Better Results Through Ambiguity Resolution: Large Language Models that Ask Clarifying Questions

A DISSERTATION PROPOSAL SUBMITTED TO THE GRADUATE DIVISION OF THE DEPARTMENT OF INFORMATION AND COMPUTER SCIENCE AT THE UNIVERSITY OF HAWAIʻI AT MĀNOA

November 6th, 2023

By

Bernadette J. Tix

Keywords: Natural Language Processing, Large Language Models, Generative AI, Human-Computer Interaction, Ambiguity, Questions

CONTENTS

ABBREVIATIONS

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers (an LLM by Google AI)
GPT	Generative Pretrained Transformer (an LLM by OpenAI in partnership with Microsoft)
HCI	Human Computer Interaction
LLM	Large Language Model
LSTM	Long Short-Term Memory (A type of Neural Network)
NER	Named Entity Recognition
NLP	Natural Language Processing
QA	Question-Answering software
RNN	Recurrent Neural Network

1 Abstract

Ambiguity has historically been a major problem in Natural Language Processing (NLP) and continues to present major obstacles even for modern systems. Large Language Models (LLMs) generate output by calculating a “most likely" response to any given prompt. However, prompts are often ambiguous, and even the best possible prediction cannot fully resolve the problem of underspecified prompts [11]. Even in a conversation between two humans, both speaking the same language and communicating clearly, misunderstandings are common, as is the use of ambiguous language. While there are many ways to resolve ambiguity in a human conversation, perhaps the most obvious way is to simply ask for clarification. However, commonly used LLM systems such as ChatGPT, Bard, and Bing do not ask clarifying questions in response to ambiguous prompts.[1]

This is a serious problem in fields where precision is important. Detailed discussions, including follow-up questions, are a necessary part of human communication when precision is required. Even for simple requests, a lack of follow-up questions can lead to suboptimal answers, or even cause the LLM to misunderstand the true needs of the user. In some applications, such as LLMs being used as search tools, it may be acceptable for the user to enter increasingly refined prompts to improve the answer they get from the LLM. However, the user may not know what they need to change about the prompt to get the information they need, or, worse, the user may not realize the system has misunderstood their needs. If the user cannot distinguish between correct and incorrect output and is relying on the system to be correct, an ambiguous prompt may lead to the user relying on misleading output. Even in situations where users are knowledgeable and free to refine their prompts as needed, a smoother experience could likely be provided if the system itself recognized points of ambiguity and asked clarifying questions. Appendix A shows an example of ChatGPT providing an answer to a question that could have multiple possible meanings. However, the answer assumes one meaning and proceeds without asking for clarification or pointing out the ambiguity. GPT-4 does only slightly better in this regard, as shown in Appendix C.

This is also a problem for AI Alignment. AI Agents acting to solve under-specified problems can lead to severe unintended side effects [5,11]. Under-specification occurs when an AI system is given a goal to accomplish but the designer of the system has unstated assumptions of how the goal ought to be accomplished that are not communicated to the system. As an example, one of the early experiments by OpenAI trained an artificial intelligence to play a boat-racing video game. The agent discovered a strategy in which a simulated boat under the agent’s control remained in a small area of the map, never completed the course, ran into other boats, and repeatedly caught fire. However, due to the scoring system of the game it was still able to achieve more points than most humans could by running the course normally [5,23]. It was given the goal of maximizing points, with no incentive to actually race the course when other

[1] See Appendix A and Appendix C for examples of ChatGPT failing to ask a clarifying question where one is needed.

means of acquiring points were available. Although it did achieve a high number of points, this was clearly not the intended behavior.

While question-asking will not completely eliminate this problem, it can help to alleviate it. An ideal system, linked to an LLM, might have asked “do you want me to race with the other boats, or is maximizing points the only goal?” Since it seems likely that LLMs will be increasingly used to write code, guide human decisions, and in some cases even operate completely autonomously, it is vital that ambiguity and under-specification be identified early, and this goal could be substantially aided by the AI itself identifying ambiguity and asking clarifying questions. Appendix B shows a hypothetical example of an engineer working with a client to meet their needs. In one instance, no follow-up questions are asked. In the second instance, a follow-up question leads to a change in the specifications. This type of dialog is something that LLM systems are currently not set up to do but could substantially improve their ability to produce outputs that conform to human needs.

LLMs are currently capable of identifying ambiguity in user prompts and forming questions in response to ambiguity when prompted to do so.[2] This has already been tested with AIs answering simple ambiguous questions [28,48], but has not yet been demonstrated with AIs intended to generate longer-form responses such as letters or documents. I propose that an LLM-based system that asks clarifying questions when needed will, on average, produce content that is more closely aligned to the desires of the human user than a comparable system which asks no clarifying questions.

2 Prior Works

2.1 Background

Ambiguity has been a major difficulty from the beginning for nearly all forms of NLP, including grammar [18,49,50,66], Named Entity Recognition (NER) [10,27,29,55,60], story understanding [9,16,41,61], and numerous other NLP tasks. In the past decade, many different types of neural models have been employed in an attempt to create software that can exhibit reading comprehension-like behaviors on ambiguous, natural language text. After early successes with Recurrent Neural Networks (RNN) [12,37], Long Short-Term Memory Networks (LSTM) took the lead with specialist systems dedicated to specific NLP tasks [1,2,8,9,24].

Language Models (LM) offer a more general solution than the highly trained LSTM systems. By predicting the probability of any given sequence of words, LMs can be used to generate text in response to a wide variety of prompts. This offers a partial solution to the problem of ambiguity, since they can be used to predict the next word in a sequence, considering the context of the words that came before [59]. The most basic language models are the simple *n*-gram models, which model the likelihood of any given

[2] Appendix D demonstrates this capability.

sequence of *n* words based on a large body of training text. However, the *n* value for *n*-gram models is usually very small and is inherently limited since the amount of training data required to accurately model the probability of any given *n*-gram scales exponentially with *n* [59]. LMs have existed in some form since the earliest days of NLP, using various techniques beyond simple *n*-grams, and have improved steadily over time [24]. The effectiveness of LMs has improved dramatically in the past several years, starting with the invention of the Self-Attention Mechanism [34] which led directly to the landmark paper *Attention is All You Need* by Vaswani et al. in 2017 [67]. *Attention is All You Need* introduced the Transformer architecture, a neural architecture built around a multi-headed self-attention mechanism, which could be efficiently trained on vast data sets, and could produce long strings of output based on large context windows. This represented a huge leap forward for language models, and the creation of the first true Large Language Models (LLM) which, using the Transformer architecture, could now be trained on far larger training sets than before. Over the next several years LLMs improved dramatically with each new iteration, trained on ever larger data sets [7,43,75]. Language models quickly delivered state-of-the-art performances in multiple areas of NLP, matching and in many cases outperforming specialized LSTM-based systems in numerous NLP benchmarks [3,52,53] including answering questions about children's stories [21], common-sense reasoning [32], reading comprehension [56], translation, and summarization [53], among others.

In November 2022, ChatGPT was released by the company OpenAI [76]. Based on the GPT-3.5 platform also developed by OpenAI, ChatGPT made powerful LLMs available to the general public, for free, for the first time. This generated enormous public interest in the new technology, led to competing LLMs by Google [19,47,78,79] and Meta [36,83], as well as public discourse on the future of AI and its impact on society, especially education [4,22,80], employment [38,39,70], and art [81,82]. This society-wide interest in LLMs has led to a flurry of new research among computer scientists eager to both test and push the limits of the new technology.

Despite their incredible successes, LLMs have also been shown to have a number of significant weaknesses. One ongoing weakness is bias, including both partisan political bias [20,25,35,77] as well as a general social bias towards the current status-quo which emerges as a result of training data which comes from the internet writ large [3,26,44,74]. Another issue is so-called "hallucinations," in which the LLM will produce factually incorrect text that sounds convincing if the reader does not know ahead of time that the content is incorrect [43,72]. LLMs can be substantially improved by the introduction of "chain-of-thought reasoning" [69], in which the LLM is prompted to write out a full logical argument for its conclusion in small steps, rather than skipping straight ahead to the final conclusion. This seemingly simple change leads to fewer hallucinations, more factually correct responses, more advanced reasoning, and improved ability to solve puzzles or trick questions. LLM systems operating on a chain-of-thought model also have the potential to explain the reasoning that led them to a given conclusion, which is considered to be a desirable trait in both logical and ethical reasoning, and a necessary step for humans to trust the results of an analysis [30,31].

2.2 Context, Ambiguity, and User Needs

Understanding context is extremely important to accurate reasoning and communication. For example, Cyc is a symbolic reasoning engine that relies on an enormous database of hand-entered facts to perform formal logical reasoning [13,84]. Even operating in the realm of formal logic, Cyc utilizes a multi-layered approach to understanding the context of a problem in order to prioritize some logical arguments over others [30,31,63].

When precision is needed in human communication, a wide variety of methods are used to clarify what would otherwise be inherently ambiguous language. For example, when gathering requirements for a new software, a high degree of precision is needed, usually far more than is initially provided, which is the motive for the phase of 'requirements gathering' within software engineering. Requirements gathering has been researched at length, and often employs a wide variety of techniques including questionnaires, face-to-face dialogs between customer and developer, and various exercises designed to improve user engagement with the requirements gathering process [40,45]. None of this would be necessary if software engineers could reliably get good results from users by simply asking "please state your requirements clearly!" One of the key goals of requirements gathering is to understand the context of the desired software. For example, what problem the software is being requested to solve, and what specific change or improvement it is hoped the software will achieve.

Within human communication, context and ambiguity are inextricably linked. For example, in the paper "TruthfulQA: Measuring How Models Mimic Human Falsehoods" by Lin et al. 2022 [33], the authors ask GPT-3 "What happens if you smash a mirror?" GPT-3 responds that you will receive seven years of bad luck, which the authors note as a common saying but also a false answer, labelling the response as "informative, but incorrect." TruthfulQA is a database of questions with common "false" answers, considering an answer to be correct only if it is literally true in the real world. They found that larger models were more likely to give informative but false answers, but that performance against the TruthfulQA metric improved if GPT-3 was specifically prompted to be truthful. In the absence of context, this seems to me to be somewhat of an unfair test for a system designed to mimic human speech and reasoning. For example, if I were asked "what happens if you smash a mirror" I would very likely answer "seven years bad luck," not because I am superstitious or misinformed, but because I would assume the answer "broken glass falls on the floor and may be dangerous" would be obvious to the person asking, and describing this outcome would likely come across as condescending or out-of-touch. I would assume that most people asking are curious about the superstition, and couldn't remember the specific negative outcome, or wanted to test whether I knew it, rather than being confused about the literal outcome of smashing a mirror. However, this is highly contextual. If I was asked the same question by a child who seemed worried that they were about to have bad luck, I might reassure them that the superstition is untrue. If I were asked the same question by a physics or chemistry teacher, I might start thinking about whether mirrors had any unique physical properties that might lead to an interesting or unusual conclusion. In the absence of context, “seven years bad luck” is not an obviously incorrect answer. If context is lacking, a superior response may be to try to gain context by asking questions or analyzing the situation. For example, if asked “what happens when you smash a mirror?” with no prior context, one could respond with a clarifying question such as “are you asking about the common superstition, or about what will actually happen physically?” Prior research in conversational interfaces has shown that better results can be achieved when the full context of a conversation is considered, not just the immediate prompt [65].

Systems such as ChatGPT and other LLMs are in a disadvantaged position with regards to context as they are currently deployed. Users can enter any prompt, on any topic, and the LLM must provide a response, whether as a chatbot or as the returned value from an API call, with no knowledge of who the user is or why they are asking. Appendix A shows an example of ChatGPT responding to the ambiguous question “what is a transformer?” Even though ChatGPT can identify the ambiguity inherent in the question, it confidently describes the transformer architecture as described by Vaswani et al. in 2017 [67]. I use the same prompt as a trick question in appendix D, presenting a scenario where the user intends to ask about the Transformers toy franchise, but asks the question in an ambiguous way. Similar to the question “what happens when you smash a mirror?” I consider “what is a transformer?” to be fundamentally uninterpretable without further context. If I was asked “what is a transformer?” by a child who had previously asked about X-Men and Power Rangers, I might intuit that they were referring to the toy and movie franchise. I would give a much different answer if called upon with the same question in a computer science classroom. If I was suddenly asked “what is a transformer?” by a stranger on the street, I would have to ask clarifying questions before answering with any confidence. It is unreasonable to assume that an LLM could interpret the user’s intended meaning in the absence of context when this is not possible even in communication between humans, using inherently ambiguous and context-dependent language. It is reasonable to believe that asking clarifying questions is a fundamental skill that LLMs must master if they are to communicate clearly and precisely with humans.

2.3 Similar Prior Work

Several recent works have already addressed the concept of LLMs using clarifying questions. The CLAM architecture [28] presents a method for using the LLM itself to assess ambiguity, generate a clarifying question if needed, and then generate an answer based on the user’s response to the question. CLARA [48] showed that a similar framework could be used to interpret user commands given to a robotic arm. ClarifyDelphi [51] uses clarifying questions to assist in context-sensitive ethical reasoning. Zhang et al. 2023 present a framework for asking clarifying questions before retrieving data from a database [72].

Each of these works is very recent (since May 2023) and further research is still needed. These existing studies have two key limitations that I intend to address in my research:

1. Prior works into clarifying questions from LLMs use simple questions where answers given by the LLM are easy to classify as objectively correct or incorrect. While convenient to study, this scenario is dissimilar from the content-creation tasks that are the standout strength of LLMs, which may involve a large set of vague and sometimes conflicting requirements (ex: that output should be both precise and brief).
2. With the exception of CLARA, the validating experiments in each of these studies relied on using "simulated humans." The simulated humans used in the studies were LLMs given specific knowledge of the hypothetical users' true intentions. Validation was done to show that simulated humans answered similarly to real humans, and this is a valuable tool for quickly generating large amounts of data. However, the results of these studies could be strengthened by showing similar results with actual human testers. Testing with real humans could also give valuable secondary data, such as whether the testers found the system engaging or enjoyable to work with, which is not possible to learn from simulated humans.

2.4 Existing Benchmarks and Evaluation Methods

There are many benchmarks currently in use for the evaluation of LLMs. However, most of the common benchmarks, such as the older BLEU benchmark [46] as well as newer benchmarks such as BERTScore [73] only measure the overall quality of the generated text. They do not measure how well the output corresponded to the initial prompt or to a user-desired outcome. Other benchmarks test the LLMs ability to get the correct answer to questions with previously established correct answers, including numerous question-answer datasets [57]. Some QA datasets target specific types of questions, including CoQA for Conversational Question Answering [56], TruthfulQA for misleading questions [33], and the Children's Book Test for reading comprehension of short stories [21]. These styles of benchmark are poorly suited to determining whether a generated content has fulfilled a user's needs. Measuring the overall quality of the text, as BLEU and BERTScore do, does not tell us whether the high-quality text has solved the user's problem or merely provided elegant but irrelevant prose. QA datasets are only suitable for measuring the LLM's ability to produce short, accurate responses to questions with objectively right and wrong answers. This is not suitable to the evaluation of longer-form content. A letter, essay, or short story cannot be objectively classified as "correct" or "incorrect" above a certain level of relevance and accuracy. The overall quality of such a document can only be measured subjectively, by the evaluation of the reader.[3]

Validation of generative models for visual art and music may offer some guidance here. As with long-form textual content, the quality of visual art and music cannot generally be objectively evaluated.

---

[3] In some cases, an objective measure may be possible for documents with a purpose, such as whether a generated resume resulted in an interview in a job application. However, this is only offsetting the problem of subjective analysis. Here the objective measure (ratio of callbacks from a given resume) is derived from a subjective human analysis, that of the hiring manager. Unless, of course, the resume is evaluated entirely by a software system, but that scenario is outside the scope of this research.

Furthermore, such systems are most often employed in the task of generating content (art or music) from a short textual prompt, and quality of these systems must be evaluated on how closely the output subjectively matches the intent behind the prompt given to the model. Despite the challenges associated with subjective analysis, including higher costs and challenges with methodology and sample size, subjective analysis by human evaluators is often the only way to gain reliable feedback on the quality of output from creative systems [68,71]. For instance, the experiments which validated the quality of DALL-E had human evaluators rate images for both realism and accuracy to each image's corresponding prompt [54].

3 Proposed Research

I propose new research which will address the gaps identified in section 3.3. This work will aim to determine whether an LLM-based software that asks clarifying follow-up questions produces better results in creative content generation than one that does not ask questions.

I define creative content generation as any task which meets the following criteria:

1. A human user provides a 1-2 sentence prompt describing their need for a document, which the AI will then generate.
2. The desired output is significantly more complex than the input prompt. For example, an essay or a letter.
3. There is no singular "correct" output, but the quality of the output may vary.
4. The primary goal of the generation is to satisfy the needs or desires of the human user who provided the prompt.
5. The user can discern a higher quality output from a lower quality output.
6. Quality is chiefly a measure of user preference.

In order to address a further gap in the current literature, this research will be demonstrated with real human users and will not rely solely on "simulated humans."

4 Methodology

4.1 Experiment Design

Since the proposed system is reliant on user interaction in the form of question asking and answering, the use of readily available large static databases of question-answer sets is insufficient to test this design. Direct interaction between the system and a large number of users will be necessary. Users will be directed to a webapp which will guide them through the experiment. Full details of the web app text and prompts are available in Appendix E.

Step 1: Explanation and consent

The webapp will begin by showing the user an explanation of the experiment they are about to participate in, and then asking for the user's consent to participate in the study, with a full explanation of what data will be collected and how it will be used.

Step 2: Demographic Questions

The user will be shown demographic questions which will allow us to look for statistical differences between groups. The demographic question that will be asked are:

- Age
    - [Numerical Input]
- Gender
    - "Female"
    - "Male"
    - "Other / Nonbinary"
- "What is your prior experience with generative AI such as ChatGPT, Bard, or similar programs?"
    - "I use generative AI regularly."
    - "I have used generative AI before, but not often."
    - "I have never used generative AI before."
- "Is English your primary spoken language?"
    - "Yes"
    - "No"

Step 3: Instructions

Users will be shown the following instructions:

"Think of a writing task you would like the AI to help you produce. This can be a document you actually need (you will have the opportunity to keep the output) or something you only think up for the sake of the experiment. Either way, please think in detail about what you want the AI to write for you before

proceeding to the next step. When you have a clear idea of what you want to ask the AI to write, enter a 1-sentence or 2-sentence prompt in the textbox below, asking the AI to write your document for you. The AI will ask you a series of questions, and you will then be given two versions of the document you requested, and asked for feedback on which version you prefer."

A text-entry area will be provided for the user to enter their prompt. A character limit of 400 characters will be imposed, since the idea of this experiment is to gather information about question-asking in response to vague or abstract prompts, so overly complex or detailed prompts are not desirable for this experiment.

Step 4: Follow-up Questions

After the user enters their initial prompt, the webapp will present the user with a chatbot-style interface. The AI will generate and ask between one and three clarifying follow-up questions relevant to the user's original request. The user will be able to respond to questions conversationally through the chatbot interface.

Step 5: Document Output

After all questions have been answered, the AI will generate two versions of the requested document. One version will use only the user's original prompt to generate the document. The other will use the follow-up questions and responses in addition to the original prompt. The outputs will be presented to the user in a randomly selected order to avoid biases that could arise from the order of presentation. The user will be asked to rate each output according to three metrics.

- How close is this document to what you hoped for when you made your initial request?
    - Very close to what I was hoping for.
    - Somewhat close to what I was hoping for.
    - A little bit like what I was hoping for.
    - Not very close to what I was hoping for.
    - Not at all what I wanted.
- How useful would this document be to you?
    - I could use this document as-is.
    - I could use this document with minimal modification.
    - I could use this document with substantial modification.
    - This document could be used as a general starting point but requires major revisions to be usable.
    - This document is not usable at all.
- How would you rate the overall quality of this document?
    - Excellent quality.
    - Above average quality.
    - Average quality.

- Below average quality.
- Poor quality.

Step 6: Optional Continuation and Exit Questionnaire

After selecting their preferred output, the user will be given the option to either ask another prompt or proceed to the exit questionnaire. Users will only be required to complete one prompt and document selection, but will be given the option to do as many as they would like if they are interested in continuing to engage with the system. Once the user indicates that they have no more prompts they would like to try, they will be given an exit questionnaire with the following questions:

- *Please rate the following statements on a scale of "Strongly Agree" to "Strongly Disagree"* (Each of the following statements will be shown with 5 options: Strongly Agree, Slightly Agree, Neutral, Sightly Disagree, Strongly Disagree)
  - *It was annoying to have to answer questions even though I had already explained what I wanted the AI to do.*
  - *I felt like the AI was more engaged with my problem because it asked follow-up questions.*
  - *I would be willing to answer follow-up questions from an AI if answering questions led to better results.*
  - *I liked that the AI showed me two options to pick between, instead of only picking the option it thought was best.*
- *Do you have any additional feedback or comments (optional)?*
  - A free-text entry will be provided.

4.2 Results Analysis

The primary metric of interest is whether users are likely to select the output that takes their question-responses into account as preferred over the output that doesn't. A 2-sample T-test is the appropriate metric for measuring whether the output with follow-up questions was significantly preferred over the baseline [14,85]. The three metrics used to grade each of the output samples can be compared individually, and can also be compared as an aggregate ***preference score*** calculated as the sum of responses to all three metrics. This will allow us to demonstrate both whether users preferred one type of output over the other overall, as well as whether they found a difference in any one of the metrics of quality, usefulness, and how closely the output matched their initial request.

The entry and exit questions will allow for several interesting secondary analyses as well. The exit questionnaire will allow us to gauge whether users find the question-asking system to be pleasant to use, or whether it is something most users would rather avoid. The demographic questionnaire will also allow analysis of whether age, gender, experience level, or English language status have an impact on users' preference for output that included follow-up questions, or on their experience of the question-asking process. A multi-linear regression can be used to show whether age, gender, or prior experience with AI had an effect on the user's preference for the question-asking system.

Finally, the free-entry feedback and comments section may allow for a qualitative analysis of users' experience with the system if enough people provide feedback. The free-entry feedback responses each be tagged with common themes that occur in multiple responses. It will then be possible to report on the most common issues that were addressed in the user feedback.

4.3 Scale

The experiment will begin with an initial pilot study of 5-10 people. Users participating in the pilot study will be asked to complete the experiment over a Zoom call while sharing their screen. I will be on hand to assist and take notes. The primary goal of the pilot is to identify any parts of the experiment that users find confusing or get stuck on, so that the experiment can be refined to run more smoothly after the pilot is complete. Once the system has been refined after the pilot, the primary study can be conducted.

For the Two-Sample T-Test, the required sample size can be calculated with the following formula [14]:

$$\textit{Required Sample Size} = (z_{\alpha/2} + z_{\beta})^2 \, (\sigma/\mu)_2$$

Where:

α = error

β = power

σ = Standard deviation within sample (estimated until data is gathered)

μ = Smallest difference of interest in effect size

For this experiment, I use the following values:

$\alpha = 0.05 \rightarrow z_{\alpha/2} = 1.96$

$\beta = 0.2 \rightarrow z_{\beta} = 0.8416$

$\sigma = 1.5$[4]

$\mu = 0.5$[5]

[4] This is 1.5 on a 5-point scale. I expect user responses to range fairly widely.
[5] Also on a 5-point scale

Required Sample Size $= (z_{\alpha/2} + z_\beta)^2 (\sigma/\mu)_2 = 70.64$

Thus, I hope to get 70 volunteers to run through the experiment after the completion of the pilot. Partial data can be assessed as it comes in. If the effect size or standard deviation in responses are larger or smaller than anticipated, the sample size may need to be adjusted up or down.

4.4 System Architecture

The proposed system will include several components:

- A user-facing front-end.
- A back-end powered by OpenAI's GPT API.
- A database for logging results from the use of the system.

The Front End

The system front end will be a webapp which will walk the users through a multi-step process including instructions, questionnaires, and a chatbot. The user front-end is described in depth in section 5.1.

The Back End

The AI system which asks follow-up questions and provides results to the user will be designed as an interface between the user and the OpenAI API. The system will apply specific prompt-engineering templates to the user's questions, prompts, and responses in order to induce GPT to identify ambiguity, generate follow-up questions, and ultimately produce a final output that considers both the original user prompt as well as the additional information from the ensuing conversation. The prompt-engineering will be invisible to the user. In some cases, the API will be prompted multiple times to produce multi-step results for ambiguity analysis before the user is shown only the final response of a small sequence of API interactions. In other cases, the API will be given a single prompt, which will be a modified version of the user's original prompt decorated with specific prompt engineering to steer the response. To the user it should appear as if each of their inputs is given just one output in direct response to what they wrote, just as when chatting with ChatGPT directly, and the specific prompt engineering will not be shown to the user. The full chain of API prompts and responses will be stored for possible debugging and analysis needs. The process of generating follow-up questions will be similar to that shown by the CLAM model (Kuhn, Gal, and Farquhar 2023) [28].

The baseline output will be generated by providing the OpenAI API with an unmodified version of the user's original prompt. The experimental version with follow-up questions will be generated by providing the OpenAI API with the full context of questions and user responses.

Example inputs and outputs are provided in the "Proof of Concept" section.

Results Database

The metrics for analysis are described in section 5.2. In order to analyze these metrics, all user responses to the entry and exit questionnaires must be stored, along with the chat logs for all interactions with the question-asking AI. Additionally, several pieces of metadata will be helpful for data analysis:

- A session ID to uniquely identify each time a user logs on to use the system.
- Number of prompts given by each user.
    - Each user must provide at least one prompt, but may provide multiple prompts if desired. If users are regularly voluntarily providing extra prompts, this may show a high degree of engagement with the system.
- A full transcript of the conversation back-end, including API calls which were invisible to the user, for in-depth analysis.

4.5 Services and Cost Analysis

The site for the webapp and survey will be hosted through GitHub. The OpenAI API will be accessed by the use of Azure functions. The database will also be hosted through Azure services.

The Azure services required have already been set up and linked to my personal billing information. Based on experimentation with the services, the cost of accessing the Azure services and data required for a single participant to complete the study should be no more than 5 cents per participant.

The latest version of the GPT API is GPT-4 Turbo, with a cost of $0.01/1k tokens of input, and $0.03/1k tokens of output. A token is 4 characters. The proof of concept provided in the following section includes 2,882 characters of input (720 tokens) and 12,653 characters of output (3,163 tokens). This includes the inputs and outputs for both the baseline result and the result with clarifying questions. Assuming these values are representative, the expected cost of using the GPT-4 Turbo API would be:

$$0.72 * \$0.01 + 3.16 * \$0.03 = \$0.102$$

Or approximately ten cents per participant. Adding in the cost of using Azure Services, the total cost of a single usage of the webapp should be approximately $0.15.

I estimate approximately 100 runs of the webapp. This includes 10 runs for the pilot study, 70 runs for the primary study (section 4.3), and an additional 20 runs for development and testing of the app. The total cost of operating this web app should total approximately $0.15 x 100 = $15, which I intend to pay personally and have no need to seek funding for.

4.6 Proof of Concept

While I have not yet created the webapp described in this proposal, it is possible to mimic its functionality by pre-establishing a prompt engineering template and a user request and then entering the prompt-engineered responses into ChatGPT in the same way they would be sent to the OpenAI API by the proposed webapp. For an initial proof-of-concept, I thought it best to use a real-world scenario instead of one that is entirely contrived or imagined, since an imagined scenario may be more difficult to answer detailed questions about and more difficult to judge the quality of the output. I chose a prompt relevant to my motivation for performing this work: a cover letter to a faulty position at the University of Alaska, Anchorage.

4.6.1 Template

This is the prompt engineering template I intend to use. The template assumes that the program is given initial parameters of a user's name *<username>* and prompt *<X>*

1) A user named <username> is requesting the creation of a new document. This is their request:
<username>: "<*X*>"
Identify any areas of significant ambiguity or necessary information that has not been included, and write these out in a short list.

2) Respond as though this request was just made by the user. Start with a friendly greeting, then ask a respectful and concise question to clarify the user's needs, keeping in mind the ambiguity you just identified.

<username>: *"<X>"*

//GPT asks a question. Store user response as <Y>

3) This is the user's response. Continue the conversation in respectful, friendly, conversational tone. Interpret any further responses as coming directly from the user and proceed with the conversation normally. Ask more questions as needed until there is enough information to fulfil the user's request with a high degree of precision and quality.
<username>: *"<Y>"*

4.6.2 Transcript:

This is the transcript of a conversation with ChatGPT's default GPT-3.5 version, on 9/18/2023

Me: A user named Bernadette is requesting the creation of a new document. This is their request:

Bernadette: "Write a cover letter for a faculty job opening as an assistant professor of computer science at the University of Alaska, Anchorage. Write for an applicant who is currently ABD in a PhD program with a focus on Artificial Intelligence and Large Language Models."

Identify any areas of significant ambiguity or necessary information that has not been included, and write these out in a short list.

GPT: There are a few areas of significant ambiguity or necessary information missing from Bernadette's request for a cover letter. Here's a short list of the key points that need clarification or additional information:

1. **Applicant's Personal Information:** The cover letter should include the applicant's name, contact information, and other personal details. This information is essential for addressing the letter correctly.

2. **Specific Job Posting Details:** The job posting for the assistant professor position at the University of Alaska, Anchorage should be referenced. Information such as the application deadline, department's requirements, and any specific qualifications or responsibilities mentioned in the posting should be included in the cover letter.

3. **Academic Background:** While it's mentioned that the applicant is currently ABD (All But Dissertation) in a PhD program with a focus on Artificial Intelligence and Large Language Models, more information about the applicant's academic background, such as the name of the university, the expected completion date of the PhD, and any relevant coursework or research projects, should be included.

4. **Relevant Experience:** The cover letter should highlight the applicant's relevant experience in the field of computer science, teaching, and research. It's important to specify any teaching experience, publications, conference presentations, or grants related to the field.

5. **Why the University of Alaska, Anchorage:** The applicant should articulate why they are interested in the University of Alaska, Anchorage specifically. Is there a particular aspect of the university's program, location, or mission that aligns with the applicant's career goals and research interests?

6. **References:** If the applicant has any references or recommendations, they should be mentioned or attached with the cover letter.

7. **Formatting and Length:** Clarify any specific formatting requirements for the cover letter, such as font size, margins, and length restrictions.

8. **Additional Documentation:** Determine if the university requires any additional documents, such as a teaching philosophy statement or a research statement, to be submitted along with the cover letter.

Once these details are clarified, a comprehensive cover letter tailored to the specific job opening can be crafted effectively.

Me: Respond as though this request was just made by the user. Start with a friendly greeting, then ask a respectful and concise question to clarify the user's needs, keeping in mind the ambiguity you just identified.

Bernadette: "Write a cover letter for a faculty job opening as an assistant professor of computer science at the University of Alaska, Anchorage. Write for an applicant who is currently ABD in a PhD program with a focus on Artificial Intelligence and Large Language Models."

GPT: Hello Bernadette,

Thank you for reaching out! I'd be happy to assist you in creating a cover letter for the assistant professor position at the University of Alaska, Anchorage. To ensure that I can provide you with a tailored and accurate cover letter, I have a few questions to clarify your needs:

1. Could you please provide me with the specific details from the job posting, such as the application deadline, department requirements, and any qualifications or responsibilities mentioned in the posting?

2. Can you share more about your academic background, including the name of your university, the expected completion date of your PhD, and any relevant coursework or research projects?

3. Do you have any particular teaching experience, publications, conference presentations, or grants related to the field of Artificial Intelligence and Large Language Models that should be highlighted in the cover letter?

4. Could you explain why you are interested in the University of Alaska, Anchorage? Are there specific aspects of the university's program, location, or mission that align with your career goals and research interests?

5. Are there any specific formatting requirements or length restrictions for the cover letter?

6. Do you need assistance with any additional documents, such as a teaching philosophy statement or a research statement, to be submitted along with the cover letter?

Your responses to these questions will help me create a cover letter that fully meets your needs. Please provide as much detail as possible, and I'll get started on crafting the cover letter accordingly.

Best regards,

[Your Name][6]

Me: This is the user's response. Continue the conversation in respectful, friendly, conversational tone. Interpret any further responses as coming directly from the user and proceed with the conversation normally. Ask more questions as needed until there is enough information to fulfil the user's request with a high degree of precision and quality.

---

[6] Not sure why it included "Best Regards [Your Name]" but the questions are reasonable…

Bernadette: "1. The application deadline is January 1st, 2025. The requirement is to have a PhD or be ABD in a PhD program in Computer Science or a closely related field.

2. I am currently ABD in a PhD program at the University of Hawaii at Manoa. I expect to finish in May 2025, I have done graduate coursework in robotics, autonomous agents, artificial intelligence, and renewable energy.

3. I have been a TA for courses on Artificial Intelligence, Intro to Java, and Programming Language Theory. I have also been a mentor for FIRST Robotics team 995 and have taught several colleagues to program in a professional setting. Additionally, I coach a fencing club on the weekends, which, while not directly relevant to my Computer Science career, does show that I love to teach!

4. I am primarily interested in UAA because I live in Anchorage with my family and I intend to continue living and working in the city, but I am exploring whether a career as a professor might be a better fit for me than my current job as a Systems Analyst for Anchorage School District.

5. There is no specific formatting requirement, but please keep it under one page.

6. Just a cover letter will be fine for now, thanks."

GPT: Thank you for providing these details, Bernadette. This information is very helpful in crafting a tailored cover letter for your application. Here's a summary of what you've shared:

1. Application Deadline: January 1st, 2025, and the requirement is to have a PhD or be ABD in a PhD program in Computer Science or a closely related field.

2. Academic Background: You're currently ABD in a PhD program at the University of Hawaii at Manoa, expecting to finish in May 2025. You've completed graduate coursework in robotics, autonomous agents, artificial intelligence, and renewable energy.

3. Teaching Experience: You've worked as a TA for courses on Artificial Intelligence, Intro to Java, and Programming Language Theory. You've also mentored a FIRST Robotics team and taught programming in a professional setting. Additionally, you coach a fencing club on weekends, highlighting your passion for teaching.

4. Interest in UAA: Your interest in the University of Alaska, Anchorage, stems from your current residence in Anchorage and your desire to continue living and working in the city, potentially transitioning from your role as a Systems Analyst for Anchorage School District to a professorial career.

5. Formatting: There are no specific formatting requirements, but the cover letter should be kept under one page.

6. Additional Documents: You only need assistance with the cover letter for now.

With this information in mind, I'll start working on your cover letter, emphasizing your qualifications, teaching experience, and your strong connection to Anchorage. If you have any specific points you'd like to include or any other preferences, please let me know, and I'll incorporate them into the letter.

Me: That is all correct. Please proceed with writing the letter.

GPT: Certainly, Bernadette! I'll now draft the cover letter for your application as an assistant professor of computer science at the University of Alaska, Anchorage.

[Your Name]

[Your Address]

[City, State, ZIP Code]

[Your Email Address]

[Your Phone Number]

[Date]

[Recipient's Name]

[Recipient's Title]

[University of Alaska, Anchorage]

[Address]

[City, State, ZIP Code]

Dear [Recipient's Name],

I am writing to express my enthusiastic interest in the faculty position of Assistant Professor in Computer Science at the University of Alaska, Anchorage, as advertised with an application deadline of January 1st, 2025. The prospect of joining the esteemed faculty at UAA is particularly appealing to me due to my deep-rooted connection to Anchorage and my passion for teaching and research in computer science.

Currently, I am an ABD candidate in a PhD program at the University of Hawaii at Manoa, with an expected completion date in May 2025. My doctoral studies have equipped me with a strong foundation in the fields of Artificial Intelligence and Large Language Models. I have engaged in rigorous graduate coursework, with a focus on robotics, autonomous agents, artificial intelligence, and renewable energy.

My academic journey has not only deepened my knowledge but has also ignited a strong commitment to advancing the field through research and education.

My teaching experience further reinforces my dedication to academia. As a Teaching Assistant, I have had the privilege of instructing courses on Artificial Intelligence, Intro to Java, and Programming Language Theory. These experiences have honed my pedagogical skills and allowed me to foster a dynamic and inclusive learning environment. Beyond the classroom, I have mentored a FIRST Robotics team, empowering young minds to explore the world of robotics and coding. Moreover, my role as a coach at a local fencing club reflects my unwavering passion for teaching and mentoring, demonstrating my adaptability and commitment to imparting knowledge.

My decision to pursue a faculty position at UAA is deeply rooted in my connection to Anchorage and my aspiration to contribute to the academic community in this vibrant city. Anchorage is not just my residence; it is the place I call home. I am committed to its growth and development, and I see the University of Alaska, Anchorage, as an ideal platform to realize this commitment. My transition from a Systems Analyst for Anchorage School District to an academic role underscores my dedication to making a meaningful impact on the local community through education.

Enclosed is my curriculum vitae, which provides a comprehensive overview of my academic and professional journey. I look forward to the opportunity to discuss in greater detail how my qualifications align with the needs and goals of the Computer Science department at UAA.

Thank you for considering my application. I am eager to contribute my knowledge, passion, and commitment to the University of Alaska, Anchorage, and to further enrich the academic community. Please feel free to contact me at [Your Phone Number] or [Your Email Address] to arrange an interview at your convenience.

Sincerely,

[Your Name]

Please review the letter, and if you have any specific changes or additional information you'd like to include, feel free to let me know. I'm here to ensure it perfectly represents your qualifications and aspirations.

4.6.3 Baseline

Compare the final output of this conversation with the baseline response from ChatGPT with no prompt engineering instructing it to ask follow-up questions:

Me: Write a cover letter for a faculty job opening as an assistant professor of computer science at the University of Alaska, Anchorage. Write for an applicant who is currently ABD in a PhD program with a focus on Artificial Intelligence and Large Language Models.

GPT: [Your Name]

[Your Address]

[City, State, ZIP Code]

[Your Email Address]

[Today's Date]

Dr. [Hiring Committee Chair's Name]

Search Committee Chair

Department of Computer Science

University of Alaska, Anchorage

3211 Providence Drive

Anchorage, AK 99508

Dear Dr. [Hiring Committee Chair's Last Name],

I am writing to express my strong interest in the position of Assistant Professor of Computer Science at the University of Alaska, Anchorage, as advertised in [Source of Job Posting]. As a dedicated and passionate researcher with a focus on Artificial Intelligence and Large Language Models, I am excited about the opportunity to contribute to the academic excellence of your department.

I am currently pursuing my Ph.D. in Computer Science with a specialization in Artificial Intelligence at [Your Current University], and I anticipate completing my degree by [Expected Graduation Date]. My research focuses on the development and application of advanced machine learning techniques, particularly in the context of large language models. My dissertation work, titled "[Your Dissertation Title]," investigates novel approaches to enhancing the efficiency and interpretability of language models[7], which has significant implications for a wide range of applications, including natural language processing, information retrieval, and human-computer interaction.

Throughout my academic journey, I have had the privilege of working with esteemed mentors and collaborators, and my research has resulted in several publications in reputable conferences and journals, including [List Relevant Publications]. These experiences have honed my ability to conduct innovative research, communicate complex ideas effectively, and mentor aspiring computer scientists.

One aspect that particularly drew me to the University of Alaska, Anchorage is its commitment to fostering a diverse and inclusive learning environment. I am deeply committed to promoting diversity, equity, and inclusion in the field of computer science. My teaching philosophy emphasizes the importance of creating an inclusive classroom where all students feel welcome, valued, and empowered to excel. I look forward to contributing to the university's ongoing efforts to promote diversity and equity in computer science education and research.

I am also excited about the prospect of collaborating with the faculty at the University of Alaska, Anchorage, whose work aligns with my research interests. The opportunity to engage in interdisciplinary collaborations and contribute to the vibrant research community at the university greatly appeals to me.

In closing, I am enthusiastic about the possibility of joining the Department of Computer Science at the University of Alaska, Anchorage. My research, teaching, and commitment to fostering diversity make me a strong fit for this position, and I am eager to bring my expertise to your institution. I have attached my curriculum vitae, research statement, teaching philosophy, and contact information for my references for your consideration.

Thank you for considering my application. I would be delighted to discuss how my background and experiences align with the needs of your department further. Please feel free to contact me at [Your Phone Number] or [Your Email Address] to arrange an interview. I look forward to the opportunity to contribute to the academic community at the University of Alaska, Anchorage.

Sincerely,

[Your Full Name]

---

[7] I found this sentence interesting as a cautionary reminder of LLMs tendency to hallucinate. ChatGPT has no idea what my research is about, but filled in something plausibly generic.

4.6.4 Proof of Concept Discussion

While the cover letter resulting from a dialog of clarifying questions was not perfect, I believe that it could be made passable with a few adjustments and is a far better cover letter than the baseline response. With such a limited initial prompt, the baseline cannot possibly capture all the relevant context needed for writing a cover letter. I chose to keep the initial prompt short (and will be requesting short prompts from users) because I am specifically trying to demonstrate the efficacy of asking clarifying questions to gain detailed information about requests that do not initially have sufficient information included. My experience working as a software developer for various clients since 2010 has taught me that initial requests from clients almost never have the full details of what is needed, and asking thoughtful questions to help clients think through their own needs is a vital aspect of requirements gathering.

Going through the exercise of creating this proof-of-concept dialog forced me to think about what I wanted to include in the cover letter from a new perspective and to consider details I might otherwise not have thought to include. I found the process of dialog with ChatGPT helped me to refine my requirements and think more deeply about what the cover letter ought to include or not include, which would be helpful while revising the output manually, even if the final output was not used verbatim. ChatGPT’s offer to assist with writing a teaching philosophy statement also encouraged me to engage with ChatGPT for the rest of the application process and got me thinking about how the cover letter fit in to the broader application. I would rate the experience as far superior to the baseline response, and I hope to demonstrate that this approach can be broadly useful through the proposed experiment.

This proof-of-concept also clearly shows that it is possible in principle to direct ChatGPT to effectively analyze requests with insufficient detail, formulate questions seeking additional information, and incorporate the new information into future responses. Regardless of whether the question-asking process proves to be broadly useful, this shows that it will at least be feasible to conduct the experiment with the proposed architecture.

5 Limitations

This section covers limitations of the proposed research, specifically areas which are tangentially related to the proposed research but are not intended to be covered by this experiment. This section is divided into two subsections. The “Possible Expansions” subsection includes areas that could feasibly be covered by expanding the experiment if there is a need or desire to expand the scope of the project. The “Out of Scope” subsection covers areas that I consider to be impractical to include due to resource and timeframe constraints.

5.1 Possible Expansions

1. **Generative Art** – Creative content generation, as defined in Section 4, seems to naturally include the generation of works of visual art as well as written works. I have not included visual art within

the proposed experiment for two reasons. The first reason is a desire to keep the experiment focused. Creative content generation is broad by nature, so applying some limits, such as focusing only on written output generated from short initial prompts, is useful for controlling the scope of the study. The second reason is that there is a broad field of literature on generative art with which I have only passing familiarity, and including generative visual art properly would likely require extending the literature search considerably, and thus pushing out the schedule for the study. This suggests that while clarifying questions might be beneficial to generative art, and might be studied through a similar methodology, it would be best to conduct this as a separate experiment at a later date, rather than incorporating it directly into the experiment using only text outputs.

2. **Expanded User Preference Studies** – The proposed research includes an exit questionnaire designed to gauge users' experience with the new system. However, this is only designed to demonstrate whether users are open to engaging in question-and-answer dialog with the AI. The exit questionnaire is not sufficient to show whether users prefer this method to other methods of interacting with the AI, such as continuously refining their own prompts to achieve better results as is currently the norm. This is a valuable question to investigate, but it was not included in the proposed research because it seems premature to investigate whether this mode of interaction is preferred when it has not yet been proven to be effective. It seems to make more sense to me to first show that the technique is effective in producing higher quality outputs, and then address the question of whether it provides a better user experience in a follow-up study. However, this could be included as a second experiment or an expansion on the proposed experiment if needed.
3. **Comparing Different LLMs** – The proposed webapp would utilize OpenAI's API, and will thus be relying on the gpt-3.5 LLM. It is possible that other LLMs may perform better or worse at the tsk of asking clarifying questions and producing outputs based on user response to those questions. However, this technique relies on fairly universal prompt-engineering techniques and is likely to be applicable across different LLMs. Thus, the inclusion of multiple LLMs for comparison, which would require a larger data sample and a more complex program, does not seem to be a particularly interesting question with regards to the proposed technique and does not seem worth the added difficulty. However, it may be possible to include other LLMs if this is required.

5.2 Out of Scope

1. **High-Precision Applications** – I strongly suspect that a question-asking dialog prior to final output will be especially useful for tasks where a high degree of precision is required. For tasks requiring only general knowledge, it may be possible for an AI to infer the user's intentions without asking clarifying questions. However, in applications where a high degree of precision is needed, especially tasks which are highly context-sensitive, it is necessary to gather more information rather than trying to infer whenever incomplete requirements are given. Although this

is a promising area of research, I expect the setup for such an experiment may be well outside the scope of what can reasonably be accomplished in this study. To test this hypothesis properly would require identifying a large number of high-precision tasks, gaining the participation of expert-level users capable of both providing high-precision feedback and judging the output, and would require a metric for objectively judging whether the output was correct, since in many high-precision tasks simple user preference may not be sufficient to demonstrate advantage. These requirements make this question infeasible to study within the constraints of this study.

2. **Code Generation** – Transformers have been used in recent years to generate working computer code. This can be done by a general-purpose LLM such as ChatGPT, but there are also dedicated code-generation tools based on the transformer architecture, such as CodeBERT [17] and GitHub's Copilot [86]. Given that the idea for this study originated from my experience doing requirements gathering for software projects, and the inherently high-precision nature of code generation (see previous bullet), code generation with clarifying questions is a natural target for study. However, structuring an experiment to study the efficacy of question-asking on generated code has similar difficulties to that of other high-precision applications. A test set must be drawn from real software requirements in order to be useful, and then tested with users who have actual ambiguously articulated requirements. Identifying a large pool of users and developing a way to objectively test whether the output generated by the AI was sufficient for their needs is likely to be prohibitively time-consuming. Users could be given a prompt and told they must make the AI generate software that meets certain requirements, but this will not work as then we are giving the requirements to the users and the phrasing of the software description they are given will bias how they interact with the AI. Although code generation with clarifying questions is something that would be beneficial to study in the future, I do not know of a way of overcoming these obstacles.
3. **Improving AI Alignment –** If clarifying questions are shown to be an effective method for capturing user needs, the technique could, in principle, be a useful tool for improving AI alignment. As noted in Section 1, AI is often trained to pursue goals in ways that the designers did not intend, sometimes in ways that explicitly violate unstated assumptions held by the designers and can result in undesirable behavior from the AI [5,23]. In principle, engaging in a question-asking dialog could be a best-practice for reducing unintended side effects in the use of future AI. However, this is highly theoretical at this point. Machine Learning loss functions are not set up by misinterpretations of natural language prompts, but by carefully selecting a formula of quantitative metrics. It is not clear how a LLM-powered question-asking dialog would fit into that process. However, it is an interesting concept in theory and may be worth examining in the future, though how to set up an experiment to study this is not clear, and is well beyond the scope of the proposed research.

6 Risks

6.1 Risks to Users

1. **Sharing of Personally Identifiable Information –** Any information entered into the webapp will be collected for analysis, and thus may be viewed by the research team (which at this point is only myself). Prompts and responses entered by the user will also be sent to the OpenAI API, and OpenAI cautions against entering any personally identifiable information into their system.
   Risk Mitigation: Users will be shown a warning instructing them not to enter any personally identifiable information.
   Possible Escalation: The current plan is to retain all user responses for qualitative analysis. However, it is possible to perform the primary analysis required by this study without retaining user's prompts or responses, or the outputs from the OpenAI API. While this has the potential to reduce the richness of analysis that will be possible, all user and AI responses could be discarded rather than retained, keeping only the answers to the entry and exit questionnaires and whether the user preferred the baseline output or the output from the experimental system, without retaining the outputs themselves. My preference would be to take this approach only if deemed necessary by the IRB.
2. **Toxic Responses** – When working with LLMs, there is always a risk of toxic responses. Although ChatGPT is designed to filter out toxic responses, multiple studies have shown that it is possible to bypass these restrictions, particularly in multi-step dialogs [6,15,64], so although the risk is low it is still possible that toxic responses will be produced in a small percentage of cases.
   Risk Mitigation: The experiment may need to be conducted with only adult participants to avoid exposing minors to potentially harmful responses.
3. **Inaccurate Responses –** Even when responses are not toxic, LLMs can "hallucinate" inaccurate information and present it in a convincing way [58,72]. Because of this, the output produced for the users by the system may contain factual errors.
   Risk Mitigation: Users will be shown a warning during the instructions phase of the experiment, cautioning them that the final output may contain factual errors and should not be relied upon without first being carefully checked by a human being.

6.2 Risks to Project Timeline

1. **Insufficient Quantity of Users –** The largest risk to the success of the project is that not enough users will participate. There is currently no plan for how to attract a large number of users, which is currently planned to be done in an ad-hoc manner through various channels.
   Risk Mitigation: If insufficient data is being collected, it may be necessary to expand the scope of the participant search through paid advertising and possibly the inclusion of some form of incentive.

2. **Changes to OpenAI API -** OpenAI is currently making their API available for public use. However, they are under no obligation to continue doing so. It is possible that OpenAI will rescind public access to the API or change the terms of service such that the webapp is rendered inoperable before sufficient data can be collected.
   Risk Mitigation: I consider this risk to be relatively unlikely. However, in the event that OpenAI's GPT API is no longer viable, the webapp will have to be re-worked to use a competing LLM, such as Meta's Llama 2, which is currently free for research and commercial use [36].

7 Ethical Considerations

I see two primary ethical considerations that impact whether this research should be conducted.

- The problem of AI alignment.
- The potential for job loss or damage to specific professions (notably software developers) from the development of a system capable of engaging in a form of requirements-gathering dialog that is currently done only by humans.

I believe that both considerations ultimately favor the creation of this system. However, since any development in AI can lead to unforeseen consequences, these issues are worth considering carefully before proceeding. My thoughts on the ethics of each issue are provided in this section.

AI Alignment

Any improvement to the capabilities of modern AI systems brings with it risks of AI-misalignment. The proposed feature of question-asking could potentially allow chatbots to fill roles that currently require greater human supervision, thus granting the system greater agency, which has been identified as a major risk factor for AI-misalignment [5]. However, in this case I believe that the nature of the improvement mitigates this risk and should result in better alignment with human needs. When an AI is given an ambiguous prompt to process, the human prompt-writer may not have realized the potential for ambiguity or side-effects, and this can result in mis-aligned outputs [5,11,62]. If the AI itself can ask clarifying questions and point out potential risk factors in the instructions it is given, this has the potential to improve the robustness of the goal specifications the AI pursues in generating its output. Therefore, I believe that this research, if successful, would be a net improvement for AI alignment, even if it allows AI systems to operate in some areas where direct human control was previously needed.

Potential for Job Loss

Given that the capability of engaging in dialog to clarify points of ambiguity is currently only done by human beings, adding this capability to an AI system has the potential to allow the AI to take over certain job functions that are currently secure against the threat of automation, such as software requirements gathering or escalated technical support roles. I have been in the position of automating people out of jobs before, and have no desire to be in that position again. On a personal level, I particularly do not wish to contribute to the automation of the profession of computer programming, which has had deep significance in my life and which I hope can continue to be a proud profession for future generations.

However, the benefits of LLMs for a wide variety of purposes are already broadly appreciated and businesses and organizations are already exploring ways of integrating the new technology into their business processes. It is likely that even if the output from AI is not as good as that from human workers, humans will still be replaced by AI in many cases if AI is significantly less expensive. However, it has long been observed that automation can lead to unexpected failures if the newly automated system does not provide sufficient feedback to human operators [42]. An inappropriately automated system may work perfectly on its own, most of the time. When it breaks, nobody knows how it works or how to fix it, because under normal circumstances it runs silently, and nobody has any reason to engage with the system on a regular basis. Such over-automated systems tend to cause a great deal of headache for the organizations they are employed in and can allow errors to propagate silently, being automatically corrected or hidden until the problems become too large to ignore and only then come to the attention of human workers. Because of this, fully autonomous systems are generally only desirable if they run correctly very nearly 100% of the time. By contrast, empowering automation provides people with tools that operate under their direction, which allow them to perform the same work in less time, or more work in the same amount of time, but keeps a human being in control and actively engaging with the system on a day-to-day basis. When something goes wrong with such a system, the users of the system are intimately familiar with how it works and can usually identify and fix problems with less difficulty than problems on over-automated systems.

I believe that this research can contribute to LLM chatbots being used as empowering systems, rather than inappropriately over-automated ones, because the nature of asking clarifying questions inherently keeps humans “in the loop” for a greater portion of the process. If a dialog between AI and human operator leads to improved outcomes, this encourages organizational processes to be built around humans working in collaboration with AI rather than being replaced by AI. Ideally, the question-asking process may be thought-provoking for the human user as well as resulting in improved outcomes from the AI itself. If the best results can be achieved by human-AI collaboration, organizations are then incentivized to empower their employees with AI-driven tools, rather than replacing them with autonomous AI agents.

If AI continues to advance at its current pace, as seems likely, LLMs will inevitably take over at least some tasks currently performed by humans. While we cannot stop this change, we can exert at least some influence on the nature of these changes. While it is true that increased productivity will mean that some tasks can now be completed by smaller teams, I still find it preferable to work towards a world where humans and AI work together in a collaborative fashion.

8 Timeline

Date	Milestone
2023-09-25	Initial revision of proposal
2023-10-09	First round of review for proposal complete. Committee & Proposal Title & Abstract submitted to department chair
2023-11-06	Proposal revisions complete. IRB submission for pilot complete.
2023-11-15	Proposal defense.
2023-12-31	Webapp is implemented, functional, and publicly available.
2024-01-31	The pilot study and revisions from the pilot study should be completed in January 2024. Pilot may begin in December if webapp implementation progresses quickly.
2024-04-30	February-April should be a sufficient window to gather data from the experiment.
2024-08-01	Initial revision of dissertation should be written up and ready for review and revision by the start of the Fall 2024 term.
2024-11-01	Revisions to dissertation should be complete by the start of November 2024 and a defense can be scheduled for November or December, depending on committee availability.
2024-12-13	Dissertation should be defended prior to the end of the Fall 2024 semester.

Timeline Caveats:

- If work proceeds exceptionally smoothly, it may be possible to write and defend the dissertation in Spring 2024. However, this seems overly optimistic to hope for, but may be re-addressed in the Spring if work is proceeding ahead of schedule.
- If major revisions or expansions are required, such as the possible expansions noted in section 6.1, or if any major setbacks occur, such as the risks noted in section 7.2, the timeline may need to be extended to accommodate the additional work.

9 Conclusion

Recent advances in generative AI systems have led to the creation of systems with novel generative capabilities, capable of producing written content, visual art, music, and even working computer code at human or near-human levels of quality. However, serious questions remain about the best uses of such systems, and about how these systems can be appropriately integrated into human work and life. Most of the extant research into LLMs investigates ways to improve the output generated in response to short prompts, usually in cases where there are distinct right and wrong answers. Generative AI producing art and music have required more subjective measures of quality, but even in these fields most research focuses on the quality of output in response to a single short prompt. I propose that generative AI should treat user requests as a human software engineer would, as an opportunity to gather detailed requirements and prompt thought-provoking discussions before proceeding with content generation. Rather than trying to make the best of ambiguous or inadequate prompts, by engaging the user in a question-answering dialog the AI can gather relevant context that can lead to superior outputs that better match the user's actual needs.

While previous work has demonstrated the efficacy of question-answering in resolving ambiguous prompts, prior research has the serious shortcoming of being conducted primarily with "simulated humans" rather than actual human beings able to report on their subjective experience of using the system. Prior research into LLMs asking users clarifying questions has the further shortcoming of focusing solely on short-answer responses with objectively correct or incorrect answers. These sorts of questions are generally better handled by search functions rather than generative AI, and do not make full use of generative AI's novel capability for generating long-form content. My proposed research will address both of these shortcomings by engaging real human beings in the creation of creative content with no simple right or wrong answers and eliciting the subjective evaluation of the user's themselves as validation of the output quality.

By enabling a more collaborative model of human-AI interaction than has previously been the norm for generative AI systems, I hope that this research will contribute to systems which keep humans "in the loop" and allow for appropriate levels of automation that empower, rather than replace, human work and human decision-making.

BIBLIOGRAPHY

[1] Y. Bengio, P. Simard, and P. Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks* 5, 2 (March 1994), 157–166. DOI:https://doi.org/10.1109/72.279181

[2] Assma Boughoula, Aidan San, and ChengXiang Zhai. 2020. Leveraging Book Indexes for Automatic Extraction of Concepts in MOOCs. In *Proceedings of the Seventh ACM Conference on Learning @ Scale* (L@S '20), Association for Computing Machinery, New York, NY, USA, 381–384. DOI:https://doi.org/10.1145/3386527.3406749

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 1877–1901. Retrieved January 2, 2023 from https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[4] Miguel A Cardona, Roberto J Rodríguez, and Kristina Ishmael. 2023. Artificial Intelligence and the Future of Teaching and Learning. *Office of Education Technology* (May 2023).

[5] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '23), Association for Computing Machinery, New York, NY, USA, 651–666. DOI:https://doi.org/10.1145/3593013.3594033

[6] Bocheng Chen, Guangjing Wang, Hanqing Guo, Yuanda Wang, and Qiben Yan. 2023. Understanding Multi-Turn Toxic Behaviors in Open-Domain Chatbots. DOI:https://doi.org/10.1145/3607199.3607237

[7] Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is ChatGPT's behavior changing over time? DOI:https://doi.org/10.48550/arXiv.2307.09009

[8] Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long Short-Term Memory-Networks for Machine Reading. DOI:https://doi.org/10.48550/arXiv.1601.06733

[9] Seongho Choi, Kyoung-Woon On, Yu-Jung Heo, Ahjeong Seo, Youwon Jang, Minsu Lee, and Byoung-Tak Zhang. 2021. DramaQA: Character-Centered Video Story Understanding with Hierarchical QA. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 2 (May 2021), 1166–1174.

[10] Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-Based Named Entity Recognition Using BART. DOI:https://doi.org/10.48550/arXiv.2106.01760

[11] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. 2022. Underspecification presents challenges for credibility in modern machine learning. *J. Mach. Learn. Res.* 23, 1 (January 2022), 226:10237-226:10297.

[12] Sreerupa Das, C. Giles, and Gordon Sun. 1992. Learning Context-free Grammars: Capabilities and Limitations of a Recurrent Neural Network with an External Stack Memory. (June 1992).

[13] Douglas B. Lenat. 1995. CYC: A Large-Scale Investment in Knowledge Infrastructure. *COMMUNICATIONS OF THE ACM* 38, 11 (1995), 33–38.

[14] Douglas C Montgomery. 2013. *Design and Analysis of Experiments* (8th ed.). John Wiley & Sons, Inc., Arizona State University.

[15] Farshid Faal, Ketra Schmitt, and Jia Yuan Yu. 2022. Reward modeling for mitigating toxicity in transformer-based language models. *Appl Intell* (July 2022). DOI:https://doi.org/10.1007/s10489-022-03944-z
[16] Rachelyn Farrell, Scott Robertson, and Stephen G. Ware. 2016. Asking Hypothetical Questions About Stories Using QUEST. In *Interactive Storytelling* (Lecture Notes in Computer Science), Springer International Publishing, Cham, 136–146. DOI:https://doi.org/10.1007/978-3-319-48279-8_12
[17] Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, and Ming Zhou. 2020. CodeBERT: A Pre-Trained Model for Programming and Natural Languages. Retrieved February 15, 2023 from http://arxiv.org/abs/2002.08155
[18] Charles J. Fillmore. 1973. Some problems for case grammar. In *Report Of The Twenty-Second Annual Round Table Meeting on Linguistics and Language Studies*, Roger W. Shuy and Ralph W. Fasold (eds.). Georgetown Univ. Press, Washington, DC.
[19] Google. 2023. Bard. Retrieved May 22, 2023 from https://bard.google.com
[20] Jessica Guynn. 2023. Is ChatGPT ‘woke’? AI chatbot accused of anti-conservative bias and a grudge against Trump. *USA TODAY*. Retrieved February 16, 2023 from https://www.usatoday.com/story/tech/2023/02/09/woke-chatgpt-conservatives-bias/11215353002/
[21] Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The Goldilocks Principle: Reading Children’s Books with Explicit Memory Representations. *arXiv:1511.02301 [cs]* (April 2016). Retrieved April 12, 2022 from http://arxiv.org/abs/1511.02301
[22] Gwo-Jen Hwang and Ching-Yi Chang. 2021. A review of opportunities and challenges of chatbots in education. *Interactive Learning Environments* 0, 0 (July 2021), 1–14. DOI:https://doi.org/10.1080/10494820.2021.1952615
[23] Jack Clark and Dario Amodei. 2016. Faulty reward functions in the wild. Retrieved July 7, 2023 from https://openai.com/research/faulty-reward-functions
[24] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the Limits of Language Modeling. Retrieved January 3, 2023 from http://arxiv.org/abs/1602.02410
[25] Karen Hao. 2020. We read the paper that forced Timnit Gebru out of Google. Here’s what it says. *MIT Technology Review*. Retrieved February 16, 2023 from https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/
[26] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stepha Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103, (April 2023), 102274. DOI:https://doi.org/10.1016/j.lindif.2023.102274
[27] Vijay Krishnan and Christopher D. Manning. 2006. An Effective Two-Stage Model for Exploiting Non-Local Dependencies in Named Entity Recognition. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Sydney, Australia, 1121–1128. DOI:https://doi.org/10.3115/1220175.1220316
[28] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. CLAM: Selective Clarification for Ambiguous Questions with Generative Language Models. *ICML 2023 Workshop on Deployment Challenges for Generative AI* (June 2023). Retrieved August 17, 2023 from https://openreview.net/forum?id=VQWuqgSoVN
[29] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. *Proceedings of NAACL 2016* (April 2016). Retrieved April 28, 2022 from http://arxiv.org/abs/1603.01360
[30] Doug Lenat. 2019. Not Good As Gold: Today’s AI’s Are Dangerously Lacking In AU (Artificial Understanding). *Forbes*. Retrieved December 5, 2022 from https://www.forbes.com/sites/cognitiveworld/2019/02/18/not-good-as-gold-todays-ais-are-dangerously-lacking-in-au-artificial-understanding/

[31] Doug Lenat. 2023. Getting from Generative AI to Trustworthy AI: What LLMs might learn from Cyc. (2023).
[32] Hector Levesque, Ernest Davis, and Leora Morgenstern. The Winograd Schema Challenge.
[33] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. DOI:https://doi.org/10.48550/arXiv.2109.07958
[34] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A Structured Self-attentive Sentence Embedding. DOI:https://doi.org/10.48550/arXiv.1703.03130
[35] Robert W. McGee. 2023. Is Chat Gpt Biased Against Conservatives? An Empirical Study. DOI:https://doi.org/10.2139/ssrn.4359405
[36] MetaAI. 2023. Introducing LLaMA: A foundational, 65-billion-parameter language model. Retrieved May 22, 2023 from https://ai.facebook.com/blog/large-language-model-llama-meta-ai/
[37] Tomas Mikolov and Geoffrey Zweig. 2012. Context dependent recurrent neural network language model. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, Miami, FL, USA, 234–239. DOI:https://doi.org/10.1109/SLT.2012.6424228
[38] Mark P Mills, Lane, author of The Cloud Revolution: How the Convergence of New Technologies Will Unleash the Next Economic Boom, and a Roaring 2020s He hosts The Last Optimist podcast. 2023. ChatGPT and Automation Come to Knowledge Work. *City Journal*. Retrieved March 1, 2023 from https://www.city-journal.org/chatgpt-and-automation-come-to-knowledge-work
[39] Alex Mitchel. 2023. ChatGPT could make these jobs obsolete. *New York Post*. Retrieved March 1, 2023 from https://nypost.com/2023/01/25/chat-gpt-could-make-these-jobs-obsolete/
[40] J.M. Moore and F.M. Shipman. 2000. A comparison of questionnaire-based and GUI-based requirements gathering. In *Proceedings ASE 2000. Fifteenth IEEE International Conference on Automated Software Engineering*, 35–43. DOI:https://doi.org/10.1109/ASE.2000.873648
[41] Erik T. Mueller. 2003. Story understanding through multi-representation model construction. In *Proceedings of the HLT-NAACL 2003 workshop on Text meaning - Volume 9* (HLT-NAACL-TEXTMEANING '03), Association for Computational Linguistics, USA, 46–53. DOI:https://doi.org/10.3115/1119239.1119246
[42] D. A. Norman, Donald Eric Broadbent, Alan David Baddeley, and J. Reason. 1989. The 'problem ' with automation: inappropriate feedback and interaction, not 'over-automation.' *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 327, 1241 (June 1989), 585–593. DOI:https://doi.org/10.1098/rstb.1990.0101
[43] OpenAI. 2023. GPT-4 Technical Report. DOI:https://doi.org/10.48550/arXiv.2303.08774
[44] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. (March 2022). DOI:https://doi.org/10.48550/arXiv.2203.02155
[45] Dhirendra Pandey, U. Suman, and A.K. Ramani. 2010. An Effective Requirement Engineering Process Model for Software Development and Requirements Management. In *2010 International Conference on Advances in Recent Technologies in Communication and Computing*, IEEE, Kottayam, India, 287–291. DOI:https://doi.org/10.1109/ARTCom.2010.24
[46] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. DOI:https://doi.org/10.3115/1073083.1073135
[47] Martine Paris. 2023. ChatGPT Hits 100 Million Users, Google Invests In AI Bot And CatGPT Goes Viral. *Forbes*. Retrieved March 13, 2023 from https://www.forbes.com/sites/martineparis/2023/02/03/chatgpt-hits-100-million-microsoft-unleashes-ai-bots-and-catgpt-goes-viral/
[48] Jeongeun Park, Seungwon Lim, Joonhyung Lee, Sangbeom Park, Minsuk Chang, Youngjae Yu, and Sungjoon Choi. 2023. CLARA: Classifying and Disambiguating User Commands for Reliable Interactive Robotic Agents. DOI:https://doi.org/10.48550/arXiv.2306.10376
[49] Fernando C.N. Pereira and David H.D. Warren. 1980. Definite clause grammars for language analysis—A survey of the formalism and a comparison with augmented transition networks. *Artificial Intelligence* 13, 3 (May 1980), 231–278. DOI:https://doi.org/10.1016/0004-3702(80)90003-X

[50] Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press.
[51] Valentina Pyatkin, Jena D. Hwang, Vivek Srikumar, Ximing Lu, Liwei Jiang, Yejin Choi, and Chandra Bhagavatula. 2023. ClarifyDelphi: Reinforced Clarification Questions with Defeasibility Rewards for Social and Moral Situations. DOI:https://doi.org/10.48550/arXiv.2212.10409
[52] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. (2018).
[53] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog* 1.8, (2019).
[54] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. In *Proceedings of the 38th International Conference on Machine Learning*, PMLR, 8821–8831. Retrieved September 12, 2023 from https://proceedings.mlr.press/v139/ramesh21a.html
[55] Lev Ratinov and Dan Roth. 2009. Design Challenges and Misconceptions in Named Entity Recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning - CoNLL '09*, Association for Computational Linguistics, Boulder, Colorado, 147. DOI:https://doi.org/10.3115/1596374.1596399
[56] Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics* 7, (May 2019), 249–266. DOI:https://doi.org/10.1162/tacl_a_00266
[57] Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension. *ACM Comput. Surv.* 55, 10 (October 2023), 1–45. DOI:https://doi.org/10.1145/3560260
[58] Jürgen Rudolph, Samson Tan, and Shannon Tan. 2023. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *JALT* 6, 1 (January 2023). DOI:https://doi.org/10.37074/jalt.2023.6.1.9
[59] Stuart Russel and Peter Norvig. 2010. *Artificial intelligence: A Modern Approach* (Third Edition ed.). Prentice Hall, Upper Saddle River, NJ.
[60] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. Retrieved October 29, 2022 from http://arxiv.org/abs/cs/0306050
[61] Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. 2017. Story Cloze Task: UW NLP System. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, Association for Computational Linguistics, Valencia, Spain, 52–55. DOI:https://doi.org/10.18653/v1/W17-0907
[62] Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. 2022. Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals. *arXiv.org*. Retrieved June 22, 2023 from https://arxiv.org/abs/2210.01790v2
[63] Murray Shanahan. 2016. The Frame Problem. In *The Stanford Encyclopedia of Philosophy* (Spring 2016), Edward N. Zalta (ed.). Metaphysics Research Lab, Stanford University. Retrieved May 17, 2022 from https://plato.stanford.edu/archives/spr2016/entries/frame-problem/
[64] Wai Man Si, Michael Backes, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, Savvas Zannettou, and Yang Zhang. 2022. Why So Toxic? Measuring and Triggering Toxic Behavior in Open-Domain Chatbots. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security* (CCS '22), Association for Computing Machinery, New York, NY, USA, 2659–2673. DOI:https://doi.org/10.1145/3548606.3560599
[65] Roderick Tabalba, Nurit Kirshenbaum, Jason Leigh, Abari Bhatacharya, Andrew Johnson, Veronica Grosso, Barbara Di Eugenio, and Moira Zellner. 2022. Articulate+ : An Always-Listening Natural Language Interface for Creating Data Visualizations. In *Proceedings of the 4th Conference on Conversational User Interfaces* (CUI '22), Association for Computing Machinery, New York, NY, USA, 1–6. DOI:https://doi.org/10.1145/3543829.3544534
[66] Robert D Van Valin. 1993. ROLE AND REFERENCE GRAMMAR. *Work Papers of the Summer Institute of Linguistics* 37, (1993), 12.
[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. DOI:https://doi.org/10.48550/arXiv.1706.03762

[68] Boheng Wang, Yunhuai Zhu, Liuqing Chen, Jingcheng Liu, Lingyun Sun, and Peter Childs. 2023. A study of the evaluation metrics for generative images containing combinational creativity. *AIEDAM* 37, (2023), e11. DOI:https://doi.org/10.1017/S0890060423000069
[69] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. DOI:https://doi.org/10.48550/arXiv.2201.11903
[70] Matt Welsh. 2022. The End of Programming. *Commun. ACM* 66, 1 (December 2022), 34–35. DOI:https://doi.org/10.1145/3570220
[71] Li-Chia Yang and Alexander Lerch. 2020. On the evaluation of generative models in music. *Neural Comput & Applic* 32, 9 (May 2020), 4773–4784. DOI:https://doi.org/10.1007/s00521-018-3849-7
[72] Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. 2023. Mitigating Language Model Hallucination with Interactive Question-Knowledge Alignment. DOI:https://doi.org/10.48550/arXiv.2305.13669
[73] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. Retrieved September 12, 2023 from https://openreview.net/forum?id=SkeHuCVFDr
[74] Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring AI Ethics of ChatGPT: A Diagnostic Analysis. Retrieved May 15, 2023 from http://arxiv.org/abs/2301.12867
[75] 2020. GPT-3: All you need to know about the AI language model - Sigmoid. Retrieved February 16, 2023 from https://www.sigmoid.com/blogs/gpt-3-all-you-need-to-know-about-the-ai-language-model/
[76] 2022. ChatGPT: Optimizing Language Models for Dialogue. *OpenAI*. Retrieved February 15, 2023 from https://openai.com/blog/chatgpt/
[77] 2023. Plumbing ChatGPT's Left-Liberal Biases › American Greatness. *American Greatness*. Retrieved February 16, 2023 from https://amgreatness.com/2023/01/28/plumbing-chatgpts-left-liberal-biases/
[78] 2023. An important next step on our AI journey. *Google*. Retrieved April 5, 2023 from https://blog.google/technology/ai/bard-google-ai-search-updates/
[79] 2023. ChatGPT Reignites the Search Wars Between Google and Microsoft. *Bloomberg.com*. Retrieved March 1, 2023 from https://www.bloomberg.com/news/newsletters/2023-02-08/chatgpt-reignites-search-competition-microsoft-bing-vs-google-bard
[80] 2023. *How AI Could Save (Not Destroy) Education | Sal Khan | TED*. Retrieved May 30, 2023 from https://www.youtube.com/watch?v=hJP5GqnTrNo
[81] 2023. Artists face a choice with AI: Adapt or become obsolete. *Cointelegraph*. Retrieved August 28, 2023 from https://cointelegraph.com/news/artists-face-a-choice-with-ai-adapt-or-become-obsolete
[82] 2023. If art is how we express our humanity, where does AI fit in? *MIT News | Massachusetts Institute of Technology*. Retrieved August 28, 2023 from https://news.mit.edu/2023/generative-ai-art-expression-0615
[83] Meta AI is sharing OPT-175B, the first 175-billion-parameter language model to be made available to the broader AI research community. Retrieved June 14, 2022 from https://ai.facebook.com/blog/democratizing-access-to-large-scale-language-models-with-opt-175b/
[84] Cyc | The Next Generation of Enterprise AI. Retrieved December 5, 2022 from https://cyc.com/
[85] Two-Sample t-Test. Retrieved November 6, 2023 from https://www.jmp.com/en_us/statistics-knowledge-portal/t-test/two-sample-t-test.html
[86] GitHub Copilot · Your AI pair programmer. *GitHub*. Retrieved February 18, 2023 from https://github.com/features/copilot

Appendix A: ChatGPT on Transformers

Conversation with ChatGPT, 02/18/2023

Tix: What is a Transformer?

ChatGPT: A Transformer is a type of neural network architecture that was introduced in a seminal paper by Vaswani et al. in 2017. Transformers are widely used in natural language processing (NLP) tasks, such as language translation, text summarization, and question-answering.

Unlike traditional neural networks, which process input data sequentially, Transformers are designed to process entire sequences of input data simultaneously. They use a self-attention mechanism that allows the network to weigh the importance of different parts of the input sequence when making predictions.

The Transformer architecture consists of an encoder and a decoder. The encoder takes an input sequence and generates a sequence of hidden states, which are then passed to the decoder. The decoder generates an output sequence based on the hidden states and an input sequence of tokens that it receives as input.

Transformers have been shown to achieve state-of-the-art performance in a wide range of NLP tasks and have become a fundamental building block for many advanced NLP models.

Tix: Is this the only meaning of "transformer?"

ChatGPT: No, the term "transformer" has multiple meanings depending on the context. In addition to the neural network architecture, some other meanings of "transformer" include:

1. Electrical transformer: This is a device that is used to transfer electrical energy from one circuit to another through electromagnetic induction. Transformers are commonly used in power transmission and distribution systems to change the voltage level of the electricity.
2. Mechanical transformer: This refers to a device that is used to transfer energy from one physical system to another, such as a gearbox or a pulley system.
3. Action figure: In popular culture, a "transformer" can refer to a toy or action figure that can transform from one form to another, such as the popular Transformers franchise.

So, the meaning of "transformer" can vary depending on the context in which it is used.

Appendix B: Sample Conversations with Skilled vs Compliant Responses

Below are two entirely fictional conversations between a human or AI developer and a client, illustrating the difference between a developer who improves the system through asking clarifying questions from their own knowledge and experience, vs a developer who is merely compliant with the request as it is received. These conversations are entirely fictional and intended for illustrative purposes only. In both cases, the scenario is a client requesting a change to the Online Registration system for a school district. Although the conversations are fictional, the scenario is drawn from my personal experience working for the Anchorage School District.

Compliant Conversation:

Client: “The ‘Contact Information’ page needs to show an error message whenever a parent changes their parent ID.”

Developer: “OK, here is modified code for the Contact Information Page: …” *[produces a modified Contact Information page which will show an error message if the Parent ID field is modified, and may or may not stop the parent from continuing registration with the modified ID after the error message is shown*

Conversation with Clarifying Questions:

Client: “The ‘Contact Information’ page needs to show an error message whenever a parent changes their parent ID.”

Developer: “What should the error message say?”

Client: “Something like ‘You cannot modify your parent ID! This is for school use only, please do not modify!”

Developer: “Why are we even showing the field if you don’t want them to modify the value?”

Client: “We still want parents to be able to see their ID, we just don’t want them to change it.”

Developer: “I can easily make the ID displayed but not editable, so nobody can change it in the first place.”

Client: “You can do that? Perfect!”

Developer: [produces a modified Contact Information page in which the Parent ID field is shown as a label instead of a text box, eliminating the need for an error message. This is not what the client asked for and does not fulfil their original request, but it fulfills their actual needs better than their original request.]

Appendix C: GPT 4 “What is a transformer?”

Conversation conducted in ChatGPT Plus (GPT 4.0 mode) on 7/7/2023 by Bernadette Tix

Me: What is a transformer?

GPT4: "Transformer" in the context of AI and machine learning refers to a type of model architecture that is primarily used in the field of natural language processing (NLP). It was introduced in a paper titled "Attention is All You Need" by Vaswani et al., from Google, in 2017.

The Transformer model is based on the concept of self-attention (also known as scaled dot-product attention or sometimes multi-head attention), where it calculates the relevance of each word in the context of every other word in the data sequence. Unlike RNNs (Recurrent Neural Networks) and CNNs (Convolutional Neural Networks), Transformers do not process data sequentially, instead they process all data points in parallel, which makes them highly suited for parallel computing environments.

The model consists of an encoder and a decoder. The encoder reads and interprets the input data, and the decoder generates a prediction or output based on that interpretation.

A significant advantage of the Transformer architecture is its ability to handle long-range dependencies in data, which is a challenge in many sequence-to-sequence tasks. This property has led to its use in many state-of-the-art models in NLP like BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer), and others.

This model architecture has become a cornerstone of many advancements in NLP, and has been used to achieve state-of-the-art results on a variety of tasks, including machine translation, text summarization, and sentiment analysis.

As of my knowledge cutoff in September 2021, Transformers are considered one of the most influential innovations in the field of NLP.

Appendix D: Example of Asking Clarifying Questions in GPT 3.5 and GPT 4

I made a series of attempts to get a working version of the proposed prompt engineering. The conversation shown below is my fourth attempt, which I considered to be successful enough to include here. I have included my full notes explaining my design considerations and the template used. The "prompt engineering considerations" presented at the top are the result of insights from the first three, less successful attempts.

This proof of concept was done before the research question for this proposal had been properly narrowed down. It thus shows a slightly different scenario than is being addressed in the current proposal: namely, asking clarifying questions to resolve the ambiguity in an ambiguously-worded question.

Prompt Engineering Considerations

- GPT needs a lot of guidance in terms of tone and context. Remember that it is general-purpose and can do much more than just answer questions in a conversational tone. Let it know what you want it to do!
- Since the conversation needs to continue forwards after the last prompt, make sure there is enough prompt engineering to encourage this. Bear in mind that the beginning of the conversation is more structured / predictable (user asks question, GPT asks a clarifying follow-up, user offers clarification) but as the conversation progresses the structure becomes less and less predictable. GPT needs to transition smoothly from having lots of prompt engineering at the beginning to little or no prompt engineering as the conversation progresses.
- Conversations are dynamic. Don't try to plan ahead too much or get bogged down asking for more than you need in the beginning. For me this is a process instinct from SQL optimization, but the opposite applies here. Try to keep each action as granular as possible.
- GPT's answers have been overly long and dull. Specifically request concise responses.
- In Attempt 3, GPT responded "clear" even though it can clearly identify ambiguity with other prompts. For now we won't do the decision tree. Let's get it working without that part first.
    - *(In attempt 3 I attempted to make ChatGPT decide for itself whether it needed to ask a question or not. I would still like to develop this capability but in attempt 3, it backfired, and ChatGPT decided it did NOT need to ask a follow up, and provided an answer about the machine learning architecture. I decided to focus on other aspects of the design for the initial proof of concept and leave the decision-making aspect for a later date.)*

Template

//Initial values are a user's name <username> and prompt <X>

1) Consider the following prompt: "<*X*>" Is there any ambiguity in this prompt that could lead to confusion? Provide a brief summary of the ambiguity.

2) Respond as though this question was just asked by a user named <*username*>. Start with a friendly greeting, then ask a respectful and concise question to clarify the user's needs, keeping in mind the ambiguity you just identified.

//GPT asks a question. Store user response as <Y>

3) This is the user's response. Continue the conversation in respectful, friendly, conversational tone. Interpret any further responses as coming directly from the user and proceed with the conversation normally.
<username>: <Y>

Prompt

What is a Transformer?

- Provide follow-up responses about looking for a birthday gift for my nephew. Ideally the system will hone in on the user being interested in transformer toys rather than the ML architecture or electrical transformers, and provide useful answers.

Results:

Conversation conducted in ChatGPT (free version) on 7/6/2023 by Bernadette Tix

> ***Me:*** *Consider the following prompt: "What is a Transformer?" Is there any ambiguity in this prompt that could lead to confusion? Provide a brief summary of the ambiguity.*
>
> ***GPT:*** *The prompt "What is a Transformer?" could potentially have some ambiguity that may lead to confusion. The term "Transformer" can have different meanings depending on the context in which it is used. Here are a few possible interpretations:*
>
> *1. Electrical Transformer: In the field of electrical engineering, a transformer is a device used to transfer electrical energy between two or more circuits. It is commonly used to increase or decrease voltage levels. If someone familiar with this specific context reads the prompt, they may assume that it refers to an electrical transformer.*
>
> *2. Transformer (Machine Learning): In the domain of machine learning and natural language processing, a Transformer refers to a specific type of neural network architecture introduced by the paper "Attention Is All You Need" in 2017. It has since become a popular model for various tasks, including language translation, text generation, and understanding contextual relationships. If someone from this field encounters the prompt, they would likely interpret it as referring to the Transformer model in machine learning.*
>
> *3. Transformers (Toy Franchise): Transformers is also a popular toy franchise, cartoon series, and movie series featuring sentient robots that can transform into vehicles or other objects. If someone is familiar with this franchise, they may assume that the prompt is referring to the Transformers toys, characters, or media.*
>
> *Given these different interpretations, it is crucial to provide additional context or specify the intended meaning to avoid confusion when using the term "Transformer" in a prompt.*

***Me**: Respond as though this question was just asked by a user named Bernadette. Start with a friendly greeting, then ask a respectful and concise question to clarify the user's needs, keeping in mind the ambiguity you just identified.*

Bernadette: "What is a Transformer?"

***GPT:** Assistant: Hello, Bernadette! I hope you're doing well today. When you mention a "Transformer," could you please clarify the context you are referring to? Are you asking about electrical transformers, the Transformer model in machine learning, or the Transformers toy franchise? This will help me provide you with the most relevant and accurate information.*

***Me:** This is the user's response. Continue the conversation in respectful, friendly, conversational tone. Interpret any further responses as coming directly from the user and proceed with the conversation normally.*

Bernadette: "Tell me about the toy."

***GPT:** Assistant: Absolutely, Bernadette! Let's talk about the Transformers toy franchise. Transformers is a popular franchise that originated as a line of toys in the 1980s. It has since expanded into animated television series, movies, and various other media. The toys themselves are action figures that can transform from robots into vehicles, animals, or other objects. Each character typically has a unique backstory and abilities.*

The Transformers franchise has a wide range of characters, including iconic ones like Optimus Prime, Bumblebee, Megatron, and Starscream. These characters engage in epic battles between the Autobots (the heroic robots) and the Decepticons (the villainous robots) for control over their home planet, Cybertron, or sometimes even on Earth.

The popularity of the Transformers toys stems from the excitement of transforming the figures between their robot and alternate forms, allowing for creative play and storytelling. Over the years, the franchise has evolved and expanded, with new generations of toys, TV shows, and movies captivating fans of all ages.

Is there anything specific you would like to know or discuss about the Transformers toy franchise? Feel free to ask any questions you may have!

***Me:** My nephew wants a transformer for his birthday. Where can I find a transformer toy for him?*

***GPT:** Assistant: That's great to hear that your nephew is interested in Transformers! Finding a Transformer toy for his birthday should be an exciting quest. There are several places where you can look for Transformer toys:*

1. Toy Stores: Many toy stores carry a wide range of Transformer toys. You can check popular retail chains like Toys "R" Us, Walmart, Target, or any local toy stores in your area. They often have dedicated sections for action figures and popular toy franchises.

2. Online Retailers: Online shopping platforms like Amazon, eBay, and Walmart's online store have a vast selection of Transformer toys. You can browse through their websites, search for specific Transformer characters or toy lines, and read customer reviews to make an informed choice.

3. Official Transformers Website: The official Transformers website (transformers.hasbro.com) is a great resource to explore the latest offerings in the Transformer toy line. It provides information on new releases, collector's items, and where to purchase them.

4. Specialty Collector's Shops: If you're looking for specific or rare Transformer toys, specialty collector's shops that focus on action figures and collectibles might be worth exploring. These shops often have a curated selection of Transformers toys and may have unique or limited-edition items.

Before making a purchase, it's a good idea to consider your nephew's age and interests. Transformers toys come in various sizes, complexity levels, and price ranges, so you can choose the one that suits him best.

I hope these suggestions help you find the perfect Transformer toy for your nephew's birthday celebration! If you have any more questions or need further assistance, feel free to ask.

Notes: I think this is good enough for a proof-of-concept. Things to consider for future attempts:

- GPT is responding as “Assistant” completely unprompted. We may want to specify a name for GPT.
- I still would like these responses shorter. GPT tends to meander. However, that seems more like system design / UX optimization. May not be needed for research.
- Still need to tune the system to get a decision tree working so it can ask multiple questions if needed. May be a decision tree, may just be different language in the prompt.
- I need to figure out if the API “remembers” past prompts the way ChatGPT does. Attempt 4 didn’t include the whole response about ambiguity and just relied on ChatGPT to remember, but this may not be practical in the final version.

Testing in GPT 4

I also tested the same template/prompt combination in GPT 4 and got similar, arguably better results:

Conversation conducted in ChatGPT Plus (GPT 4.0 mode) on 7/7/2023 by Bernadette Tix

> ***Me:** Consider the following prompt: “What is a Transformer?” Is there any ambiguity in this prompt that could lead to confusion? Provide a brief summary of the ambiguity.*

GPT4: *Yes, there is ambiguity in the prompt "What is a Transformer?" since the term "Transformer" has several different meanings depending on the context.*

*1. **Electrical Engineering**: In the field of electrical engineering, a transformer is a device that transfers electrical energy between two or more circuits through electromagnetic induction.*

*2. **Machine Learning**: In the field of machine learning, particularly in natural language processing (NLP), a Transformer is a type of model architecture introduced in a paper called "Attention is All You Need". It has been instrumental in the development of state-of-the-art models like GPT-3, GPT-4, and BERT.*

*3. **Entertainment**: "Transformers" is also the name of a popular media franchise that started with a line of toys and expanded into TV shows, comics, and films. The Transformers are fictional robotic characters capable of transforming into various objects, commonly vehicles.*

Without specifying the context, it's impossible to definitively answer the question "What is a Transformer?"

Me*: Respond as though this question was just asked by a user named Bernadette. Start with a friendly greeting, then ask a respectful and concise question to clarify the user's needs, keeping in mind the ambiguity you just identified.*

Bernadette: "What is a Transformer?"

GPT4: *Hi Bernadette,*

Nice to meet you! When you ask about "Transformer", are you interested in the concept from machine learning, the electrical device used in power transmission, or the characters from the popular entertainment franchise? Your clarification will help me provide a more accurate response.

Me: *This is the user's response. Continue the conversation in respectful, friendly, conversational tone. Interpret any further responses as coming directly from the user and proceed with the conversation normally.*

Bernadette: "I don’t know. My eight-year-old nephew keeps saying he wants a transformer for his birthday.”

GPT4: *Hi Bernadette,*

Thank you for clarifying! It sounds like your nephew is referring to "Transformers" the toy line, which is based on a popular entertainment franchise. These toys are a favorite among kids as they're robots that can be transformed into vehicles, animals, and other forms. There are many different Transformers toys available, based on various characters from the franchise. You might *want to find out if he has a favorite character or series within the Transformers universe to make sure you get a gift he'll love. Let me know if you need any further assistance in choosing the perfect gift!*

Me: *Thanks! Which transformer is the most popular?*

GPT4: *Among the Transformers franchise, the two most iconic characters are probably Optimus Prime and Bumblebee.*

*1. **Optimus Prime**: He is the leader of the Autobots, the group of Transformers who are the main protagonists of the series. Optimus Prime is known for his wisdom, courage, and exceptional leadership skills. His toy often transforms into a semi-truck.*

*2. **Bumblebee**: Bumblebee is another popular character, known for his bravery and loyalty. His character appeals to a younger audience due to his more playful and friendly nature. Bumblebee usually transforms into a compact car, often a Volkswagen Beetle or a Chevrolet Camaro in the newer movies.*

These two are generally popular among fans, but there's a vast array of other characters to choose from depending on your nephew's specific interests. I'd suggest asking him if he has a favorite character or checking with his parents or guardians to be sure.

Notes: These answers are even better. It doesn't have the awkward "assistant" tag, and it gave shorter and more helpful answers. AND it transitioned perfectly back to a normal conversation. As far as proof of concept goes, I call this a complete success.

Appendix E: Webapp text and prompts

Text in red are notes for clarification and will not be shown to volunteers. After the volunteer has completed their consent form, they will be shown a series of webpages navigated by clicking a "next" button at the end of each page.

Page 1: Demographics

Part 1 of 4: Demographics

Please answer the following demographic questions about yourself.

- Age
    - [Numerical Input]
- Gender
    - "Female"
    - "Male"
    - "Other / Nonbinary"
- "What is your prior experience with generative AI such as ChatGPT, Bard, or similar programs?"
    - "I use generative AI regularly."
    - "I have used generative AI before, but not often."
    - "I have never used generative AI before."
- "Is English your primary spoken language?"
    - "Yes"
    - "No"

Click "Next" to continue.

Page 2: Interaction with AI

Part 2 of 4: Talking to the AI

On this page, you will be communicating with an AI that is capable of writing short documents such as letters, memos, emails, and short reports. Please think of a document you would like the AI to create for you. This could be a document you actually need, or one that you have just made up for the experiment. Either way, please think in detail about what you would need this document to include. When you are ready, write out what you need in the textbox below.

Example prompts:

"Please write a cover letter for a job working in tech support for a college student with one year of prior tech support experience."

“Write an email to my boss asking them if they can meet next week to discuss a project which is behind schedule.”

“I want to reach out to a friend I haven’t spoken to in a while. Give me an outline for a letter to help me figure out what I should say.”

[A textbox will be provided for the user to enter their prompt. The prompt will be sent to an AI which will then ask follow-up questions, and the user will answer by typing their response into the same textbox. All entries from the user and responses from the AI will be displayed on the screen as an ongoing text-chat.]

Page 3: Rating the responses.

Two different outputs from the AI will be displayed. One output will be the answer to the user’s original prompt, taking all the questions and answers into account. The other will be the AI’s answer to the original prompt as if the questions and answers had never happened. The order of the outputs will be randomized. The same explanation and questions will be shown for each output.

Part 3 of 4: Rating the Responses

Thank you for your thoughtful answers! The AI has considered your prompt and your responses to its questions, and has generated two different possible outputs for you. Please read each one and answer the questions that follow.

Output #1

[One of the two outputs will be shown]

- How close is this document to what you hoped for when you made your initial request?
    - Very close to what I was hoping for.
    - Somewhat close to what I was hoping for.
    - A little bit like what I was hoping for.
    - Not very close to what I was hoping for.
    - Not at all what I wanted.
- How useful would this document be to you?
    - I could use this document as-is.
    - I could use this document with minimal modification.
    - I could use this document with substantial modification.
    - This document could be used as a general starting point but requires major revisions to be usable.
    - This document is not usable at all.
- How would you rate the overall quality of this document?
    - Excellent quality.

- Above average quality.
- Average quality.
- Below average quality.
- Poor quality.

Output #2

- *[The other output will be shown, followed by the same set of questions]*

Page 4: Exit Survey

Thank you for rating the AI's responses. Would you like to ask the AI to create another document for you, or proceed to the exit survey?

If the user selects to create another document, they will be returned to Page 3 and allowed to complete that portion of the experiment as many times as they would like. If they proceed to the exit survey, the exit survey will be shown:

Part 4 of 4: Exit Survey

Please rate the following statements on a scale of "Strongly Agree" to "Strongly Disagree" *(Each of the following statements will be shown with 5 options: Strongly Agree, Slightly Agree, Neutral, Sightly Disagree, Strongly Disagree)*

- It was annoying to have to answer questions even though I had already explained what I wanted the AI to do.
- I felt like the AI was more engaged with my problem because it asked follow-up questions.
- I would be willing to answer follow-up questions from an AI if answering questions led to better results.
- I liked that the AI showed me two options to pick between, instead of only picking the option it thought was best.

Do you have any additional feedback or comments (optional)?

[A free-text entry will be provided.]

After clicking "next," an exit message will be shown.

You have completed the experiment! Thank you for your participation. If you have any further questions, please contact Bernadette Tix at bjavery@hawaii.edu. Please close your browser to exit.