Towards AI-Assisted Protocol Analysis in Design Research: Automating Question Labelling with GPT-4 According to Eris' (2004) Taxonomy

Ahmed Shahriar Sakib, Ada Hurst, Frank Safayeni
University of Waterloo, Canada
adahurst@uwaterloo.ca

This study explores the potential of large language models (LLM)-based tools, specifically GPT-4 – a state-of-the-art language processing model - to assist in the analysis of verbal protocols of design. We focus on Eris' taxonomy, a well-established framework that classifies questions asked by participants in a design-focused task according to three broad categories: low-level, deep reasoning, and generative design questions. Using a large dataset of pre-classified questions from design review meetings, a series of experiments test GPT-4's capability in the categorization task and evaluate how different factors influence its precision. Results indicate that GPT-4 matches performance by human coders – a promising result for design researchers who can benefit from this tool with little prior natural language processing expertise. Overall, findings offer insights into the strengths and limitations of LLMs in this context and suggest directions for future research into the use of LLM-based tools in qualitative analyses of design activity.

Introduction

Protocol analysis [1] is a widely utilized method for studying cognitive processes. It holds particular significance in the context of research into design activity, where it has been used widely to study people's thinking and behavior while engaged in design tasks, as well as for evaluating the effect of supports and interventions on the design process. For example, recent

Design Computing and Cognition DCC'24. J.S. Gero (ed),
pp. xx-yy.

reviews describe broad use of protocol analysis in studying conceptual design [2], problem framing [3], and design learning [4].

At a high level, the approach involves three main activities: capturing video and audio recording of participants engaged in a design activity, 2) transcribing of participants' verbalizations, and 3) segmenting and coding the transcripts to generate data which is analyzed to extract insight. All three stages have traditionally presented challenges for researchers [5], however, over the decades, technological developments have significantly improved the efficiency and effectiveness of this process. For example, the ubiquity of digital cameras and the ease with which virtual meetings can be held and recorded have facilitated more and better data collection. Similarly, digital transcription services, which are increasingly AI-driven, have significantly lowered the cost and time of producing high-quality transcripts. Finally, advances in natural language processing (NLP) have enabled new and more efficient ways to analyze verbal protocols of design (e.g., [6], [7], [8]).

More recently, Large Language Models (LLMs) have shown promise for revolutionizing the way researchers analyze and interpret textual data [9], [10]. Powered by advanced NLP techniques, LLMs can comprehend and generate human-like text, offering researchers a versatile means of exploring and understanding qualitative information. These models, often pre-trained on vast corpora of diverse language data, can capture intricate linguistic nuances, making them particularly adept at uncovering underlying patterns, sentiment, and context within qualitative data [11].

As we delve deeper into the landscape of language models, one standout example is ChatGPT (Chat Generative Pre-trained Transformer), a conversational variant of LLMs built by OpenAI. Its ability to engage in dynamic and context-aware conversations opens new avenues for researchers to interact with and explore qualitative data in a more conversational manner. ChatGPT relies on a pre-trained generative language model that generates responses by identifying patterns provided by the user as input [12]. Hence, the language model essentially assigns probabilities to each word within a vocabulary that could follow a specified input sequence. These word embeddings [13] are developed through artificial neural networks, learning a probability distribution from provided texts in an unsupervised manner, i.e., without the need for additional human input or labeling. The production of the output sequence considers the tokens from the input sequence, their positions, and the previously generated output. This process is referred to as autoregressive generation [14]. Large sets of training data are required to train these LLMs. As these models have grown in sophistication and depth, their ability to generate coherent and contextually relevant text has advanced significantly.

GPT variants (e.g., GPT-3.5, GPT-4) are now being tested in various inductive and deductive qualitative analyses applications with moderate success. For example, Hamilton et al. [15] explore the use of ChatGPT for performing thematic analysis on interview data and found some overlap between human and AI-generated themes. Similarly, De Paoli [16] show how GPT3.5 can be used to partially replicate the steps in inductive thematic analysis prescribed by Braun and Clarke [17]. Xiao et al. [18] explore LLM for deductive analyses and use GPT-3 in conjunction with an expert-drafted codebook to complete a coding task and report "fair to substantial agreement" with expert results. Early reports suggest that AI performs better on deductive than inductive analyses [19].

Aims and scope

This research is broadly motivated by the goal of determining the extent to which LLMs can assist in the process of analyzing verbal protocols of design. This study, specifically, aims to address this aim in the context of one existing framework for analyzing verbal protocols of design: Eris' [20] question asking taxonomy. Eris' interest in studying question asking in design is based on the premise that question asking during designing influences how designers think, including how they think creatively, make decisions, and learn [20, p. 11]. His working definition of a question in a design context is "*a verbal utterance related to the design tasks at hand that demands an explicit verbal and/or nonverbal response*" [20, p. 36]. His taxonomy builds on prior existing taxonomies of question asking, primarily work by Lehnert [21] and Graesser [22], [23] and classifies questions into one of three broad categories:

- Low-Level Questions (LLQs), where the asker is seeking clarification or obtaining missing information (e.g., "*Is this manual pushing or motor pushing?*")
- Deep Reasoning Questions (DRQs), where the asker is seeking to establish causal explanations of phenomena (e.g., "*How was the candy being ejected from your machine?*") The premise of DRQs, as in the case of LLQs, is that the answer to the question exists and is known by the question receiver.
- Generative Design Questions (GDQs), where the asker is seeking to generate different alternative known and unknown options for how to address a goal or obstacle (e.g., "*How do we animate this*?").

Within each of these broader categories, there are several clearly defined sub-categories (see appendix in [24]). This taxonomy provides a useful lens

for uncovering patterns of convergent (LLQ, DRQ) and divergent (GDQ) thinking in design and has been widely used as a coding scheme of verbal protocols of design, to study various phenomena and processes including idea generation [24], design reviews [25] [26], peer feedback [27], and the role of expertise in feedback [28]. Further, the deductive nature of the coding task was judged to be well-suited to the GPT tool.

Within this scope, the research question this study seeks to answer is: *How accurately can GPT-4 (latest GPT variant in January 2024) classify questions in verbal protocols of design according to Eris' (2004) taxonomy?*

Method

Dataset

The study employed a dataset from a prior publication [28]. In that research, teams of two to three students enrolled in a graduate engineering design program met with their tutor weekly over six weeks to discuss a term-long design project, where they were challenged to design an enhanced automatic candy-wrapping machine for a small manufacturer of sweets. Audio recordings of each session were transcribed with an AI-based tool (Otter) and manually checked to ensure quality. Then, two research assistants independently identified and coded questions in the transcripts using Eris' [20] taxonomy. Any coding disagreements were resolved with the assistance of a third, more experienced, researcher. The dataset (Table1) consisted of a total of 31 transcribed tutoring sessions, with each speaker's utterances labelled and timestamped. In addition, all question utterances (2203 in total) were identified and coded (labelled) as LLQ, DRQ, or GDQ.

Dataset

At the time of this study (January 2024), GPT-4 was the latest iteration in OpenAI's lineup of generative pre-trained transformers, representing an improvement not only in its scale but also in its underlying architecture and capabilities. The experiments were carried out utilizing the official Python Library for the OpenAI chat completion API [29]. GPT-4 provides text outputs in response to its inputs which are also referred to as "prompts". The prompt is divided into two segments: system message and user message. The system message is used to set the behavior or context for the conversation. It helps set the tone for the interaction and provides high-level context that the model can use to better understand the user's intent or the desired outcome. On the other hand, user messages are the inputs or prompts

given by the user to the model. GPT-4 Turbo [30] model was used for all the experiments.

Table 1 Dataset characteristics. For each session (S) and group (G), we include session length in mins and (words), and distribution of questions by category – LLQ, DRQ, GDQ. E.g., the cell S1, G1 denotes first session details for group 1: 67.50 minutes, 6472 words, and the question distribution – 95 LLQ, 9 DRQ and 25 GDQ

	S1	**S2**	**S3**	**S4**	**S5**	**S6**
G1	67.50 (6472) [95, 9, 25]	38.03 (1945) [45,7, 11]	20 (255) [5, 1, 5]	94.93 (10299) [131,11, 44]	72.90 (4852) [77, 8, 23]	n/a
G2	65.37 (19369) [51, 0, 32]	72.22 (21112) [55, 2, 36]	66 (18991) [50, 5, 42]	83.93 (28539) [70,4,36]	73.92 (9191) [49, 5, 20]	n/a
G3	45.80 (10712) [54, 6, 13]	23.58 (2346) [27,1, 8]	38.22 (5187) [36,6,18]	31.20 (3838) [26,8,15]	30.07 (3110) [29, 4, 11]	n/a
G4	50.45 (26600) [60, 15, 12]	66.92 (21760) [68,1, 22]	66.58 (8994) [64, 15, 8]	58.27 (4048) [45,15,4]	72.17 (21551) [58, 13, 25]	n/a
G5	99.15 (13690) [29, 1, 28]	50.52 (3364) [25,5, 10]	66.13 (3392) [23, 2, 13]	78.10 (8241) [26,3,17]	70.45 (14274) [22,14, 19]	68.68 (7454) [17,2,19]
G6	42.62 (2023) [38, 5, 2]	38.30 (1878) [17,1, 13]	22.82 (269) [11, 2, 1]	10.75 (382) [12,2,2]	11.75 (559) [7, 1, 2]	n/a

The system message (shown below) provided an overview of the chosen taxonomy, including both high-level definitions of each of the major categories (LLQ, DRQ, GDQ), as well as definitions and examples of each of the sub-categories, copied from Appendix 1 in [24].

You are an expert in protocol analysis of design activity. You can perform analysis of the textual data and are able to label question utterances according to Eris' (2004) taxonomy. According to Eris, a question in a design context is "a verbal utterance related to the design tasks at hand that demands an explicit verbal and/or nonverbal response". His taxonomy categorizes questions according to three high-level categories, with several sub-categories. The categories with examples are shown below:

Low Level Questions
Low-level questions are primarily information-seeking questions and they are formulated when the questioners want clarification about a given topic/event or are trying to obtain missing information. Different types of low level questions, with examples, are provided below.
<LLQ types and examples from Appendix 1 of Cardoso et al., 2016>

Deep Reasoning Questions (DRQ)

Low-level (LLQ) and Deep Reasoning Questions (DRQ) share the common premise that a specific answer, or a specific set of answers, exists. As the purpose of these questions is either to seek for information (i.e. low level questions) or to establish causal explanations of phenomenon (i.e. deep reasoning questions), they facilitate convergent thinking processes. Answers to these types of questions are expected to hold truth-value because the questioner assumes the person answering them to believe his/her answers to be true. Different types of Deep Reasoning Questions, with examples, are provided below.
<DRQ types and examples from Appendix 1 of Cardoso et al., 2016>

Generative Design Questions (GDQ)
Questions that are raised in design situations can operate quite differently from low-level or deep reasoning questions. Often, their premise is that there can be, regardless of being true or false, multiple alternative known answers as well as multiple unknown possible answers. The questioner's intention is to disclose the alternative known answers, and to generate the unknown possible ones. Such questions are characteristic of divergent thinking, where the questioner attempts to move away from the facts to the possibilities that can be generated from them. There are five GDQ categories:
<GDQ types and examples from Appendix 1 of Cardoso et al., 2016>

The system was further fine-tuned through the *seed* and *temperature* settings. The seed setting [31] ensures that the system will make a best effort to sample deterministically, such that repeated requests with the same seed and parameters should return the same result. However, determinism is not guaranteed due to the frequent updates made to model configurations and system settings by OpenAI. The temperature setting [32] controls the variation in the output text generation. A higher temperature (e.g., 0.8) leads to more random and diverse responses, while a lower temperature (e.g., 0.2) produces more deterministic and focused outputs. The seed was set to a constant number across all the experiments and the temperature was set to 0 so that the model would adhere closely to the input instructions and generate output with minimal variation. The user prompts varied with each experiment and expected outcomes.

Tasks

A series of experiments were conducted to gradually determine the effectiveness of GPT-4 in labelling questions. Figure 1 describes the overall workflow comprising the four experiments, including high-level objectives and main findings. Experiment 1 aimed to determine a baseline performance of the classification task, with and without a training set. Experiment 2 aimed to clarify the impact of the size of the training set on performance. Given the probabilistic nature of GPT's performance, Experiment 3 examined the impact of running experiment multiple times. Finally, in a series of three parts (a-c), Experiment 4 explored the impact of context surrounding a question on classification performance. Each experiment and their results are detailed in the following section.

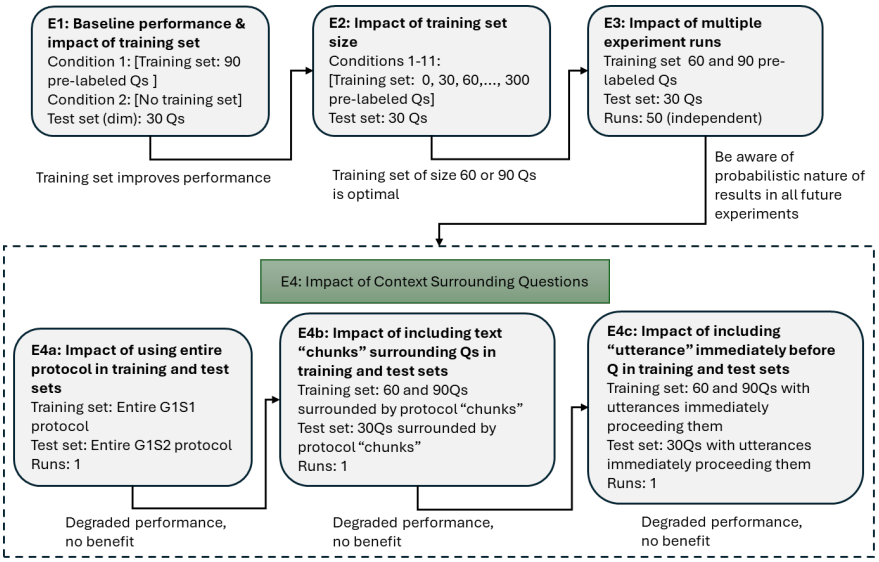

Fig. 1 Overview of experiments (E)

Experiments and results

Experiment 1

Experiment 1 aimed to provide a baseline of GPT-4's effectiveness in labelling questions and of the value of training data in this task. In the first condition, a training set was given consisting of 90 stand-alone human-labelled questions, uniformly distributed between LLQ, DRQ, and GDQ types. In the second condition, no labelled examples (training set) were provided. The classification task was conducted on a new set of 30 question utterances (testing set). Testing set questions were randomly selected from the data, and uniformly distributed between LLQ, DRQ, and GDQ types. Both sets were sampled from all the sessions combined. The prompt was:

> *Classify each of the questions below, delimited by with triple backticks, using the taxonomy proposed by Eris. Label each question with one of the three categories: Low-level questions, Deep Reasoning Questions, or Generative Design Questions. Delimit the label with triple backticks. State your reasoning for the assigned label. Format the result as a markdown table.*
> ``` *<test set>* ```
> *To help you categorize the questions above, here are some examples delimited by triple backticks, each line contains an example that has two segments - question and category separated by colon (:)*
> *<training set>*

Table 2 contrasts the results of the labelling task under the two conditions (without and with training) with the human-labelled set. GPT-4 generated labels aligned more closely with the human labels when a training set was provided (83% alignment) compared to when it was not (60% alignment). As such, all future labelling tasks (Experiments 2 to 4) use a training set.

Table 2 Experiment 1 results. Labels assigned by: "H" → Human; "GPT/woT" → GPT-4, without a training set; "GPT/wT" → GPT-4, with a training set

Testing set: question number and utterance	H	GPT /woT	GPT /wT
1. So how do you know how much do?	DRQ	LLQ	DRQ
2. …what are the possible contributors?	GDQ	GDQ	GDQ
3. Why would you put that on there if you could swap in the new roll in a minute?	DRQ	DRQ	DRQ
4. What about cleats, right cleated conveyor cleats?	GDQ	GDQ	GDQ
5. But so it's full 15 or whatever, probably right?	LLQ	LLQ	LLQ
6. Could you take the candy and slide it on the wrapper?	LLQ	GDQ	GDQ
7. Why would I do that?	DRQ	DRQ	DRQ
8. How they wrap candy currently?	DRQ	LLQ	DRQ
9. How does the very last one behave?	GDQ	LLQ	DRQ
10. But this whole thing has to rotate, right?	LLQ	LLQ	LLQ
11. So when they're wrapping in one of these candies, have you seen the whole process?	LLQ	LLQ	LLQ
12. Yeah, but I think, isn't it the same?	GDQ	LLQ	LLQ
13. So is the main assumption right now that the wrapper will stick?	LLQ	LLQ	LLQ
14. How are you going to intersection things that are exactly the same…?	GDQ	GDQ	GDQ
15. How are you going to keep rolling and pulling it?	GDQ	GDQ	GDQ
16. How was the candy being ejected from your machine.	DRQ	LLQ	DRQ
17. Can I do this stuff?	LLQ	LLQ	LLQ
18. Makes sense? Right?	LLQ	LLQ	LLQ
19. What kind of processes out there now?	DRQ	LLQ	DRQ
20. Because these paper will come on to the stage right?	LLQ	LLQ	LLQ
21. …Why did you put grid?	DRQ	DRQ	DRQ
22. What function is providing?	DRQ	LLQ	LLQ
23. What's gonna happen?	GDQ	DRQ	GDQ
24. So what's happening here at the end of this conveyor belt, tell me a little bit more …What would happen in terms of how this thing gets pushed in the plastic?	DRQ	LLQ	DRQ
25. …What is the purpose of this box?	DRQ	LLQ	LLQ
26. And is this manual pushing? Or .. a motor pushing?	LLQ	LLQ	LLQ
27…How are you going to then finalize the closure?	GDQ	GDQ	GDQ
28. So how do we animate this?	GDQ	GDQ	GDQ
29. Or is it something that we can outsource to a company?	GDQ	LLQ	GDQ
30. Do you want to a machine to produce 15 to 20 pieces of candy per min and currently they're doing what is it?	LLQ	LLQ	LLQ

We further scrutinized the five questions (6, 9, 12, 22, and 25) for which the GPT-4-assigned labels didn't match the human label. For those questions, the authors believe that the GPT-4-assigned labels are correct when considering the question in isolation, as written. The discrepancy may be due to several reasons, beyond a simple deficiency in the tool or a mistake in the labelling by the human. In making their judgement, the human coders are also considering the larger context of the conversation, including what is discussed prior to the question, and the answer that immediately follows it. The impact of context is further explored in Experiment 4.

Experiment 2

Experiment 2 sought to determine the effect of the size of the training set on the accuracy of labelling by GPT-4. The labelling task described in Experiment 1 was repeated under different conditions of the size of training set, which was varied from 0 (i.e., no training) to 300 pre-labelled questions, in increments of 30. The testing set was kept constant and identical to the one in Experiment 1. Note that the cases of training sets of sizes 0 and 90 correspond to the exact cases tested in Experiment 1. Table 3 presents the results of the labelling task for each question, and for each training-set size condition. Overall, it appears that while a training set improves alignment with human-generated labels, no accuracy improvements are achieved when the size is increased past 90 questions.

Table 3 Experiment 2 results

	Size of training set, in number of pre-labelled questions										
	0	**30**	**60**	**90**	**120**	**150**	**180**	**210**	**240**	**270**	**300**
Alignment (%)	60	67	83	83	80	83	83	73	70	73	83

Experiment 3

Although utilizing a set *seed* and *temperature* is expected to produce consistent output, GPT-4 responses can still differ due to the inherent nondeterministic nature of the model [33]. Experiment 3 sought to determine the sensitivity of the results across multiple "runs" of the experiment. Using the training set sizes of 60 and 90, as determined from Experiment 2, the labelling task was repeated once again, independently run 50 times, for each case. Table 4 presents the aggregated labelling results, by question type. For example, a score of 0.5 indicates that GPT-4 assigned that label in half of the 50 runs. Any GPT- 4 labels that do not align with the human label are highlighted in red. In the case of "split" labels, if one of the two labels matches the human label, they are highlighted in orange.

Table 4 Experiment 3 results

	Human label			GPT, training w/60 Qs			GPT, training w/90 Qs		
	LLQ	**DRQ**	**GDQ**	**LLQ**	**DRQ**	**GDQ**	**LLQ**	**DRQ**	**GDQ**
Q1	1			1			1		
Q2			1			1			1
Q3		1			1			1	
Q4			1			1			1
Q5	1			1			1		
Q6	1				0.24	0.76			1
Q7		1			1			1	
Q8		1			1		1		
Q9			1		1			1	
Q10	1			0.86	0.14		1		
Q11	1				1		1		
Q12			1	1			1		
Q13	1			0.62	0.38		1		
Q14			1		0.48	0.52			1
Q15			1		0.48	0.52			1
Q16		1			1			1	
Q17	1			1			1		
Q18	1			1			1		
Q19		1			1		1		
Q20	1			1			1		
Q21		1			1			1	
Q22		1		0.66	0.34		1		
Q23			1		0.34	0.66		1	
Q24		1			1			1	
Q25		1		0.66	0.34		0.78	0.22	
Q26	1				1		1		
Q27			1		0.48	0.52			1
Q28			1			1			1
Q29			1			1			1
Q30	1			1			1		

The results demonstrate the probabilistic nature of the labelling, with many questions being labelled differently on different runs. Of note is that the uncertainty is much larger in the task with the smaller training set (resulting in 9 questions with split labels) compared to the task with the larger training set (resulting in only 1 question with a split label). It was thus determined that while the labelling accuracy was the same with the training sets of 60 and 90 in Experiment 2, this experiment demonstrated that the larger training set of 90 pre-labelled questions provides more robust results across multiple runs of the experiment.

Experiment 4

In all experiments described so far, the labelling task was focused on questions isolated from their context in the conversation. Yet, human coders use the context of the conversation to make judgements for the appropriate labels, including both the discussion before a question is posed, and the response immediately after. The goal of Experiment 4 was to determine whether GPT-4 can also use context in the labelling task, and whether this results in labels that are more similar to those assigned by humans.

Experiment 4a

In this experiment, the entire conversation in the session was used for context in both training and testing sets. The training set consisted of the entire transcript of one complete session (G1, S1), with the questions in the transcript identified, and their human-assigned labels provided. The testing set consisted of the entire transcript of a separate session (G1, S2), with the questions (but not their labels) identified. The training and test set were selected from the same group of participants (G1), so that GPT-4 could learn not only from the labelled examples, but also the participants' speaking styles. The exact user prompt in Experiment 4a was:

> *Classify all the questions identified in the conversation below, using the taxonomy proposed by Eris. Label each question with one of the three categories: Low-level questions, Deep Reasoning Questions, or Generative Design Questions, and present the results in JSON format that contains only questions and labels. Each question is identified and kept inside a square bracket `[]` at the end of an utterance. Each line (separated by newline character "\n") represents an utterance from a person in the conversation. There can be multiple questions from a single utterance. Label as you go through the conversation. In determining an appropriate label, consider the context of the conversation, including what has been discussed before the question, and the response that follows it. Here is the conversation -*
> ```<test set>```
> *To help you classify the questions, an example is given below which is a similar discussion between multiple subjects on the same topic. In this example, each line contains an utterance from a person in the conversation. The lines are separated by the newline character "\n". The appropriate question is identified and kept inside a square bracket and a label can be found on that question inside parentheses. If there are multiple questions from a single utterance, then those questions and associated labels are listed inside the square brackets separated by a comma. The example -*
> ```<training set>```

The experiment produced unexpected results. Although the prompt has clear instructions to label all the identified questions in the testing set, GPT-4 failed to do so. Moreover, some of the GPT-4-labelled questions in the result were mixed up with the questions provided in the training set.

Experiment 4b

Next, to address the limitations of the previous attempt, a different experiment was designed. The idea in the new experiment was to provide GPT-4 a smaller "chunk" of context surrounding the question. It was hoped that this chunk of context would improve the labelling accuracy.

In the new experiment, the training set consisted of examples of "chunks" of text from different sessions. Each chunk contained a labelled question, with a specified amount of text both immediately before and after it. Initially, it was believed that including one to two utterances before and after the question would be appropriate. An utterance was defined as any text uttered by a participant, located between two turn-taking activities in the conversation. However, it was later determined that given the large variety of the length of utterances (which could vary from one word to several paragraphs in length), a different approach was needed. Thus, instead, approximately 50-80 words were included both before and after each question in such a way that complete sentences were included.

Two different versions of the training set were used: one with 60, and one with 90 pre-labelled questions. In both cases, these were the same 60/90 questions used in Experiment 3. The testing set comprised of similarly constructed chunks of text, enveloping the same 30 questions used in Experiments 1-3. The following user prompt was used:

> *Classify the question within the 30 conversation segments listed below, using the taxonomy proposed by Eris. Label each question with one of the three categories: Low-level questions, Deep Reasoning Questions, or Generative Design Questions—and present the results in JSON format that contains both questions and labels. Each question is identified and kept inside a square bracket `[]` at the end of an utterance. Each line represents an utterance from a person in the conversation. In determining an appropriate label, consider the context of the conversation, including what has been discussed before the question, and the response that follows it.*
> ``` *<test set>* ```
> *To help you classify the questions, here are some examples. 60 discussion segments between multiple subjects separated using <example> tag are provided below. In these examples, each line contains an utterance from a person in the conversation. The appropriate question is identified and kept inside a square bracket just like before followed by the label for that question inside parentheses.*
> *<training set>*

In both cases with training set sizes of "chunks" surrounding 60 and 90 questions, 14 to 16 of the 30 GPT-4 generated labels were incorrect, resulting in a significantly degraded performance compared to the case when no context "chunks" around the questions was provided.

Experiment 4c

Given the challenges in the previous two attempts, the experimental task was further paired down. Specifically, Experiment 4c sought to determine if the label accuracy could be improved by providing GPT-4 the utterance that immediately follows the question. The reasoning was that when humans perform the labelling task, they are trained to use the answer to a question for important clues about the intention of the question.

Using the same training sets (sizes 60 and 90) that were used in Experiment 2, the labelling task was repeated once again on the same test set of 30 questions, with one key difference: for both the training and test sets, GPT-4 was also provided with the utterance that immediately followed the question, which was presumed to contain the response to the question. An utterance is defined here exactly as described in Experiment 4b. The following user prompt was used:

> *Classify the question within the 30 conversation segments listed below, using the taxonomy proposed by Eris. Label each question with one of the three categories: Low-level questions, Deep Reasoning Questions, or Generative Design Questions—present the results in JSON format that contains both questions and labels. Also, State your reasoning for the assigned label. Each question is identified and kept inside a square bracket `[]`. Each line (separated by newline character "\n") represents an utterance from a person in the conversation. In determining an appropriate label for the question, consider the utterance that follows it in the conversation, which is provided in the following line.*
> ```<test set>```
> *To help you classify the questions, here are some examples. 90 discussion segments between multiple subjects are provided below which are separated using <example> tag. In each example, the first line contains the question kept inside a square bracket just like before followed by the label for that question inside parentheses. The second line provides the utterance that follows the question in the conversation. The lines are separated by the newline character "\n".*
> *<training set>*

The experiment results are presented in Table 5, which contrasts the label assignments in this experiment ("GPT, answer provided") to those in Experiment 2 ("GPT, answer NOT provided"), as well as to the human-assigned labels. It is observed that supplying GPT-4 with the answer to the question results in worse alignment with human-generated labels, compared to not providing the answer. While increasing the training set size from 60 to 90 questions results in improved alignment (50% to 60%), it still falls significantly below the performance when the question answers are not provided (83% alignment).

Table 5 Experiment 4c results

	Human	GPT, answer NOT provided		GPT, answer provided	
		60 training	90 training	60 training	90 training
Q1	DRQ	DRQ	DRQ	LLQ	LLQ
Q2	GDQ	GDQ	GDQ	LLQ	GDQ
Q3	DRQ	DRQ	DRQ	DRQ	DRQ
Q4	GDQ	GDQ	GDQ	LLQ	LLQ
Q5	LLQ	LLQ	LLQ	LLQ	LLQ
Q6	LLQ	GDQ	GDQ	LLQ	GDQ
Q7	DRQ	DRQ	DRQ	DRQ	DRQ
Q8	DRQ	DRQ	DRQ	LLQ	LLQ
Q9	GDQ	DRQ	DRQ	DRQ	DRQ
Q10	LLQ	LLQ	LLQ	LLQ	LLQ
Q11	LLQ	LLQ	LLQ	LLQ	LLQ
Q12	GDQ	LLQ	LLQ	LLQ	LLQ
Q13	LLQ	LLQ	LLQ	LLQ	LLQ
Q14	GDQ	GDQ	GDQ	LLQ	GDQ
Q15	GDQ	GDQ	GDQ	LLQ	GDQ
Q16	DRQ	DRQ	DRQ	LLQ	LLQ
Q17	LLQ	LLQ	LLQ	LLQ	LLQ
Q18	LLQ	LLQ	LLQ	LLQ	LLQ
Q19	DRQ	DRQ	DRQ	LLQ	LLQ
Q20	LLQ	LLQ	LLQ	LLQ	LLQ
Q21	DRQ	DRQ	DRQ	DRQ	DRQ
Q22	DRQ	LLQ	LLQ	LLQ	LLQ
Q23	GDQ	GDQ	GDQ	LLQ	DRQ
Q24	DRQ	DRQ	DRQ	LLQ	LLQ
Q25	DRQ	LLQ	LLQ	LLQ	LLQ
Q26	LLQ	LLQ	LLQ	LLQ	LLQ
Q27	GDQ	GDQ	GDQ	LLQ	GDQ
Q28	GDQ	GDQ	GDQ	GDQ	GDQ
Q29	GDQ	GDQ	GDQ	GDQ	GDQ
Q30	LLQ	LLQ	LLQ	LLQ	LLQ
Alignment (%)	n/a	83	83	50	60

Discussion

As LLMs become more sophisticated and ubiquitous, researchers have rightly begun to ponder how ChatGPT and other technologies will shape qualitative research [34]. This research was motivated by the broader aims of exploring how LLMs could be used in analyzing verbal protocols of design. Specifically, in this paper, our research question was to determine the extent to which the most recent variant of ChatGPT (GPT-4) could accurately classify question utterances according to Eris's [20] taxonomy.

The study took advantage of a large pre-existing dataset comprised of a large corpus of over 2200 human-classified question utterances, which also included the complete transcripts of the dialogues in which the questions were uttered. The study was comprised as a series of careful experiments that sought to understand the nuances of using GPT-4 for performing this classification. The following subsections highlight study findings, including limitations, implications and opportunities for future research.

Classification accuracy and sensitivity

Experiment 1 tested GPT-4's baseline performance in this classification task. GPT-4 was provided with a system message that described the classifications scheme (the meaning of the LLQ, DRQ, and GDQ categories and the sub-categories within each, with examples). The classification task consisted of assigning each of the randomly selected 30 questions to one of the three categories (but not to their subcategories). Without any further training, GPT-4 achieved an alignment of 60% with the human-assigned categories. However, when given 90 additional examples of human-labelled questions (or training data), GPT-4's performance reached an alignment of 83%. Importantly, where there was misalignment, there was ambiguity in the correct classification for the question. For example, the two different human coders had disagreed about the correct category in the initial coding, or the current authors disagreed with the human-assigned category. Thus, GPT-4 arguably provided more accurate labels than the human coders. Experiment 2 sought to further test GPT-4's sensitivity to the training set size. Results from this experiment suggest that GPT-4 can perform this task and maintain adequate alignment with human labels with a fairly limited training dataset (a mere 60 pre-labelled questions). Experiment 3 examined the sensitivity of the classification across a different dimension – the number of "runs" of a classification task. Indeed, the results demonstrated the probabilistic nature of GPT-4's labelling, with many questions being labelled differently on different runs. Consistency across different runs was improved with a larger training set.

Four additional "sensitivity" aspects could be further investigated in future studies. First, while we experimented with different sizes of the training set (up to 300 pre-labelled questions), it would be useful to explore how the results might vary with larger-sized testing sets (beyond the 30 that were used in all experiments in this study). Second, given GPT-4's sensitivity to prompt changes, we ponder how varying the order of labelled questions in training data affects results across different runs (e.g., in Experiment 3). Third, future studies could test sensitivity to the origin of the training set data. In this study, training and test data were sourced from the

same original dataset (though without overlap), but how would the results differ if the training set originated from a different study? Finally, although the classification task was only tested on the three broader question categories, it would be useful to investigate GPT-4's performance at the sub-category level.

Overall, the results of Experiments 1-3 suggest that GPT-4 can classify questions according to Eris' taxonomy with impressive accuracy in terms of alignment with the human-sourced labels. Importantly, GPT-4 completed the task (labelling 30 new questions in the test set) with varying runtime from half a minute to a few minutes, depending on the total input size and output formatting. This efficiency stands in contrast to the time-intensive nature of protocol analysis, especially when dealing with large datasets.

The role of context in the question classification task

When human coders perform the identification and classification of questions, they are instructed to take into consideration both the question and the context in which the question is asked for clues [20, p. 99]. This includes what the participants have discussed earlier and how the question receiver responds to the question. Experiment 4 explored the impact of providing additional context to GPT-4 (varying from providing the entire session transcript to just the utterance immediately following the question) on the alignment with human-determined labels.

The results indicate that providing any additional context not only does not improve but can also degrade performance. The model failed to label questions and even confused questions from the training and testing sets. This observation of the relationship between context window and performance aligns with recent findings [35], [36], which indicate that long input contexts are not utilized properly by current LLMs and ChatGPT. Curry et al.'s [35] experience mirrors our own: ChatGPT seems to "*conflate and modify data, likely owing to its probabilistic approach to synthesis and its goal to produce content, regardless of the input*".

This finding prompts the need for further investigation into the factors contributing to this decrease in effectiveness. Additionally, future research could examine the concept of "context" in the context of NLP. Defining context with precision is imperative for advancing the understanding of its influence on model performance. For instance, if using the raw session transcript as context might confuse GPT-4, a future study could investigate whether summarizing the context before the classification task (perhaps through a separate GPT-4 task) could enhance performance.

A related expansion of the classification task involves training GPT-4 to identify question utterances, in addition to classifying them according to the

taxonomy. This is also a context-specific task, as question utterances are not always clearly denoted in transcripts with a question mark; the context around an utterance (and if reviewing the video recording, the tone) informs the human researcher if an utterance is posed as a question.

Broader considerations in using LLMs in design protocol analysis

Recent advancements in AI, specifically LLMs, show great potential but using them in design research require consideration that LLMs are not inherently generalizable and need custom training to be task-specific, as evidenced by prior studies [18] and our own. Despite improvements in context awareness and nuanced text generation, LLMs still struggle with precise contextual understanding, especially with specialized language across different disciplines.

While GPT-4 is a highly advanced LLM, it often "hallucinates," producing plausible but incorrect or irrelevant information, which could degrade information quality [37]. To address this issue, researchers suggest a range of strategies, including feedback mechanisms, accessing external information, and early refinement during LLM development. Some notable works are Retrieval-Augmented Generation [38], Knowledge Retrieval [39], and CoNLI [40]. Prompt engineering is also crucial, as evidenced by our study's use of specifically developed prompts based on official guidelines [41]. However, these may not always be the most effective. As noted by Wei et al. [42], there is potential to refine prompts further to elicit better responses from LLMs, suggesting that more effective prompt versions could yield superior results. Also, using LLMs like GPT for qualitative analysis may introduce bias as they learn language from their training data [43]. This can lead to biased insights, or ignoring marginalized groups' needs [44], potentially resulting in designs that exclude or alienate some users.

Design researchers looking to use GPT in other protocol analysis applications need to also consider that these models are frequently updated (with no detailed logs), enhancing performance but complicating the replicability and comparability of research over time. Proprietary algorithms and datasets restrict independent verification and pose challenges in assessing biases and ethical concerns on training materials. Finally, the GPT-4 API is not free to use and could be unaffordable for researchers with limited resources, especially when dealing with very large datasets. Since GPT is not open source, it may lead some researchers to prefer alternative open-source LLMs.

Conclusion

A series of experiments have demonstrated that GPT-4 can effectively categorize stand-alone question utterances using Eris' question-asking taxonomy. The study underscores the potential of LLMs to enhance qualitative analysis in design research, by reducing the time and resources required for manual coding, while requiring minimal NLP expertise on the part of the researcher. The findings highlight both the strengths and limitations of LLMs, suggesting that while they hold promise, further refinement and understanding of their application in complex analytical tasks such as verbal protocol analysis are necessary.

References

1. Ericsson KA, Simon HA (1984) Protocol analysis: Verbal reports as data, revised edition. The MIT Press
2. Hay L, Duffy AH, McTeague C, Pidgeon LM, Vuletic T, Grealy M (2017) A systematic review of protocol studies on conceptual design cognition: Design as search and exploration. Des Sci 3:e10
3. Kelly N, Gero JS (2022) Reviewing the concept of design frames towards a cognitive model. Des Sci 8:e30
4. Litster G, Hurst A (2021) Protocol analysis in engineering design education research: observations, limitations, and opportunities. Stud Eng Ed, 1(2): pp.
5. Sarkar P, Chakrabarti A (2013) A support for protocol analysis for design research. Des Issues, 29(4): 70-81
6. Gero JS, Milovanovic J (2023) The situatedness of design concepts: Empirical evidence from design teams in engineering. Proc Des Soc, 3: 3503-3512
7. Nespoli OG, Hurst A, Gero JS (2021) Exploring tutor-student interactions in a novel virtual design studio. Des Stud 75:101019
8. Chandrasegaran, S, Salah A, Lloyd P (2023) Constructing design activity in words: Exploring linguistic methods to analyse the design process. Des Stud 86: 101182
9. Mogavi R, Deng C, Kim J, Zhou P, Kwon Y, Metwally A, Tlili A, Bassanelli S, Bucchiarone A, Gujar S, Nacke L (2024) ChatGPT in education: A blessing or a curse? A qualitative study exploring early adopters' utilization and perceptions. Comp Hum Behav: Art Hum, 2(1): 100027
10. Bano M, Zowghi D, Whittle J (2023) Exploring qualitative research using LLMs. arXiv preprint arXiv:2306.13298
11. Zhao W, et al. (2023) A survey of large language models. arXiv preprint arXiv:2303.18223

12. Ray PP (2023) ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems 3:121-154
13. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781
14. Jurafsky D, Martin JH (2023) Speech and language processing: An introduction to natural language processing, computational linguistics, and speech cognition. 3rd Ed draft: https://web.stanford.edu/~jurafsky/slp3/ [Accessed 23 Jan 2024]
15. Hamilton L, Elliott D, Quick A, Smith S, Choplin V (2023) Exploring the use of AI in qualitative analysis: A comparative study of guaranteed income data. Int J Qual Meth, 22
16. De Paoli S (2023) Performing an inductive thematic analysis of semi-structured interviews with a Large Language Model: An exploration and provocation on the limits of the approach. Soc Sci Comp Rev 0(0)
17. Braun V, Clarke V (2006) Using thematic analysis in psychology. Qual Res Psych 3(2): 77-101
18. Xiao Z, Yuan X, Liao QV, Abdelghani R, Oudeyer PY (2023) Supporting qualitative analysis with Large Language Models: Combining codebook with GPT-3 for deductive coding. Comp Proc 28th Int Conf Intelligent User Interfaces 75–78
19. Siiman L, Rannastu-Avalos M, Pöysä-Tarhonen J, Häkkinen P, Pedaste M (2023) Opportunities and challenges for AI-assisted qualitative data analysis: An example from collaborative problem-solving discourse data. Int Conf Innov Tech Learn 87-96
20. Eris O (2004) Effective inquiry in engineering design. Kluwer Academic Publishers
21. Lehnert GW (1978) The process of question answering. Lawrence Erlbaum Associates: Hillsdale, New Jersey
22. Graesser A, Lang K, Horgan D (1988) A taxonomy for question generation. Questioning Exchange 2(1): 3-15
23. Graesser A, Person N (1994) Question asking during tutoring. American Ed Res J, 31(1):104-137
24. Cardoso C, Badke-Schaub P, Eris O (2016) Inflection moments in design discourse: How questions drive problem framing during idea generation. Des Stud 46: 59-78
25. Cardoso C, Eriş O, Badke-Schaub P, Aurisicchio M (2014) Question asking in design reviews: how does inquiry facilitate the learning interaction? 10th Design Thinking Res Symp, Purdue University
26. Hurst A, Duong C, Flus M, Litster G, Nickel J, Dai A (2021) Evaluating peer-led feedback in asynchronous design critiques: A question-centered approach. ASEE Virt Annual Conf Content Access
27. Cardoso C, Hurst A, Nespoli OG (2020) Reflective inquiry in design reviews: The role of question-asking during exchanges of peer feedback. Int J Eng Ed 36(2): 614-622

28. Hurst A, Lin S, Treacy C, Nespoli OG, Gero JS (2023) Comparing academics and practitioners Q&A tutoring in the engineering design studio. Proc Des Soc 3: 997-1006.
29. OpenAI (2023) Text Generation - OpenAI API [Online]. Available: https://platform.openai.com/docs/guides/text-generation/chat-completions-api. [Accessed 10 Jan 2024]
30. OpenAI (2023) Models - OpenAI API [Online]. Available: https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo. [Accessed 10 Jan 2024]
31. OpenAI (2023) Text Generation - OpenAI API [Online]. Available: https://platform.openai.com/docs/guides/text-generation/reproducible-outputs. [Accessed 10 Jan 2024]
32. OpenAI (2023) API Reference - OpenAI API [Online]. Available: https://platform.openai.com/docs/api-reference/chat/create#chat-createtemperature. [Accessed 10 January 2024]
33. OpenAI (2023) How to make your completions outputs consistent with the new seed parameter | OpenAI Cookbook [Online]. Available: https://cookbook.openai.com/examples/reproducible_outputs_with_the_seed_parameterpro. [Accessed 10 Jan 2024]
34. van Manen M (2023) What does ChatGPT mean for qualitative health research? Qual Health Res 33(13): 1135-1139
35. Curry N, Baker P, Brookes G (2024) Generative AI for corpus approaches to discourse studies: a critical evaluation of ChatGPT. Ap Corp Ling 4(1)
36. Liu N, et al. (2024) Lost in the middle: How language models use long contexts. Trans Assoc Comp Ling 12:157-73.
37. Achiam J, et al (2023) Gpt-4 technical report. arXiv preprint: arXiv:2303.08774
38. Lewis P, et al (2020) Retrieval-augmented generation for knowledge intensive NLP tasks. Adv Neural Inf Proc Sys 33:9459-74.
39. Varshney N, Yao W, Zhang H, Chen J, Yu D (2023) A stitch in time saves nine: Detecting and mitigating hallucinations of LLMs by validating low-confidence generation. arXiv preprint arXiv:2307.03987
40. Lei D, Li Y, Wang M, Yun V, Ching E, Kamal E (2023) Chain of natural language inference for reducing Large Language Model ungrounded hallucinations. arXiv preprint arXiv:2310.03951
41. OpenAI (2023) Prompt Engineering - OpenAI API [Online]. Available: https://platform.openai.com/docs/guides/prompt-engineering. [Accessed 10 April 2024]
42. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, Le QV, Zhou D (2022) Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Proc Sys 35:24824-37.
43. Perez E, et al. (2023). Discovering language model behaviors with model-written evaluations. In Findings of the Assoc for Comp Ling 13387–13434, Toronto, Canada.
44. Hu T, Kyrychenko Y, Rathje S, Collier N, van der Linden S, Roozenbeek J. (2023) Generative language models exhibit social identity biases. arXiv preprint arXiv:2310.15819.