Argument & Computation -1 (2024) 1–51
DOI 10.3233/AAC-230014
IOS Press
CORRECTED PROOF

Evaluating large language models' ability to generate interpretive arguments

Zaid Marji [*] and John Licato
Computer Science and Engineering, University of South Florida, FL, USA

Abstract. In natural language understanding, a crucial goal is correctly interpreting open-textured phrases. In practice, disagreements over the meanings of open-textured phrases are often resolved through the generation and evaluation of *interpretive arguments*, arguments designed to support or attack a specific interpretation of an expression within a document. In this paper, we discuss some of our work towards the goal of automatically generating and evaluating interpretive arguments. We have curated a set of rules from the code of ethics of various professional organizations and a set of associated scenarios that are ambiguous with respect to some open-textured phrase within the rule. We collected and evaluated arguments from both human annotators and state-of-the-art generative language models in order to determine the relative quality and persuasiveness of both sets of arguments. Finally, we performed a Turing test-inspired study in order to assess whether human annotators can tell the difference between human arguments and machine-generated arguments. The results show that machine-generated arguments, when prompted a certain way, can be consistently rated as more convincing than human-generated arguments, and to the untrained eye, the machine-generated arguments can convincingly sound human-like.

Keywords: Interpretive arguments, persuasion, large language models, generative language models

1. Introduction

Natural languages are optimized for flexibility and expressiveness in order to adapt to ever-changing human needs. The expressiveness of language allows for new ideas to be expressed, and the flexibility of language allows for the application of statements to new contexts. This flexibility is especially apparent when it comes to the specification of rules, such as those seen in laws, contracts, codes of conduct, and so on – a certain amount of flexibility in the allowed interpretations of terms in rules can ensure that the rule can be applied to new scenarios which may not have been conceived of by the rule writers.

This feature of the language used by rule systems has been referred to as "open-texturedness" [9, 41]. A natural language phrase is *open-textured* when there "always remains a set of (perhaps remote) possibilities under which there would be no right answer to the question of whether it applies" [3]. For example, consider the rule: "A teacher must behave in a manner that brings respect to their school and their profession." The phrase 'brings respect' in this context is open-textured, as it is not very clear what this expression entails in all situations, and there always remains a set of circumstances where its applicability is an open question. It is implausible to exhaustively list an exception-free accounting of all possible scenarios and conditions that can be considered instances of this open-textured term, and such an attempt would inevitably limit the scope of the rule, making it inflexible in the face of unanticipated situations [7,9,18,19,29,31,40]. In order to decide whether a specific action conforms to or violates this rule, a process of interpretation must take place. One of the techniques to interpret texts is

[*] Corresponding author. E-mail: zaidm@usf.edu.

to generate different arguments for and against specific interpretations and assume that the interpretation that is supported by the strongest arguments is the correct interpretation. Such arguments are called *interpretive arguments* [19,22,23,33,35,44,45].

Formally, interpretive arguments are of the form: "If expression E occurs in document D, E has a setting of S, and E would fit this setting of S by having interpretation $\mathcal{I}$, then E ought to be interpreted as $\mathcal{I}$" [35]. Although a significant body of literature in the argument space has studied the nuances of interpretive argumentation and open-texturedness, particularly as they are used in the legal field [2,17,43], little to no work exists examining how well state-of-the-art AI can generate and assess interpretive arguments. This is especially timely given the increased capabilities of generative language models, which are already being employed to generate and assess arguments in general [10]. In fact, the use of open-textured terms is a necessary and unavoidable feature of regulatory and legal language [7,9,18,19,29,31,40]. Thus, any robust account of automated rule-following must address how to handle them. Furthermore, it has been argued that artificially intelligent systems, particularly in rule-governed domains, must be able to carry out human-like interpretive reasoning, otherwise they will not be able to follow human laws or act in accordance with human values [15,16].

We, therefore, focus specifically on interpretive argumentation in this paper. Our starting point is the game *Aporia*, created specifically to elicit interpretive argumentation so that it could be collected and studied [24]. We curated a set of rules from the code of ethics of various professional organizations and a set of associated scenarios that are ambiguous with respect to some open-textured phrase within the rule. Those scenarios were selected to allow for good arguments to be made on both sides of the issue (so that the issue would not be one-sided) and to allow both sides to plausibly win the argument. We also collected and evaluated arguments from both human annotators and OpenAI's GPT-3 in order to find out the relative quality and persuasiveness of both sets of arguments. Figure 1 shows the general structure of the dataset created for this task.[1] We carried out human studies where human annotators evaluated 5

Profession: Professional economic developer
Rule: Professional economic developers shall maintain in confidence the affairs of any client, colleague or organization and shall not disclose confidential information obtained in the course of professional activities.
Ambiguous Phrase: Maintain confidence in client affairs
Scenario: A professional economic developer has been working with a client who is considering relocating their business to the developer's city. The client has been very secretive about their plans and has only shared information with the developer on a need-to-know basis. The developer has been given explicit permission to share information about the client with their colleagues if it is necessary to work on the project. However, the developer then learns that the client is considering relocating their business because they are engaged in illegal activities which could negatively impact the city. The developer then shared this information with their colleagues as he believed they needed to know this information to be informed of the legal issues and reputational damage working with this client might entail. However, this information was not required to perform the job that the professional was hired to do. This caused damage to the client's project and reputation. Many of the client's projects were canceled, leading to several lost job opportunities in the city.
Stance: Compliance
Argument: A confidentiality agreement should be considered valid only if it is lawful and does not impose dangerous ethical consequences. The professional economic developer acted in compliance with the rule by sharing confidential information with colleagues when he believed it was necessary to inform them of the legal and reputational risks of working with the client. Clients should not be able to use confidentiality agreements to cover illegal or dangerous activities, and the professional economic developer's actions were necessary to protect the public and his colleagues.

Fig. 1. An exemplar of an interpretive argument. In this case, the argument is that the rule should be interpreted in such a way that the scenario will be considered an instance of the concept represented by the ambiguous phrase.

[1] The dataset and code can be found at: https://github.com/Advancing-Machine-Human-Reasoning-Lab/llm-evaluation.

different arguments for a specific scenario and a specific stance (i.e., arguments for either compliance or non-compliance with the given rule) for a number of scenario-stance pairs on a 5-point Likert scale. Finally, we performed a simplified version of the Turing test in order to assess if human annotators can tell the difference between human arguments and machine-generated arguments. Further details will be discussed in the **Methodology** section.

Our experimental setup was designed to answer the following research questions:

RQ1 How well can state-of-the-art large language models (LLMs) argue interpretively, and which prompting methods generate the best interpretive arguments?

RQ2 How do human- and machine-generated interpretive arguments compare?

RQ3 Can people distinguish between human- and machine-generated interpretive arguments?

The contributions in this paper include:

- An experimental design and associated surveys that can be used to assess and compare the persuasiveness of interpretive arguments generated by LLMs.
- The experimental design is used to evaluate OpenAI's GPT-3 performance, and can be used to assess future LLMs in order to evaluate improvements and identify potential regressions.
- A dataset of human interpretive arguments for and against compliance of the aforementioned scenarios with respect to the ambiguous rule.
- A dataset of machine-generated interpretive arguments using different LLMs for the same set of scenarios.
- A dataset of human evaluations of all the aforementioned interpretive arguments.
- Datasets for testing annotators' ability to tell the difference between human- and machine-generated interpretive arguments.
- Our findings demonstrate that interpretive arguments generated by SOTA LLMs are rated as more convincing than (non-expert) human interpretive arguments.
- Our findings demonstrate that non-expert human annotators, with no prior exposure to human and machine-generated arguments, are unable to distinguish between the two. This suggests that SOTA LLMs exhibit a level of fluency and human-like diction that convincingly mirrors humans.

2. Background

Studying interpretive argumentation with Aporia. Aporia [24] is a game that pits players against each other in a competition to argue whether an open-textured rule applies to a scenario. The game is designed to be fun in order to encourage eager participation and obtain useful datasets of interpretive arguments. The Aporia game was developed with the express purpose of being a framework to automate reasoning about, and iteratively reducing, ambiguity in open-textured text. This framework is also designed to provide structure to the process of generating training data for interpretive argumentation through human gameplay or automated text generation.

The game can be played by any group of three or more people. It is recommended to be played with six players. The game is played in rounds. At the end of each round, points will be awarded to the winner. In each round, two players are randomly chosen to play against each other, with a third player designated as a judge. The gameflow is as follows:

(1) Each round starts with a tuple consisting of:
   - Profession
   - Rule that members of that profession are expected to follow. The rule must contain at least one open-textured phrase, which introduces ambiguity in its interpretation.
   - Scenario describing an action taken by a member of that profession. The scenario is crafted to leverage the ambiguity of the associated rule.

(2) Player 1 chooses which side to argue for (stance); either the professional acted in a way that complies with the rule, or their actions were non-compliant.
(3) Player 1 provides an argument for their chosen stance
(4) Player 2 provides a counterargument
(5) Judge declares the winner
(6) Judge provides an explanation for their decision

Figure 2 shows an overview of Aporia as a framework. The players (including the judge) take as inputs the information for the current round of the game and generate arguments, judgments, and explanations

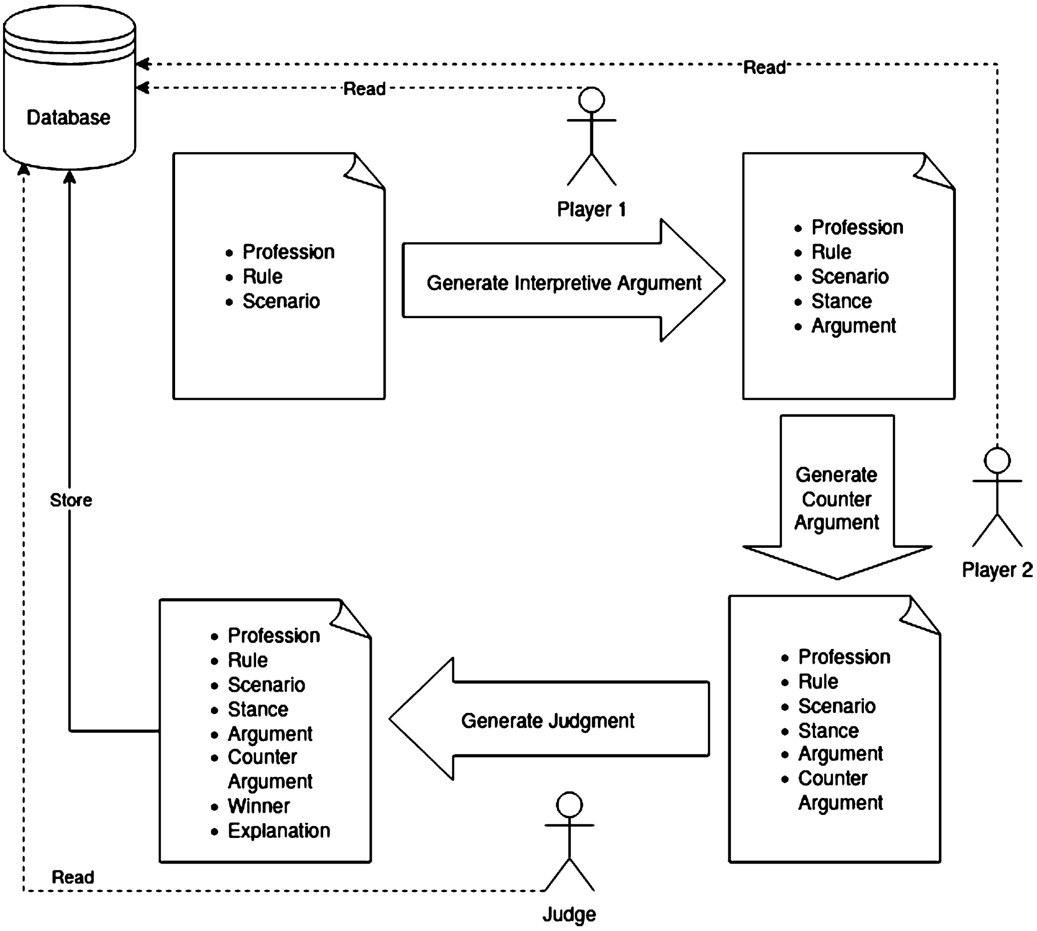
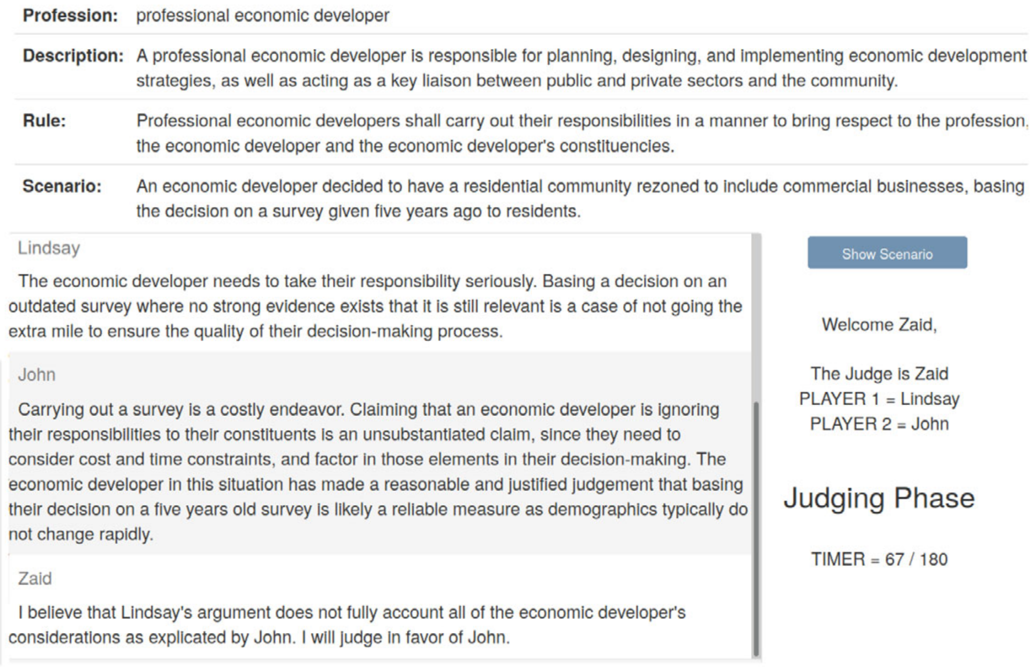


Fig. 2. Aporia as a framework.



Profession:	professional economic developer
Description:	A professional economic developer is responsible for planning, designing, and implementing economic development strategies, as well as acting as a key liaison between public and private sectors and the community.
Rule:	Professional economic developers shall carry out their responsibilities in a manner to bring respect to the profession, the economic developer and the economic developer's constituencies.
Scenario:	An economic developer decided to have a residential community rezoned to include commercial businesses, basing the decision on a survey given five years ago to residents.

Lindsay

The economic developer needs to take their responsibility seriously. Basing a decision on an outdated survey where no strong evidence exists that it is still relevant is a case of not going the extra mile to ensure the quality of their decision-making process.

John

Carrying out a survey is a costly endeavor. Claiming that an economic developer is ignoring their responsibilities to their constituents is an unsubstantiated claim, since they need to consider cost and time constraints, and factor in those elements in their decision-making. The economic developer in this situation has made a reasonable and justified judgement that basing their decision on a five years old survey is likely a reliable measure as demographics typically do not change rapidly.

Zaid

I believe that Lindsay's argument does not fully account all of the economic developer's considerations as explicated by John. I will judge in favor of John.

Welcome Zaid,

The Judge is Zaid
PLAYER 1 = Lindsay
PLAYER 2 = John

Judging Phase

TIMER = 67 / 180

Fig. 3. An example of Aporia after a round is complete.

as necessary to complete one training example. The players may optionally have access to previous game rounds that allow them to review existing data that they may use to improve their current arguments. Access to previous rounds is especially useful for automated argument generation, as they may be used for few-shot learning or other strategies to improve argument generation over time. Figure 3 shows a partial screenshot of the Aporia game in action.

In the paper introducing this game [24], a human study was conducted where human participants were invited to play the game using a set of rules and scenarios collected in a previous human study [19]. The data shows a clear preference for participants to argue for compliance (ie. the professional's actions were compliant with the rule), and a tendency for arguments in favor of compliance to win. These findings indicate that the scenarios were not perfectly balanced. A balanced scenario in this context means that both sides of the argument are equally plausible, leading participants to choose to defend either side with roughly similar likelihood and having a similar chance to win the argument on either side. Moreover, many scenarios had issues where the source of the ambiguity in the scenarios was not directly caused by the ambiguity in the open-textured language of the rule. For example, many scenarios lacked essential facts about the scenario, which was the primary source of ambiguity. Those issues were present in some of the arguments that the players used. For those reasons, we decided in the present study to use the same set of rules from the original dataset, discard the provided scenarios, and generate new scenarios while paying closer attention to balance in order to avoid the aforementioned issues.

In another human study that used this game and its associated dataset, the impact of certain cognitive biases on persuasion [6] was studied where the empirical results show that human participants tended to rate the *same* argument as more persuasive if the participants were led to believe that a woman or a White person wrote the argument. These results indicate that cognitive biases may have non-negligible effects on persuasion.

Argument generation using LLMs. Large Language Models (LLMs) are statistical models (typically, artificial neural networks using the Transformer architecture [39,49]) trained on vast amounts of textual data collected from various sources. Specifically, generative LLMs are trained to predict the next word (or, more accurately, the next token) following a certain excerpt of text. Given a sequence of words (tokens), the LLM will produce a pseudo-probability distribution over all the possible next tokens. The next token is then selected based on the token with the highest pseudo-probability or using a weighted sampling strategy. Using this approach iteratively, LLMs are used to generate text by feeding them an initial excerpt of text (commonly called a *prompt*) and repeatedly predicting the next token, one token at a time. A textual response can be constructed by iterating this process until a special token indicating the end of the response is generated.

Large language models can effectively capture patterns, structures, and knowledge in the training data. When prompted with questions or tasks, they generate responses based on the learned knowledge and patterns [14,20,32,47]. This led to a new trend in learning paradigms that has emerged since the advent of generative LLMs called 'pre-train, prompt, predict' [20]. One of the main advantages of this paradigm is that it substantially reduces the need for large task-specific datasets that are required in the supervised learning paradigm. This paradigm is commonly referred to as *prompt-based learning*. Due to the knowledge embedded in LLMs, they were found to perform well on many tasks without any task-specific training data. These recent advancements in the learning paradigm were essential to enable this work, as the supervised learning paradigm has been prohibitively challenging, as the scarcity of training data on interpretive argumentation and the costs associated with collecting such training data in vast quantities using human studies or other means has stifled the ability to produce high-quality interpretive arguments. Prompt-based learning has been instrumental in overcoming those challenges and has enabled huge advancements in this area of research.

When machines are able to perform tasks without being specifically trained on those tasks, the machine is said to perform *zero-shot learning*. For example, to solve a sentiment analysis problem in a zero-shot setting, we can prompt an LLM (such as OpenAI's GPT-3 [4]) with the text: "I have been stuck in traffic for over an hour. I feel so ___". We presume that the model will generate an appropriate sentiment as a completion for that prompt or, alternatively, the completion will be restricted to a set of predetermined outputs. One-shot and few-shot learning paradigms [4,36] use similar prompts that include one or more solved examples to condition the model on the task at hand [46]. For example, the following prompt might be used to predict the rating of a movie from a textual review:

- **Review:** Interesting plot. Recommended.
  Rating: 5 stars
- **Review:** Boring!!!!
  Rating: _______

Prompt-based learning can be done using manual template engineering by using manually created templates [20,28,36]. Other approaches can learn such prompts automatically, including discrete prompts that typically correspond to natural language phrases [8,12,20,42], or continuous prompts that use sequences of embeddings directly that do not necessarily correspond to embeddings that occur in natural languages [20,21,50].

A recent trend in prompt-based learning that shows promise is explanation-based prompting [11,20], which started with chain-of-thought prompting [48] and many styles of prompting inspired by it, such as self-ask [30] and maieutic prompting [13]. For example, a chain-of-thought prompt may be a question-answering task, where the task is presented in a few-shot learning fashion such that the provided examples do not simply answer the question, and instead provide reasonable step-by-step inferences leading to the correct conclusion. Such a style of prompts will encourage the generative language model to provide explanations before committing to an answer to the question, which was experimentally found to improve accuracy, especially on arithmetic questions [48]. The original paper provides this example: "Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?". The direct answer was "A: The answer is 27." which is incorrect. However, using chain-of-thought the answer was: "A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9." This example illustrates how an LLM's ability to obtain a correct answer may be heavily affected by the prompting style that is used, thus motivating the second half of our first research question (**RQ1**). The self-ask [30] prompting style encourages the LLM to ask itself some questions that will help it answer the original question. The original paper provides this example: "Q: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?", and instead of immediately answering the question, the LLM is encouraged to ask itself sub-questions such as: "Q: How old was Theodor Haecker when he died?" and "How old was Harry Vaughan Watkins when he died?". By identifying those intermediate sub-questions and answering them before answering the initial question, LLMs have been found to improve their accuracy on many question-answering tasks. On the other hand, the maieutic prompting [13] style is inspired by the maieutic method of Socrates. The maieutic method uses abductive and recursive explanations to identify and consequently reduce inconsistencies in the explanations generated. Again, this method has been found to improve accuracy on question-answering tasks, providing further evidence that teaching LLMs to explain or reason through their answers improves their reliability.

In this paper, two of the most capable language models from OpenAI at the time of running this experiment were used for the main set of experiments. GPT-3 [4] is used to generate scenarios and arguments, as shown in the rest of the paper, while ChatGPT [25] is only used for scenario generation. ChatGPT (a chatbot powered by GPT-3.5) is not a fully documented model, as the parent company has not published any academic papers or any other resource that includes full details of how GPT-3.5 was trained. However, based on technical reports from the parent company, it is likely instruction-tuned [27] using reinforcement learning from human feedback (RLHF) [5] to follow a variety of written instructions. Since GPT-4 [26] was yet to be released at the time of our experimentation, our present work focuses on GPT-3. However, we report briefly on a small follow-up experiment designed to show that the lessons we learn from the present work about prompting styles are still applicable to GPT-4 and other state-of-the-art generative language models in Section 4.

The Turing test. The Turing Test, first introduced by British mathematician and computer scientist Alan Turing in 1950, remains one of the most influential concepts in the study of artificial intelligence. In his seminal paper, "Computing Machinery and Intelligence," [38] Turing posited a scenario in which a human evaluator would engage in natural language conversations with both a human and a machine designed to generate human-like responses. If the evaluator could not reliably distinguish which participant was the machine based on the conversation alone, then the machine would be considered to have demonstrated human-like intelligence.

The original Turing test sets a relatively high threshold for a machine to pass the test, as a conversation with a machine is an interactive setting that allows the human evaluator to probe the limits of the understanding of the machine of the subject of the conversation. However, not all tests designed to tell humans and machines apart follow the same setup. For example, some CAPTCHA[2] systems use text

[2] CAPTCHA stands for *Completely Automated Public Turing Test To Tell Computers and Humans Apart*.

and image recognition as a form of a Turing test [37]. Hence, in the context of this paper, we address our third research question (**RQ3**) using experiments that we refer to as *Turing test-inspired*.

3. Methodology and results

The experiment was carried out in 5 stages as follows:

- **Stage 1:** Curating scenarios
- **Stage 2:** Collecting and evaluating human arguments
- **Stage 3:** Generating and evaluating machine arguments
- **Stage 4:** Turing-inspired tests
- **Stage 5:** Additional comparisons of machine and human arguments

3.1. Overview and rationale

The overall goals of the methodology are summarized in the aforementioned research questions. The individual stages and the sequence in which they were carried out were designed to provide the necessary data to answer the research questions or to provide the necessary data for the subsequent stages. In order to answer **RQ1**, we need to evaluate arguments generated using different LLMs and different prompt designs, using both zero-shot and few-shot learning styles. We ran some surveys in order to rate the different arguments (Stage 3) and each argument was rated independently. However, in order to run this study, we need some rules and scenarios that are used as the subjects of the interpretive arguments. We already had a dataset of ambiguous rules for this purpose. However, we needed to create a set of high-quality scenarios suitable for this task. This is where Stage 1 comes in. Additionally, we need some human arguments for two reasons. First, we needed some rated human arguments in order to address **RQ2** (Stage 2) and **RQ3** (Stage 4). We also used the human arguments for few-shot learning purposes, and therefore, the collected human arguments and their associated ratings (Stage 2) were also necessary for generating machine arguments. In order to address **RQ3**, we need to run a few Turing tests (Stage 4). Finally, in Stage 5, we ran some additional surveys to corroborate and validate the results we got in the earlier stages, as well as detect if there are any potential biases that the human annotators may have that lead them to favor some arguments due to the perception that an argument was human- or machine-generated.

3.2. Stage 1: Curating scenarios

Our goal at this stage was to curate a set of ambiguous scenarios for use in the subsequent stages. For this purpose, we started with a dataset of rules collected from the codes of ethics of various professional organizations, following [24]. However, we found some systematic issues with the scenarios used in that publication (see background section (§2)), so we decided to generate a new set of scenarios for this study.

When this experiment was conducted, GPT-3 and ChatGPT were the most advanced LLMs available. We used four different sources for the scenarios at this stage:

(1) GPT-3[3] generated scenarios
(2) ChatGPT[4] generated scenarios

[3] OpenAI's *text-davinci-003*.
[4] OpenAI's *ChatGPT January 9 2023 Version*.

(3) Human–machine collaboration using GPT-3
(4) Human–machine collaboration using ChatGPT

The machine-generated scenarios have been produced without human interference or oversight. On the other hand, human–machine collaboration means that a human was directly involved in the production of that set of scenarios. Specifically, for GPT-3, we used a multi-step *introspective* prompting style inspired by the self-ask prompting style [30] and similar chain-of-thought prompting styles. We used a 9-step prompt that starts by asking GPT-3 to identify an open-textured phrase in a given rule, generate multiple competing interpretations, generate arguments and counterarguments for each side, and iteratively improve those arguments. The prompt design has been developed through an iterative process of prompting the model, analyzing its outputs, identifying weaknesses, and either tweaking the prompt or adding additional intermediate prompts. Since the prompt got large at times, it was split into multiple steps, such that each step focused on one or a few goals. The full text of the prompt is provided in Appendix A.3. In the human–machine collaboration mode, a human reviewed, edited, or replaced the generated completions, including the intermediate steps leading to the final scenario.

Similarly, some of the scenarios were generated in collaboration with ChatGPT. This means that ChatGPT was instructed to generate appropriate scenarios. ChatGPT was given instructions to follow while generating the scenarios. Those instructions were typically provided in the first few messages in a conversation with the chatbot. The two most important sets of instructions given to ChatGPT are the guidelines for judging a scenario (Fig. 4) and the steps it needs to follow (Fig. 5). Those instructions were created through an iterated process of trial and error. Many of those instructions were added to guide the chatbot while generating the scenarios by manually reviewing the chatbot's responses and attempting to address some of the common issues that arose in the manual review process, which is part of the human–machine collaboration effort.

In total, 42 scenarios were generated using the aforementioned methods. A survey was created (using Qualtrics[5]) to assess those scenarios and was given to graduate students of our research lab. Figure 6 shows an example of a question in this survey. In total, 7 different annotators provided feedback, which was used to select 16 scenarios for the subsequent stages. The annotators were provided the same guidelines given to ChatGPT for reference. The scenarios were selected based on the recommendation of the

(1) The scenario should be ambiguous with respect to the interpretation of the rule.
(2) The scenario should provide good arguments for both sides of the debate.
(3) The scenario should be specific about the circumstances of the scenario and not rely on unspoken assumptions.
(4) The scenario should focus on the actions of the professional, not their intentions.
(5) The scenario should not be too long, as it will be part of a survey that users will have to read.
(6) The main source of disagreement about the scenario should be the ambiguity inherent in the language of the rule.
(7) The scenario should be reviewed and improved if necessary.
(8) Thinking, considering, or contemplating about doing something that violates the rule does not mean they violated the rule.
(9) Having good intentions, or thinking about doing something good in the future does not mean they are following the rule either.
(10) The ambiguous phrase should be a specific phrase, typically no more than three or four words, that can be interpreted in multiple ways in the context of the rule.

Fig. 4. The annotation guidelines that were given to ChatGPT for generating scenarios.

[5] https://www.qualtrics.com/

(1) The profession and rule are provided.
(2) Identify an open-textured phrase in the rule that can be interpreted in multiple ways.
(3) Mention at least two possible competing interpretations
(4) Generate a scenario that highlights the contention of the different interpretations.
(5) Think of possible arguments for compliance
(6) Think of possible arguments for non-compliance
(7) Evaluate whether the scenario is ambiguous and whether the main source of disagreement is the ambiguity inherent in the language of the rule.
(8) On a scale of 1-5, where 1 means it is too easy and 5 means it is too difficult. Rate how easy it is to argue for compliance.
(9) On a scale of 1-5, where 1 means it is too easy and 5 means it is too difficult. Rate how easy it is to argue for non-compliance.
(10) Think of possible counterarguments for the arguments for compliance and non-compliance
(11) Consider whether the scenario relies on unspoken assumptions and if it focuses on actions rather than intentions.
(12) Reflect on the scenario and see if it can be improved or if there are any missing information that could make the scenario more clear.
(13) Finalize the scenario and present it for review.

Fig. 5. The steps that ChatGPT was instructed to follow for generating scenarios.

annotators, where the provided scenarios met the guidelines and were believed to be balanced based on the total feedback we received at this stage.

3.3. Stage 2: Collecting and evaluating human arguments

At this stage, we had 16 scenarios from the previous stage. Our goal was to collect 10 different arguments for each scenario: 5 arguments in favor of compliance and 5 in favor of non-compliance. We created a survey with 10 questions each and used Amazon Mechanical Turk[6] to recruit human annotators. Figure 7 shows an example of a question in the survey. Participants had to meet the following criteria to participate:

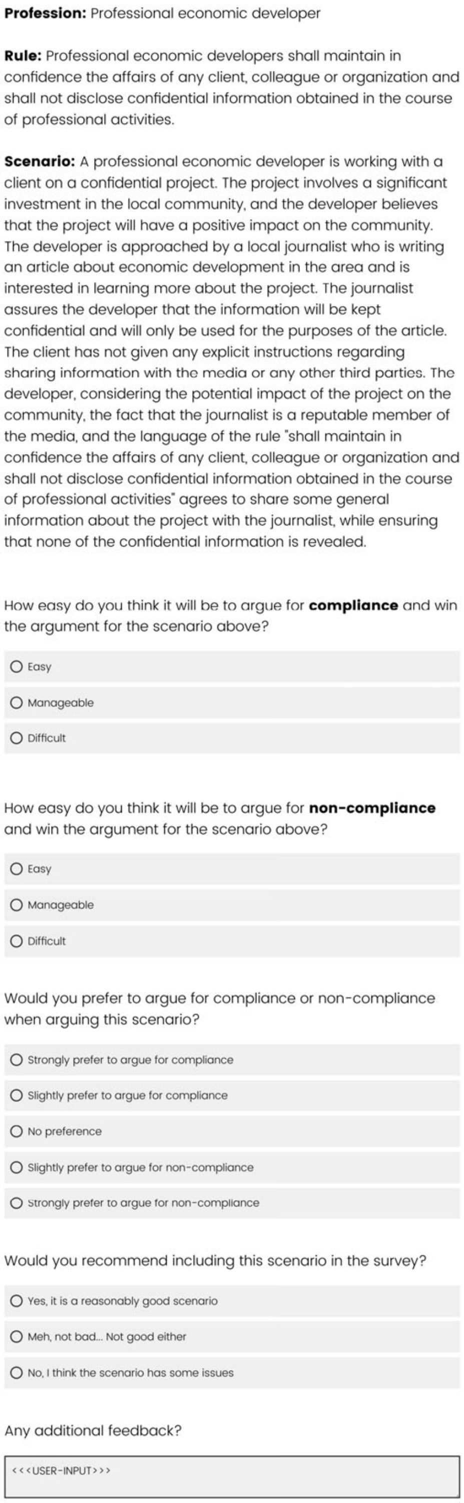
- Participants must have at least 50 completed and approved tasks on the platform.
- Participants must have an approval rate of 92% or better.

The collected arguments were reviewed for quality, bad faith, or low effort, and the offending submissions were removed. Such submissions were identified by manually reviewing a small sample of arguments from each user. Reasons for rejecting a submission include:

- Submissions that were determined to be complaints or statements about the task itself rather than arguments made in the spirit of the task. These include rants about the scenario (e.g., "This is preposterous, I cannot believe we're asked to argue in favor of (non-)compliance!"), or very short responses (e.g., "That's clearly unethical!").
- Arguing for the wrong stance (ie. arguing for compliance when asked to argue for non-compliance or vice versa).
- Very poor grammar that makes the submission ineligible for inclusion.

Only one submission fit the above criteria and was rejected for one or more of the above reasons. We also looked for arguments that were consistently rated as "This argument is meaningless, irrelevant, or

[6] https://requester.mturk.com



Profession: Professional economic developer

Rule: Professional economic developers shall maintain in confidence the affairs of any client, colleague or organization and shall not disclose confidential information obtained in the course of professional activities.

Scenario: A professional economic developer is working with a client on a confidential project. The project involves a significant investment in the local community, and the developer believes that the project will have a positive impact on the community. The developer is approached by a local journalist who is writing an article about economic development in the area and is interested in learning more about the project. The journalist assures the developer that the information will be kept confidential and will only be used for the purposes of the article. The client has not given any explicit instructions regarding sharing information with the media or any other third parties. The developer, considering the potential impact of the project on the community, the fact that the journalist is a reputable member of the media, and the language of the rule "shall maintain in confidence the affairs of any client, colleague or organization and shall not disclose confidential information obtained in the course of professional activities" agrees to share some general information about the project with the journalist, while ensuring that none of the confidential information is revealed.

How easy do you think it will be to argue for **compliance** and win the argument for the scenario above?

○ Easy

○ Manageable

○ Difficult

How easy do you think it will be to argue for **non-compliance** and win the argument for the scenario above?

○ Easy

○ Manageable

○ Difficult

Would you prefer to argue for compliance or non-compliance when arguing this scenario?

○ Strongly prefer to argue for compliance

○ Slightly prefer to argue for compliance

○ No preference

○ Slightly prefer to argue for non-compliance

○ Strongly prefer to argue for non-compliance

Would you recommend including this scenario in the survey?

○ Yes, it is a reasonably good scenario

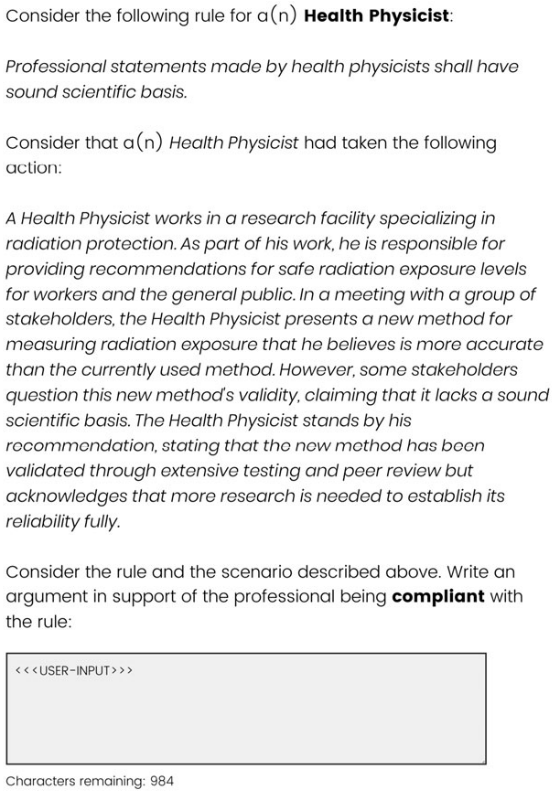
○ Meh, not bad... Not good either

○ No, I think the scenario has some issues

Any additional feedback?

<<<USER-INPUT>>>

Fig. 6. A survey question at stage 1.



Consider the following rule for a(n) **Health Physicist**:

Professional statements made by health physicists shall have sound scientific basis.

Consider that a(n) *Health Physicist* had taken the following action:

A Health Physicist works in a research facility specializing in radiation protection. As part of his work, he is responsible for providing recommendations for safe radiation exposure levels for workers and the general public. In a meeting with a group of stakeholders, the Health Physicist presents a new method for measuring radiation exposure that he believes is more accurate than the currently used method. However, some stakeholders question this new method's validity, claiming that it lacks a sound scientific basis. The Health Physicist stands by his recommendation, stating that the new method has been validated through extensive testing and peer review but acknowledges that more research is needed to establish its reliability fully.

Consider the rule and the scenario described above. Write an argument in support of the professional being **compliant** with the rule:

<<<USER-INPUT>>>

Characters remaining: 984

Fig. 7. A survey question at stage 2.

nonsense" during the evaluation phase. However, we found no such submissions, which indicated that the manual screening process successfully excluded all such arguments.

Additionally, we used an online tool[7] to detect any submissions that are likely to be generated using a GPT variant. We found, through informal tests, that this tool and similar tools frequently produced false positives and false negatives, which aligns with recent work on AI-generated text detectors [34], suggesting that such tools, at best, can only perform marginally better than random at distinguishing AI- from human-written text. However, some form of detection was necessary, so we only considered rejecting cases where the detector reported that two or more responses were AI-written text with a 95% confidence or higher. See our discussion in the conclusion section (§6) on the topic of bot detection for further discussion of this issue. We accepted submissions with one positive result in order to reduce the chances of rejecting legitimate submissions. This procedure identified only one submission where the participant was believed to have used an AI text generator in their submissions.

After collecting the human arguments, we had 16 scenarios, with 5 human arguments in favor of compliance and 5 human arguments in favor of non-compliance each, for a total of 160 human arguments. Our next goal was to evaluate the persuasiveness of the collected arguments by collecting at least 5 different annotations for each collected argument. For this purpose, we created a survey with two question formats. In the first question format, users were given a profession, a rule, a scenario, a stance, and an argument. They were asked to rate the argument on a Likert scale [1] from 1–5, going from "Very con-

[7] https://openai-openai-detector.hf.space/

vincing" to "Very unconvincing" with an additional choice that reads "This argument is meaningless, irrelevant, or nonsense". This additional option allows participants to indicate that there is an issue with the provided argument and that a standard Likert scale does not apply. This allowed us to detect if an argument was written in bad faith or did not conform to our expectations such as arguing for the wrong stance. Users were given 5 different arguments for the same scenario-stance pair. In the second question format, for each scenario-stance pair, users were asked to order the same arguments from the preceding Format 1 questions from most convincing to least convincing. Each user was given 5 randomly chosen scenario-stance pairs to annotate for a total of 25 questions of the first format and 5 questions of the second format. Refer to Appendix A.1 for the question templates used in those surveys.

In order to further ensure submission quality, a red-flags quality control system was developed that reviews the submissions for quality control signals that indicate issues with the submission. This system was developed by manually reviewing a small sample of submissions and tweaking the sensitivity of the criteria to ensure that submissions that were clearly of good quality or poor quality were classified correctly and iteratively refining the criteria until the automated quality assessment matched the results of the manual review. A random or low-effort submission is expected to generate 10 or more red flags. Any items with 4 or more red flags were removed from the dataset. The red flags are based on 3 types of quality control signals, each with their own subset of items. The major types of red flags were as follows:

- **Time-based:** For example, any question submitted within 6 seconds or less generates a flag.
- **Order-based:** The order provided in the second question format had to be more or less consistent with the individual ratings of the arguments. For example, an argument rated as "Very convincing" is expected to be ranked higher than an argument rated as "Unconvincing". Within some tolerance, if this expectation is not met, a flag is raised.
- **Consensus-based:** The first two types only look at individual submissions. However, this third type considers the consistency of the ratings among the participating population. The median (and deviation) rating of each argument is noted for the subset of responses that generated no flags for time-based and order-based quality indicators. Within some tolerance, submissions that consistently rate arguments outside the median range raise a flag.

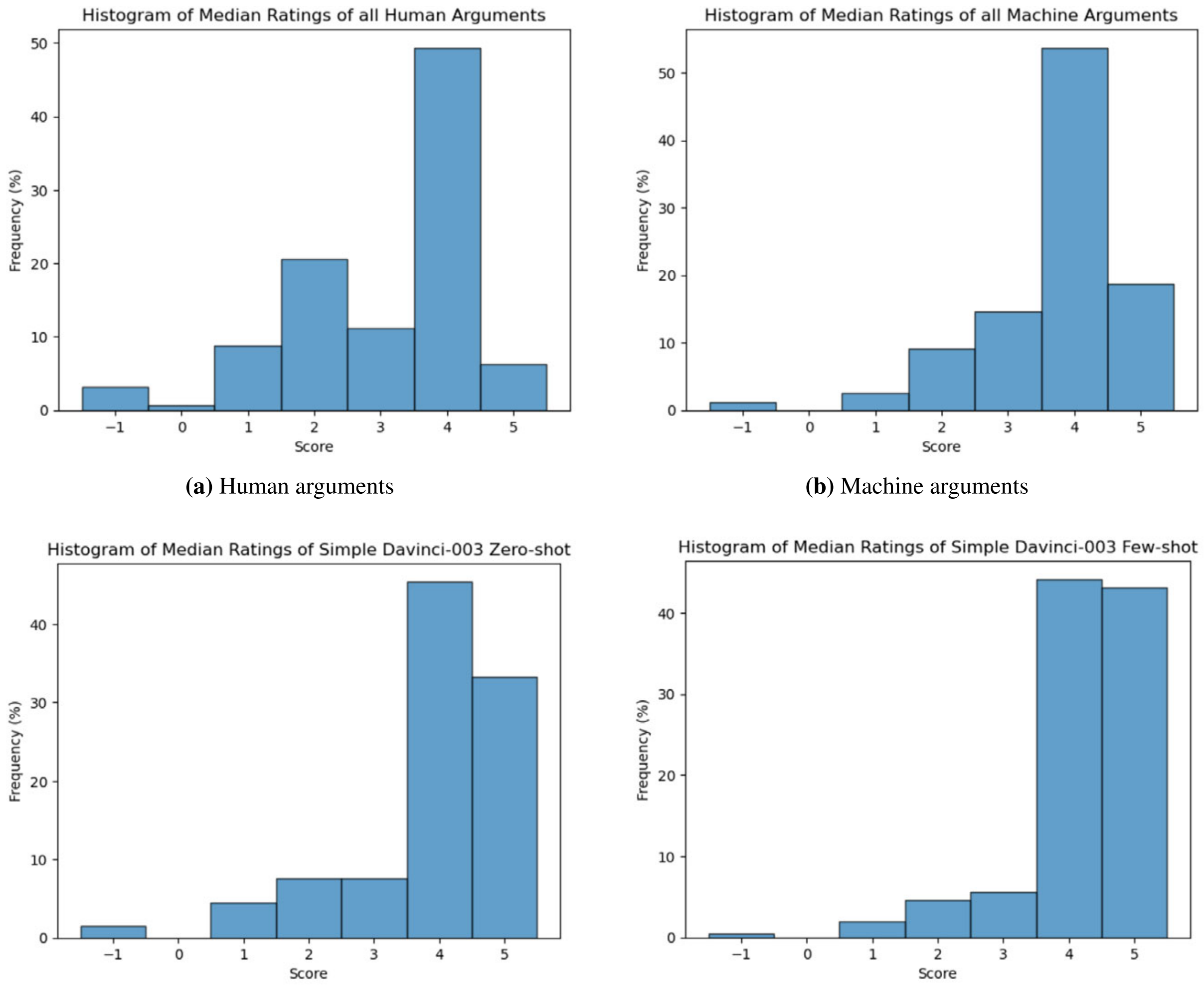
For additional details about the quality control system, please refer to the GitHub page linked earlier.

3.3.1. Results

In order to evaluate the collected human arguments, we found the median rating of each human argument. Responses were scored as follows: "Very convincing" (5), . . . , "Very unconvincing" (1), "This argument is meaningless, irrelevant, or nonsense" (-1). The median rating of the five available annotations is calculated for each argument. The median rating was chosen to represent the persuasiveness of each argument in order to exclude any outliers compared to the majority of the votes. Subsequently, the average for all human arguments is calculated. Similarly, the standard deviation is also calculated using the median ratings of all the human arguments. The average median for all human arguments is 3.06 with a standard deviation of 1.35. The item-weighted average (ie. where all arguments are given equal weight) is 2.91. Figure 8(a) shows the distribution of the median ratings of all human arguments.

3.4. Stage 3: Generating and evaluating machine arguments

At this stage, we had 16 scenarios with 5 human arguments in favor of compliance and 5 human arguments in favor of non-compliance for each scenario. All of those arguments had 5 different annotations



(a) Human arguments

(b) Machine arguments

(c) Simple Davinci-003 ZS arguments

(d) Simple Davinci-003 FS arguments

Fig. 8. Histograms of median ratings.

of their persuasiveness. Our goal at this stage was to generate 20 different arguments for each scenario using LLMs: 10 arguments in favor of compliance and 10 in favor of non-compliance.

We used GPT-3 models of different sizes, prompt designs, and both zero-shot and few-shot prompting. Table 1 lists all the different arrangements that were used in this experiment. These choices were selected in order to answer **RQ1**. Specifically, we used two prompt designs. The first is a simple prompt copied directly from the question format in Stage 2 (Appendix A.4). For the few-shot learning version of this prompt, we used 3-shot learning. The number of exemplars was chosen based on the smallest context size of the available models. *text-curie-001* model has a context size of 2000 tokens, while the *text-davinci-003* model has a context size of 4000. Therefore, the output of the simple prompts needed to remain below 2000 tokens. Three exemplars was the maximum number that could be consistently included in the prompt without exceeding the maximum token limit. For each argument, we first excluded all human arguments that match the profession of the argument in question. The remaining arguments were sorted from the highest quality to the lowest quality (based on the evaluations from Stage 2), and the 6

Table 1

Summary of the models and prompts used to generate arguments in stage 3

Prompt type	Shots	Model	Count
Simple	0	davinci-003	1
Simple	0	curie-001	1
Simple	3	davinci-003	1
Simple	3	curie-001	1
Introspective	0	davinci-003	1
Introspective	0	davinci-002	1
Introspective	2	davinci-003	2
Introspective	2	davinci-002	2
Total			10

best arguments were selected. From this set, we randomly selected a subset of 3 arguments for use as exemplars and presented them as part of the prompt in a random order. The exemplars selected for each argument generation are part of the dataset.

The second prompt design is a 10-step introspective design that allows the LLM to formulate and improve its arguments iteratively. The introspective prompt pipeline is based on a multitude of sources for inspiration. First of all, with analogy to the chain-of-thought prompting and similar approaches, these prompts break down the task of interpretive argumentation into discrete tasks, which, in concert, are designed to generate quality arguments. The individual questions and steps were inspired by multiple sources. Many publications provide systems that are targeted at human arguers in order to generate good arguments systematically. For example, some authors propose critical questions that a good interpretive argument must address [17,43,45]. These types of questions can be used to guide an LLM in producing high-quality arguments. The LLM is repeatedly asked to elaborate on critical aspects of its response, attempt to find weaknesses or omissions in its arguments, and iteratively improve those arguments as it follows the steps in the introspective pipeline. The full text of the introspective prompting pipeline is provided in Appendix A.5. The prompts used in this prompting style are significantly larger than the ones used in the simple prompting style. For this reason, the *text-curie-001* context size of 2000 tokens was not a suitable choice for the introspective prompts, especially for the few-shot learning setting. That is why we chose only to use the models with the largest context size.

For the few-shot learning version of this prompt, we used 2-shot learning. Similar to the simple prompting style, the number of exemplars was made to ensure the prompt fits in the context size of 4000, the context size of the largest models. We have manually curated 4 exemplars that span all 10 steps of this prompt design. No profession occurs more than once in this set of exemplars. Figure 1 shows an argument generated in favor of compliance using this prompt design. For each argument, we first excluded all exemplars that match the profession of the argument in question. From the set of the remaining exemplars, we randomly selected a subset of 2 arguments for use and presented them as part of the prompt in a random order. The exemplars selected for each argument generation are part of the dataset.

Those two prompt designs were chosen to help answer **RQ1**, namely, whether an introspective prompt design could yield better arguments compared to the simpler prompt design. We also used both zero-shot and few-shot prompts to answer the research question of whether few-shot learning could yield better arguments compared to a zero-shot learning design. Finally, we used models of different sizes. For the simple prompts, we used *text-davinci-003* and *text-curie-001*; while for the introspective prompts, we used *text-davinci-003* and *text-davinci-002*. As mentioned earlier, this experiment was carried out before GPT-4 was released; see Section 4 for a small follow-up experiment carried out using GPT-4.

At this point, we had 16 scenarios with 5 human arguments in favor of compliance and 5 human arguments in favor of non-compliance, as well as 10 machine arguments in favor of compliance and 10 machine arguments in favor of non-compliance for a total of 320 machine-generated arguments. All of the human arguments had 5 different annotations of their persuasiveness. Our goal was to evaluate the persuasiveness of the machine-generated arguments by collecting at least 5 different annotations for each of the machine-generated arguments. For this purpose, we created a survey using a format identical to the survey in Stage 2. Each question contained 5 arguments; therefore, we had double the number of questions at this stage compared to the previous one. Each scenario-stance pair had to be included twice with a different subset of five questions from the ten that were generated. We split the arguments based on their source deterministically, and hence, the Format 2 questions required participants to order arguments for one of the two available subsets consistently. We included the same number of questions in each survey as Stage 2, and therefore, we required (roughly) twice the number of participants to fill out the survey. Everything was identical to Stage 2 in terms of the surveys, except that the source of the arguments was machine-generated, which the participants were not made aware of as we did not make any claims about the source of the arguments in this stage or the previous one.

3.4.1. Results

Using the same scoring scheme as the one in Stage 2, the average median rating for all machine-generated arguments is 3.63 with a standard deviation of 1.06. The item-weighted average is 3.45. Figure 8(b) shows the distribution of the median ratings for all machine arguments. These results indicate that human annotators believe that machine-generated arguments were consistently more persuasive than human-written arguments. Table 2 shows a detailed breakdown of all the experiments we carried out. In order to verify the statistical significance of those results, we conducted one-way ANOVA for all of the machine-generated arguments. We got an F-statistic of 30.5 and $p < 0.01$, which shows a significant difference between the group means. Appendix A.7 shows the Tukey HSD pairwise comparisons of all the means.

Examining those results, we find that few-shot learning clearly outperforms the zero-shot learning prompts which matches our expectations. We also find that the *text-curie-001* model using 3-shot learning and simple prompts is roughly equivalent to the performance of the human arguments. The same

Table 2

Average median rating of each source of arguments from stages 2 and 3

Prompt type	Shots	Model	Average rating*	StdDev
Simple	0	davinci-003	4.13	0.57
Simple	0	curie-001	2.58	1.53
Simple	3	davinci-003	4.39	0.45
Simple	3	curie-001	3.02	1.12
Introspective	0	davinci-003	3.38	0.88
Introspective	0	davinci-002	3.38	0.75
Introspective	2	davinci-003	4.07	0.78
Introspective	2	davinci-002	3.64	0.90
Machine			**3.63**	**1.06**
Human			**3.06**	**1.35**

* Higher is better

Table 3

Average median order of each source of arguments from stage 3 (subset 1)

Prompt type	Shots	Model	Average order*	StdDev
Simple	0	davinci-003	2.11	0.76
Simple	3	curie-001	3.94	0.83
Introspective	0	davinci-003	3.59	0.72
Introspective	2	davinci-003	2.66	1.01
Introspective	2	davinci-002	2.95	0.82

* Lower is better

Table 4

Average median order of each source of arguments from stage 3 (subset 2)

Prompt type	Shots	Model	Average order	StdDev
Simple	0	curie-001	4.02	0.91
Simple	3	davinci-003	2.02	0.69
Introspective	0	davinci-002	3.67	0.71
Introspective	2	davinci-003	2.50	1.24
Introspective	2	davinci-002	3.11	0.72

model with zero-shot learning is the only model that underperforms compared to the human results. We also notice that the introspective prompt design underperforms compared to the simpler prompting approach. Overall, the machine-generated arguments outperform the human arguments. Tables 3 and 4 show the average median for all the machine-generated arguments. Since the argument sources were split deterministically, those tables show the results of each of the two subsets independently. The arguments that were labeled as most convincing (top) got a score of 1, while the ones labeled as least convincing (bottom) got a score of 5. The results show that the relative ordering of all sources of arguments is consistent with the average median rating in Table 2.

3.5. *Stage 4(a): Blind Turing test*

At this stage, we had 16 scenarios with 5 human arguments and 10 machine arguments for each stance. All of those arguments had 5 annotations of their persuasiveness. Our goal at this stage was to test the human annotator's ability to tell the difference between human and machine arguments. For this purpose, we created a survey with two question formats. In the first format, users were presented with a profession, a rule, a scenario, a stance, and an argument. They were asked to guess as to whether the argument was written by a human or generated by a machine. In the second format, we presented users with two arguments instead of one, informed them that one of those arguments was written by a human and the other was generated by a machine, and they had to guess which one was written by a human. Refer to Appendix A.1 for the question templates used in those surveys.

We refer to this as a Turing test-inspired variant, adapted specifically for interpretive arguments. As discussed earlier, the Turing test at the conceptual level is a test to evaluate a machine's level of intelligence by assuming that if a machine could perform certain tasks in a way that is indistinguishable from humans performing the same task, then that machine can be reasonably said to have at least the same level of intelligence as humans. The original Turing test [38] allows humans to interact with the machine to stress test the machine's ability to pass tailored tests to the satisfaction of the human testers. In our simplified version, such an interaction is absent, and we are simply asking human annotators to tell human and machine arguments apart.

Table 5

Blind Turing test results from stage 4(a)

Prompt type	Shots	Model	Accuracy%	Avg score	StdDev	t-tests
Simple	0	davinci-003	55	0.12	1.51	-
Simple	0	curie-001	61	0.27	1.55	-
Simple	3	davinci-003	51	0.16	1.42	Pass
Simple	3	curie-001	47	0.02	1.26	Pass
Introspective	0	davinci-003	38	−0.15	1.33	Pass
Introspective	0	davinci-002	34	−0.36	1.30	Pass
Introspective	2	davinci-003	50	−0.08	1.44	Pass
Introspective	2	davinci-002	52	0.21	1.33	Pass
Overall			52	0.15	1.43	Pass

Table 6

Blind Turing test confusion matrix of format 1 questions

Source	Guessed as machine	Not sure	Guessed as human
Machine	46%	3%	51%
Human	61%	6%	33%

3.5.1. *Results*

Table 5 shows a detailed breakdown of all the human annotators' accuracies on the Blind Turing test; if the supplied response was correct (ie. guessed as "Very likely [correct guess]" or "Somewhat likely [correct guess]"), the item was counted as correctly labeled, while incorrect or "Not sure" responses were counted as incorrectly labeled. The results also show that in the Blind Turing test, annotators correctly identified the human arguments only 52% of the time, which is basically no better than random chance. Responses were also scored as follows: "Very likely [correct guess]" (2), ..., "Very likely [incorrect guess]" (−2), and the average scores are shown in the table as well. Table 6 shows the confusion matrix detailing how each source of arguments was classified based on a 3 categories classification of 'Machine', 'Human', and 'Not sure'.

In order to verify the statistical significance of those results, we conducted two one-tailed one-sample t-tests. For the first t-test, we defined the null hypothesis as the true mean is larger than or equal to 0.5. Recall that an annotation of "Not sure" is scored as 0, while an annotation of "Somewhat likely [correct guess]" is scored as 1. Therefore, a mean value larger than or equal to 0.5 indicates that participants could reliably identify the source of an argument. The alternative hypothesis is that the true mean is less than 0.5, which means that participants were unable to identify the source of the arguments correctly. If the test shows a statistically significant result, the Turing test is marked as a 'Pass' indicating that the machine passed the Turing test. If the test is inconclusive, a second t-test is performed. For the second t-test, we defined the null hypothesis as the true mean is less than or equal to 0. The alternative hypothesis is that the true mean is larger than 0, which means that participants were able to identify the source of the arguments correctly. If the test shows a statistically significant result, the Turing test is marked as a 'Fail', indicating that the machine failed the Turing test. Otherwise, the result is inconclusive and is marked with '-'. This procedure allows us to look for evidence that the annotators were unable to guess the correct source of the arguments reliably. The "Not sure" option is scored as a 0, and for the purposes of the first t-test, we define a mean in the range of −0.5 to 0.5 to correspond to this option. In order to pass the Turing test, participants have to consistently choose either the "Not sure" option or any of the incorrect guesses. If we cannot show that the Turing test was passed, we attempt to look for evidence that the machine failed the Turing test, and we look for evidence that the annotators were able to make better than random guesses, hence a true mean above 0. If we find no evidence to support either hypothesis, we conclude that the result is inconclusive.

In the first one-tailed one-sample t-test, the overall sample size is 459. We find that the mean is 0.15, the standard deviation is 1.43, the t-statistic is -6.17, $p < 0.05$, the effect size is -0.25, and since the mean is below 0.5, we conclude that overall GPT-3 has passed the Blind Turing test. Additionally, Table 6 shows the confusion matrix of the Format 1 questions. The results show that the participants had a tendency to label human arguments as machine arguments, as 61% of the human arguments were labeled as machine arguments. On the other hand, machine arguments were (roughly) equally likely to be labeled as machine or human arguments, as only 46% of the machine arguments were labeled correctly, and 51% were labeled as human arguments.

3.6. *Stage 4(b): Guided Turing test*

The blind Turing test was designed to evaluate whether machine-generated arguments were similar enough to human arguments in order to substitute for human-provided arguments in applications where the source of the argument is irrelevant or of low importance. However, could this be a consequence of the human participants' lack of expertise in the task? I.e., is it the case that humans could *learn* to distinguish human- from machine-generated interpretive arguments? Our guided version of the test was therefore designed to evaluate whether machine-generated arguments had distinctive features differentiating them from human arguments that human annotators could learn.

This stage was carried out in an identical manner to the blind version of the test with one notable modification: We first provided users with 8 examples of human arguments and 8 examples of machine arguments as part of the survey instructions. Users were provided with a downloadable text file to refer to the provided examples when needed, rather than relying entirely on their memory. Everything else was identical to the previous stage.

3.6.1. *Results*

Table 7 shows a detailed breakdown of all the human annotators' accuracies on the Guided Turing test. The results also show that in the Guided Turing test, annotators correctly identified the human arguments 62% of the time, and the average response score was 0.53 with a standard deviation of 1.44. In the first one-tailed one-sample t-test, the overall sample size is 463. In the first t-test, we find no statistically

Table 7

Guided Turing test results from stage 4(b)

Prompt type	Shots	Model	Accuracy%	Avg score	StdDev	t-tests
Simple	0	davinci-003	72	0.84	1.17	Fail
Simple	0	curie-001	38	−0.16	1.48	Pass
Simple	3	davinci-003	63	0.68	1.40	Fail
Simple	3	curie-001	60	0.50	1.40	Fail
Introspective	0	davinci-003	74	0.77	1.25	Fail
Introspective	0	davinci-002	57	0.34	1.44	-
Introspective	2	davinci-003	61	0.51	1.47	Fail
Introspective	2	davinci-002	71	0.76	1.40	Fail
Overall			62	0.53	1.44	Fail

Table 8

Guided Turing test confusion matrix of format 1 questions

Source	Guessed as machine	Not Sure	Guessed as human
Machine	64%	2%	34%
Human	59%	6%	35%

Table 9

Average median rating of each source of arguments when compared to a human argument at stage 5

Prompt type	Shots	Model	Average rating	StdDev
Simple	0	davinci-003	1.32	1.15
Simple	0	curie-001	−0.29	1.51
Simple	3	davinci-003	1.21	1.09
Simple	3	curie-001	0.24	1.53
Introspective	0	davinci-003	0.41	1.19
Introspective	0	davinci-002	0.15	1.43
Introspective	2	davinci-003	0.94	1.23
Introspective	2	davinci-002	0.48	0.95
Overall			0.80	1.20

significant evidence that the machine passed the Guided Turing test. So we ran the second t-test, and found the mean is 0.53, the t-statistic is 9.05, $p < 0.05$, the effect size is 0.37, and concluded that overall the machine failed the Guided Turing test. Additionally, Table 8 shows the confusion matrix of the Format 1 questions. The results show that the participants had a tendency to label most arguments as machine arguments, as 64% of the machine arguments and 59% of the human arguments were labeled as machine arguments.

3.7. Stage 5: Additional comparisons of machine and human arguments

At this stage, we wanted to perform some additional validation on our results. One of the crucial questions we attempted to address at this stage is whether or not annotators had a bias favoring either the human or machine arguments as such. For this reason, we decided to run the survey where we tell the users that an argument was human- or machine-generated, and observe how consistent the results were between the direct comparisons performed in Stage 4 (second question in the second format) and the indirect comparisons performed in Stages 2 and 3. Refer to Appendix A.1 for the question template used in this survey.

3.7.1. Results

For this stage, we found the median rating of all human–machine argument pairs included in the survey. Responses were scored as follows: "Machine argument is significantly more convincing" (2), . . . , "Human argument is significantly more convincing" (−2).

The average median for all the evaluated pairs is 0.80 with a standard deviation of 1.20. Table 9 shows a detailed breakdown of all the experiments. In order to verify the statistical significance of those results, we conducted one-way ANOVA for all of the machine-generated arguments. The sample size is 575, and we got an F-statistic of 11.2 and $p < 0.01$, which shows a significant difference between the group means.

Table 10

Summary of the models and prompts used to generate arguments

Prompt type	Shots	Model	Count
Simple	0	gpt-4-0314	1
Simple	3	gpt-4-0314	1
Introspective	0	gpt-4-0314	1
Introspective	2	gpt-4-0314	2
Total			5

4. Experiment using GPT-4

GPT-4 [26] has been shown to outperform GPT-3 in virtually every measure available, and it is reasonable to assume that the ability to generate persuasive interpretive arguments is among those measures. Although GPT-4 was not available at the time of our experimentation, it would be worthwhile to see whether the lessons we learned above about which prompting styles generate the best interpretive arguments. We, therefore, report on a small experiment we carried out to do just that.

Using the same procedure in Stage 3, we generated 10 different arguments for each scenario using GPT-4 (specifically, *gpt-4-0314*): 5 arguments in favor of compliance and 5 in favor of non-compliance. We used the same prompt designs and both zero-shot and few-shot prompting. Table 10 shows the different arrangements that were used in this experiment.

However, unlike the previous set of experiments, which were carried out on Amazon Mechanical Turk, we used the Prolific[8] platform to carry out this experiment. Additionally, we imposed some restrictions on participants that were not enforced in the previous set of experiments. Namely:

- Participants are from the US or UK
- Participants must have a graduate degree
- Participants are balanced based on gender
- Participants must have completed at least 50 tasks on the platform
- Participants must have an acceptance rate of 92% or better

After we carried out the experiment, we had 16 scenarios with 5 arguments in favor of compliance and 5 arguments in favor of non-compliance for a total of 160 arguments generated using GPT-4. All arguments were evaluated for their persuasiveness by collecting at least 3 different annotations for each of the arguments. The surveys were in a format identical to the surveys in Stages 2 and 3.

We report two sets of results. One using a filtered subset of annotations, and another using an unfiltered subset. The filtered subset uses the red flags system explained in Section 3.4, while disabling the *Consensus-based* red flags subset. Those flags were disabled because they are based on the consensus of multiple annotations for each argument. Since those filters rely on the availability of at least 5 annotations that are validated through the other filtration criteria, not enough data was collected to allow those filters to be reliable. The other types of red flags evaluate submissions independently of other submissions and, hence, are not affected by the amount of available submissions. Since we still use a threshold of a maximum of 3 flags to accept a submission, the filtration process is slightly less strict compared to Stage 3 filters. Additionally, in Stage 3 we collected at least 5 annotations for all arguments *after* applying the filters. However, for this smaller experiment, we only have 3 annotations for each argument in the unfiltered set. Hence, some arguments have less than 3 annotations available with the filters.

[8] https://www.prolific.co/

Table 11

Average median rating of each source of arguments

Prompt type	Shots	Model	Average rating (filtered)	StdDev	Average rating (unfiltered)	StdDev
Simple	0	gpt-4-0314	3.72	0.94	3.91	0.72
Simple	3	gpt-4-0314	3.80	0.77	3.86	0.74
Introspective	0	gpt-4-0314	3.40	1.14	3.58	0.88
Introspective	2	gpt-4-0314	3.65	0.97	3.55	0.84
Overall			3.64	0.98	3.69	0.82

Table 12

Average median order of each source of arguments

Prompt type	Shots	Model	Average order (filtered)	StdDev	Average order (unfiltered)	StdDev
Simple	0	gpt-4-0314	2.92	1.24	2.73	1.17
Simple	3	gpt-4-0314	3.05	1.14	2.56	1.04
Introspective	0	gpt-4-0314	2.92	1.39	3.03	1.30
Introspective	2	gpt-4-0314	2.99	1.26	3.20	1.19

4.0.1. Results

Using the same scoring scheme as the one in Stage 3, the average median for all GPT-4 arguments in the unfiltered set is 3.69 with a standard deviation of 0.82, and the item-weighted average is 3.51. For the filtered subset, the average median is 3.64 with a standard deviation of 0.98, and the item-weighted average is 3.63. Table 11 shows a detailed breakdown of all the experiments we carried out. In order to verify the statistical significance of those results, we conducted one-way ANOVA for all of the machine-generated arguments and got an F-statistic of 4.59 and $p < 0.01$ for the unfiltered set, which shows a significant difference between the group means. Appendix A.8 shows the Tukey HSD pairwise comparisons of all the means. Unfortunately, the results for the filtered subset do not show statistical significance, which is why we report both sets of results. Table 12 shows the relative ordering of the different sources of arguments and is consistent with the average median ratings from Table 11.

Examining those results, we find that the introspective prompt design underperforms compared to the simpler prompting approach. The results do not show a statistically significant difference between few-shot and zero-shot prompts (see Appendix A.8). The filtered subset shows results consistent with the pattern that few-shot learning outperforms zero-shot learning. In contrast, the unfiltered subset shows the opposite pattern with very small differences between the means. Again, since those differences are not statistically significant, we refrain from making any conclusions. However, it seems that the filters have some influence on the results. The filtered data is more consistent with the results of the experimental protocol developed for Stage 3. It should be considered more reliable as it excludes submissions that may represent lower effort or lower quality submissions. Nonetheless, additional experimentation is warranted to find out if the chat versions of LLMs that are instruction-tuned or optimized using reinforcement learning from human feedback (RLHF) might affect the performance of the LLMs in the few-shot learning setting on this task or the prompting style used which has been developed with the non-chat LLMs in mind.

Note that the ratings obtained in this experiment should not be interpreted to be directly comparable to the results in Stage 3 due to the fact that this experiment has been conducted on a different platform using different participation criteria, which influences the characteristics of the population of the participants. Therefore, we do not use those results to compare the performance of GPT-4 to the performance of GPT-3 in the previous experiments.

5. Discussion

We can now address our three primary research questions.

RQ1: How well can state-of-the-art LLMs argue interpretively, and which prompting methods generate the best interpretive arguments?. The results from Stage 3, as summarized in Table 2, show that the best performing model is the largest model offered by OpenAI, namely *text-davinci-003*, which produced the best results using a simple prompting scheme combined with a few-shot learning approach. The model obtained an average median rating of 4.39 with a standard deviation of 0.45, which indicates that most annotators rated the generated arguments as either 'Convincing' (4) or 'Very Convincing' (5). This indicates that most annotators believed the model generated arguments of decent quality. The results also show that larger and newer models consistently and measurably outperformed smaller or older models.

The results also show that using few-shot learning yields substantially better results compared to zero-shot learning for all the models tested. This result is what we expected and demonstrates the value the training exemplars bring to this task.

We also observe that the introspective prompt design underperforms compared to the simpler prompting approach. While we believe that an introspective approach similar to the one used in those experiments may be of value, those results indicate that creating a prompting pipeline that outperforms the simple prompting approach is a non-trivial task and requires an iterative approach of design, test, and analysis. Since the introspective approach used in this set of experiments was the first attempt at such a prompt design, and did not benefit from any substantial feedback or analysis, we argue that those results are only a measure of the specific template used in this experiment and should not be used to conclusively dismiss similar prompting styles that are developed with the benefit of an iterative approach that incorporates additional feedback and analysis.

RQ2: How do human- and machine-generated interpretive arguments compare?. The results from Stage 2 and 3, as summarized in Table 2, show that the largest model offered by OpenAI, namely *text-davinci-003*, consistently produced arguments that were rated as more persuasive as compared to the human arguments. This is strong evidence in favor of LLMs compared to non-expert human arguers. Using *text-curie-001* (which is a variant of GPT-3 that is advertised by OpenAI as a smaller, faster, cheaper, and capable alternative) produced arguments that were rated as equally persuasive as their human counterparts when using a few-shot learning approach while also rated as less persuasive compared to human arguments when using a zero-shot learning approach. This means that an LLM that matches or exceeds the performance of the *text-curie-001* variant of GPT-3 can produce arguments that are competitive with non-expert human arguments.

All of those results are corroborated by the results from Stage 5, as summarized in Table 9, which produced consistent results using a different experimental setup.

RQ3: Can people distinguish between human- and machine-generated interpretive arguments?. The results from Stage 4(a), as summarized in Table 5, show that participants could not identify the source of an argument correctly at a rate that is statistically significantly better than random chance in the blind version of the Turing test. Conversely, the results from Stage 4(b), as summarized in Table 7, show that participants were able to correctly identify the source of an argument for all models except for the *text-curie-001* using a simple prompt with zero-shots. Recall from Stage 3 that this model was the only one that underperformed compared to human arguments. This result implies that at least part of the reasoning used by the human annotators is the persuasiveness of the argument, where participants (seemingly) concluded based on the examples provided to them that the better arguments are usually the machine-generated arguments, and hence were more likely to incorrectly conclude that the arguments generated using *text-curie-001* using the simple prompt with zero-shots were human arguments. This provides some limited evidence that the persuasiveness of arguments was an important signal that most annotators relied on when selecting the source of an argument.

6. Conclusion

In this paper, we discussed the experimental design and the associated surveys and methods that were used to answer the research questions discussed earlier. Specifically, our goal was to develop experimental protocols to assess the persuasiveness of machine-generated arguments and to compare arguments generated using different SOTA LLMs to each other and human arguments. We also sought to find out if non-expert human annotators could distinguish human- and machine-generated arguments with and without prior exposure to such arguments. Our findings show that (non-expert) human annotators consistently rated machine-generated arguments as more convincing (when using LLMs beyond a certain size) than human arguments. The results also show that non-expert human annotators could not tell the human- and machine-generated arguments apart if they were not provided prior information about such arguments. This result indicates that machines are able to produce human-like text. The results also show that when provided with some examples of human- and machine-generated arguments, the annotators had a slightly improved ability to tell the two sources of arguments apart.

Those results are very encouraging for the future of automated interpretive argumentation as they demonstrate that the SOTA LLMs are capable of performing interpretive reasoning and interpretive argumentation with reasonable competence. Note that the comparisons carried out in those experiments were in comparison with non-expert human participants. A more thorough analysis of the available data is planned for future publication. Moreover, we recommend carrying out this experimental protocol while recruiting expert arguers for both the generation of human arguments and the evaluation of both the human and machine arguments. Insights from experts can definitely create a more accurate, reliable, and trustworthy assessment of the capabilities of current LLMs.

The use of this experimental protocol to compare the performance of different LLMs can be an essential tool to assess the capabilities of LLMs accurately. Interpretive argumentation is probably more challenging than many tasks that LLMs are tested on, making those assessments more meaningful for complex linguistic evaluation and making sure that newer LLMs are, in fact, improving in performance on some challenging tasks. Evaluating the performance of LLMs on the task of interpretive argumentation is not a trivial task. Automating such assessments is very difficult; hence, including human annotators is essential for this task. This experimental protocol provides clear guidance on how to carry out such evaluations.

Crowdsourced data in the post-ChatGPT age. An important point was raised by one of this manuscript's reviewers, on which we should now say a few words. At the time of our primary data collection for this proposed work, OpenAI's ChatGPT was barely a few months old and did not have the widespread adoption it has today. For this reason, we felt that our method of filtering out AI-generated participant responses – imperfect as it may be – was sufficient for our purposes. We also switched platforms mid-study from Amazon Mechanical Turk to Prolific, as we felt the latter had better internal quality controls and put more effort into ensuring that participants were not using AI tools. We present our methodology in sufficient detail that it can be replicated by those who seek to see whether the same results hold when carried out under stricter conditions (e.g., entirely in-person), and we encourage others to do so.

However, we feel it is important to observe that the tools and prompting methods available for carrying out online crowdsourced data collections are evolving at such a rapid pace that the entire future of such work is facing a crisis. Particularly in text-heavy work such as the present argumentation study, it is difficult to see how sufficient quality control can be upheld in future studies if administered remotely. We make this observation here in the hope that it inspires future work into solving this serious problem.

Future work. Another important future goal is to develop the introspective prompting approach further. Chain-of-thought prompting and similar approaches illustrate the value of allowing an LLM to reason in a step-by-step fashion, breaking the problem into smaller problems, and reflecting on its outputs. Using existing literature that may be used to guide or teach human arguers to produce good arguments could be used to encourage an LLM to use the same techniques to produce high-quality arguments. The introspective approach did not yield better results compared to the simple prompting approach in our experiments, which is counter-intuitive to what we suspected would be the case. However, considering the variability of performance that different prompting styles cause, we believe it is worthwhile to determine whether different prompts may yield better results. Possible reasons for the introspective prompting approach to underperform include:

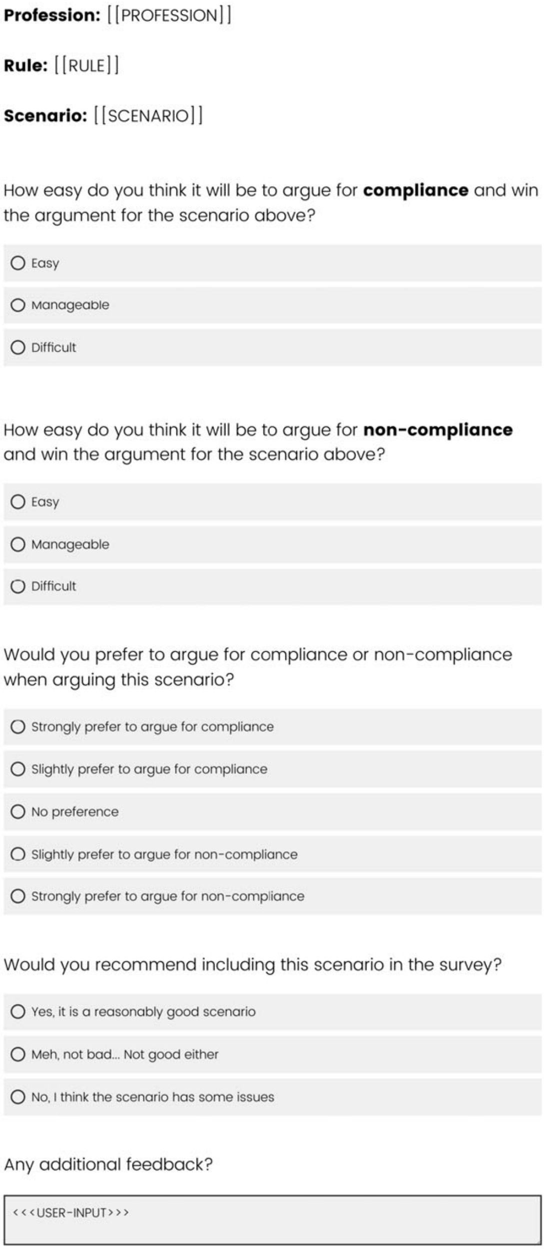
- The introspective approach requires more powerful LLMs to perform well. It is possible that LLMs need to cross some capability threshold in order for this approach to become viable.
- The introspective approach requires additional iterative improvements and redesign along with evaluations in order to find out what works and what does not.
- Since the introspective design relies on a pipeline consisting of a long sequence of steps, early reasoning defects in the output of an LLM, or the accumulation of errors over the consecutive steps, may negate any positive effects that the introspective design introduces. This intersects with the first point, that better LLMs are necessary, such that fewer reasoning defects are introduced during the execution of the pipeline.

Future work includes qualitative analysis to supplement and further explain the quantitative results. Some goals for the qualitative analysis include:

- Identify some of the factors that affect the quality of the generated arguments.
- Analysis of argument quality measures (other than convincingness) for human- and machine-generated arguments, such as Cogency, Relevance, Clarity, and Defeasibility.
- Identify major factors leading to reduced performance when using introspective prompting.

Appendix

A.1. Survey templates



Profession: [[PROFESSION]]

Rule: [[RULE]]

Scenario: [[SCENARIO]]

How easy do you think it will be to argue for **compliance** and win the argument for the scenario above?

○ Easy

○ Manageable

○ Difficult

How easy do you think it will be to argue for **non-compliance** and win the argument for the scenario above?

○ Easy

○ Manageable

○ Difficult

Would you prefer to argue for compliance or non-compliance when arguing this scenario?

○ Strongly prefer to argue for compliance

○ Slightly prefer to argue for compliance

○ No preference

○ Slightly prefer to argue for non-compliance

○ Strongly prefer to argue for non-compliance

Would you recommend including this scenario in the survey?

○ Yes, it is a reasonably good scenario

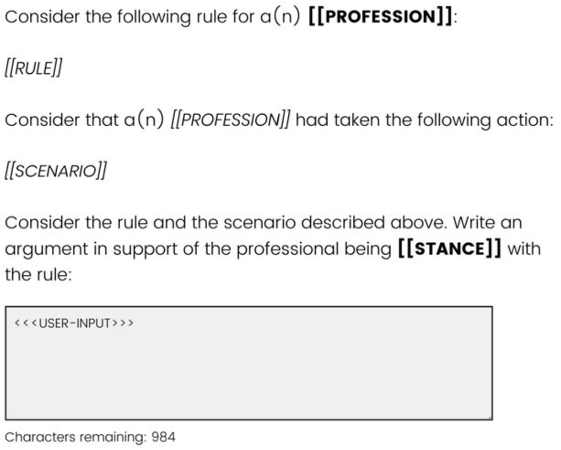
○ Meh, not bad... Not good either

○ No, I think the scenario has some issues

Any additional feedback?

<<<USER-INPUT>>>

Fig. 9. Template of survey questions at stage 1.



Consider the following rule for a(n) **[[PROFESSION]]**:

[[RULE]]

Consider that a(n) *[[PROFESSION]]* had taken the following action:

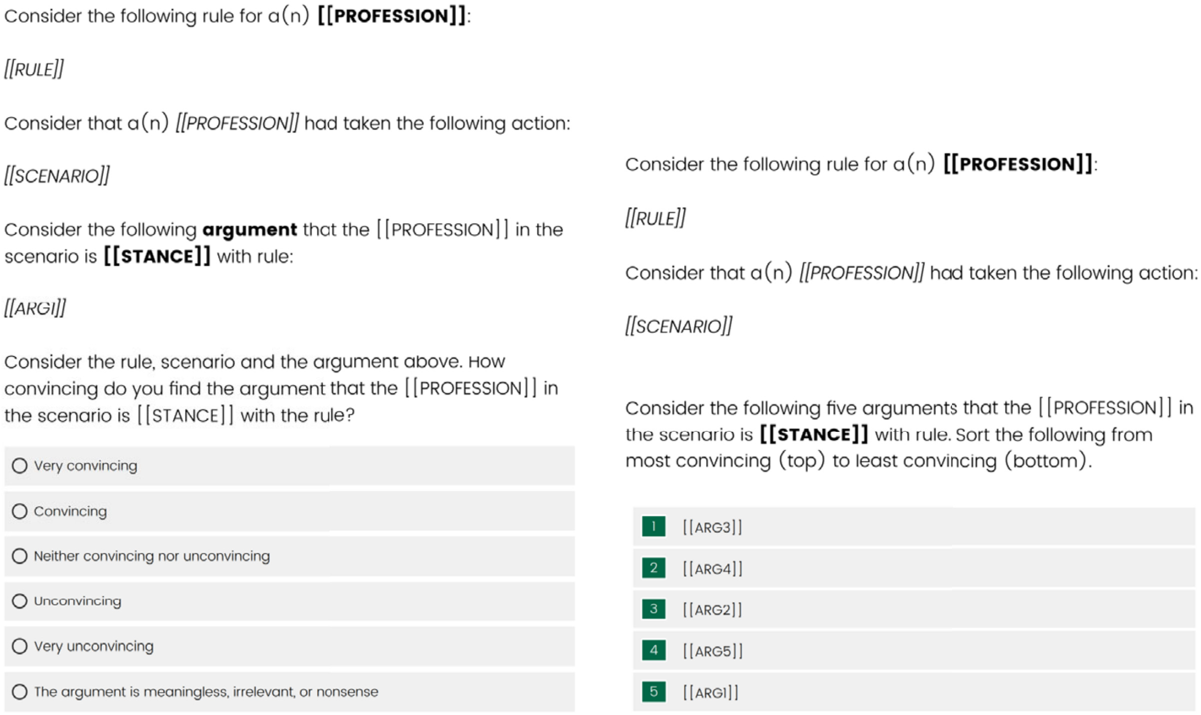
[[SCENARIO]]

Consider the rule and the scenario described above. Write an argument in support of the professional being **[[STANCE]]** with the rule:

<<<USER-INPUT>>>

Characters remaining: 984

Fig. 10. Template of survey questions at stage 2 for collecting human arguments.



Consider the following rule for a(n) **[[PROFESSION]]**:

[[RULE]]

Consider that a(n) *[[PROFESSION]]* had taken the following action:

[[SCENARIO]]

Consider the following **argument** that the [[PROFESSION]] in the scenario is **[[STANCE]]** with rule:

[[ARG1]]

Consider the rule, scenario and the argument above. How convincing do you find the argument that the [[PROFESSION]] in the scenario is [[STANCE]] with the rule?

○ Very convincing

○ Convincing

○ Neither convincing nor unconvincing

○ Unconvincing

○ Very unconvincing

○ The argument is meaningless, irrelevant, or nonsense

(a) Format 1

Consider the following rule for a(n) **[[PROFESSION]]**:

[[RULE]]

Consider that a(n) *[[PROFESSION]]* had taken the following action:

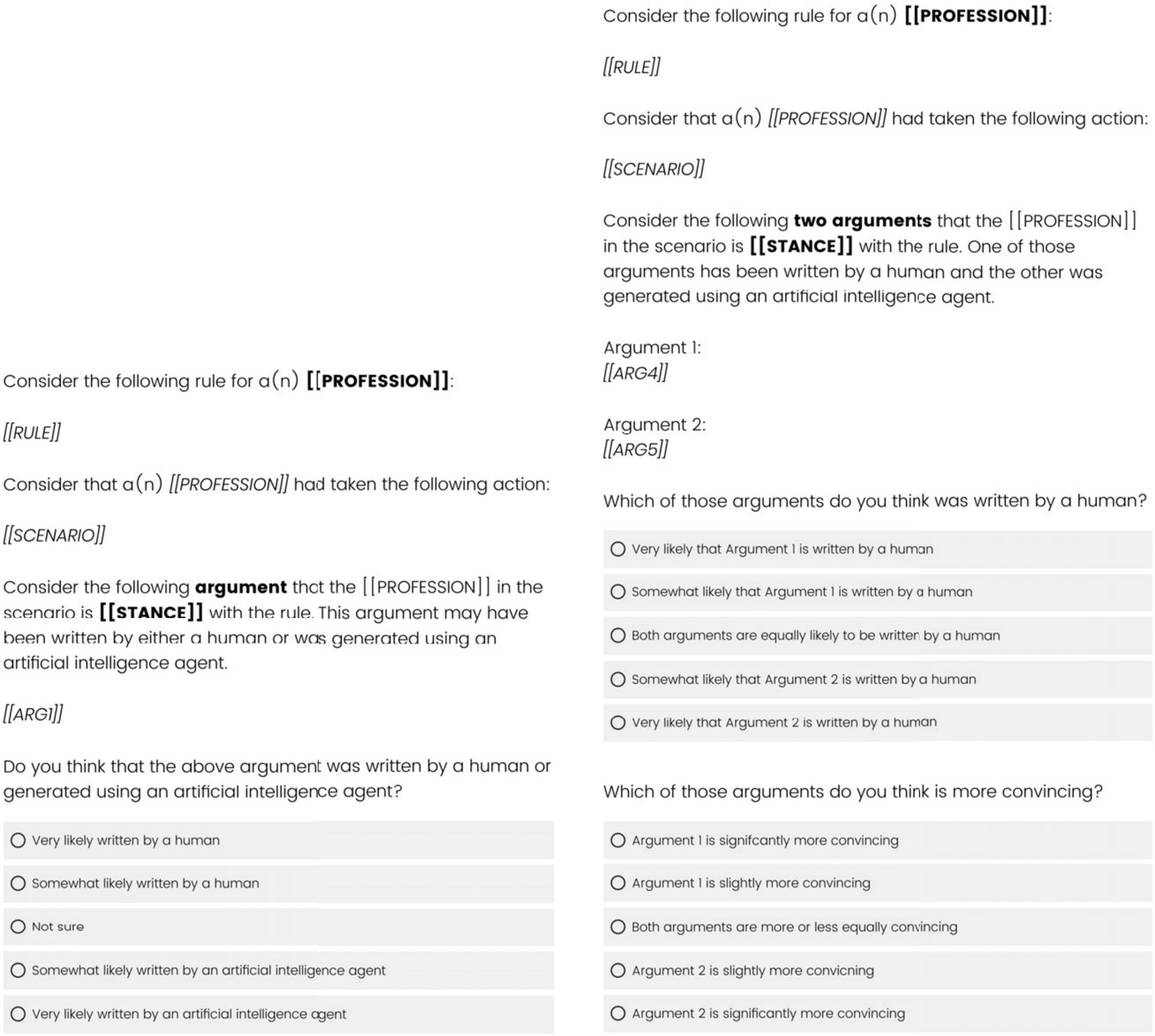
[[SCENARIO]]

Consider the following five arguments that the [[PROFESSION]] in the scenario is **[[STANCE]]** with rule. Sort the following from most convincing (top) to least convincing (bottom).

1	[[ARG3]]
2	[[ARG4]]
3	[[ARG2]]
4	[[ARG5]]
5	[[ARG1]]

(b) Format 2

Fig. 11. Templates of survey questions at stages 2 and 3.



Consider the following rule for a(n) **[[PROFESSION]]**:

[[RULE]]

Consider that a(n) *[[PROFESSION]]* had taken the following action:

[[SCENARIO]]

Consider the following **argument** thct the [[PROFESSION]] in the scenario is **[[STANCE]]** with the rule. This argument may have been written by either a human or was generated using an artificial intelligence agent.

[[ARG1]]

Do you think that the above argument was written by a human or generated using an artificial intelligence agent?

○ Very likely written by a human

○ Somewhat likely written by a human

○ Not sure

○ Somewhat likely written by an artificial intelligence agent

○ Very likely written by an artificial intelligence agent

(a) Format 1

Consider the following rule for a(n) **[[PROFESSION]]**:

[[RULE]]

Consider that a(n) *[[PROFESSION]]* had taken the following action:

[[SCENARIO]]

Consider the following **two arguments** that the [[PROFESSION]] in the scenario is **[[STANCE]]** with the rule. One of those arguments has been written by a human and the other was generated using an artificial intelligence agent.

Argument 1:
[[ARG4]]

Argument 2:
[[ARG5]]

Which of those arguments do you think was written by a human?

○ Very likely that Argument 1 is written by a human

○ Somewhat likely that Argument 1 is written by a human

○ Both arguments are equally likely to be written by a human

○ Somewhat likely that Argument 2 is written by a human

○ Very likely that Argument 2 is written by a human

Which of those arguments do you think is more convincing?

○ Argument 1 is signifcantly more convincing

○ Argument 1 is slightly more convincing

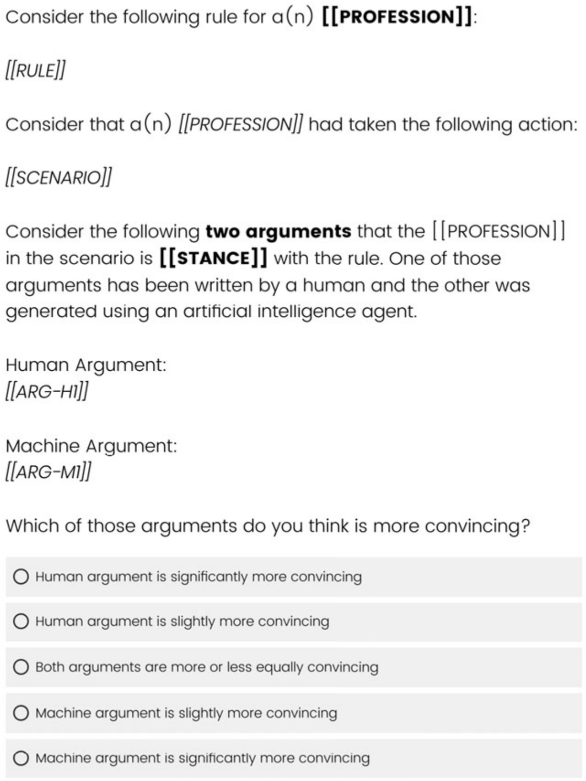
○ Both arguments are more or less equally convincing

○ Argument 2 is slightly more convicning

○ Argument 2 is significantly more convincing

(b) Format 2

Fig. 12. Template of survey questions at stage 4.



Consider the following rule for a(n) **[[PROFESSION]]**:

[[RULE]]

Consider that a(n) *[[PROFESSION]]* had taken the following action:

[[SCENARIO]]

Consider the following **two arguments** that the [[PROFESSION]] in the scenario is **[[STANCE]]** with the rule. One of those arguments has been written by a human and the other was generated using an artificial intelligence agent.

Human Argument:
[[ARG-H1]]

Machine Argument:
[[ARG-M1]]

Which of those arguments do you think is more convincing?

○ Human argument is significantly more convincing

○ Human argument is slightly more convincing

○ Both arguments are more or less equally convincing

○ Machine argument is slightly more convincing

○ Machine argument is significantly more convincing

Fig. 13. Template of survey questions at stage 5.

A.2. *Legend for prompt design listings*

Variables are enclosed in two brackets. There are two types of variable access: "Recall a variable" denoted as `[[R:variable-name]]`, and "Generate a variable" denoted as `[[G:variable-name]]`. A variable is generated and set using GPT-3 text completion. Variables may be recalled from an externally set value, for example `[[R:PROFESSION]]` refers to the profession for the item being generated, while `[[R:PHRASE]]` refers to the value generated using GPT-3 using a previous use of `[[G:PHRASE]]`. Multi-step prompts are evaluated sequentially. The sequence ===== is used to separate the different prompts, and each step starts with the text `TEMPLATE X:` indicates the step number where `X` stands for the step number. The arrows ↪ indicate that a line has wrapped at the end of page.

A.3. Prompt template for GPT-3 scenario generation

```
TEMPLATE 1:

Read the following rule with respect to the given profession, and
 ↪  answer the following questions:

Profession: [[R:PROFESSION]]
Rule: [[R:RULE]]

Q: Which phrase in the rule is ambiguous and likely to be open to
 ↪  different interpretations? List five different ways in which
 ↪  that phrase could be interpreted.
A: "[[G:PHRASE]]"

1. [[G:INTERPRETATION:1]]
2. [[G:INTERPRETATION:2]]
3. [[G:INTERPRETATION:3]]
4. [[G:INTERPRETATION:4]]
5. [[G:INTERPRETATION:5]]

=====

TEMPLATE 2:

Read the following rule with respect to the given profession, and
 ↪  answer the following questions:

Profession: [[R:PROFESSION]]
Rule: [[R:RULE]]

Q: Which phrase in the rule is ambiguous and likely to be open to
 ↪  different interpretations? List five different ways in which
 ↪  that phrase could be interpreted.
A: ``[[R:PHRASE]]"

1. [[R:INTERPRETATION:1]]
2. [[R:INTERPRETATION:2]]
3. [[R:INTERPRETATION:3]]
4. [[R:INTERPRETATION:4]]
5. [[R:INTERPRETATION:5]]
```

```
Q: Consider the different interpretations of the ambiguous phrase
 ↪  mentioned above. For each of the provided interpretations,
 ↪  provide an example of a clear and concrete real-world scenario
 ↪  where that interpretation would lead to contradicting
 ↪  conclusions about whether the rule is followed or not in
 ↪  relation to one of the other interpretations. Explain how each
 ↪  scenario is likely to cause controversy around the
 ↪  professional's conduct.
A:

1. [[G:CONFLICT:1]]

2. [[G:CONFLICT:2]]

3. [[G:CONFLICT:3]]

4. [[G:CONFLICT:4]]

5. [[G:CONFLICT:5]]

Q: Consider the different scenarios discussed above. Which of these
 ↪  scenarios is most likely to happen in a real-world scenario and
 ↪  be most likely to cause controversy? Why?
A: The most likely scenario to happen in the real world is scenario
 ↪  [[G:INDEX]], where [[G:CONTROVERSY]]

Q: Consider that the phrase "[[G:PHRASE-P]]" can be interpreted to
 ↪  mean "[[G:INTERPRETATION-P:1]]" or to mean
 ↪  "[[G:INTERPRETATION-P:2]]". Write a clear and concrete
 ↪  real-world scenario with realistic details where
 ↪  [[G:CONTROVERSY-P]]

Ambiguous Scenario:

[[G:SCENARIO]]

=====

TEMPLATE 3:

Read the following rule and scenario, and answer the following
 ↪  questions:

Profession: [[R:PROFESSION]]
Rule: [[R:RULE]]
```

```
Ambiguous Scenario: [[R:SCENARIO]]

Q: Which phrase in the rule is ambiguous and likely to be open to
 ↪  different interpretations?
A: "[[R:PHRASE]]"

Q: Why might the professional in the provided scenario be considered
 ↪  to have violated the rule?
A: [[G:WHY-VIOLATED]]

Q: Why might the professional in the provided scenario be considered
 ↪  to be following the rule?
A: [[G:WHY-FOLLOWED]]

Consider all the reasons for why the professional may be considered
 ↪  to be following or having violated the rule. Write an argument
 ↪  clearly explaining and summarizing why the professional should
 ↪  be considered as compliant with the rule.

Argument: [[G:ARG-FOR]]

Consider all the reasons for why the professional may be considered
 ↪  to be following or having violated the rule. Write an argument
 ↪  clearly explaining and summarizing why the professional should
 ↪  be considered as non-compliant with the rule.

Counter Argument: [[G:ARG-AGAINST]]

Q: Can we improve the previous argument to take account of the
 ↪  counter argument? (yes/no). If the answer is yes, how?
A: Yes, [[G:IMPROVE-FOR]]

Consider the previous argument. Restate an improved version of the
 ↪  previous argument that explains why the professional economic
 ↪  developer in the provided scenario is in compliance with the
 ↪  given rule. Consider the counter argument and the improvements
 ↪  suggested above. Include all considerations that are relevant to
 ↪  support the desired conclusion while keeping the argument short
 ↪  and concise.

Argument: [[G:ARG-FOR]]

=====

TEMPLATE 4:
```

```
Read the following rule and scenario, and answer the following
 ↪  questions:

Profession: [[R:PROFESSION]]
Rule: [[R:RULE]]
Ambiguous Scenario: [[R:SCENARIO]]

Q: Which phrase in the rule is ambiguous and likely to be open to
 ↪  different interpretations?
A: "[[R:PHRASE]]"

Argument for compliance: [[R:ARG-FOR]]

Q: What are some good reasons [[G:Q-REASONS-FOR]]?
A: [[G:REASONS-FOR]]

Q: Why [[G:Q-WHY-FOR]]?
A: [[G:WHY-FOR]]

Consider the previous argument. Restate an improved version of the
 ↪  previous argument that explains why the professional economic
 ↪  developer in the provided scenario is in compliance with the
 ↪  given rule. Consider the counter argument and the improvements
 ↪  suggested above. Include all considerations that are relevant to
 ↪  support the desired conclusion while keeping the argument short
 ↪  and concise.

Argument: [[G:ARG-FOR]]

=====

TEMPLATE 5:

Read the following rule and scenario, and answer the following
 ↪  questions:

Profession: [[R:PROFESSION]]
Rule: [[R:RULE]]
Ambiguous Scenario: [[R:SCENARIO]]

Q: Which phrase in the rule is ambiguous and likely to be open to
 ↪  different interpretations?
A: "[[R:PHRASE]]"
```

```
Argument for compliance: [[R:ARG-FOR]]

Q: Can we improve the ambiguous scenario to take account of the
 ↪  arguments for and against compliance to make the scenario more
 ↪  ambiguous about whether the professional is compliant or
 ↪  non-compliant with the rule? (yes/no). If the answer is yes,
 ↪  how?
A: Yes, [[G:IMPROVE-FOR]]

Consider the ambiguous scenario. Also consider the arguments for and
 ↪  against compliance, as well as the improvements to the scenario
 ↪  suggested in the previous question. Restate an improved version
 ↪  of ambiguous scenario such that deciding between compliance and
 ↪  non-compliance is more difficult.

Improved Ambiguous Scenario: [[G:SCENARIO]]

=====

TEMPLATE 6:

Read the following rule and scenario, and answer the following
 ↪  questions:

Profession: [[R:PROFESSION]]
Rule: [[R:RULE]]
Ambiguous Scenario: [[R:SCENARIO]]

Q: Which phrase in the rule is ambiguous and likely to be open to
 ↪  different interpretations?
A: "[[R:PHRASE]]"

Q: Why might the professional in the provided scenario be considered
 ↪  to have violated the rule?
A: [[G:WHY-VIOLATED]]

Q: Why might the professional in the provided scenario be considered
 ↪  to be following the rule?
A: [[G:WHY-FOLLOWED]]

Consider all the reasons for why the professional may be considered
 ↪  to be following or having violated the rule. Write an argument
 ↪  clearly explaining and summarizing why the professional should
 ↪  be considered as non-compliant with the rule.
```

```
Argument: [[G:ARG-AGAINST]]

Consider all the reasons for why the professional may be considered
 ↪  to be following or having violated the rule. Write an argument
 ↪  clearly explaining and summarizing why the professional should
 ↪  be considered as compliant with the rule.

Counter Argument: [[G:ARG-FOR]]

Q: Can we improve the previous argument to take account of the
 ↪  counter argument? (yes/no). If the answer is yes, how?
A: Yes, [[G:IMPROVE-AGAINST]]

Consider the previous argument. Restate an improved version of the
 ↪  previous argument that explains why the professional economic
 ↪  developer in the provided scenario is non-compliant with the
 ↪  given rule. Consider the counter argument and the improvements
 ↪  suggested above. Include all considerations that are relevant to
 ↪  support the desired conclusion while keeping the argument short
 ↪  and concise.

Argument: [[G:ARG-AGAINST]]

=====

TEMPLATE 7:

Read the following rule and scenario, and answer the following
 ↪  questions:

Profession: [[R:PROFESSION]]
Rule: [[R:RULE]]
Ambiguous Scenario: [[R:SCENARIO]]

Q: Which phrase in the rule is ambiguous and likely to be open to
 ↪  different interpretations?
A: "[[R:PHRASE]]"

Argument for non-compliance: [[R:ARG-AGAINST]]

Q: What are some good reasons [[G:Q-REASONS-AGAINST]]?
A: [[G:REASONS-AGAINST]]

Q: Why [[G:Q-WHY-AGAINST]]?
A: [[G:WHY-AGAINST]]
```

```
Consider the previous argument. Restate an improved version of the
 ↪  previous argument that explains why the professional economic
 ↪  developer in the provided scenario is non-compliant with the
 ↪  given rule. Consider the counter argument and the improvements
 ↪  suggested above. Include all considerations that are relevant to
 ↪  support the desired conclusion while keeping the argument short
 ↪  and concise.

Argument: [[G:ARG-AGAINST]]

=====

TEMPLATE 8:

Read the following rule and scenario, and answer the following
 ↪  questions:

Profession: [[R:PROFESSION]]
Rule: [[R:RULE]]
Ambiguous Scenario: [[R:SCENARIO]]

Q: Which phrase in the rule is ambiguous and likely to be open to
 ↪  different interpretations?
A: "[[R:PHRASE]]"

Argument for non-compliance: [[R:ARG-AGAINST]]

Q: Can we improve the ambiguous scenario to take account of the
 ↪  arguments for and against compliance to make the scenario more
 ↪  ambiguous about whether the professional is compliant or
 ↪  non-compliant with the rule? (yes/no). If the answer is yes,
 ↪  how?
A: Yes, [[G:IMPROVE-AGAINST]]

Consider the ambiguous scenario. Also consider the arguments for and
 ↪  against compliance, as well as the improvements to the scenario
 ↪  suggested in the previous question. Restate an improved version
 ↪  of ambiguous scenario such that deciding between compliance and
 ↪  non-compliance is more difficult.

Improved Ambiguous Scenario: [[G:SCENARIO]]

=====
TEMPLATE 9:


Read the following rule and scenario, and answer the following
 ↪  questions:

Profession: [[R:PROFESSION]]
Rule: [[R:RULE]]

Q: Which phrase in the rule is ambiguous and likely to be open to
 ↪  different interpretations?
A: "[[R:PHRASE]]"

Ambiguous Scenario: [[R:SCENARIO]]

Consider the ambiguous scenario. Restate a more concise version of
 ↪  ambiguous scenario while maintaining all the information present
 ↪  in it.

Shortened Ambiguous Scenario: [[G:SCENARIO]]
```

A.4. Prompt template for simple argument generation

```
Consider the following rule for a(n) [[R:PROFESSION]]:

[[R:RULE]]

Consider that a(n) [[R:PROFESSION]] had taken the following action:

[[R:SCENARIO]]

Consider the rule and the scenario described above. Write an
 ↪  argument that the professional is [[R:STANCE]] with the rule:

Argument: [[G:ARGUMENT]]
```

A.5. Prompt template for multi-step introspective argument generation

```
TEMPLATE 1:

Read the following rule and scenario with respect to the given
 ↪  profession, and answer the following questions:

Profession: [[R:PROFESSION]]
Rule: [[R:RULE]]
Scenario: [[R:SCENARIO]]

Q: Which phrase in the rule is the source of confusion in this
 ↪  scenario?
A: "[[G:PHRASE]]"

=====

Read the following rule and scenario with respect to the given
 ↪  profession, and answer the following questions:

Profession: [[R:PROFESSION]]
Rule: [[R:RULE]]
Scenario: [[R:SCENARIO]]

Q: Which phrase in the rule is the source of confusion in this
 ↪  scenario?
A: "[[R:PHRASE]]"

Q: How can this phrase be interpretted in order for the professional
 ↪  in the scenario above to be considered compliant with the rule?
A: The phrase can be interpreted to mean that [[G:INTR-FOR]]

Q: How can this phrase be interpretted in order for the professional
 ↪  in the scenario above to be considered non-compliant with the
 ↪  rule?
A: The phrase can be interpreted to mean that [[G:INTR-AGAINST]]

=====

TEMPLATE 3:

Read the following rule and scenario with respect to the given
 ↪  profession, and answer the following questions:
```

```
Profession: [[R:PROFESSION]]
Rule: [[R:RULE]]
Scenario: [[R:SCENARIO]]

Q: Which phrase in the rule is the source of confusion in this
 ↪  scenario?
A: "[[R:PHRASE]]"

Q: How can this phrase be interpretted in order for the professional
 ↪  to be considered compliant with the rule?
A: The phrase can be interpreted to mean that [[R:INTR-FOR]]

Q: How can this phrase be interpretted in order for the professional
 ↪  to be considered non-compliant with the rule?
A: The phrase can be interpreted to mean that [[R:INTR-AGAINST]]

Q: Write an argument explaining why the rule should be interpreted
 ↪  as being compatible with the professional's actions.
Argument for compliance: [[G:I-ARG-FOR]]

Q: Write an argument explaining why the rule should be interpreted
 ↪  as being incompatible with the professional's actions.
Argument for non-compliance: [[G:I-ARG-AGAINST]]

=====

TEMPLATE 4:

Read the following rule and scenario with respect to the given
 ↪  profession, and answer the following questions:

Profession: [[R:PROFESSION]]
Rule: [[R:RULE]]
Scenario: [[R:SCENARIO]]

Q: Which phrase in the rule is the source of confusion in this
 ↪  scenario?
A: "[[R:PHRASE]]"

Q: How can this phrase be interpretted in order for the professional
 ↪  to be considered compliant with the rule?
A: The phrase can be interpreted to mean that [[R:INTR-FOR]]

Q: How can this phrase be interpretted in order for the professional
 ↪  to be considered non-compliant with the rule?
```

```
A: The phrase can be interpreted to mean that [[R:INTR-AGAINST]]

Q: Write an argument explaining why the rule should be interpreted
 ↪  as being compatible with the professional's actions.
Argument for compliance: [[R:I-ARG-FOR]]

Q: Write an argument explaining why the rule should be interpreted
 ↪  as being incompatible with the professional's actions.
Argument for non-compliance: [[R:I-ARG-AGAINST]]

Consider all the reasons for why the professional may be considered
 ↪  to be following or having violated the rule including the
 ↪  arguments above. Write an argument clearly explaining and
 ↪  summarizing why the professional should be considered as
 ↪  compliant with the rule.

Argument for compliance: [[G:ARG-FOR-1]]

Consider all the reasons for why the professional may be considered
 ↪  to be following or having violated the rule including the
 ↪  arguments above. Write a counter argument clearly explaining and
 ↪  summarizing why the professional should be considered as
 ↪  non-compliant with the rule.

Counter Argument: [[G:C-ARG-AGAINST]]

Q: Can we improve the previous argument in favor of compliance by
 ↪  taking into account the counter argument? (yes/no). If the
 ↪  answer is yes, how?
A: Yes, [[G:IMPROVE-FOR]]

Consider the previous argument. Restate an improved version of the
 ↪  previous argument that explains why the professional in the
 ↪  provided scenario is compliant with the given rule. Consider the
 ↪  counter argument and the improvements suggested above. Include
 ↪  all considerations that are relevant to support the desired
 ↪  conclusion while keeping the argument short and concise.

Argument for compliance: [[G:ARG-FOR-2]]

=====

TEMPLATE 5:
```

```
Read the following rule and scenario with respect to the given
 ↪  profession, and answer the following questions:

Profession: [[R:PROFESSION]]
Rule: [[R:RULE]]
Scenario: [[R:SCENARIO]]

Q: Which phrase in the rule is the source of confusion in this
 ↪  scenario?
A: "[[R:PHRASE]]"

Q: How can this phrase be interpretted in order for the professional
 ↪  to be considered compliant with the rule?
A: The phrase can be interpreted to mean that [[R:INTR-FOR]]

Q: How can this phrase be interpretted in order for the professional
 ↪  to be considered non-compliant with the rule?
A: The phrase can be interpreted to mean that [[R:INTR-AGAINST]]

Q: Write an argument explaining why the rule should be interpreted
 ↪  as being compatible with the professional's actions.
Argument for compliance: [[R:I-ARG-FOR]]

Q: Write an argument explaining why the rule should be interpreted
 ↪  as being incompatible with the professional's actions.
Argument for non-compliance: [[R:I-ARG-AGAINST]]

Consider all the reasons for why the professional may be considered
 ↪  to be following or having violated the rule including the
 ↪  arguments above. Write an argument clearly explaining and
 ↪  summarizing why the professional should be considered as
 ↪  non-compliant with the rule.

Argument for non-compliance: [[G:ARG-AGAINST-1]]

Consider all the reasons for why the professional may be considered
 ↪  to be following or having violated the rule including the
 ↪  arguments above. Write a counter argument clearly explaining and
 ↪  summarizing why the professional should be considered as
 ↪  compliant with the rule.

Counter Argument: [[G:C-ARG-FOR]]
```

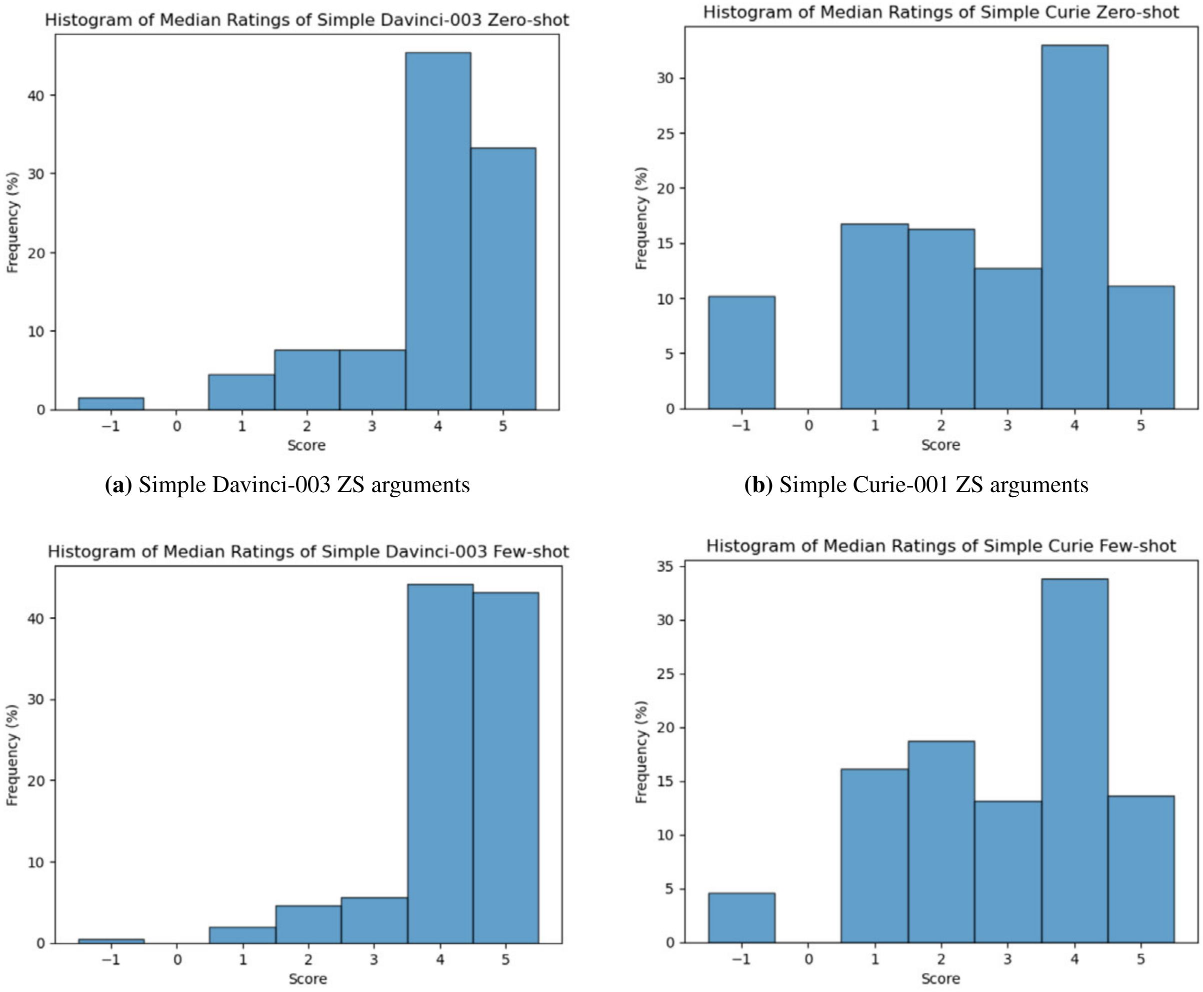
```
Q: Can we improve the previous argument in favor of non-compliance
 ↪  by taking into account of the counter argument? (yes/no). If the
 ↪  answer is yes, how?
A: Yes, [[G:IMPROVE-AGAINST]]

Consider the previous argument. Restate an improved version of the
 ↪  previous argument that explains why the professional in the
 ↪  provided scenario is non-compliant with the given rule. Consider
 ↪  the counter argument and the improvements suggested above.
 ↪  Include all considerations that are relevant to support the
 ↪  desired conclusion while keeping the argument short and concise.

Argument for non-compliance: [[G:ARG-AGAINST-2]]

=====

TEMPLATE 6:

Read the following rule and scenario with respect to the given
 ↪  profession, and answer the following questions:

Profession: [[R:PROFESSION]]
Rule: [[R:RULE]]
Scenario: [[R:SCENARIO]]

Q: Which phrase in the rule is the source of confusion in this
 ↪  scenario?
A: "[[R:PHRASE]]"

Consider the following two possible interpretations of the ambiguous
 ↪  phrase above:
Intepretation supporting compliance: [[R:I-ARG-FOR]]
Intepretation supporting non-compliance: [[R:I-ARG-AGAINST]]

Also, consider the following two arguments:
Argument for compliance: [[R:ARG-FOR-2]]
Argument for non-compliance: [[R:ARG-AGAINST-2]]

Q: What are some of the positive future consequences that might
 ↪  entail if the interpretation supporting compliance is accepted?
A: [[G:CONQ-FOR-P]]

Q: What are some of the negative future consequences that might
 ↪  entail if the interpretation supporting compliance is accepted?
A: [[G:CONQ-FOR-N]]
```

```
Q: What are some of the positive future consequences that might
 ↪  entail if the interpretation supporting non-compliance is
 ↪  accepted?
A: [[G:CONQ-AGAINST-P]]

Q: What are some of the negative future consequences that might
 ↪  entail if the interpretation supporting non-compliance is
 ↪  accepted?
A: [[G:CONQ-AGAINST-N]]

=====

TEMPLATE 7:

Read the following rule and scenario with respect to the given
 ↪  profession, and answer the following questions:

Profession: [[R:PROFESSION]]
Rule: [[R:RULE]]
Scenario: [[R:SCENARIO]]

Q: Which phrase in the rule is the source of confusion in this
 ↪  scenario?
A: "[[R:PHRASE]]"

Consider the following two arguments:
Argument for compliance: [[R:ARG-FOR-2]]
Argument for non-compliance: [[R:ARG-AGAINST-2]]

Also, consider the following potential consequences for each
 ↪  argument:
Positive consequences for accepting compliance: [[R:CONQ-FOR-P]]
Negative consequences for accepting compliance: [[R:CONQ-FOR-N]]
Positive consequences for accepting non-compliance:
 ↪  [[R:CONQ-AGAINST-P]]
Negative consequences for accepting non-compliance:
 ↪  [[R:CONQ-AGAINST-N]]
```

```
Consider the previous two arguments and their potential
 ↪  consequences. Restate an improved version of the argument that
 ↪  explains why the professional in the provided scenario is
 ↪  compliant with the given rule. Consider the counter argument and
 ↪  the improvements suggested above. Include all considerations
 ↪  that are relevant to support the desired conclusion while
 ↪  keeping the argument short and concise.

Argument for compliance: [[G:ARG-FOR-3]]

=====

TEMPLATE 8:

Read the following rule and scenario with respect to the given
 ↪  profession, and answer the following questions:

Profession: [[R:PROFESSION]]
Rule: [[R:RULE]]
Scenario: [[R:SCENARIO]]

Q: Which phrase in the rule is the source of confusion in this
 ↪  scenario?
A: "[[R:PHRASE]]"

Argument for compliance: [[R:ARG-FOR-3]]

Consider the argument for compliance. Restate a more concise version
 ↪  of the argument while maintaining the salient points of the
 ↪  argument.

Shortened Argument: [[G:SHORT-ARG-FOR]]

=====

TEMPLATE 9:

Profession: [[R:PROFESSION]]
Rule: [[R:RULE]]
Scenario: [[R:SCENARIO]]

Q: Which phrase in the rule is the source of confusion in this
 ↪  scenario?
A: "[[R:PHRASE]]"
```

```
Consider the following two arguments:
Argument for compliance: [[R:ARG-FOR-3]]
Argument for non-compliance: [[R:ARG-AGAINST-2]]

Also, consider the following potential consequences for each
 ↪  argument:
Positive consequences for accepting compliance: [[R:CONQ-FOR-P]]
Negative consequences for accepting compliance: [[R:CONQ-FOR-N]]
Positive consequences for accepting non-compliance:
 ↪  [[R:CONQ-AGAINST-P]]
Negative consequences for accepting non-compliance:
 ↪  [[R:CONQ-AGAINST-N]]

Consider the previous two arguments and their potential
 ↪  consequences. Restate an improved version of the argument that
 ↪  explains why the professional in the provided scenario is
 ↪  non-compliant with the given rule. Consider the counter argument
 ↪  and the improvements suggested above. Include all considerations
 ↪  that are relevant to support the desired conclusion while
 ↪  keeping the argument short and concise.

Argument for non-compliance: [[G:ARG-AGAINST-3]]

=====

TEMPLATE 10:

Read the following rule and scenario with respect to the given
 ↪  profession, and answer the following questions:

Profession: [[R:PROFESSION]]
Rule: [[R:RULE]]
Scenario: [[R:SCENARIO]]

Q: Which phrase in the rule is the source of confusion in this
 ↪  scenario?
A: "[[R:PHRASE]]"

Argument for non-compliance: [[R:ARG-AGAINST-3]]

Consider the argument for non-compliance. Restate a more concise
 ↪  version of the argument while maintaining the salient points of
 ↪  the argument.

Shortened Argument: [[G:SHORT-ARG-AGAINST]]
```

A.6. Histograms of all median ratings of all arguments for all LLMs as rated in stage 3



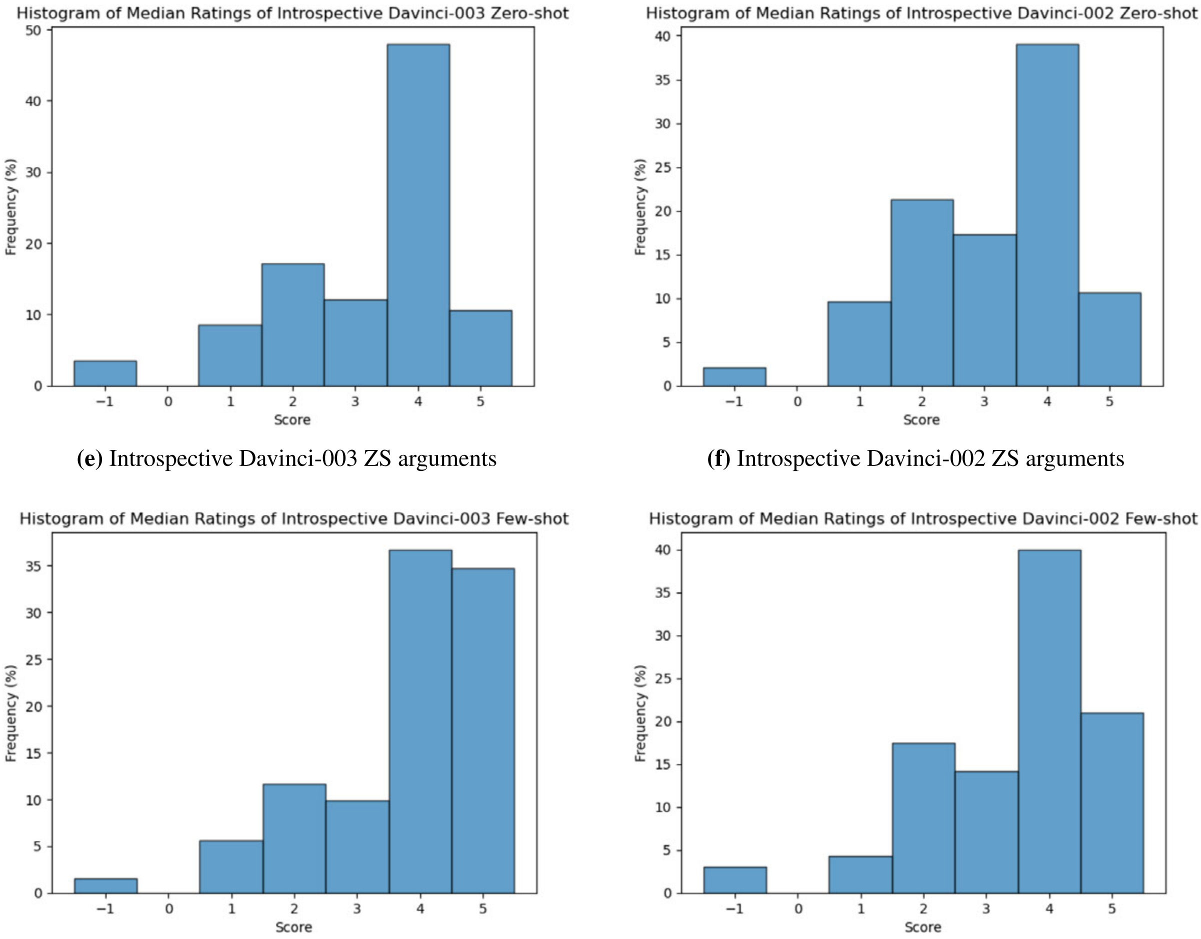
(a) Simple Davinci-003 ZS arguments

(b) Simple Curie-001 ZS arguments

(c) Simple Davinci-003 FS arguments

(d) Simple Curie-001 FS arguments

Fig. 14. Histograms of median ratings (part 1).



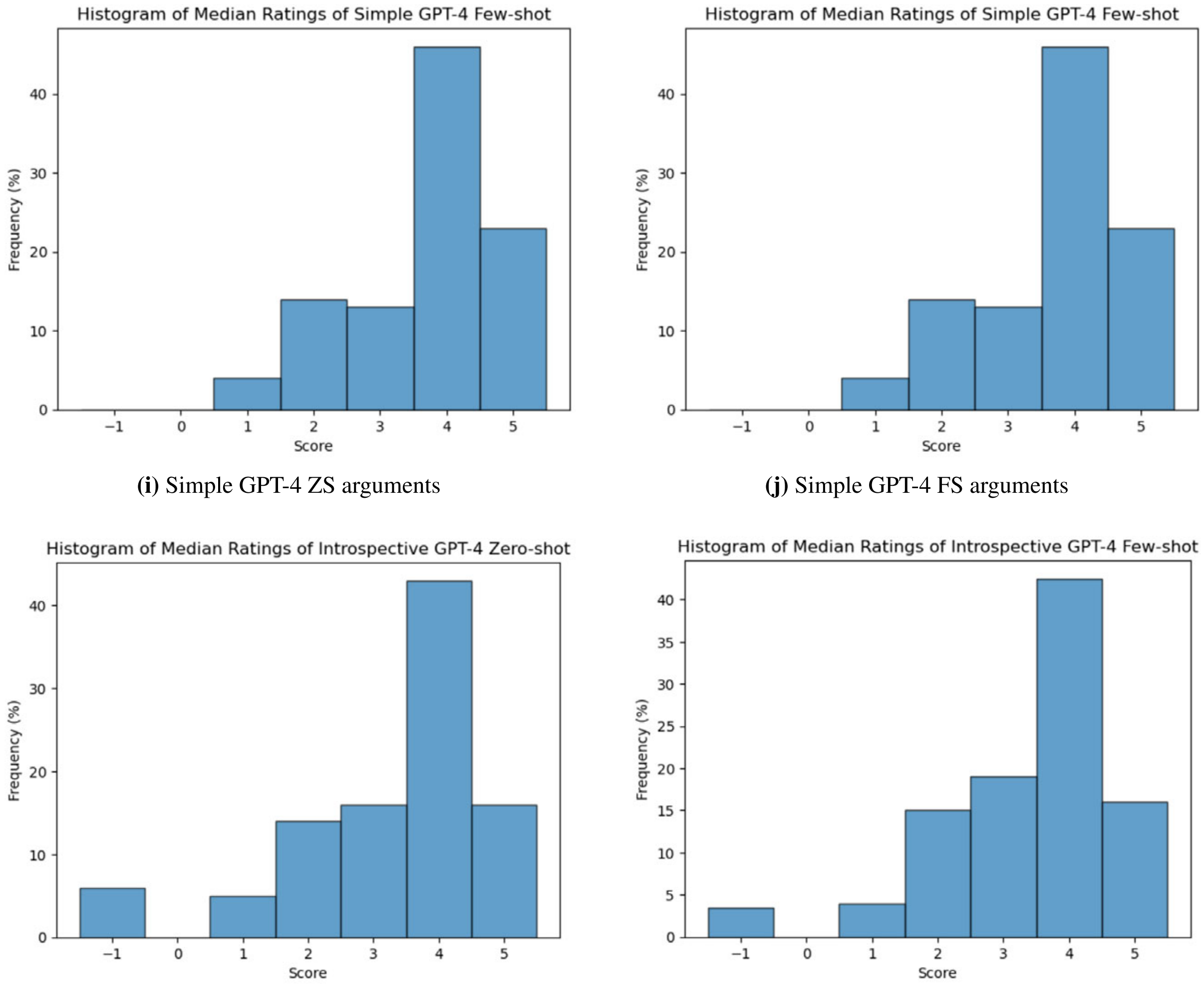
(e) Introspective Davinci-003 ZS arguments

(f) Introspective Davinci-002 ZS arguments

(g) Introspective Davinci-003 FS arguments

(h) Introspective Davinci-002 FS arguments

Fig. 14. Histograms of median ratings (part 2).



(i) Simple GPT-4 ZS arguments

(j) Simple GPT-4 FS arguments

(k) Introspective GPT-4 FS arguments

(l) Introspective GPT-4 FS arguments

Fig. 14. Histograms of median ratings (part 3).

A.7. Tukey HSD pairwise comparison of the results in Table 2

See Table 13.

A.8. Tukey HSD pairwise comparison of the results in Table 11 for the unfiltered subset

See Table 14.

Table 13

Multiple comparison of means – Tukey HSD, FWER = 0.05

group1	group2	meandiff	*p*-adj	Lower	Upper	Reject*
introspective_davinci2_fs	introspective_davinci2_zs	−0.3212	0.1312	−0.6857	0.0433	False
introspective_davinci2_fs	introspective_davinci3_fs	0.3342	0.0153	0.0368	0.6315	True
introspective_davinci2_fs	introspective_davinci3_zs	−0.2309	0.5334	−0.5948	0.1330	False
introspective_davinci2_fs	simple_curie_fs	−0.5188	0.0004	−0.8827	−0.1549	True
introspective_davinci2_fs	simple_curie_zs	−0.7882	0.0000	−1.1527	−0.4237	True
introspective_davinci2_fs	simple_davinci3_fs	0.7600	0.0000	0.3955	1.1245	True
introspective_davinci2_fs	simple_davinci3_zs	0.4560	0.0037	0.0921	0.8198	True
introspective_davinci2_zs	introspective_davinci3_fs	0.6554	0.0000	0.2909	1.0199	True
introspective_davinci2_zs	introspective_davinci3_zs	0.0903	0.9981	−0.3302	0.5108	False
introspective_davinci2_zs	simple_curie_fs	−0.1976	0.8456	−0.6181	0.2230	False
introspective_davinci2_zs	simple_curie_zs	−0.4670	0.0177	−0.8881	−0.0459	True
introspective_davinci2_zs	simple_davinci3_fs	1.0812	0.0000	0.6602	1.5023	True
introspective_davinci2_zs	simple_davinci3_zs	0.7772	0.0000	0.3567	1.1977	True
introspective_davinci3_fs	introspective_davinci3_zs	−0.5651	0.0001	−0.9290	−0.2012	True
introspective_davinci3_fs	simple_curie_fs	−0.8530	0.0000	−1.2168	−0.4891	True
introspective_davinci3_fs	simple_curie_zs	−1.1224	0.0000	−1.4869	−0.7579	True
introspective_davinci3_fs	simple_davinci3_fs	0.4258	0.0096	0.0613	0.7903	True
introspective_davinci3_fs	simple_davinci3_zs	0.1218	0.9722	−0.2421	0.4857	False
introspective_davinci3_zs	simple_curie_fs	−0.2879	0.4285	−0.7079	0.1321	False
introspective_davinci3_zs	simple_curie_zs	−0.5573	0.0015	−0.9778	−0.1368	True
introspective_davinci3_zs	simple_davinci3_fs	0.9909	0.0000	0.5704	1.4114	True
introspective_davinci3_zs	simple_davinci3_zs	0.6869	0.0000	0.2669	1.1069	True
simple_curie_fs	simple_curie_zs	−0.2694	0.5205	−0.6900	0.1511	False
simple_curie_fs	simple_davinci3_fs	1.2788	0.0000	0.8583	1.6993	True
simple_curie_fs	simple_davinci3_zs	0.9747	0.0000	0.5548	1.3947	True
simple_curie_zs	simple_davinci3_fs	1.5482	0.0000	1.1272	1.9693	True
simple_curie_zs	simple_davinci3_zs	1.2442	0.0000	0.8237	1.6647	True
simple_davinci3_fs	simple_davinci3_zs	−0.3040	0.3559	−0.7246	0.1165	False

* *reject* = *True* implies that the specified mean difference is statistically significant.

Table 14

Multiple comparison of means – Tukey HSD, FWER = 0.05

group1	group2	meandiff	*p*-adj	Lower	Upper	Reject
introspective_gpt4_fs	introspective_gpt4_zs	−0.1050	0.9087	−0.5095	0.2995	False
introspective_gpt4_fs	simple_gpt4_fs	0.3250	0.1641	−0.0795	0.7295	False
introspective_gpt4_fs	simple_gpt4_zs	0.4450	0.0245	0.0405	0.8495	True
introspective_gpt4_zs	simple_gpt4_fs	0.4300	0.0837	−0.0371	0.8971	False
introspective_gpt4_zs	simple_gpt4_zs	0.5500	0.0135	0.0829	1.0171	True
simple_gpt4_fs	simple_gpt4_zs	0.1200	0.9112	−0.3471	0.5871	False

References

[1] G. Albaum, The Likert scale revisited, *Market Research Society. Journal.* **39**(2) (1997), 1–21. doi:10.1177/147078539703900202.

[2] T. Bench-Capon, Open texture and argumentation: What makes an argument persuasive?, in: *Logic Programs, Norms and Action: Essays in Honor of Marek J. Sergot on the Occasion of His 60th Birthday*, A. Artikis, R. Craven, N.K. Çiçekli, B. Sadighi and K. Stathis, eds, Springer, 2012.

[3] S. Blackburn, *Oxford Dictionary of Philosophy*, Oxford University Press, 2016.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., Language models are few-shot learners, *Advances in neural information processing systems* **33** (2020), 1877–1901.

[5] P.F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg and D. Amodei, Deep reinforcement learning from human preferences, in: *Advances in Neural Information Processing Systems*, 2017, pp. 4299–4307.

[6] L. Fields, Z. Marji and J. Licato, Of a different persuasion: Perception of minority status and persuasive impact, in: *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 44, 2022.

[7] J. Franklin, Discussion paper: How much of commonsense and legal reasoning is formalizable? A review of conceptual obstacles, *Law, Probability and Risk* **11**(2–3) (2012), 225–245. doi:10.1093/lpr/mgs007.

[8] T. Gao, A. Fisch and D. Chen, Making pre-trained language models better few-shot learners, 2020, arXiv preprint arXiv:2012.15723.

[9] H.L.A. Hart, *The Concept of Law*, Clarendon Press, 1961.

[10] M. Hinton and J.H.M. Wagemans, How persuasive is Ai-generated argumentation? An analysis of the quality of an argumentative text produced by the Gpt-3 Ai text generator, *Argument and Computation* **14**(1) (2023), 59–74. doi:10.3233/AAC-210026.

[11] J. Huang and K.C.-C. Chang, Towards reasoning in large language models: A survey, 2022, arXiv preprint arXiv:2212.10403.

[12] Z. Jiang, F.F. Xu, J. Araki and G. Neubig, How can we know what language models know?, *Transactions of the Association for Computational Linguistics* **8** (2020), 423–438. doi:10.1162/tacl_a_00324.

[13] J. Jung, L. Qin, S. Welleck, F. Brahman, C. Bhagavatula, R.L. Bras and Y. Choi, Maieutic prompting: Logically consistent reasoning with recursive explanations, 2022, arXiv preprint arXiv:2205.11822.

[14] T. Kojima, S.S. Gu, M. Reid, Y. Matsuo and Y. Iwasawa, Large language models are zero-shot reasoners, *Advances in neural information processing systems* **35** (2022), 22199–22213.

[15] J. Licato, Automated ethical reasoners must be interpretation-capable, in: *Proceedings of the AAAI 2022 Spring Workshop on "Ethical Computing: Metrics for Measuring AI's Proficiency and Competency for Ethical Reasoning"*, 2022.

[16] J. Licato, War-gaming needs argument-justified AI more than explainable AI, in: *Proceedings of the 2022 Advances on Societal Digital Transformation (DIGITAL) Special Track on Explainable AI in Societal Games (XAISG)*, 2022.

[17] J. Licato, L. Fields and Z. Marji, Resolving open-textured rules with templated interpretive arguments, in: *European Conference on Argumentation*, 2022.

[18] J. Licato and Z. Marji, Probing formal/informal misalignment with the loophole task, in: *Proceedings of the 2018 International Conference on Robot Ethics and Standards (ICRES 2018)*, 2018.

[19] J. Licato, Z. Marji and S. Abraham, Scenarios and recommendations for ethical interpretive AI, in: *Proceedings of the AAAI 2019 Fall Symposium on Human-Centered AI*, Arlington, VA, 2019.

[20] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi and G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, 2021, arXiv preprint arXiv:2107.13586.

[21] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang and J. Tang, GPT understands, too, 2021, arXiv preprint arXiv:2103.10385.

[22] F. Macagno, D. Walton and G. Sartor, Pragmatic maxims and presumptions in legal interpretation, *Law and Philosophy* **37**(1) (2018), 69–115, ISSN 1573-0522. doi:10.1007/s10982-017-9306-4.

[23] D.N. MacCormick and R.S. Summers, *Interpreting Statutes: A Comparative Study*, Routledge, 1991.

[24] Z. Marji and J. Licato, Aporia: The argumentation game, in: *Proceedings of the Third Workshop on Argument Strength*, 2021.

[25] OpenAI, Introducing ChatGPT, 2022.

[26] OpenAI, GPT-4 technical report, 2023.

[27] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C.L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike and R. Lowe, *Training Language Models to Follow Instructions with Human Feedback*, 2022.

[28] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A.H. Miller and S. Riedel, Language models as knowledge bases? 2019, arXiv preprint arXiv:1909.01066.

[29] H. Prakken, On the problem of making autonomous vehicles conform to traffic law, *Artificial Intelligence and Law* **25**(3) (2017), 341–363, ISSN 1572-8382. doi:10.1007/s10506-017-9210-0.

[30] O. Press, M. Zhang, S. Min, L. Schmidt, N.A. Smith and M. Lewis, Measuring and narrowing the compositionality gap in language models, 2022, arXiv preprint arXiv:2210.03350.

[31] R. Quandt and J. Licato, Problems of autonomous agents following informal, open-textured rules, in: *Human–Machine Shared Contexts*, W.F. Lawless, R. Mittu and D.A. Sofge, eds, Academic Press, 2020.

[32] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever et al., Language models are unsupervised multitask learners, *OpenAI blog* **1**(8) (2019), 9.

[33] A. Rotolo, G. Governatori and G. Sartor, Deontic defeasible reasoning in legal interpretation: Two options for modelling interpretive arguments, in: *Proceedings of the 15th International Conference on Artificial Intelligence and Law, ICAIL '15*, ACM, New York, NY, USA, 2015, pp. 99–108. ISBN 978-1-4503-3522-5. doi:10.1145/2746090.2746100.

[34] V.S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang and S. Feizi, 2023, Can AI-generated text be reliably detected?

[35] G. Sartor, D. Walton, F. Macagno and A. Rotolo, Argumentation schemes for statutory interpretation: A logical analysis, in: *Legal Knowledge and Information Systems. (Proceedings of JURIX*, Vol. 14, 2014, pp. 21–28.

[36] T. Schick and H. Schütze, Exploiting cloze questions for few shot text classification and natural language inference, 2020, arXiv preprint arXiv:2001.07676.

[37] V.P. Singh and P. Pal, Survey of different types of CAPTCHA.

[38] A.M. Turing, *Mind, Mind* **59**(236) (1950), 433–460. doi:10.1093/mind/LIX.236.433.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser and I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* **30** (2017).

[40] J.J. Vecht, Open texture clarified, *Inquiry* **0**(0) (2020), 1–21. doi:10.1080/0020174X.2020.1787222.

[41] F. Waismann, *The Principles of Linguistic Philosophy*, St. Martins Press, 1965.

[42] E. Wallace, S. Feng, N. Kandpal, M. Gardner and S. Singh, Universal adversarial triggers for attacking and analyzing NLP, 2019, arXiv preprint arXiv:1908.07125.

[43] D. Walton, F. Macagno and G. Sartor, *Statutory Interpretation: Pragmatics and Argumentation*, Cambridge University Press, 2021.

[44] D. Walton, G. Sartor and F. Macagno, An argumentation framework for contested cases of statutory interpretation, *Artificial Intelligence and Law* **24** (2016), 51–91. doi:10.1007/s10506-016-9179-0.

[45] D. Walton, G. Sartor and F. Macagno, Statutory interpretation as argumentation, in: *Handbook of Legal Reasoning and Argumentation*, G. Bongiovanni, G. Postema, A. Rotolo, G. Sartor, C. Valentini and D. Walton, eds, Springer, Netherlands, Dordrecht, 2018, pp. 519–560. ISBN 978-90-481-9452-0. doi:10.1007/978-90-481-9452-0_18.

[46] Y. Wang, Q. Yao, J.T. Kwok and L.M. Ni, Generalizing from a few examples: A survey on few-shot learning, *ACM computing surveys (csur)* **53**(3) (2020), 1–34. doi:10.1145/3365000.

[47] J. Wei, M. Bosma, V.Y. Zhao, K. Guu, A.W. Yu, B. Lester, N. Du, A.M. Dai and Q.V. Le, Finetuned language models are zero-shot learners, 2021, arXiv preprint arXiv:2109.01652.

[48] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le and D. Zhou, Chain of thought prompting elicits reasoning in large language models, 2022, arXiv preprint arXiv:2201.11903.

[49] W.X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong et al., A survey of large language models, 2023, arXiv preprint arXiv:2303.18223.

[50] Z. Zhong, D. Friedman and D. Chen, Factual probing is [MASK]: Learning vs. learning to recall, 2021, arXiv preprint arXiv:2104.05240.