

Article

Prompt Engineering for Knowledge Creation: Using Chain-of-Thought to Support Students' Improvable Ideas

Alwyn Vwen Yen Lee ^{*}, Chew Lee Teo and Seng Chee Tan 

National Institute of Education, Nanyang Technological University, Singapore 637616, Singapore

* Correspondence: alwyn.lee@nie.edu.sg

Abstract: Knowledge creation in education is a critical practice for advancing collective knowledge and fostering innovation within a student community. Students play vital roles in identifying gaps and collaborative work to improve community ideas from discourse, but idea quality can be suboptimal, affected by a lack of resources or diversity of ideas. The use of generative Artificial Intelligence and large language models (LLMs) in education has allowed work on idea-centric discussions to advance in ways that were previously unfeasible. However, the use of LLMs requires specific skill sets in prompt engineering, relevant to the in-context technique known as Chain-of-Thought (CoT) for generating and supporting improvable ideas in student discourse. A total of 721 discourse turns consisting of 272 relevant question–answer pairs and 149 threads of student discourse data were collected from 31 students during a two-day student Knowledge Building Design Studio (sKBDS). Student responses were augmented using the CoT approach and the LLM-generated responses were compared with students' original responses. Findings are illustrated using two threads to show that CoT-augmented inputs for the LLMs can generate responses that support improvable ideas in the context of knowledge creation. This study presents work from authentic student discourse and has implications for research and classroom practice.

Keywords: prompt engineering; knowledge creation; knowledge building; chain-of-thought; large language models



Citation: Lee, A.V.Y.; Teo, C.L.; Tan, S.C. Prompt Engineering for Knowledge Creation: Using Chain-of-Thought to Support Students' Improvable Ideas. *AI* **2024**, *5*, 1446–1461. <https://doi.org/10.3390/ai5030069>

Academic Editors: Luis Javier Garcia Villalba and Arslan Munir

Received: 4 July 2024

Revised: 8 August 2024

Accepted: 14 August 2024

Published: 16 August 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The use of conventional large language models (LLMs) has generated a massive amount of interest across multiple disciplines and fields, with abilities ranging from answering simple questions to the solving of complex problems and generation of codes [1]. In the field of education, knowledge creation is an important and critical suite of practices by which a community advances its collective knowledge and is closely related to innovation [2]. Although students may not be on par with scientists in their ability to create new knowledge, they can be put on a knowledge building trajectory and start developing similar practices such as identifying major gaps in a community's knowledge, learning to set up goals, and identifying ways to address and work together to improve a community's ideas over time [3]. In recent times, technological advancements have supported students in knowledge building discourse via an online discourse platform [4] (Knowledge Forum) that supports knowledge creation and knowledge building, where students can collaboratively work on ideas, and it also serves as an archive of the community's ideas as students develop and improve their ideas over time.

Although student efforts in knowledge creation have been quite well-researched and analyzed over the years along with a greater understanding of students' responses [5,6], the quality of responses remains a critical issue because it is dependent on a multitude of factors, including innate factors such as motivation, along with the availability and diversity of ideas that excite students in furthering ideas through discourse. At times, students' responses are either lacking or not at the level that some educators would expect to achieve

in certain learning environments. The emergence of generative AI, especially with the increased accessibility of Generative Pre-trained Transformer (GPT)-related technologies such as ChatGPT, gives rise to a new technological tool that enables educators and even students to work on their questions and ideas in ways that were previously not possible.

However, to utilize and converse effectively with the LLMs, users must be proficient with prompt engineering, essentially crafting instructions to LLMs for customizing outputs and the enhancement of the models' capabilities. The key to prompt engineering is to design and craft prompts that will be used for downstream tasks, in guiding pre-trained models to perform desired tasks and achieve required performance goals [7]. It has been particularly difficult to use LLMs with challenging tasks such as arithmetic reasoning, commonsense reasoning, and symbolic reasoning [8], especially if the goal is to work with commonsense-reasoning-related tasks. Methods to guide model learning include fine-tuning pre-trained models and in-context prompting as means to guide models to perform desired tasks.

Instead of the traditional route of fine-tuning LLMs using gradient updates [9], which is a costly and time-consuming process, an increasing number of researchers prefer the use of in-context prompting with LLMs to achieve improved success using several input-output exemplars [10]. The use of prompts to guide model learning [11], also referred to as prompt learning, can enable language models to perform few-shot or zero-shot learning and adapt to new scenarios with minimally labelled data [12]. This paper focuses on understanding the use of prompt engineering for knowledge creation and is guided by the following research question: "How can the approach of Chain-of-Thought [13] (CoT) be applied to support students' improvable ideas in a knowledge creation environment?"

We hypothesize that the Chain-of-Thought (CoT) approach in prompt engineering will provide for a more elaborate and insightful way of supporting students' ideas in a knowledge creation environment, therefore improving students' capacity for generating and improving ideas while enhancing the overall quality of the knowledge building discourse facilitated by the use of LLMs. The importance of this development lies in how a prompting-only approach can uncover greater possibilities of working without large training datasets and the ability of a single checkpoint model to perform many tasks without a loss of generality. With the use of Chain-of-Thought prompting for knowledge creation and commonsense reasoning, this work intends to contribute to the field of applied AI for educational purposes.

2. Background and Literature

2.1. Emergence of Generative AI and LLMs

Since the 1960s, generative AI has been present in the form of chatbots and has been growing in prominence within the past decade due to the development of generative adversarial networks (GANs) in 2014 by Goodfellow and colleagues [14]. GANs were seen to be a disruptive development, considering their enormous potential in mimicking distributions of data, the synthesis of new content, and the ability to create real-world objects across multiple domains. While there were potential benefits and impacts on society, there were also negative connotations with fake media content such as deepfake videos, which blurred the lines between actual recorded data and digitally constructed content, making it difficult to determine authenticity [15].

These forms of information disorders do not detract from the large numbers of benefits that were provided via the use of another family of neural networks that use a transformer architecture, namely Generative Pre-trained Transformer (GPT) [10]. This key advancement in AI has helped to power applications such as ChatGPT, which rose to fame towards the end of 2022 and presented itself as a much more powerful and accessible system that can eventually steer research and development to artificial general intelligence and a system that can solve human-level problems. On the education front, it has already been used for teaching and learning with prospects and challenges [16,17], designs and practices [18], and

strategies [19] that are crafted around the responsible use of ChatGPT, which is a critical aspect for encouraging wider use by educators and students.

2.2. Role and Various Approaches of Prompt Engineering

Prompt engineering, as enabled by in-context learning and prompting, gained prominence after the introduction of ChatGPT, since the main method of input and interaction with the LLM is to issue text to it to garner a response, or technically an instruction to the LLM. When communicating with an LLM such as ChatGPT, prompts—including questions, claims, and statements—are used and users would expect an accurate and adequate answer.

As it is with how people respond to each other in real life, the prompts can have varying levels of clarity and specificity. Essentially, good prompts that elicit good responses from LLMs need to fulfil two criteria: clarity and specificity [11]. Clear prompts contain a simple vocabulary and straightforward language that is unambiguous and does not contain technical jargon. Prompts with specificity require the provision of the necessary contextual information for the LLM to answer the question or to fulfil the given task.

Prompting also takes on many forms and, among the various approaches, standard prompting is considered the basic approach and is often used for benchmarking LLM performances. Several approaches exist, including the following:

- Zero-shot. This entails simply feeding a natural language description of the task to the model without exemplars and requests for a result;
- One-shot. A single exemplar is provided to the model so that it can provide a result similar to what is intended and is often better than a model without an exemplar;
- Few-shot. A set of high-quality demonstrations or exemplars are introduced to the model at inference time, consisting of both input–output pairs for a target task, essentially to get the model to provide a result that better considers the human intention and criteria, with performance that is often better than zero-shot and one-shot attempts.

Table 1 presents several examples of what prompting looks like when students ask questions with a zero-shot, one-shot, or few-shot approach.

Table 1. Examples of zero-shot, one-shot, and few-shot prompting in question-and-answer format based on topics that students ask (screenshots of inputs and outputs can be found in Figure A1 of Appendix A).

Approaches	Format	Input	Output
Zero-shot	<Question> ?	Answer True or False: Question: Will recycling alone reverse global warming?	False
One-shot	<Question> ?	Answer which appliance: Question: Which appliance consumes more electricity in same period of time, laptop or desktop?	Water heater
	<Answer> <Question> ?	Answer: Desktop Question: Which appliance consumes more electricity in same period of time, water heater or fridge?	
Few-shot	<Question> ?	Answer good or bad idea: Question: Can we keep appliances on standby when not using them?	Good idea.
	<Answer>	Answer: Bad idea.	
	<Question> ?	Question: Maybe we should use less plastics if they become trash in the ocean?	
	<Answer>	Answer: Good idea.	
	<Question> ?	Question: Should all renewable energy sources be solar?	
	<Answer>	Answer: Bad idea.	
	<Question> ?	Question: Is it better to encourage eating more vegetables to reduce greenhouse gas emissions?	

It should be noted that the question-and-answer format need not be strictly used and the format of prompting depends on the nature of the tasks, which can be classification

tasks that give exemplars that demonstrate a particular task. The last example in Table 1 with the few-shot example demonstrates that ideas can be largely classified by the LLM.

Beyond in-context prompting, there are other extensions of prompting with LLMs, such as fine-tuning [20], automation [21,22], and the customization of generated outputs. There are also published catalogs [11] documenting frameworks for documenting patterns for structuring prompts to solve a range of problems and adaptations to different domains, with patterns that have been used successfully to improve outputs from LLMs.

2.3. Chain-of-Thought (CoT)

Complicated reasoning processes that often involve intricate ideas and plans will likely require intermediate steps that lead to a consensus and final solutions, which few-shot prompting is able to demonstrate to a certain extent.

A recent study [13] presented a promising approach called Chain-of-Thought (CoT) prompting, which was able to unlock the promising potential of reasoning in large language models by introducing a series of intermediate reasoning steps. In essence, this method of prompting encourages the model to reason similarly to how the prompt was written as a series of steps. This form of prompting has several benefits over standard prompting: it is able to build on and incorporate intermediate steps into current few-shot learning tasks, thereby improving the model’s improvement on complex problems; it enhances the way in which results can be interpreted and therefore presented with greater insights into model behavior and decision-making, with more transparency on and a greater understanding of how it generates outputs; and, more importantly, it can outperform standard prompting across arithmetic reasoning, commonsense reasoning, and symbolic reasoning tasks.

Among these three reasoning tasks (arithmetic reasoning, commonsense reasoning, and symbolic reasoning tasks), and as explained and exemplified in the context of CoT [13], the study in this paper focuses on the commonsense reasoning tasks that align with the nature of knowledge creation. We mapped related processes together to advance the way in which prompt engineering can be conducted for knowledge creation.

On the one hand, in general, commonsense reasoning, which is key for interacting with the world, is still beyond the reach of most natural language understanding systems [23]. Commonsense reasoning pertains to the answering of questions that involve our understanding of everyday situations and making inferences based on general knowledge. Within the domain of knowledge creation, this touches on how learners can pose questions related to one’s authentic experiences, raise real problems from these experiences, and create and improve ideas to help resolve problems. On the other hand, Table 2 shows an example of commonsense reasoning that can be resolved using an LLM, which justifies the foray into the use of prompt engineering and LLMs for reasoning and knowledge creation.

Table 2. Example of commonsense reasoning using an LLM (a screenshot of the input and output can be found in Figure A2 of Appendix A).

Question: If John wants to go to a place where he can enjoy playing with water, where would he go to? Answer choices: (a) an aquarium (b) the ocean (c) nearby pond (d) a reservoir (e) his bathtub
Answer: (b) the ocean

There are several desirable properties for facilitating reasoning in LLMs and these properties were identified in Wei et al.’s paper [13] as follows:

1. Firstly, models should be allowed to break down multi-step problems into intermediate steps, which can be implemented with more resources if required;
2. Next, the CoT as an interpretable window into the behavior of the model can allow the model to decide how best to support an answer;
3. CoT reasoning is used for arithmetic, commonsense reasoning, and symbolic manipulation, with potential application to tasks that humans can solve via language;

4. CoT reasoning can be readily elicited from sufficiently large off-the-shelf language models via examples of CoT sequences in exemplars of few-shot prompting.

As with other forms of prompting, CoT prompting can also be conducted in the following two ways, similar to those shown in Section 2.2:

- Zero-shot CoT. Natural language statements are used to encourage the model to generate reasoning chains;
- Few-shot CoT. The model is prompted with a few high-quality demonstrations or exemplars to generate a result. This approach was used in this study to eventually lead to the final answer as the benefits are more pronounced for complicated reasoning tasks with larger language models.

By including several demonstrations of CoT as exemplars in few-shot prompting, this combination of actions can aid LLMs to generate chains of thought in response. To demonstrate this, we asked the LLM a question related to human sustainability in terms of the overcrowding of usable spaces. Table 3 shows an example of what standard prompting looks like in comparison to how the provision of exemplars for CoT prompting will look.

Table 3. Comparing standard prompting against Chain-of-Thought (CoT) prompting (screenshots of inputs and outputs can be found in Figure A3 of Appendix A).

Approaches	Format	Input	Output
Standard prompting	<Question> ?	Question: John wants to enjoy places that are less crowded. Where can he go to? Answer choices: (a) a shopping mall (b) a park (c) a train station (d) a corridor (e) by the roadside	Answer: (a) a shopping mall
	<Answer>	Answer: (b) a park	
CoT prompting <i>(italicized)</i>	<Question> ?	Question: John likes to visit crowded areas. Where can he go to? Answer choices: (a) a shopping mall (b) a park (c) a train station (d) a corridor (e) by the roadside	Answer: <i>The answer should be a place with more people. From the given choices, a shopping mall, a train station, and possibly a park during peak times would have more people compared to a corridor or the roadside. So, the answer could be either (a) a shopping mall or (c) a train station.</i>
	<Answer provided with CoT>	Question: John wants to enjoy places that are less crowded. Where can he go to? Answer choices: (a) a shopping mall (b) a park (c) a train station (d) a corridor (e) by the roadside	
	<Question> ?	Answer: <i>The answer should be a place with lesser people. From the given choices, a park has lesser people and is more enjoyable than a corridor or the roadside. So the answer is (b) a park</i>	
	<Question> ?	Question: John likes to visit crowded areas. Where can he go to? Answer choices: (a) a shopping mall (b) a park (c) a train station (d) a corridor (e) by the roadside	

From the responses in Table 3, the standard prompt was able to provide a likely correct answer, and it might be debatable whether a train station is more crowded than the provided answer of a shopping mall. Standard prompting was able to ensure that there is a reasonable answer, but the reasoning was not made clear to the user.

With CoT prompting, it is apparent that a straight direct answer was possible but not given due to the limited information that in-context prompting can provide. However, the value of CoT prompting provides additional reasons that will help the LLM user better decide what might be a more suitable answer, which, in this case, is either a shopping mall or a train station, depending on peak periods. This is a demonstration of how the provision of CoT for prompting can significantly improve the ability of the LLM to conduct complex reasoning.

As part of this extended literature review, we note several enhancements and extensions that can be used to further enhance the current CoT approach’s reasoning performance, one of which is existing work to replace the original naïve greedy decoding strategy

used in CoT prompting with self-consistency [24]. This technique leverages the intuition of how there may be multiple diverse ways of thinking and reasoning that will lead to a single unique correct answer. This development banks on the rationale that, when there is more deliberate thinking and analysis required for a problem, there will likely be a greater diversity of reasoning paths that lead to an answer [25]. Hence, this development has been noted as a possible extension for enhancing the way in which CoT can be improved to assist with the processes of knowledge creation and knowledge building. However, we also highlight that this development is not the focus of this paper.

2.4. Knowledge Creation and Knowledge Building

Complicated reasoning processes that often involve intricate ideas and plans will likely require intermediate steps that lead to the convergence and creation of newer knowledge. Before this can be achieved, there needs to be an alignment of education with the needs of knowledge creation in the digital age. Knowledge building is akin to knowledge creation that is practiced in research laboratories, focusing on value to the individual, community, and society, evolving with theory and pedagogy over the years, and making knowledge creation processes available to school-aged students [26].

Knowledge building, as a principle-based pedagogy [27] and suite of practices, allows students to work on big questions (e.g., Does a reduction in single-use plastics make a difference to climate change?) and collaboratively develop and improve ideas through epistemic agency, improvable discourse, and ideas. By having students' ideas placed within a public space such as the Knowledge Forum [4], it is then possible for students to visualize and seek out diverse views and ideas and build on each other's ideas as part of community efforts to further advance collective knowledge. This sustained design implementation research is currently a systematic approach that can help reconstruct educational practices to establish self-improving systems for continual alignments in knowledge creation, with implications for developing self-improving systems and communities that leverage technology for realigning education in knowledge creation [26].

3. Materials and Methods

3.1. CoT Design for Knowledge Creation

Based on current practices where students seek out each other's ideas or reference them from other sources, the use of generative AI with an LLM provides an alternative and emergent source of information that can support students in improving their ideas. Using the CoT prompting method described in Section 2.3 and through the lens of knowledge creation with the use case situated in a knowledge building environment to support improvable ideas, the following actions can be taken to design CoT for knowledge creation:

1. Students' questions, as part of knowledge building, can be broken down into multiple intermediate steps to be implemented as prompts;
2. The CoT can be sufficiently contextualized and supplemented with information to provide a sufficient window of interpretation and, therefore, allow room for the ideas to be fully fleshed out in the generated output;
3. CoT reasoning can be used for commonsense reasoning as part of the alignment with the needs of knowledge creation in the knowledge building environment;
4. CoT reasoning can be elicited in the LLM by including examples of CoT sequences into the exemplars of few-shot prompting.

Several few-shot exemplars for CoT prompting are illustrated in Table 4 for knowledge creation, and these can be used to obtain CoT responses from the LLM.

Table 4. Few-shot exemplars showing how we can add Chain-of-Thought prompting (italicized) to existing questions and responses for knowledge creation.

Question 1: If we want to reduce carbon emissions, what should we stop doing? Answer choices: (a) using natural gas (b) using fossil fuels (c) using wind energy (d) using solar energy
Answer 1: <i>The answer must be able to reduce carbon emissions. Of the above choices, fossil fuels contribute the most carbon emissions.</i> So, the answer is (b) using fossil fuels.
—Separate question-and-answer—
Question 2: Which way of travelling to school is more environmentally friendly and sustainable? Answer choices: (a) by private electric car (b) by public diesel bus (c) by public electric taxi (d) by public electric trains
Answer 2: <i>The answer must consider public transport is more economical than private transport and then the seating size of transport. Of the above choices, public transport that is electric vehicle is more environmentally friendly and sustainable.</i> So, the answer is (d) by public electric trains.

In this study, we observed how CoT reasoning can be used with commonsense reasoning for knowledge creation by reformatting actual students’ questions from authentic learning settings with CoT prompting and then inputting them into an LLM before comparing the outputs with the actual students’ responses to determine the utility of CoT prompting. The resulting comparisons and outputs will inform educators and students on how they may use CoT to aid their journeys in knowledge building and improve students’ ideas with LLMs.

3.2. Settings, Participants, and Setup

3.2.1. Context and Nature of Student Questions

This study followed a quasi-experimental design, with data obtained from an authentic learning environment that was designed and implemented as a student Knowledge Building Design Studio (sKBDS). The theme was centered on sustainability and the event was held in November 2021. An ethics review was approved and clearance from the university was sought before we collected and analyzed data. A total of 31 participating students from primary and secondary schools (between Grades 5 and 10) underwent two days of online facilitation and exploration to investigate real-world problems relevant to the theme and their questions to each other were posted on the Knowledge Forum. The nature of the student questions was a mixture of natural science and social science, ranging from simple questions to questions with justifications and explanations as students sought answers to their queries.

3.2.2. Extraction of Question–Answer Pairs

Question–answer pairs that students posted on the Knowledge Forum involved students asking and answering questions on a topic and were predominantly used for clarifications of a discussed topic, the review of a recently completed activity, or as part of preparations for presentations. A complete question–answer pair entails a query that is contextualized to the sKBDS’s theme on sustainability (e.g., “How did carbon come about?”) and not solely casual talk (e.g., “What drinks? Bubble tea?”), followed by a suitable response from another participant in the sKBDS. The timing of the reply was asynchronous and limited to responses garnered from the two days of facilitation.

As part of the data-cleaning process and to fulfil the purposes of this study, only complete question–answer pairs were considered for use due to the need for the LLM outputs to be compared with original human answers that allowed us to discuss the utility of CoT. Out of the 721 discourse turns that were collected during the two-day sKBDS event, 272 complete question–answer pairs were identified to be relevant and usable data, forming 149 threads of discussion. In this paper, we illustrate two threads of complete

question–answer pairs in the findings to showcase the potential utility of CoT to support improvable ideas in the context of knowledge creation and knowledge building.

3.2.3. Extraction of Question–Answer Pairs

Through this study, we aimed to better understand the utility of CoT by selecting complete scientific and social science question–answer pairs and augmenting them with CoT few-shot exemplars before sending them as inputs to the LLM. As shown in the examples in Tables 3 and 4, the proposed enhancements will allow LLM-generated outputs to be compared against existing human answers. As part of the method, two or more question–answer pairs need to be consecutive and relevant to each other for comparison, and the responses that follow need not be from the same respondent, since the goal is to assess the quality of answers and differences between an LLM-generated answer and the original human answer. Figure 1 shows how the responses can be compared against each other. The Results section presents the differences and comparisons between the LLM-generated output and students’ original answers.

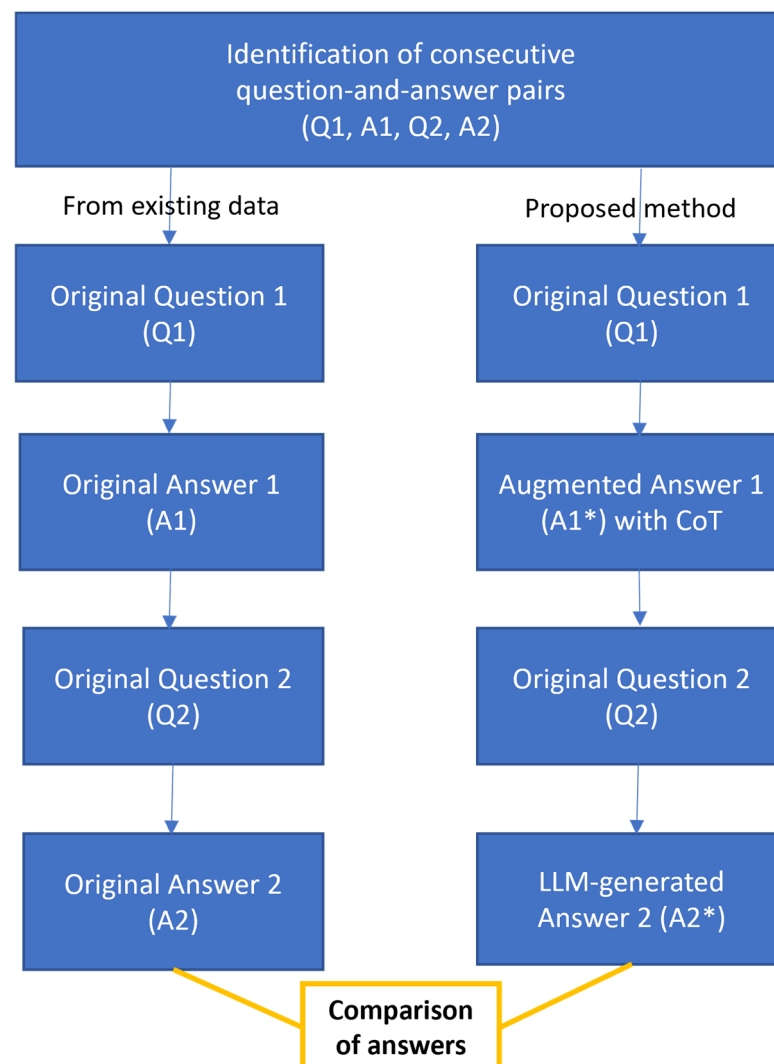


Figure 1. Comparison of answers A2 and A2*, provided respectively by humans (left track) and generated by the LLM using augmented exemplars with CoT few-shot prompting (right track).

3.2.4. Choice of Large Language Model (LLM)

GPT-3.5 Turbo was chosen as the LLM to be used in this study. Although similar to other LLMs such as BERT and PaLM, GPT’s good performance across most tasks and

subjects made it the preferred choice. Additionally, the ease of access to GPT3.5 Turbo via an application programming interface (API) is another reason why it was chosen for implementation. Initial testing with emergent models (e.g., GPT4, GPT4o) showed responses with a level of quality similar to expectations and the proposed method in this study is expected to perform well for knowledge creation, even with the versions of LLMs coming in the near future.

4. Results

A total of 721 discourse turns were scrutinized and 272 question–answer pairs were found to contain content that was relevant to the discussion on sustainability. This indicated that 79% of the student postings were relevant and the other discourse turns were either trivial or casual talk and were thus not analyzed. Among the relevant question–answer pairs, 149 threads of student discussion with at least thread depth 3 were found, and the longest thread at depth 9 had three different participants.

Two example threads of complete question–answer pairs are shown to illustrate how CoT can be used to support improvable ideas in the context of knowledge creation and knowledge building. Both illustrations showcase the sequence of existing data from students' discourse and how responses can be augmented to provide an enhanced and potentially improvable set of responses that are generated by the LLM. The final outputs are compared against each other, and discussions are carried out in Section 5. Figure 2 shows the first of the two illustrations, with an augmented answer constructed with CoT (A1*) serving as the input to the LLM to generate an improved response (A2*). Figure 3 shows the second illustration, where an original response has some semblance of CoT, but was still augmented with the intended use of CoT, and the findings show the additional impact of the CoT approach in helping to improve existing ideas and sustain the development of ideas.

4.1. First Illustration from Empirical Data

The full thread shown in Illustration 1 (Figure 2) can be found in Figure A4 of Appendix A. The LLM-generated response (A2*) is a longer response with more details than the original human answer (A2), and it is fully presented below.

“One advantage of eating plant-based meat is its potential to have a lower environmental impact compared to traditional animal-based meats. Plant-based meats are often produced using fewer resources and generating fewer greenhouse gas emissions than raising and processing animals for meat. This can contribute to reducing the overall carbon footprint associated with food production. Additionally, plant-based meats can offer a similar taste and texture to traditional meats, making them a viable option for individuals who want to reduce their meat consumption for health, ethical, or environmental reasons without completely giving up the familiar flavors and dishes they enjoy”.

4.2. Second Illustration from Empirical Data

The full thread shown in Illustration 2 (Figure 3) can be found in Figure A5 of Appendix A. The LLM-generated response (A2*) is also a longer response with more details than the human-generated response (A2), and the former is fully presented below.

“Certainly, creating an environment where animals and humans can coexist harmoniously is indeed possible, but it requires conscious effort and consideration for both parties. Here's one way to achieve this: Promoting wildlife conservation and habitat restoration: By protecting and restoring natural habitats, humans can provide animals with the space and resources they need to thrive. This involves preserving natural landscapes, planting native vegetation, and creating wildlife corridors that allow animals to move freely between different areas. When ecosystems are healthy, they can support a variety of species, including humans, while minimizing conflicts. Additionally, raising awareness about the importance of biodiversity and the role each species plays in the ecosystem can foster a greater sense of empathy and understanding among humans. Educating people about responsible behavior

around wildlife, such as not feeding them or encroaching on their territory, can help reduce negative interactions”.

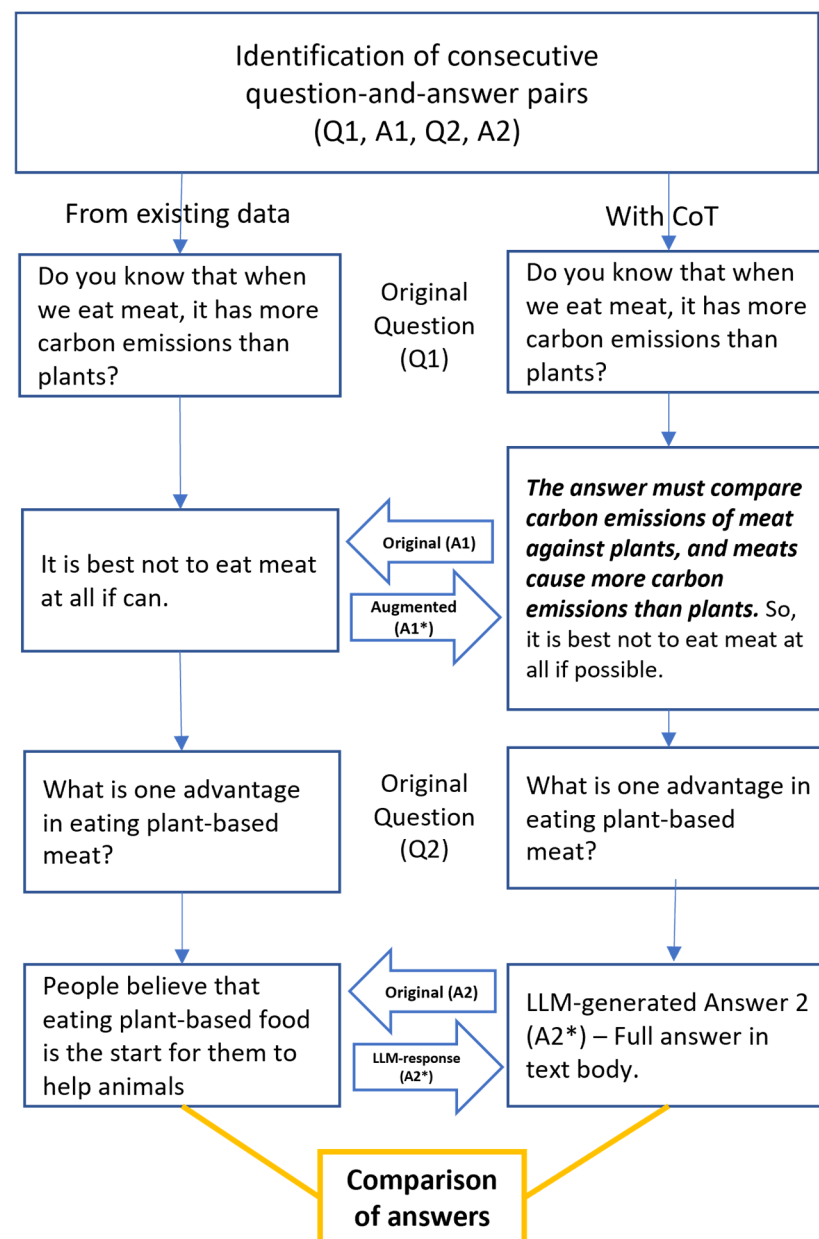


Figure 2. Illustration 1 comparing human answers (A2) against the LLM response (A2*) generated from the CoT-augmented (italicized) answer (A1*) as the input.

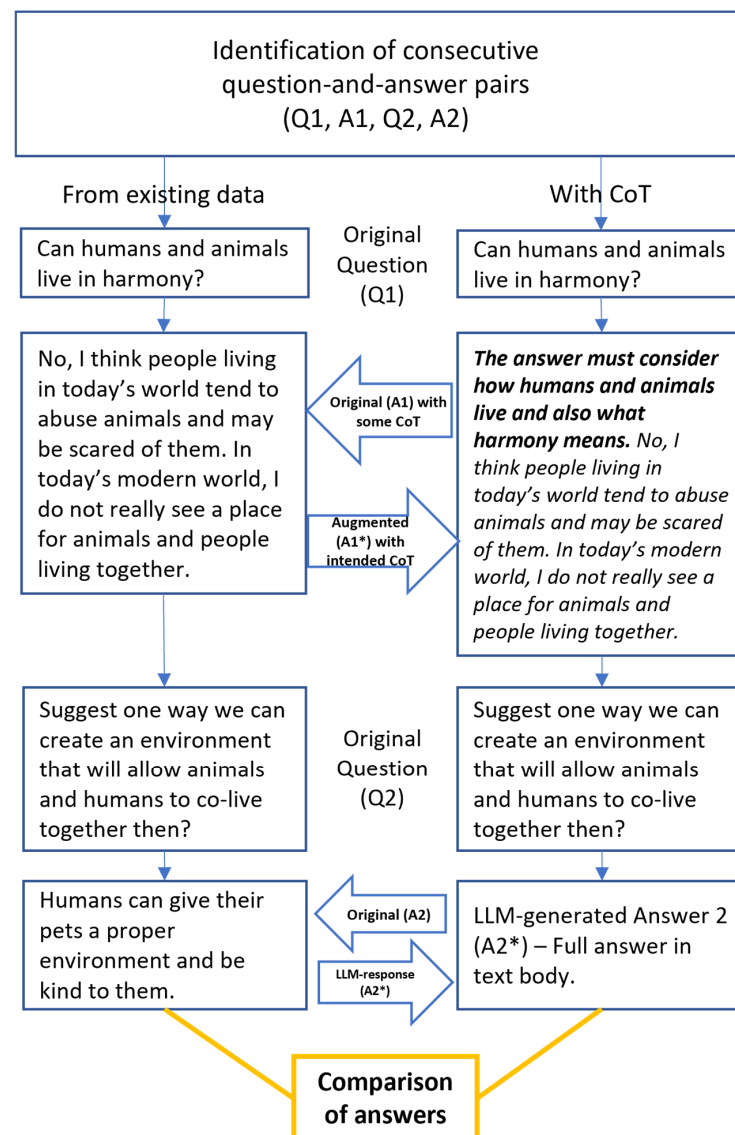


Figure 3. Illustration 2. Signs of CoT in the original response (A1) but differences still observed between the human answer (A2) and the LLM response (A2*) due to the use of the CoT-augmented (italicized) answer (A1*) as the input.

5. Discussion and Limitations

CoT prompting has been explored as a viable alternative to elicit multi-step reasoning behavior in LLMs for knowledge creation purposes. For the specific domain of commonsense reasoning that is aligned with the knowledge creation and knowledge building domains, the findings show that the focus is not just on achieving commonsense goals, but also obtaining an improvement in the quality of LLM-generated content and ideas, with an increase in the diversity of ideas that allow students to improve their work. This is evident from our findings, and the following discussion will explain the differences between the two illustrations.

In the first illustration, on the one hand, we see the human-generated answer by a student as a direct answer that is simple and kickstarts a subsequent discussion. It is, however, lacking in content and substance in provoking deeper thought that may help others improve their ideas. On the other hand, the LLM-generated response is not only an elaborate and more detailed answer, but also one that used the CoT exemplars to guide the response and provide reasons on why plant-based meats produce fewer greenhouse gas emissions. The LLM also generated an additional consequence and more descriptions,

although these were not warranted, but it was also likely a feature that the LLM decided to include in its output.

In the second illustration, when additional exemplars with CoT prompting were included in the original answer A1, the LLM-generated output shows the significant difference that CoT exemplars can make in encouraging the LLM to generate a substantially more diverse and better-quality output. In addition to providing more descriptions to address the question, the resulting output became more predictive and even prescriptive in nature, with suggestions extending to how educating people about responsible behavior around wildlife can help reduce negative interactions.

This study has presented just two illustrations out of the other valid interactions that students generated throughout the study. What has been showcased is that the linguistic nature of CoT reasoning makes it generally applicable, similarly to what Wei et al. [13] showed in demonstrating the reasoning capabilities of generative AI. However, by adopting the use of CoT prompting as part of the synthesis of better answers to improve students' ideas and work, we are potentially seeing a re-tuning of developments to match the emerging needs of students within a knowledge society [17]. With this course of action, CoT prompting can be used beyond the original reasoning processes to generate a broader and diversified range of responses that students may not be able to provide without generative AI and for knowledge creation and knowledge building purposes.

There are admittedly some limitations that may not be fully resolvable at this time. Some may feel that CoT is supposed to emulate the thought processes of human reasoners, but this is still not a certainty, since the neural networks currently being used in LLMs can acquire knowledge but may not be able to reliably conduct reasoning [28]. After all, the LLMs were created by training them on large troves of data to produce human-like responses similar to natural language queries, with no guarantee of correct reasoning paths, which may potentially lead to both correct and wrong answers.

Further, even though we have provided examples from actual student discourse data, it is unlikely that most students will be able to inquire with CoT exemplars when conducting few-shot prompting, but it is possible for students to be taught accordingly on how to use the LLM in an appropriate manner and for the intended purpose. This, however, also means that using CoT to augment students' inquiries may be very taxing and have to be manually performed at this point, while scaling of few-shot CoT prompting may require the maturing of the technology or a rethink of how few-shot processes can eventually become zero-shot generalization that can be performed as part of deep reinforcement learning. Last, but not least, CoT was able to be used due to LLMs and it is possible that the use case may still currently be prohibitive in other real-time applications that use smaller models due to limitations of cost and other material resources.

6. Conclusions

This study has provided a perspective on how students and even educators can tap into LLMs to support knowledge creation and knowledge building processes via the use of few-shot exemplars augmented into CoT prompting. Although prompts were manually crafted around students' authentic inquiries, the results show that it is also in these authentic learning environments that CoT prompting generates improved responses that are more insightful than students' responses, which are limited in terms of content and scope.

In general, it is possible but uncommon that some students may be able to conduct questioning that is already similar to few-shot CoT prompting in order to aid learning. Therefore, we propose through this study that the general approach of using CoT for inquiries can be part of the necessary literacy and scaffolding required to compose and generate questions that are critical for knowledge creation and knowledge building purposes.

Author Contributions: Conceptualization, A.V.Y.L., C.L.T. and S.C.T.; data curation, A.V.Y.L. and C.L.T.; formal analysis, A.V.Y.L.; investigation, A.V.Y.L. and C.L.T.; methodology, A.V.Y.L.; resources, A.V.Y.L. and C.L.T.; software, A.V.Y.L.; validation, A.V.Y.L., C.L.T. and S.C.T.; visualization, A.V.Y.L.;

writing—original draft, A.V.Y.L.; writing—review & editing, A.V.Y.L., C.L.T. and S.C.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Personal Data Protection Act, and approved by the Institutional Review Board of Nanyang Technological University (IRB-2019-10-034, approved October 2019).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets presented in this article are not readily available because of intellectual property, privacy, and ethics regulations stipulated by the university's IRB. Requests to access the datasets should be directed to the corresponding author.

Acknowledgments: The views expressed in this paper are those of the authors and do not necessarily represent the views of the host institution. The research team would like to thank the instructors and student participants involved in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

The following are screenshots from prompts that were reported and used in this paper, consolidated into several figures based on the chronological order of mention.

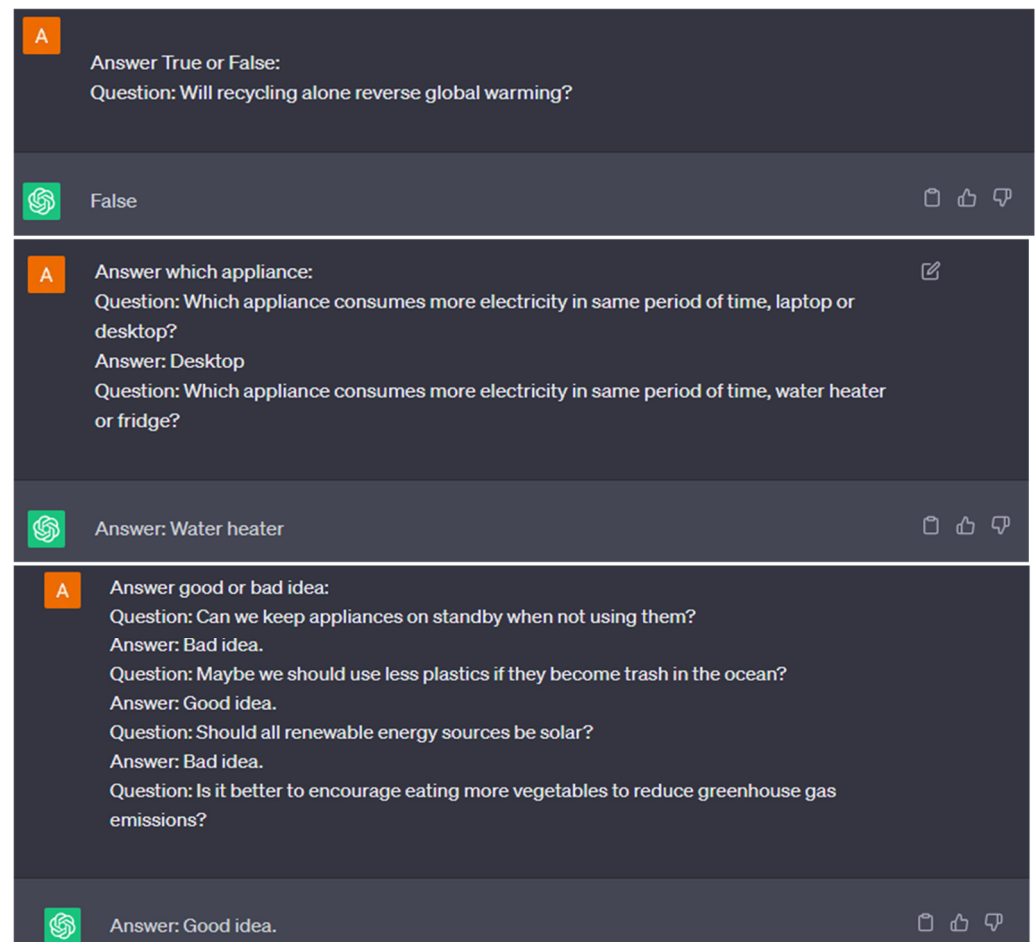


Figure A1. Responses from the LLM based on zero-shot (**top**), one-shot (**middle**), and few-shot (**bottom**) prompting.

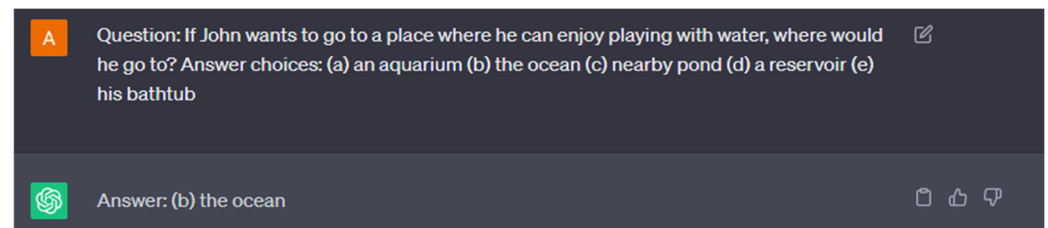


Figure A2. LLM output from the commonsense reasoning task.

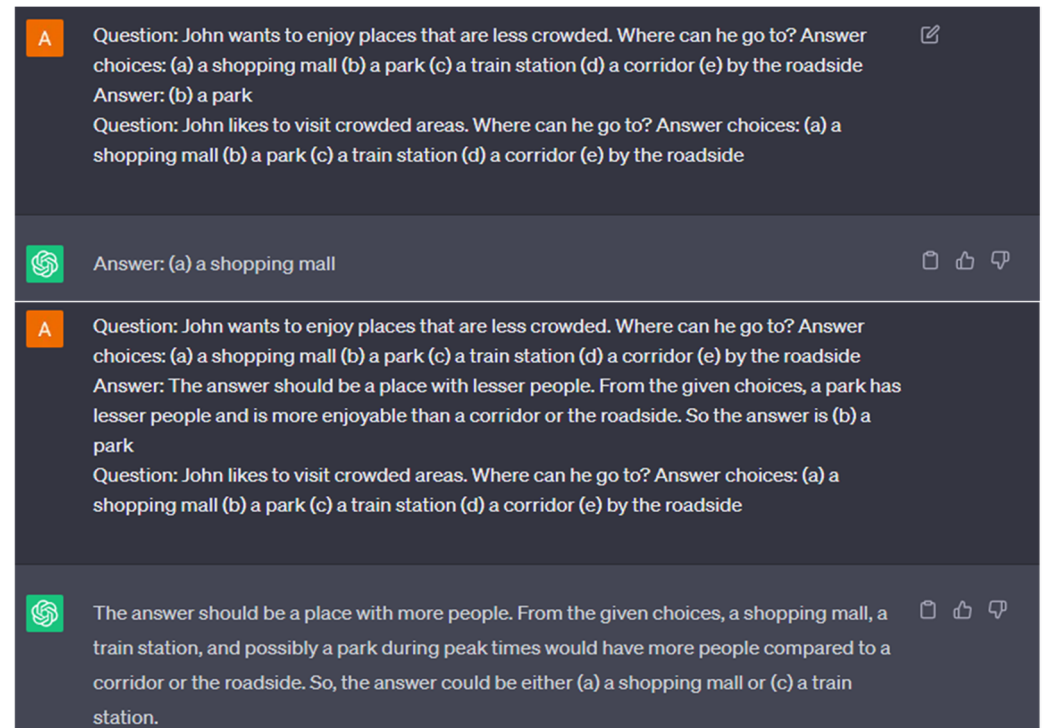


Figure A3. Comparison of the LLM output using standard prompting (**top**) against CoT prompting (**bottom**).

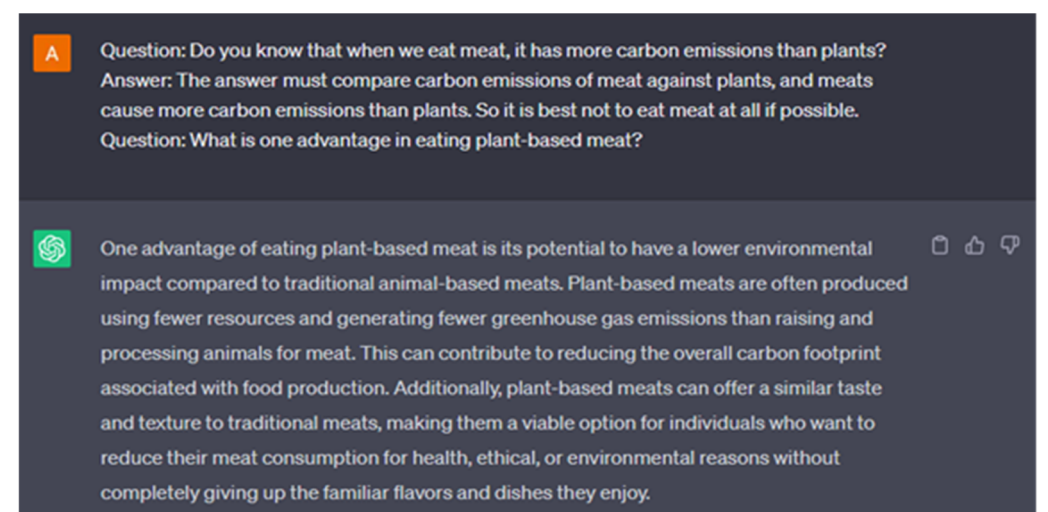


Figure A4. Illustration 1's LLM-generated output after CoT prompting was augmented in the first answer.

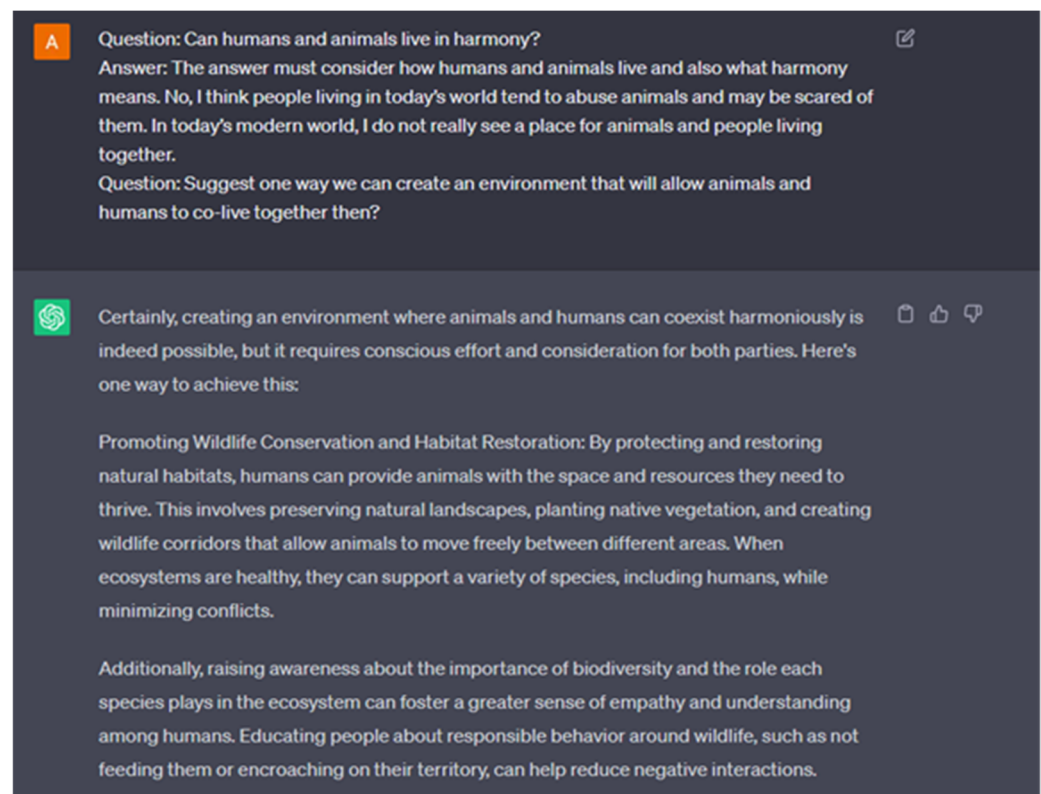


Figure A5. Illustration 2's more elaborate LLM-generated output after additional CoT exemplars were used in prompting.

References

- Hadi, M.U.; Qureshi, R.; Shah, A.; Irfan, M.; Zafar, A.; Shaikh, M.B.; Akhtar, N.; Wu, J.; Mirjalili, S. A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage. *TechRxiv*, 2023; preprint. [\[CrossRef\]](#)
- Paavola, S.; Lasse, L.; Kai, H. Models of innovative knowledge communities and three metaphors of learning. *Rev. Educ. Res.* **2004**, *74*, 557–576. [\[CrossRef\]](#)
- Bereiter, C.; Marlene, S. Learning to work creatively with knowledge. In *Powerful Learning Environments: Unravelling Basic Components and Dimensions*; De Erik, C., Lieven, V., Noel, E., Jeroen Van, M., Eds.; Pergamon Press: Oxford, UK, 2003; pp. 55–68.
- Scardamalia, M. CSILE/Knowledge forum®. In *Education and Technology: An Encyclopedia*; ABC-CLIO: Santa-Barbara, CA, USA, 2004; pp. 183–192.
- Hakkarainen, K.; Paavola, S.; Kangas, K.; Seitamaa-Hakkarainen, P. Social perspectives on collaborative learning: Toward collaborative knowledge creation. In *International Handbook of Collaborative Learning*; Cindy, H.-S., Clark, C., Carol, C., O'Donnell, A., Eds.; Routledge: New York, NY, USA, 2013; pp. 57–73.
- Lee, A.V.Y.; Tan, S.C. Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *J. Learn. Anal.* **2017**, *4*, 76–101. [\[CrossRef\]](#)
- Liu, V.; Lydia, B.C. Design guidelines for prompt engineering text-to-image generative models. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022; pp. 1–23.
- Rae, J.W.; Borgeaud, S.; Cai, T.; Millican, K.; Hoffmann, J.; Song, F.; Aslanides, J.; Henderson, S.; Ring, R.; Young, S.; et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv* **2021**, arXiv:2112.11446. [\[CrossRef\]](#)
- Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. Training verifiers to solve math word problems. *arXiv* **2021**, arXiv:2110.14168. [\[CrossRef\]](#)
- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
- White, J.; Fu, Q.; Hays, S.; Sandborn, M.; Olea, C.; Gilbert, H.; Elnashar, A.; Spencer-Smith, J.; Schmidt, D.C. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv* **2023**, arXiv:2302.11382. [\[CrossRef\]](#)
- Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* **2023**, *55*, 1–35. [\[CrossRef\]](#)
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837.
- Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [\[CrossRef\]](#)

15. Westerlund, M. The emergence of deepfake technology: A review. *Technol. Innov. Manag. Rev.* **2019**, *9*, 39–52. [[CrossRef](#)]
16. Opara, E.; Mfon-Ette Theresa, A.; Aduke, T.C. ChatGPT for teaching, learning and research: Prospects and challenges. *Glob. Acad. J. Humanit. Soc. Sci.* **2023**, *5*, 33–40.
17. Lee, A.; Vwen, Y. Staying ahead with generative artificial intelligence for learning: Navigating challenges and opportunities with 5Ts and 3Rs. *Asia Pac. J. Educ.* **2024**, *44*, 81–93. [[CrossRef](#)]
18. Lee, A.V.Y.; Tan, S.C.; Teo, C.L. Designs and practices using generative AI for sustainable student discourse and knowledge creation. *Smart Learn. Environ.* **2023**, *10*, 59. [[CrossRef](#)]
19. Halaweh, M. ChatGPT in education: Strategies for responsible implementation. *Contemp. Educ. Technol.* **2023**, *15*, ep421. [[CrossRef](#)] [[PubMed](#)]
20. Lester, B.; Al-Rfou, R.; Constant, N. The power of scale for parameter-efficient prompt tuning. *arXiv* **2021**, arXiv:2104.08691. [[CrossRef](#)]
21. Zhang, Z.; Zhang, A.; Li, M.; Smola, A. Automatic chain of thought prompting in large language models. *arXiv* **2022**, arXiv:2210.03493. [[CrossRef](#)]
22. Shum, K.; Diao, S.; Zhang, T. Automatic prompt augmentation and selection with chain-of-thought from labeled data. *arXiv* **2023**, arXiv:2302.12822. [[CrossRef](#)]
23. Talmor, A.; Herzig, J.; Lourie, N.; Berant, J. CommonsenseQA: A question answering challenge targeting commonsense knowledge. *arXiv* **2018**, arXiv:1811.00937. [[CrossRef](#)]
24. Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S. Self-consistency improves chain of thought reasoning in language models. *arXiv* **2022**, arXiv:2203.11171. [[CrossRef](#)]
25. Evans, J. Intuition and reasoning: A dual-process perspective. *Psychol. Inq.* **2010**, *21*, 313–326. [[CrossRef](#)]
26. Tan, S.C.; Chan, C.; Bielaczyc, K.; Ma, L.; Scardamalia, M.; Bereiter, C. Knowledge building: Aligning education with needs for knowledge creation in the digital age. *Educ. Technol. Res. Dev.* **2021**, *69*, 2243–2266. [[CrossRef](#)]
27. Scardamalia, M. Collective cognitive responsibility for the advancement of knowledge. In *Liberal Education in a Knowledge Society*; Open Court: Chicago, IL, USA, 2002; pp. 67–98.
28. Liu, X.; Wu, Z.; Wu, X.; Lu, P.; Chang, K.W.; Feng, Y. Are llms capable of data-based statistical and causal reasoning? Benchmarking advanced quantitative reasoning with data. *arXiv* **2024**, arXiv:2402.17644. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.