviewers and analysts will reduce subjectivity and increase rigor. This approach will provide more insights that can be used to design research assistants.

Our prototype design is effective and allows for assisting with the experiment design phase using a conversation. There are multiple ways to build on top of this research. Future research could focus on other less-researched steps, such as result analysis or discussion, or implement other user requirements. Other possibilities are making an alternative design for the assistant or performing more rigorous tests and evaluations on this design. As observed in our experiments, we recommend not including the literature in the outline component and including the user in the element identification. The evaluation for our prototype did not involve real users, so this is an option for future research and will provide important insights into usability and user experience. If this option is chosen, we recommend having an intuitive interface and making the goal of each step clear to the user. As well as summarizing or showing the main output of each component to make the conversation less cluttered. Different models as the backbone for the system are also interesting to look at to find out if our results are generalizable. A more powerful model can have fewer of the observed limitations and provide better reasoning and experiment designs.

## 6 Conclusion

This research provides insights into current research assistants and what users desire from them. Using a literature study and interviews with researchers, we identify that the experiment design phase is the hardest to perform and assist with. The main request from participants is factual and logical output, providing trustworthy assistance. It must also be an efficient process compared to alternatives. A conversation is a fitting approach for a research assistant, allowing for discussion and feedback, mimicking the iterative nature of research. An ideal assistant supports the entire research process, from ideation to writing, and is able to gather and use literature. Since the experiment design phase has the most potential for new insights, we develop a prototype for this with a conversation that mimics the research process from the results of the user study. The conversation allows for the exchange of knowledge and discussion. Providing summaries generated from the methodology and design sections of papers to the system enhances performance with fitting knowledge. Our prototype design, which splits the task into multiple components, is a fitting approach and performs well.

## 7 Acknowledgments

I would like to thank the participants for taking the time to participate in this study. I also want to thank my supervisors for providing assistance and feedback throughout the thesis. Finally, I thank my family and friends for their support and reviews of the text.

## References


Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Creativity and Cognition*, page 413–425. ACM.

Joshua Ashkinaze, Julia Mendelsohn, Li Qiwei, Ceren Budak, and Eric Gilbert. 2024. How ai ideas affect the creativity, diversity, and evolution of human ideas: Evidence from a large, dynamic experiment.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2025. Researchagent: Iterative research idea generation over scientific literature with large language models.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578.

Virginia Braun and Victoria Clarke. 2021. *Thematic Analysis : A Practical Guide*. SAGE Publications Ltd, London :.

Alexia Cambon, Brent Hecht, Benjamin Edelman, Donald Ngwe, Sonia Jaffe, Amy Heger, Mihaela Vorvoreanu, Sida Peng, Jake Hofman, Alex Farach, Margarita Bermejo-Cano, Eric Knudsen, James Bono, Hardik Sanghavi, Sofia Spatharioti, David Rothschild, Daniel G. Goldstein, Eirini Kalliamvakou, Peter Cihon, Mert Demirer, Michael Schwarz, and Jaime Teevan. 2023. Early llm-based tools for enterprise information workers likely provide meaningful boosts to productivity. Technical Report MSR-TR-2023-43, Microsoft.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023. Benchmarking large language models in retrieval-augmented generation.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. 2024. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery.

Martin Danler, Werner O. Hackl, Sabrina B. Neururer, and Bernhard Pfeifer. 2024. Quality and effectiveness of AI tools for students and researchers for scientific literature review and analysis. *Studies in health technology and informatics*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey.

A. Ghafarollahi and M. J. Buehler. 2024a. Protagents: Protein discovery via large language model multi-agent collaborations combining physics and machine learning.

Alireza Ghafarollahi and Markus J. Buehler. 2024b. Sciagents: Automating scientific discovery through multi-agent intelligent graph reasoning.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R D Costa, José R Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. 2025. Towards an ai co-scientist.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Va-

sic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Monika Guhan, Divyavarsini Venkatesan, and Suganthan Chandramohan. 2023. A survey on analyzing the effectiveness of ai tools among research scholars in academic writing and publishing. *International Journal Of Advance Research And Innovative Ideas In Education*.

Siyuan Guo, Cheng Deng, Ying Wen, Hechang Chen, Yi Chang, and Jun Wang. 2024. Ds-agent: Automated data science by empowering large language models with case-based reasoning.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2024. Mlagentbench: Evaluating language agents on machine learning experimentation.

Tal Ifargan, Lukas Hafner, Maor Kern, Ori Alcalay, and Roy Kishony. 2024. Autonomous llm-driven research from data to human-verifiable research papers.

Liqiang Jing, Zhehui Huang, Xiaoyang Wang, Wenlin Yao, Wenhao Yu, Kaixin Ma, Hongming Zhang, Xinya Du, and Dong Yu. 2025. Dsbench: How far are data science agents to becoming data science experts?

Hao Kang and Chenyan Xiong. 2025. Researcharena: Benchmarking large language models' ability to collect and organize information as research agents.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Ruochen Li, Teerth Patel, Qingyun Wang, and Xinya Du. 2024. Mlr-copilot: Autonomous machine learning research based on large language models agents.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, Daniel McFarland, and James Zou. 2023. Can large language models provide useful feedback on research papers? a large-scale empirical analysis.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2023. Agentbench: Evaluating llms as agents.

Kyle Lo, Joseph Chee Chang, Andrew Head, Jonathan Bragg, Amy X. Zhang, Cassidy Trier, Chloe Anastasiades, Tal August, Russell Authur, Danielle Bragg, Erin Bransom, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Yen-Sung Chen, Evie Yu-Yen Cheng, Yvonne Chou, Doug Downey, Rob Evans, Raymond Fok, Fangzhou Hu, Regan Huff, Dongyeop Kang, Tae Soo Kim, Rodney Kinney, Aniket Kittur, Hyeonsu Kang, Egor Klevak, Bailey Kuehl, Michael Langan, Matt Latzke, Jaron Lochner, Kelsey MacMillan, Eric Marsh, Tyler Murray, Aakanksha Naik, Ngoc-Uyen Nguyen, Srishti Palani, Soya Park, Caroline Paulic, Napol Rachatasumrit, Smita Rao, Paul Sayre, Zejiang Shen, Pao Siangliulue, Luca Soldaini, Huy Tran, Madeleine van Zuylen, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Marti A. Hearst, and Daniel S. Weld. 2023. The semantic reader project: Augmenting scholarly documents through ai-powered interactive reading interfaces.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodriques, and Andrew D. White. 2023. Paperqa: Retrieval-augmented generative agent for scientific research.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. 2024. Discoverybench: Towards data-driven discovery with large language models.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim,

Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Vishakh Padmakumar and He He. 2024. Does writing with language models reduce content diversity?

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback.

Linlu Qiu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, and Xiang Ren. 2024. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement.

Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. 2024. Scifibench: Benchmarking large multimodal models for scientific figure interpretation.

Tuukka Ruotsalo, Jaakko Peltonen, Manuel J. A. Eugster, Dorota Głowacka, Patrik Floréen, Petri Myllymäki, Giulio Jacucci, and Samuel Kaski. 2018. Interactive intent modeling for exploratory search. *ACM transactions on office information systems*, 36(4):1–46.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. 2025. Agent laboratory: Using llm agents as research assistants.

Yijia Shao, Yucheng Jiang, Theodore A. Kanell, Peter Xu, Omar Khattab, and Monica S. Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers.

Tim van Dam, Maliheh Izadi, and Arie van Deursen. 2023. Enriching source code with contextual data for code completion models: An empirical study.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024. Large language models for automated open-domain scientific hypotheses discovery.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models.

Shujian Zhang, Chengyue Gong, Lemeng Wu, Xingchao Liu, and Mingyuan Zhou. 2023. Automl-gpt: Automatic machine learning with gpt.

Yuxiang Zheng, Shichao Sun, Lin Qiu, Dongyu Ru, Cheng Jiayang, Xuefeng Li, Jifan Lin, Binjie Wang, Yun Luo, Renjie Pan, Yang Xu, Qingkai Min, Zizhao Zhang, Yiwen Wang, Wenjie Li, and Pengfei Liu. 2024. Openresearcher: Unleashing ai for accelerated scientific research.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. Hypothesis generation with large language models. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, page 117–139. Association for Computational Linguistics.

## A Model & Machine Setup

The model we use for the prototype is the Llama 3.1 8B Instruct model. This is the largest open source model that can run on our machine. The model parameters are set to be deterministic to allow for rerunning the experiments. The maximum token length of the output is 2048, except for the combiner step, which can reach this limit and is thus increased to 4096. The repetition penalty is only used for the summarization task and is set at 1.1 to prevent the model from infinitely repeating itself. The machine is an Nvidia A10 GPU, which has 24 GB of VRAM, which is the limiting factor for this research. The model is quantized to 4 bits, reducing the size to around 6 GB. The model has a context window of 128k tokens, However, every 3.5k tokens take up 1 GB of space on the machine, making the effective max context window 65k. We calculate that the average paper is 14k tokens using the literature used for the experiments. We limit the literature and system prompts to 40k, leaving 25k for the rest of the conversation. To better utilize this 40k, we summarize each paper, allowing for more literature in the input at the cost of possibly losing important information. Each round of the conversation is roughly 1.0k - 1.5k tokens, meaning we can have around 15 - 25 rounds. This would not be enough rounds if the entire prototype were one conversation. But, with the structure of the prototype consisting of different components, we can feed the final output of each component to the next without the rest of the conversation, allowing for adequate rounds.

## B Existing Research Assistants Study

Table 4: Overview of research assistants' capabilities and functionalities

| Name | Knowledge Origin | Personalization | Library | Research Phase | Functionalities | Transparent |
|---|---|---|---|---|---|---|
| Jenni | Papers | Steer | Save, Upload | Writing | Extracting, Selecting, Chat | Y |
| Storm | Internet | Steer | - | Ideation | Extracting, Summarizing | Y |
| Co Storm | Internet | Steer, Iterate | - | Ideation, Hypothesis | Extracting, Summarizing, Chat | Y |
| ORKG Ask | Papers | Filter | - | Ideation | Extracting, Summarizing | Y |
| Semantic Reader | - | - | Save | Ideation | Extracting | - |
| Notebook LM | Uploaded | - | Upload | Ideation, Hypothesis, Experiment Design | Extracting, Summarizing, Selecting, Chat | Y |
| Connected Papers | Papers | - | Save | Ideation | - | - |
| Typeset io | Papers | Filter, Iterate | Save, Upload | Ideation, Hypothesis | Extracting, Summarizing, Selecting, Chat | Y |
| Elicit | Papers | Filter, Iterate | Upload | Ideation, Hypothesis | Extracting, Summarizing, Selecting, Chat | Y |
| Scite | Papers | Filter | - | Ideation | Extracting, Summarizing | Y |
| Scholarcy | Uploaded | - | - | Ideation | Extracting, Summarizing | - |
| ChatGPT | Training, Uploaded | Steer, Iterate | - | Ideation, Hypothesis, Experiment Design, Writing | Extracting, Summarizing, Chat | Y [8] |

Explanation of categories in the top row of the table.

- **Knowledge Origin -** Highlights where the knowledge the assistant has come from
- **Personalization -** Personalization options for literature retrieval
- **Library -** The ability to store literature documents
- **Research Phase -** The phase the research assistant (in)directly assists with
- **Functionalities -** The main functionalities of the assistant
- **Transparent -** Highlights if the assistant shows references

Additional explanation of terms for the personalization, library, and functionalities categories.

- **Personalization**
- Steer - The ability to provide a system prompt or limited clarifications
- Iterate - The ability to chat indefinitely
- Filter - The ability to filter where information should be gathered from
- **Library**
- Save - The ability to save interesting documents to your profile for easier use later
- Upload - The ability to upload your own documents or literature manually
- **Functionalities**
- Extracting - The research assistant can extract interesting or fitting sentences from the literature directly
- Summarizing - The research assistant can summarize or paraphrase information
- Selecting - The ability to select which specific documents are used for the task
- Chat - The ability to have a conversation between the user and the research assistant that is not simply to answer more questions but have discussions or build towards a goal

# C System Prompts

## C.1 Summarization Prompt

Summarization Prompt

```
You are a research assistant whose focus lies on assisting with the process of
designing an experiment for a specific research topic.
Your current task is to summarize a related work provided by the user.
The goal of this step is to shrink down the size of the related work for later
use but retain all information.
Do note that the topic is for the research of the user and the related work is
an existing research, only the related work needs to be summarized.
The user provided a topic of the research:

{topic}

This is the related work:

{paper}

Please now make a summary of the related work.
You must preserve all important details and insights from the related work so
that the summary can be used in future steps instead of the full paper.
Focus on extracting all important information from the related work that can
later be of use for research topic.
Do not include any information of the research topic from the user in the
summary only include information from the related work.
It is important to immediately start your response with the title and make a
complete and cohesive story instead of a list or sections.
```

### C.2 Outline Prompt

Outline Prompt

```
You are a research assistant whose focus lies on assisting the user throughout
the research process.
This first step is to help refine the outline of the research by defining all
important research aspects.
You should also take an active role when needed by asking the user questions
for clarification.
Additionally you should also be critical and point out any mistakes or flaws.
These are summaries of related works which you can use as guidelines or
inspiration:

{summaries}

This is the topic provided by the user which will need to be transformed to a
full research outline:

{topic}

This is the research outline template which needs to be filled in:

Goal:
Research Questions:
Hypothesis:
Scope:

The Goal is the main objective of goal of the research.
the Research Questions are the main questions that need to be answered to
fulfill the goal.
The Hypothesis is the expected outcome of the research.
The Scope is the specification of where the research is contained in, the goal
can be broad but this defines the limits.

The outline does not need to go in depth about specific elements or future
steps in the research, the focus is to have a general outline of the current
topic and idea.
If it is not possible to fill in parts of the outline with the user provided
input you must not fill it in under any circumstance.
If you cannot fill something in you should ask questions that the user will
answer in a followup message to fill the gaps in.
Your response must always contain all outline sections and all questions should
be at the end of your response.
```

### C.3 Procedure Prompt

Procedure Prompt

```
You are a research assistant whose focus lies on assisting the user throughout
the research process.
With the help of the research outline it is your task to list all the steps
needed to perform the experiment from start till end.
It is important to not add or remove steps unless specified by the user or
automated feedback.
You will also be provided summaries of related works which can contain relevant
information or examples.

Summaries:

{summaries}

Research Outline:

{outline}

Now for this research outline please design the procedure of the experiment.
Make sure it is complete and has all the steps from start till end and follows
the defined scope of the research outline.
It is better too have too many steps since it is easier to remove steps than it
is to come up with missing steps.
The procedure must achieve the goal of the research and answer all research
questions with the help of the experiment.
Additionally keep in mind the scope of the research outline whilst designing
the procedure.
```

### Procedure Review Prompt

```
You are a critical reviewer for a research conference.
It is your job to specifically review experiment designs and provide critical
feedback.
You will be given summaries of related works, a research outline, and the
experiment procedure.
The goal of this task is to provide feedback on the experiment procedure so it
can be improved in a later step.
These are summaries of related works

{summaries}

This is the research outline that the experiment procedure aims to research:

{outline}

This is the experiment procedure on which you need to provide the feedback:

{procedure}

Please now provide a review for the experiment procedure.
The feedback must mainly focus on the steps themselves but if applicable you
can also provide feedback on the described details of each step.
All the steps together must make a complete and cohesive experiment procedure.
And the details and descriptions of each step must be fitting.
Also focus your attention to make sure everything from the research outline is
included in the experiment procedure and all research questions can be answered.
Your answer must not contain an experiment procedure but only your review.
```

### Procedure Update Prompt

```
Here is automated feedback on the experiment procedure, please take a look at
it and update the experiment procedure.
Add steps, change the order of steps, and update the details of each steps
according to the feedback.
Make sure to respond with only the experiment procedure again for your next
response.
The feedback:

{feedback}
```

### C.4 Elements Prompt

The *{element}* is one of the following: datasets, models, metrics, baselines, assumptions, limitations.

Element Prompt

```
You are a research assistant whose focus lies on assisting the user throughout
the research process.
It is your task to identify what specific {element} this research will use in
the experiment procedure.
You will be provided summaries of related work, the research outline, and the
experiment procedure for which you need to identify the {element}.
If it is already mentioned in the research outline use that, otherwise it is
your task to select the specific {element}.

Summaries:

{summaries}

Research Outline:

{outline}

Experiment Procedure:

{procedure}

Now list the {element} that this experiment procedure will use.
Make sure the {element} are reasonable and feasible and not too many.
Focus on fitting and specific {element} for the experiment procedure that are
logical to the research outline.
Make sure you output the specific {element} and not simply the criteria it
needs to adhere to.
```

Element Review Prompt

```
You are a critical reviewer for a research conference.
It is your job to specifically review experiment designs and provide critical
feedback.
You will be given summaries of related works, a research outline, and the
experiment procedure.
The goal of this task is to provide feedback on the selected {element} and
evaluate if they fit in the experiment procedure.
These are summaries of related works

{summaries}

This is the research outline that the experiment procedure aims to research:

{outline}

Experiment Procedure:

{procedure}

This {element} on which you need to provide the feedback:

{generated element}

Please now provide a review for on the selected {element} for in the experiment
procedure.
The {element} must be logical and fit in the experiment procedure as well as
link to the research outline.
Your answer must only contain your review.
```

Element Update Prompt

```
Here is automated feedback on selected {element}, please take a look at it and
update the {element}.
Make sure to respond with only the updated {element} again for your next
response and not mention the changes you made.
The feedback:

{feedback}
```

### C.5 Combiner Prompt

Combiner Prompt 1

```
You are a research assistant whose focus lies on assisting the user throughout
the research process.
This is the last step in finalising the experiment design.
You will be provided summaries of related work, the research outline, the
experiment procedure, and the selected elements.
The selected elements consist of datasets, models, metrics, and baselines.
The goal of this step is to add all selected elements to the experiment
procedure.

Summaries:

{summaries}

Research Outline:

{outline}

Experiment Procedure:

{procedure}

Selected Elements (datasets, models, metrics, baselines):

{datasets}

{models}

{metrics}

{baselines}

Now copy all selected elements inside the appropriate steps of the experiment
procedure.
Include all parts of the selected elements into the experiment procedure, do
not leave anything out.
Do not spread out the parts of one element over too many steps, everything from
one element should be included in the same step of the experiment procedure.
Make sure to only output the steps of the updated experiment procedure.
```

Combiner Prompt 2

```
Now that you added the elements you must also shorten and add these 2 sections
below the experiment procedure.

{assumptions}

{limitations}

Respond with the experiment procedure and the added 2 sections.
```

## D Section Experiment Output Examples (Ideation

Procedure Example Normalization Step

**Design Sections**

```
...
**Step 4: Normalization of Ideas**

* Normalize all research ideas generated by human participants and LLMs to
  ensure that they are comparable and indistinguishable.
* Remove any identifying information, such as participant IDs or model names,
  from the ideas.
* Use a standardization process to ensure that all ideas are in the same format
  and have the same structure.
* This will ensure that when reviewing later, it is not possible to
  differentiate between human and model ideas.
...
```

**Result Sections**

```
...
**Step 4: Normalization of Ideas**

* All research ideas generated by participants and LLMs will be normalized to
  ensure that they are comparable and indistinguishable.
* Normalization will involve removing any identifying information, such as
  author names or model names, and converting the ideas into a standardized
  format.
* The normalized ideas will be stored in a separate dataset for further
  analysis.
...
```

**Full Papers**

```
...
**Step 4: Normalization of Ideas**

* Normalize all research ideas generated by participants and the LLM to ensure
  that they are presented in a consistent format.
* Remove any identifying information, such as the participant's name or the
  model's name, from the ideas.
* Use a standardized format for presenting the ideas, such as a specific
  font or layout.
* This will ensure that when reviewing the ideas later, it is not possible to
  differentiate between human and model ideas.
...
```

Procedure Example Data Analysis Step

**Design Sections**

```
...
**Step 6: Data Analysis**

* Analyze the data collected from the study, including the quality, relevance,
  and potential biases of research ideas generated by human participants and
  LLMs.
* Analyze the reviews from the human reviewers, including the frequency and
  types of feedback provided.
* Evaluate the effectiveness of human reviewers in evaluating research ideas
  generated by LLMs and human participants.
* Compare the performance of human participants and LLMs in generating research
  ideas.
* Evaluate the quality, relevance, and potential biases of research ideas
  generated by human participants and LLMs.
* Identify any differences or similarities between the two groups.
* Provide detailed description of the statistical methods used and the expected
  outcomes.
...
```

**Result Sections**

```
...
**Step 6: Data Analysis and Statistical Methods**

* The data collected from Steps 2-5 will be analyzed using statistical methods,
  including correlation and regression analysis.
* The analysis will investigate the relationships between the variables,
  including the comparison of LLM-generated research ideas to human-generated
  research ideas.
* The expected outcomes of the analysis will be specified, including the
  comparison of the LLM-generated research ideas to the human-generated
  research ideas.
...
```

**Full Papers**

```
...
**Step 6: Data Analysis**

* Collect and analyze the data from the research idea generation task and
  the review task.
* Compare the research ideas generated by participants and the LLM to determine
  if there are any differences in terms of quality, accuracy, and usefulness.
* Evaluate the research ideas generated by participants and the LLM using a set
  of predefined evaluation criteria.
* Use statistical methods to analyze the data and determine if there are any
  significant differences between the research ideas generated by participants
  and the LLM.
...
```

Element Example Metrics 1/2

**Design Sections**

```
**Metrics**

1. **Quality of Research Ideas**: Measured by the number of research ideas that
   are relevant, well-structured, and address a clear research question.
2. **Relevance of Research Ideas**: Measured by the number of research ideas
   that are relevant to the field of natural language processing and address a
   current research gap.
3. **Novelty of Research Ideas**: Measured by the number of research ideas that
   are novel and address a current research gap.
4. **Impact of Research Ideas**: Measured by the potential impact of the
   research ideas generated by human participants and large language models,
   including their potential to contribute to the field of natural language
   processing.
5. **Inter-rater Agreement**: Measured by the agreement between human reviewers
   and the large language model in evaluating research ideas, including the
   frequency and types of feedback provided.
```

**Result Sections**

```
**Metrics**

**Creativity**: A measure of how new and unique the research ideas are,
   considering their originality and novelty.
2. **Relevance**: A measure of how well the research ideas align with the field
   of NLP and the research questions being addressed.
3. **Contribution**: A measure of the potential of the research ideas to
   contribute to the field of NLP.
4. **Clarity**: A measure of how clear and well-structured the research ideas
   are.
5. **Coherence**: A measure of how well the research ideas are organized and
   connected.
6. **Diversity**: A measure of the variety of topics and approaches being
   addressed.
7. **Relationship strength**: A measure of the strength of the relationships
   between the research ideas generated by participants and LLMs.
8. **Quality**: A measure of the overall quality of the research ideas,
   considering their clarity, coherence, and overall presentation.
```

Element Example Metrics 2/2

**Full Papers**

**Metrics**

1. **Research Idea Quality**: Measured by the number of well-defined research questions, the clarity of the research objectives, and the relevance of the research to the field of NLP.
2. **Originality of Research Ideas**: Measured by the number of novel and innovative research ideas generated by the LLM.
3. **Usefulness of Research Ideas**: Measured by the potential impact of the research ideas on the field of NLP, the feasibility of the research, and the potential for the research to address real-world problems.
4. **Research Idea Novelty**: Measured by the number of novel and innovative research ideas generated by the LLM.
5. **Research Idea Relevance**: Measured by the relevance of the research ideas to the field of NLP and the potential impact of the research on the field.
6. **Impact of Research Ideas**: Measured by the potential impact of the research ideas on the field of NLP, the potential for the research to address real-world problems, and the potential for the research to lead to new discoveries or innovations.
7. **LLM's Ability to Identify Research Gaps**: Measured by the LLM's ability to identify gaps in existing research and generate research ideas that address these gaps.
8. **LLM's Ability to Generate Research Ideas that Address Real-World Problems**: Measured by the LLM's ability to generate research ideas that address real-world problems and have the potential to lead to new discoveries or innovations.
9. **Agreement between Human and LLM Reviews**: Measured by the percentage of reviews that are identical or similar between human and LLM reviews.
10. **Quality of Reviews**: Measured by the accuracy, completeness, and usefulness of the reviews generated by the LLM.

## E Component Experiment Output Examples (Ideation)

Summarization Component Experiment - Example Element Models

**Complete Prototype Design**

** Models **

1. **GPT-4**: A state-of-the-art LLM that will be used to generate research ideas.
2. **LLaMA-3.1 405b**: A large language model that will be used to generate research ideas.
3. **Text-embedding-ada-002**: A sentence embedding model used in the CoQuest system.
4. **Dense Passage Retriever (DPR)**: A bi-encoder architecture used in the Retrieval-Augmented Generation (RAG) model.
5. **BART**: A generator used in the RAG model to condition on the retrieved documents and the input to generate the output.

**Summarization Component Removed**

** Models **

1. **BERT (Bidirectional Encoder Representations from Transformers)**: A large language model that will be fine-tuned for research idea generation. BERT will be used to generate research ideas in the field of natural language processing.
2. **Human Experts**: Participants who are experts in the field of natural language processing will be used to generate research ideas, review generated ideas, and provide feedback on the quality and relevance of the ideas.
3. **Random Forest Model**: A machine learning model that will be used to analyze the review data from human participants and to identify patterns and correlations in the data.
4. **Support Vector Machine (SVM)**: A machine learning model that will be used to classify the generated research ideas as relevant or irrelevant, and to identify the features that distinguish between the two classes.

Outline Component Experiment - Example Procedure Idea Evaluation Step

**Complete Prototype Design**

```
...
**Step 5: Human Reviewer Evaluation**

* A subset of human reviewers will be recruited to evaluate the research ideas
  generated by LLMs and human participants.
* Human reviewers will be provided with the same evaluation criteria used by
  participants.
* Human reviewers will be asked to review and provide feedback on the research
  ideas generated by both LLMs and human participants, including suggestions
  for improvement and potential areas of concern.
* Provide detailed description of the human reviewers' demographics and their
  experience in evaluating research ideas.
...
```

**Outline Component Removed**

```
...
**Step 5: Idea Evaluation**

* Define the evaluation criteria for the research ideas, including relevance,
  novelty, and feasibility.
* Ensure that the evaluation criteria are comprehensive and capture the scope
  of the research topic.
* Have the expert reviewers evaluate both the human-generated and agent-
  generated research ideas using the evaluation criteria.
...
```

Review Component Experiment - Example Procedure Data Analysys Step

**Complete Prototype Design**

```
...
**Step 6: Data Analysis**

* Analyze the data collected from the study, including the quality, relevance,
  and potential biases of research ideas generated by human participants and
  LLMs.
* Analyze the reviews from the human reviewers, including the frequency and
  types of feedback provided.
* Evaluate the effectiveness of human reviewers in evaluating research ideas
  generated by LLMs and human participants.
* Compare the performance of human participants and LLMs in generating research
  ideas.
* Evaluate the quality, relevance, and potential biases of research ideas
  generated by human participants and LLMs.
* Identify any differences or similarities between the two groups.
* Provide detailed description of the statistical methods used and the expected
  outcomes.
...
```

**Review Component Removed**

```
...
**Step 6: Data Analysis**

* Analyze the data collected from the study, including the quality, relevance,
  and potential biases of research ideas generated by human participants and
  LLMs.
* Compare the performance of human participants and LLMs in generating research
  ideas.
* Evaluate the effectiveness of human reviewers in evaluating research ideas
  generated by LLMs.
...
```

Combined Components Experiment - Example Procedure Writing Step

**Complete Prototype Design**

```
...
**Step 8: Writing the Research Paper**

* Write a comprehensive research paper that summarizes the main findings of the
  study.
* Discuss the limitations and potential biases of the study.
* Provide recommendations for future research in this area.
* Summarize the implications of the study's findings for the use of LLMs in
  research ideas generation.
* Discuss the potential benefits and drawbacks of using LLMs for research ideas
  generation.
* Provide detailed description of the potential applications and implications
  of the study's findings.
* Ensure that the results are accessible and understandable to a broad audience.
* Provide detailed description of the dissemination plan.
* Share the findings with stakeholders, including researchers, policymakers,
  and industry professionals.
* Publish the research paper in a reputable academic journal.
```

**Procedure, Element, Combiner Components Merged**

```
...
**Step 10: Writing the Research Paper**

* Write a research paper that summarizes the findings of the study, including
  the results of the comparison between human and LLM-generated research ideas,
  the analysis of the quality, relevance, and potential biases of the research
  ideas generated by human participants and the LLM, and the results of the LLM
  review of the research ideas.
* Discuss the implications of the study for the use of large language models in
  research ideas generation and the potential benefits and drawbacks of using
  LLMs in this context.
* Provide a clear and concise summary of the study's findings and
  recommendations for future research.
```