10th International Conference on Artificial Intelligence and Soft Computing

Evaluating Persona Prompting for Question Answering Tasks

Carlos Olea, Holly Tucker, Jessica Phelan, Cameron Pattison, Shen Zhang, Maxwell Lieb, Doug Schmidt, Jules White

Department of Computer Science, Vanderbilt University, Nashville, TN, United States

Abstract. Using large language models (LLMs) effectively by applying prompt engineering is a timely research topic due to the advent of highly performant LLMs, such as ChatGPT-4. Various patterns of prompting have proven effective, including chain-of-thought, self-consistency, and personas. This paper makes two contributions to research on prompting patterns. First, we measure the effect of single- and multi-agent personas in various knowledge-testing, multiple choice, and short answer environments, using a variation of question answering tasks known as as "openness." Second, we empirically evaluate several persona-based prompting styles on 4,000+ questions. Our results indicate that single-agent expert personas perform better on high-openness tasks and that effective prompt engineering becomes more important for complex multi-agent methods.

Keywords: Prompt Engineering, Large Language Models, Question Answering.

1 Introduction

Large language models (LLMs) have received significant attention since the release of the GPT3 model [1], most notably after the ChatGPT chat-aware LLM achieved widespread popularity [2]. This attention stems from the unprecedented performance of LLMs on a variety of tasks, not only in the role of chatbots but also in other complex tasks, such as generating user-requested code [3]; improving the efficiency and effectiveness of clinical, educational, and research work in healthcare [4]; and producing educational exercises [5]. In this context, an increasingly relevant focus of LLM application is *prompt engineering*, which is an emerging discipline that structures interactions with LLM-based computational systems to solve complex problems via natural language interfaces [6].

The pattern catalog presented in [6] and the chain-of-thought strategy in [7] are notable methods and strategies for prompt engineering. A common strategy for LLM manipulation and prompt design is the *Persona* pattern, which a structured prompt engineering approach that embeds a consistent and predefined set of characteristics, behaviors, or traits to guide the model's responses in a manner aligned with a specific personality or identity. Previous studies have evaluated LLM capability to replicate the big five personality traits [8] and generated workflows for creating personas [9]. Despite this previous work, however, the *Persona* pattern

remains relatively opaque due to the inherent variance between LLM responses, the multi-faceted nature of creating a persona for LLM use, and the variety of problems and interactions that may be encountered while using personas.

This paper further elucidates the effects of personas when using LLMs for question answering tasks. In particular, we measure the performance of various methods of minimal single and multi-agent[1] persona prompting on several datasets of multiple choice and short answer questions. These questions vary in difficulty from simple primary school science questions to in-depth understanding of medical phenomena, as well as detailed financial and life advice. To facilitate these measurements we test four different styles of persona prompting on each dataset: (1) a control style containing only the response format instructions, (2) a hand-written minimal persona, (3) a per-question self-generated single-agent persona, and (4) a per-question self-generated multi-agent roundtable persona set inspired by the work in [10], as well as the work in [11][12][13]. We test each of these methods on over 4,000 questions residing in 9 datasets.

This paper presents the following contributions to research on prompt engineering and LLM evaluation:

- An axis of comparison for question answering tasks labeled "openness" that measures the number of possible correct answers for a given question, as well as the number of ways correct answers can be expressed.
- An evaluation of the effects of single-agent persona prompting on question answering tasks, showing an increase in performance that is commensurate with increased "openness" from dataset to dataset.
- An evaluation of multi-agent prompting styles with minimal ancillary techniques, highlighting the importance of effective prompt engineering when dealing with more complex prompting styles and systems.
- An evaluation of the ChatGPT-4 and ChatGPT-3.5-turbo models on 2,000+ questions across 9 datasets, evaluating their performance with a control prompting style, as well as the handling and impact of personas on each model.

This paper is organized as follows: Section 2 contains detailed information on our prompting styles, test sets, openness and experimental method; Section 3 contains our results and resultant conclusions; Section 4 compares our research with related work; and Section 5 presents concluding remarks and future work.

2 Methodology

This section summarizes the various prompting styles we tested, describes the datasets we tested upon, introduces the concept of "openness" used in this study, and explains our experiment design.

Table 1. Prompt Styles

Control	
You are taking a test. Provide your answers by responding only with the number of the appropriate answer for the presented question.	
You are taking a test. Provide your answers by responding with one to two sentences.	
Minimal Persona	
Act as an incredibly intelligent researcher that is an expert at problem solving, common sense reasoning, and strategy. You are taking a test. Provide your answers by responding only with the number of the appropriate answer for the presented question	
Act as an incredibly intelligent researcher that is an expert at problem solving, common sense reasoning, and strategy. You are taking a test. Provide your answers by responding with one to two sentences.	
Single-Agent Auto-Generated Persona	
persona generator	Describe a detailed persona of an expert who would be able to answer the following question including their background and a detailed description of their capabilities and qualifications:[question is inserted here]
You are taking a test. Act as the persona provided and provide your answers by responding only with the number of the appropriate answer for the presented question	
You are taking a test. Act as the persona provided and provide your answers by responding with one to two sentences.	
Roundtable Auto-Generated Persona Multi-Agent	
roundtable admin initial	You are taking a test. Provide your answers by responding with the number of the appropriate answer for the presented question as well as your reasoning for choosing it.
roundtable expert	You are [insert expert name].You are assisting the administrator in taking a test by offering useful critique and information. Provide feedback on the most recent answer given by the administrator, as well as their reasoning and offer suggested changes if you think the answer is incorrect, as well as your reasoning why. Pay attention to the feedback of any other experts and correct any incorrect information or suggestions. ((Be succinct and only suggest answers that are provided by the question. Do not provide overly long feedback. Do not exceed 1500 characters in your response))
roundtable admin revisor	You are taking a test. Revise the previous answer according to the feedback provided by the experts you are collaborating with. ((You are not allowed to change the answers to the question, only the choice of answer you make.))
roundtable admin decider	You are taking a test. Decide the best answer given the feedback and revisions that have been made. ((Provide your answers by responding only with the number of the appropriate answer for the presented question.))
roundtable admin decider	You are taking a test. Decide the best answer given the feedback and revisions that have been made. ((Provide your answers by responding with one to two sentences.))
persona generator	You are an expert at creating useful personas. You create detailed personas of useful experts for answering the questions you are given including their background and a detailed description of their capabilities and qualifications. ((When you return the personas, be sure to separate them with a sequence of two newlines, followed by 5 dashes, followed by two newlines. For example: Persona description 1 — Persona description 2))

2.1 Prompting Styles Evaluated

This paper examined four prompting styles that varied in persona usage, number of agents, and generation of personas, including (1) single-agent non-persona prompting (the control), (2) single-agent static minimal persona prompting, (3) single-agent auto-generated persona prompting, and (4) multi-agent prompting. Descriptions of each prompting style follow, with examples shown in Table 1.

Single-agent non-persona prompting (the control) only provided the LLM with information needed to elicit an appropriate answer response in the proper format, *i.e.*, multiple choice or short answer. Likewise, **single-agent minimal persona prompting** was a prompting method identical to the control style, with the addition of a short, unchanging persona description preceding the instructions for test taking. This persona was given several positive modifiers and qualifications, such as being highly intelligent and an expert at reasoning and strategy.

Single-agent auto-generated persona prompting consisted of first prompting the LLM to generate an expert persona for the given question. The LLM received precise instructions about the capabilities and background of the generated persona and was also given instructions on general formatting. This generated persona was then used for answering the same question with the same formatting instructions as the previous two styles.

Multi-agent auto-generated persona prompting consisted of prompting the LLM to generate multiple expert personas for the given question. Each persona was assigned to a unique agent and the question was first given to an "administrator" LLM instructed to generate an answer to the question (similar to single-agent non-persona prompting), but was also instructed to provide reasoning for its selection of a given answer. After this answer was generated, each expert persona was provided with the current log of responses (including any messages from the administrator or other experts) and instructed to make revisions to the answer and provide its reasoning for the revisions.

After comments from each expert were acquired, the administrator was prompted to make changes according to the expert feedback if changes were deemed necessary and output a revised answer, thus completing one round. Although this process could be performed for a configurable number of rounds, or all data presented in this paper the number of rounds and experts was set to two unless otherwise specified. At the end of all the rounds, the administrator was prompted to make a final answer decision. A visual representation of this roundtable process is shown in Figure 1. This process was intended as a minimal implementation of each prompting style and/or format with few ancillary prompt engineering techniques, relying primarily on the use of personas. In particular, a few techniques were used to enhance the outcome of each prompting method to examine the effects of each in isolation.

[1] In this paper, an "agent" refers to a unique actor, *i.e.*, an LLM with its own chat history and assigned persona.

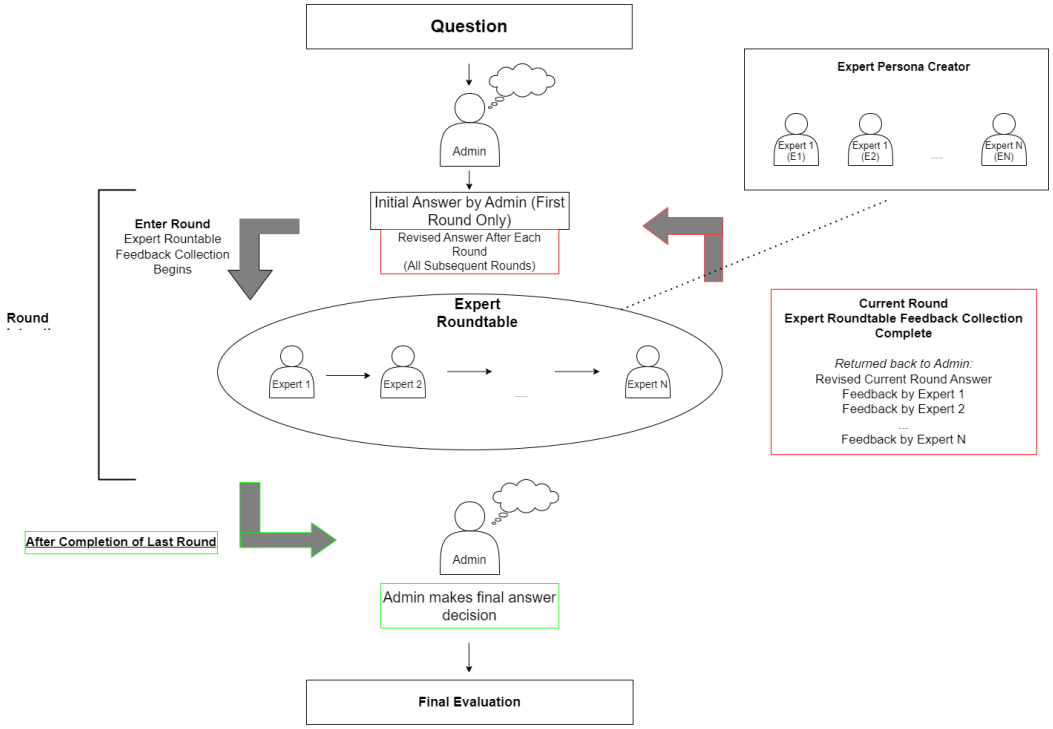

Fig. 1. Visualization of the Roundtable Process

2.2 Datasets Applied

We used a varied set of question-answering datasets in evaluating the use of the various persona prompting styles described in Section 2.1. These datasets varied across multiple axes, such as question subject, format, answer format, difficulty, etc. One core characteristic we assign we refer to as "open-endedness" (referred to simply as "openness" below), which we define as a composite of (1) the number of correct answers and (2) the number of unique ways a correct answer could be expressed.[2] For the intended answer format of each dataset we assigned a level of perceived openness since we observed that the main difference in how well different persona application strategies worked stemmed from how open they were, rather than other factors.

The datasets utilized in our experiments are summarized below, along with an associated rating of their openness.

commonsenseQA: Low. This dataset contains multiple choice questions that require "common knowledge" to answer. This set only has one "correct" answer for each multiple choice question, although the distractor answers (*i.e.*, answers that are wrong, but sound plausible) can occasionally be interpreted as possible correct answers. As a result, the dataset is rated low in openness, which explains—at least

[2] This definition aligns with the dichotomy presented in [14] that highlights multiple choice vs. open-ended questions.

partially—the consistently low scores on this particular set across all prompting styles, as shown in Table 6.

ai2-arc: Low. This dataset contained multiple choice questions at the grade-school level, which were split into "easy" and "challenge" questions. This set was rated as "low" in openness since it was multiple choice, as well as asking questions with only one acceptable answer, thereby scoring low on both axes for determining openness. Testing for this set was done exclusively on the challenge dataset to ensure that questions were sufficiently hard to show any possible differences in persona prompting style.

Measuring Massive Multitask Language Understanding: Low. This dataset contained multiple choice questions on a variety of subjects. Similar to **ai2-arc**, this dataset was rated as "low" in openness since it contained multiple choice questions with only one acceptable answer. It was also chosen because it was used for the initial evaluation of the GPT4 model in related work [15].

Databricks-dolly-15k: Medium-High. This mixed short answer dataset contained open question answering, closed question answering, general question answering, summarizing, brainstorming, information extraction, creative writing, and classification. Some tasks in this dataset (such as summarizing, information extraction, and classification) have low openness, as do sections of the question answering that have short and clearly defined answers, as shown in Table 2. These tasks are rated low since the correct answers are quite short (often one word or name) and lack expressiveness, as well as having one distinct viable answer. Other tasks (such as brainstorming, creative writing, and sections of open question answering) have high openness. Closed and general question answering lie somewhere in between. Given the spread of possible tags, this dataset was split into multiple tag groupings to test subsets of the data. The "Databricks subset" in Section 3 contains only the brainstorming and creative writing tags.

FinTalk-19k: Very High. FinTalk is a collection of reddit conversations related to finance and financial advice. Entries are split into three sections: (1) the question, (2) context around the question, and (3) the response. This dataset is rated as very high in open-endedness as there are feasible acceptable answers and many ways to express each answer.

LongForm: High. LongForm is a set of varied tasks generated from varied English sources, such as C4 and Wikipedia. Tasks include instruction following, board game explanation, open-ended question answering, and advice giving. This set is rated as high in openness as many (though not all) of the questions have multiple possible short answers.

PubMedQA: Low-Medium. PubMedQA is a question answering dataset derived from the abstracts of publications on PubMed.com. This dataset receives a variable openness score because there are two separate answer keys: (1) multiple choice where the options are yes, no, or maybe and (2) a short answer. The dataset

Table 2. Question Samples by Openness Rating

Dataset	Question	Answer	Openness Rating
Databricks	Which is a species of fish? Tope or Rope	Tope	Low
Ai2_arc	George wants to warm his hands quickly by rubbing them. Which skin surface will produce the most heat?	Dry palms	Low
Commonsense_QA	The forgotten leftovers had gotten quite old, he found it covered in mold in the back of his what?	Refrigerator	Low
Databricks	What is a verb?	A verb is an action word that describes an activity. Examples of verbs are the following: swimming, hiking, biking, roller skating, or skydiving. All of these single and combination word examples are tied to an activity that any living thing can do. For example, a dog can bike ride with a human when the human is pedaling the bike. Verb use is not limited to humans or people only but applies to all living things.	Medium
PubMed_QA	Are group 2 innate lymphoid cells (ILC2s) increased in chronic rhinosinusitis with nasal polyps or eosinophilia?	As ILC2s are elevated in patients with CRSwNP, they may drive nasal polyp formation in CRS. ILC2s are also linked with high tissue and blood eosinophilia and have a potential role in the activation and survival of eosinophils during the Th2 immune response. The association of innate lymphoid cells in CRS provides insights into its pathogenesis.	Medium

consists of a medical question, as well as context that contains information relevant to a question (the abstracts of the aforementioned publications). An example of an entry for this dataset can be found in Table 2. This dataset is considered to have ab openness rating of medium. Although the answer can be expressed several ways when using the short answer key, answers still fundamentally consist of yes, no, or maybe with little variability. This concept can be extended somewhat to the factual descriptions of medical phenomena.

2.3 Experiment Design

Testing was performed by selecting a dataset and style or set of styles to test, as well as the size of the testing sample. After selecting the datasets, styles and sample size, a random sample of the given size was selected for each style and each style was tested in order. At least 200 questions were rated for every unique combination of dataset, style and model (ChatGPT-3.5-turbo or ChatGPT-4). Some combinations underwent more evaluations, such as ai2_arc with 600 each, as well as FinTalk and the Databricks subset with 500 and 400, respectively.

For an individual question, the question was formatted, sent to the LLM with the given style, and the response was processed to identify a properly formatted response to the question. For multiple choice questions grading was done by direct comparison to the answer key for a given question provided by a dataset. For short answer questions, answers were evaluated using a judge LLM, specifically ChatGPT-4.

The judge LLM was instructed to compare the semantic similarity of two statements via the provision of a numeric score from one to ten. It was then given 5 examples of ratings with a static correct answer and varying answers to grade from scores 1-10. Evaluation of short answer questions was performed by providing the LLM judge with the dataset provided correct answer, the generated answer, and the aforementioned examples as part of the system message. The return message from the judge was then parsed for the score, of which scores of 6 or higher were considered "correct." However, comparisons of judge-compatible datasets were analyzed more by numeric question grade average than by a correct/incorrect method as the multiple choice dataset variants were.

Testing results were recorded and output containing the question, LLM response, roundtable dialog, correct response and evaluation (containing both correct/incorrect and numeric grade if applicable).

2.4 Consistency Evaluation

Although literature regarding the usefulness of LLM self-evaluation exists [16], the consistency and reliability of numeric evaluations assigned by LLMs is still an issue given their tendency to output differing ratings for the same input. As a result, we limited the scope of the evaluation to semantic similarity between two entries (correct and LLM answer), as well as tested the variance of answers for GPT-4. This testing was performed by having the LLM judge evaluate 100 answer-key

Table 3. Question Samples by Openness Rating (continued)

Dataset	Question	Answer	Openness Rating
LongForm	Describe Redux Data Structures.	Redux Data Structures is a library of reducer makers. Reducer makers help create common reducers like counters, maps, lists (queues, stacks), sets, etc. Most application states can be built by combining a handful of these standardized building blocks. Redux Data Structures was developed for Redux, but does not depend on it. It can actually be used with any reactive state container, even a custom one; Redux Data Structures doesn't have any dependency. and Todos are identified by an id property, used as a key in the todos map (and the completetedTodos set). [Answer continues for an additional 5-10 sentences]	High
Databricks	What are some unique curtain tie backs that you can make yourself?	There are many items you can use to make a curtain tie back. Some ideas for this include a chain, a circle and pin, jute rope, wooden beaded rope, a necklack or bracelet, a door knob, a length of a leather belt, macrame rope, or a string of artificial flowers.	High
FinTalk	What are some job ideas for a 62-year-old man with limited skills and qualifications, low confidence, and a failed business?	Truck driving can be a lucrative career and can provide the opportunity to see different parts of the country.	Very High
LongForm	Hey, I can't find some books I've read. What do I do? Respond in 5 sentences.	The more titles you rate the better your recommendations. Rate them by moving the slider with 10 being the best and 0 the worst. Here we chose some books you might have read. Hint: Can't find your books? Click below at 'Rate More Books' or use the upper search box.	Very High

pairs generated from the PubMed and LongForm datasets. Each answer-key pair was evaluated 5 times each, with the mean variance between all 100 pairs being 0.56 (where scores were assigned on a scale from 0-9).

3 Analysis of Results

This section reviews the results of our testing, including scores of each prompting style and model on each dataset. We also discuss the implications of the results for each of the prompting styles evaluated.

3.1 Multiple Choice Test Results

Multiple choice test sets included CommonsenseQA, PubMedQA, and ai2-arc. The results for each style on each set are shown in Table 4 and Table 5. As expected,

Table 4. Multiple Choice ChatGPT-3.5 Turbo Results Across 200 Questions by Percentage

	commonsense_qa	ai2_arc	mmlu	pubmed_qa
Control	0.68	0.93	0.76	0.69
Fixed simple persona	0.67	0.92	0.78	0.68
auto-generated persona	0.56	0.91	0.76	0.71
Roundtable	0.63	0.92	0.80	0.65

Table 5. Multiple Choice ChatGPT-4 Results Across 200 Questions by Percentage

	commonsense_qa	ai2_arc	mmlu	pubmed_qa
Control	0.73	0.94	0.92	0.8
Fixed simple persona	0.79	0.94	0.91	0.81
auto-generated persona	0.71	0.94	0.91	0.81
Roundtable	0.8	0.93	0.85	0.72

performance on each dataset and of each style varied considerably based on the model that was used (either ChatGPT-3.5-turbo or ChatGPT-4). Most notably, the multi-agent style tested on these datasets (referred to as "roundtable") often performed worse that the control. The most notable observation, however, is the apparent lack of difference between styles when powered by ChatGPT-4, particularly for the ai2_arc dataset which had a sample size of 700 questions, unlike the other entries in the table.

Performance rankings between style and model choices were inconsistent, though there was a general increase when switching form ChatGPT3.5-turbo to ChatGPT-4. However, one common issue in incorrect answers provided by roundtable prompting was the increased opportunity for hallucination.[3] Occasionally, one expert persona would hallucinate information that was relevant to a question. Other expert personas would then often accept that information as true and agree to an answer or suggest a change to an answer based on hallucinated information.

[3] A hallucination occurs when an LLM fabricates false information and presents or utilizes it as if it were true.

These results suggest that in prompts of simple knowledge extraction the types of persona usage we tested (and perhaps persona and multi-agent usage in general) has a negligible impact in the accuracy of the output. This outcome may be especially prevalent for more capable models, such as ChatGPT-4. For less capable models, such styles can be detrimental and present more opportunities for hallucination.

3.2 Short Answer Test Results

The short answer category generally had higher openness than the multiple choice category. However, there was considerable variation in openness within this category. We therefore discuss the medium and high openness datasets separately.

FinTalk, LongForm, and subsets of databricks were the high openness datasets we considered. Figure 2 shows the difference in mean score between each non-control prompting style and the control, and also includes a split by the model used (ChatGPT-3.5-turbo or ChatGPT-4).

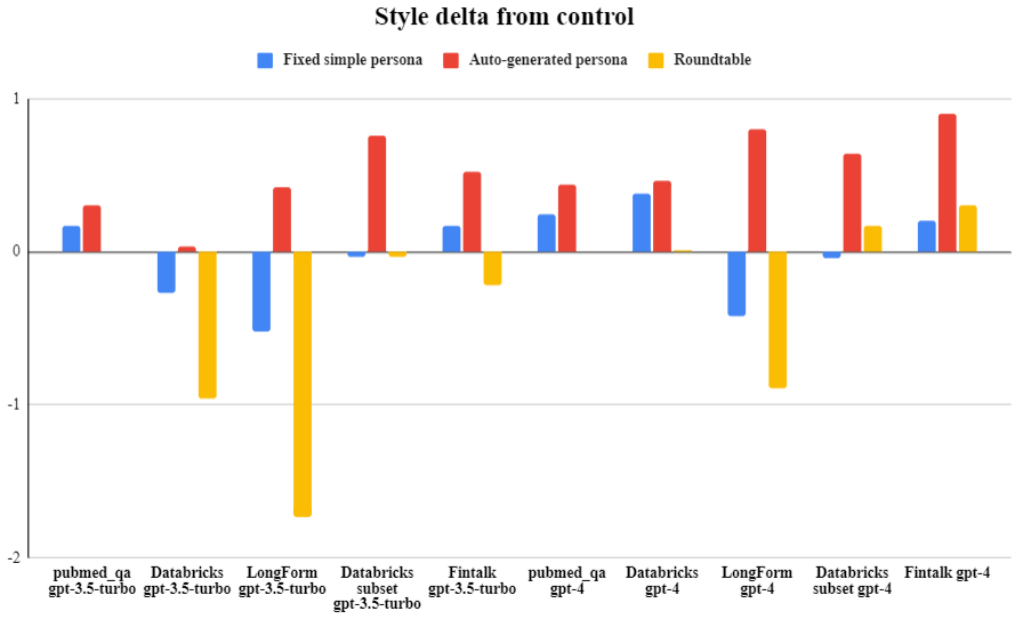

Fig. 2. Score Differences from Control for Each Prompting Style on Datasets with Free Response Answer Keys.

The auto-generated expert personas in the high openness datasets consistently scored a mean score of 0.3-0.9 above the control. The FinTalk dataset had the highest mean score since its questions have answers that are highest in openness (*e.g.*, financial/work advice answers, where both acceptable answers and ways of expressing them are plentiful). The mean scores of the auto-generated expert personas for the LongForm and databricks datasets were not as high above the control, corresponding to the level of "openness" of each dataset, respectively.

In the interest of creating more high openness test sets, we split the databricks test set into a subset of high openness questions containing the "brainstorming" and "creative writing" questions. When testing specifically on these subsets, the difference in mean score increased from 0.03 to 0.76 for ChatGPT-3.5-turbo and from 0.46 to 0.64 for ChatGPT-4. The mean scores of all styles tested on short answer questions are shown in Figure 3.

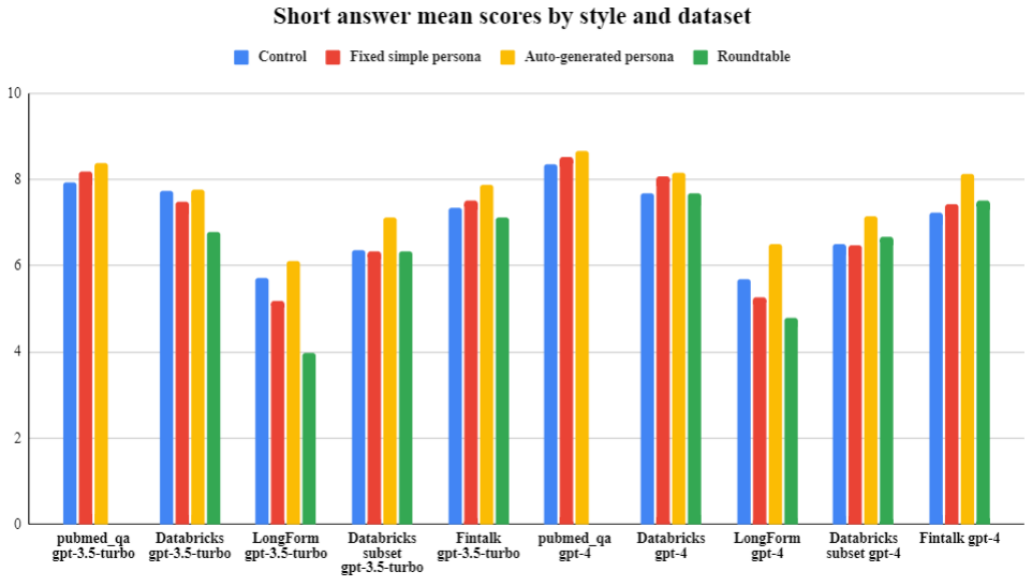

Fig. 3. Mean Score for Each Prompting Style on Datasets with Free Response Answer Keys.

One notable difference between the datasets is the relatively low scores across the board on the LongForm dataset, which stems in part from three main factors:

- The prompt we used constrained the answer to only one or two sentences. Though some questions overwrote this (such as the question at the bottom of Table 3) many of the "correct" answers that were compared could be multiple paragraphs long. In contrast, the "correct" answers for other datasets, such as FinTalk and pubmed_qa, were often three to five sentences.
- Some question and answer pairs can involve highly technical and specific questions (which can be paragraphs long themselves) and answers. Those answers could often be a unique individual's perspective, findings, or knowledge that might be hard to surmise from the situation without the same highly specialized knowledge.
- The dataset was partially generated by LLMs, thereby compounding the problem of hallucinations or otherwise incorrect responses and interpretations (again shown by the final example in Table 3).

3.3 Roundtable Prompting Test Results

As shown in the data presented in Tables 6 and 7, despite also using auto-generated personas the roundtable prompting style did not experience equivalent gains to the single-agent auto-generated personas. Unlike single-agents, there are at least two additional axes on which multi-agent prompting styles vary:

Table 6. Short Answer ChatGPT-3.5 Results Across 200 Questions by Mean Grade

	LongForm	Databricks	Databricks subset	FinTalk	pubmed_qa
Control	5.70	7.74	6.35	7.34	7.94
Fixed simple persona	5.18	7.47	6.32	7.51	8.18
Auto-generated persona	6.12	7.77	7.11	7.86	8.38
Roundtable	3.97	6.78	6.32	7.12	N/A

Table 7. Short Answer ChatGPT-4 Results Across 200 Questions by Mean Grade

	LongForm	Databricks	Databricks subset	FinTalk	pubmed_qa
Control	5.69	7.68	6.5	7.22	8.36
Fixed simple persona	5.27	8.06	6.46	7.42	8.53
Auto-generated persona	6.49	8.14	7.14	8.12	8.66
Roundtable	4.8	7.69	6.67	7.52	N/A

- **Organization of communication between agents**. In our implementation, each agent receives the output of each agent that spoke before them in the current round. This format could be substituted for one where agents only see the input from the last round, only see certain experts input either from the current or last round, or only see the answer they are evaluating and the revisions between each round.
- **The method for decision**. Voting is a common way for decision making in machine learning and ensemble models, where each agent votes for an answer and the highest voted decision is accepted. Instead we use a decider agent that receives all labeled logs of discussions between the "expert" and "revisor" agents.

Lastly, we specifically utilize rounds of discussion between the agents. When these rounds are combined with the variable number of agents, a significant number of possible configurations of a multi-agent system are created, above and beyond the personas they utilize.[4] Two key prompt engineering techniques were not utilized

[4] An exhaustive explanation of multi-agent structure effects and an exploration of a significant number of additional configurations to identify a highly performant configuration for the datasets presented are beyond the scope of this paper and are discussed as future work in Section 5.

in this method: (1) examples for reasoning and (2) voting. The lack of these two prompt engineering techniques (which is present in some of the implementations of multi-agent and roundtable methods in [10][12][13]) is likely a key ingredient in creating a high-functioning multi-agent roundtable system.

3.4 Evaluating Persona Strategies and Multi-Agent Systems in LLM Applications

We observe several trends in our analysis of the results. Due to the substantial size of the space of possible prompting styles, however, there are a number of questions that our analysis did not address. Below, we first discuss what we conclude from the data we recorded and analyzed and then address questions that require further analysis and experimentation.

Summary of findings. We first concluded that for open-ended, short answer questions there are advantages to using auto-generated (expert) personas versus simply asking the question using a minimal LLM prompt or a minimal persona prompt. This advantage is particularly pronounced in areas like advice giving, brainstorming, and creative writing. For advice giving, this benefit may stem from the prevalence of question answering and advice seeking sites (such as Reddit and Quora) in the training sets of LLMs, particularly ChatGPT. Moreover, we conclude that single-agent personas have a substantially smaller effect on the ability to answer less open-ended questions, such as multiple choice tests or simple yes/no answers to questions regarding text summarization or comprehension.

Next, we concluded that for answering multiple choice and short answer questions, as well as advice giving, brainstorming, and other open-ended tasks, there is little-to-no advantage to using the multi-agent roundtable method, particularly with two-to-three experts and one decider LLM to make the final decision. This finding to stems from the reasons outlined in Section 3.3, as well as other choices related to information flow, prompt structure, and decision processes.

Without using the two elements (examples and voting) highlighted in Section 3.3—and without specific anti-hallucination safeguards—roundtable discussions foster other avenues for hallucination. Despite this tendency, however, multi-agent implementations have been applied successfully in many domains that require experts to handle distinct and mostly separate tasks [17]. We therefore attribute this lack of performance to the simplicity of our implementation.

Topics for further investigation. The following are topics that the results of our analysis did **not** provide conclusive data for, and thus should be examined further. We focused on the efficacy of complex hand-crafted personas. However, the control in this experiment did not use any persona, but instead simply provides instructions for the completion of tasks for each dataset. Similarly, the fixed persona in this experiment is simple, and thus did not serve as an adequate comparison for a complex, extensive persona.

Our analysis also did not provide conclusive evidence on the performance of multi-agent systems for various tasks, aside from the specific multi-agent system we used for testing. However, this multi-agent implementation does support one more conclusion: as systems of interaction in LLMs become more complex (*e.g.*, by being deployed as multi-agent systems), they require additional support structures and advanced prompt engineering to ensure their effective use. As shown in previous work [10][11][12][13], if these techniques are applied systematically they can confer additional benefits versus performing the same level of prompt engineering for a single-agent, albeit requiring additional resources and time.

4 Related Work

Our work serves as a continuation and further evaluation of related work [18] that proposes "Expert prompting" and evaluates via a ChatGPT-4 backed judge, utilizing the judging method adapted from [19]. Other related work has been presented on creating performant multi-agent and roundtable prompting systems. For example, [10] introduces Solo Performance Prompting (SPP) utilizing a single agent but multiple personas to address knowledge and reasoning intensive tasks (these tasks can often have considerable overlap with the "high openness" described in this paper). Likewise, [11] presents a roundtable-style multi-agent system that serves as a contrast to the minimal multi-agent system implemented in this paper, and presents the opposite end of one of our core findings regarding the necessity of effective prompt engineering for multi-agent systems. In addition, [12] and [13] present multi-agent debate-centered solutions, with the latter opting for a judge to manage the debate between agents.

Other related work has focused on utilizing and evaluating personas. For example, [9] presents a workflow for creating personas that is build atop work by the same author [20][21] to evaluate how well LLMs perform thematic analysis of qualitative data. Work in the field of personas has also evaluated the ability of personas to act authentically like humans, *e.g.*, [8] investigates this topic by manually creating personas to embody the Big Five personality traits and having each persona complete a Big Five Inventory (BFI) test, as well as a story writing task. They find consistency between LLMs demonstrated personality, BFI scores, and writing, indicating an adequate capacity to embody these traits. Likewise, related work has focused on combating bias in LLMs, in which personas have also had a role. For instance, [22] uses auto-generated personas as a dataset to evaluate the intrinsic bias that a LLM may have by examining key identifiers for a given persona description.

Considerable related work has focuses on classifying question answering tasks. As found in [14], it is possible that multiple choice and short answer/open-ended questions measure different aspects of comprehension processes. Alternative methods of question classification, such as [23] and [24], are broader and and more

granular. Implementing these classifications for each dataset's questions is beyond the scope of this paper but represent possible future work.

Finally, related work has evaluated the efficacy of LLMs on particular datasets, ranging from cryptology [25] to medical exams [26]. There have also been evaluations of specific LLMs on a spread of datasets, such as [15], which analyzes the performance of ChatGPT-4 on several datasets, including the mmlu dataset included in our study. Our work differentiates from related work by combining the testing large amounts of data and datasets together with testing multiple different persona prompting types, as well as presenting a characteristic of question classification for question answering tasks.

5 Concluding Remarks

This paper analyzed and evaluated the performance of persona-based prompting strategies on various datasets, as well as identified a core differentiator between these datasets. We first identified a characteristic of questions known as "open-endedness" or "openness" that is defined as the number of possible correct answers, as well as the number of possible ways to express those same correct answers. Next, we evaluated our 9 datasets according to these and other characteristics. Subsequently, we identified and evaluated four separate prompting strategies for question answering (single-agent non-persona prompting (the control), single-agent static minimal persona prompting, single-agent auto-generated persona prompting and multi-agent prompting). Finally, we analyzed the results of these evaluations and present our findings.

An exhaustive explanation of multi-agent structure effects and an exploration of a significant number of additional configurations to identify a highly performant configuration for the datasets presented are beyond the scope of this paper, however this is an appropriate topic for future work. Particularly the optimization of the number of agents, decision structures and information flow. Similarly, there exist many methods of granular classification that may give a higher resolution understanding of the effects of question format and content on LLM and persona performance. This is also a pertinent topic for exploration.

The following is a summary of lessons learned from the research presented in this paper:

- **Single-agent personas are measurably more effective for higher openness questions**. Single-agent expert personas correctly answer higher openness questions more often than the other tested methods.
- **Effective prompt engineering becomes more important for complex multi-agent methods**. Multi-agent methods implemented without ancillary prompt engineering and ensemble techniques perform no better than minimal prompting with no personas across all levels of openness, thus showing that though multi-agent methods can confer sizeable benefits, they require notably more engineering and processing resources.

- **Prompt strategies should be measured objectively**. We learned that the efficacy of prompting strategies needs more investigation using objective measures, which motivates future research on quantifying the impact of various prompt engineering techniques and formats of LLM usage for different tasks.
- **Classifying questions and tasks requires thorough high-level analysis methods**. Our findings indicate the need for a more nuanced classification of questions and tasks, which leads to future work focusing on a detailed analysis of LLM performance to further examine the effectiveness of different prompting methods.
- **Understanding model interactions is essential**. Our research highlights the importance of increasing model transparency to move away from "black box" interactions, suggesting future efforts should be directed at understanding model-input interactions for more effective LLM utilization for a variety of tasks.

References

1. Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. Minds and Machines, 30:681–694, 2020.
2. Dinesh Kalla and Nathan Smith. Study and analysis of chat gpt and its impact on different fields of study. International Journal of Innovative Science and Research Technology, 8(3), 2023.
3. https://github.com/KillianLucas/open-interpreter
4. Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. Nature medicine, 29(8):1930–1940, 2023.
5. Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. Evaluating reading comprehension exercises generated by llms: A showcase of chatgpt in education applications. In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 610–625.
6. Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv preprint arXiv:2302.11382, 2023.
7. Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. Active prompting with chain-of-thought for large language models. arXiv preprint arXiv:2302.12246, 2023.
8. Hang Jiang, Xiajie Zhang, Xubo Cao, and Jad Kabbara. Personallm: Investigating the ability of large language models to express big five personality traits. arXiv preprint arXiv:2305.02547, 2023.
9. Stefano De Paoli. Improved prompting and process for writing user personas with llms, using qualitative interviews: Capturing behaviour and personality traits of users. arXiv preprint arXiv:2310.06391, 2023.
10. Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. arXiv preprint arXiv:2307.05300, 2023.
11. Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. arXiv preprint arXiv:2309.13007, 2023.
12. Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. arXiv preprint arXiv:2305.14325, 2023.

13. Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi- agent debate. arXiv preprint arXiv:2305.19118, 2023.
14. Yasuhiro Ozuru, Stephen Briner, Christopher A Kurby, and Danielle S McNamara, "Comparing comprehension measured by multiple-choice and open-ended questions," Canadian Journal of Experimental Psychology/Revue canadienne de psychologie experimentale, 67(3):215, 2013.
15. OpenAI et al. Gpt-4 technical report, 2023.
16. Jie Ren, Yao Zhao, Tu Vu, Peter J Liu, and Balaji Lakshminarayanan. Self-evaluation improves selective generation in large language models. arXiv preprint arXiv:2312.09300, 2023.
17. Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation, 2023.
18. Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. Expertprompting: Instructing large language models to be distinguished experts. arXiv preprint arXiv:2305.14688, 2023.
19. Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An opensource chatbot impressing gpt-4 with 90%* chatgpt quality.
20. Stefano De Paoli. Can large language models emulate an inductive thematic analysis of semistructured interviews? An exploration and provocation on the limits of the approach and the model. arXiv preprint arXiv:2305.13014, 2023.
21. Stefano De Paoli. Writing user personas with large language models: Testing phase 6 of a thematic analysis of semi-structured interviews. arXiv preprint arXiv:2305.18099, 2023.
22. Myra Cheng, Esin Durmus, and Dan Jurafsky. Marked personas: Using natural language prompts to measure stereotypes in language models. arXiv preprint arXiv:2305.18189, 2023.
23. Xin Li and Dan Roth. Learning question classifiers. In COLING 2002: The 19th International Conference on Computational Linguistics.
24. Alaa Mohasseb, Mohamed Bader-El-Den, and Mihaela Cocea. Question categorization and classification using grammar based approach. Information Processing Management, 54(6):1228–1243, 2018.
25. Vasily Mikhalev, Nils Kopal, and Bernhard Esslinger. Evaluating gpt-4's proficiency in addressing cryptography examinations. Cryptology ePrint Archive, 2023..
26. Maciej Rosol, Jakub S Gasior, Jonasz Laba, Kacper Korzeniewski, and Marcel Mlynczak. Evaluation of the performance of gpt-3.5 and gpt-4 on the polish medical final examination. Scientific Reports, 13(1):20512, 2023