# Enhancing AI Responses through Effective Prompt Engineering Strategies for Large Language Models

**Dr. S. Lakshmi** [0000-0003-2553-778X]
Faculty of Science and Humanities, SRM Institute of Science and Technology, Kattankulathur-603203, Chennai, Tamil Nadu, India

**Abstract:** Prompt engineering is one of the main parts which utilizes the power and capacity of large language models ie., LLMs efficiently. Large Language models (LLMs) are generally used to generate human-like text, summarization, solving the problems in various fields, understanding the language and translating the language and so on. The potential of LLMs is utilized by creation of effective prompt which is called as prompt engineering through the inputs are given properly. AI models are used for collecting the relevant information about the query. Extracting the relevant responses from various artificial intelligent models by almost all types of people from researchers to school going children. The challenge lies in crafting prompts which reduce ambiguity and give proper direction to the LLM for getting the desired responses. When we concentrate on prompts by adding the important words or using some key words, we can show better results which reflects the role of prompting techniques clearly. A technical document can be prepared by using a few short prompting techniques and creative writing and storytelling can be done effectively by using some key words in the prompts itself. LLMs such as chatGPT3, chatGPT3.5, chatGPT4, Gemini and other models are trained on huge volumes of data which can produce and generate human-like text easily. The conditional prompts allow the users to use some specific keys for extracting the information on the iterative refinement process can also be used to extract information from prompt engineering. The quality of LLM results is evaluated by using relevance, coherence, creativity and specificity. This work explores the strategies and methods of prompt engineering that could enhance the performance and reliability of the LLMs such as few-shot prompting, role assignment and prompt chaining. Effective prompt engineering is the foremost technique to maximize the utility of large language models in various applications. Advanced techniques such as control tokens and multimodal prompts that combine the text with other modalities such as images for optimizing the results of prompt engineering. Retrieval Augmented Generation gets queries from prompts and try to get relevant information from various sources such as search engines or knowledge graphs. Hencs, RAG extends the LLMs by incorporating external knowledge for enriching the model's responses. The most popular prompt engineering approaches are CoT, ToT, self-consistency and reflection played a major role. Prompt design and engineering are critical and the innovation in the Automatic Prompt Engineering (APE) would dominate in the near future. This work explores the effective utilization of Large Language Models for creating effective prompts for optimizing the responses so that we can solve complex problems easily and can reach better results in a stipulated time.

**Keywords**: LLMs, Prompt Engineering, AI Models, Optimization, Responses, Chat GPT and Gemini

## 1. Introduction

An important milestone in Artificial Intelligence is Natural Language Understanding and Processing. The development of Transformer model by Vaswami et al in 2017[1] leads the development of Generative pre-trained Transformer (GPT series) which is the benchmark in AI world. A lot of research work on prompting from the pre-trained models for extracting the required knowledge or response[2]. Large scale language models produce excellent performance in various natural language processing tasks especially in content creation. Tech giants and research institutes are doing several projects using large language models by implementing various strategies and approaches. Generally, language models are trained on general corpus or domain specific corpus such as webpages, books, magazines and journals. Then the models are fine turned for doing some specific tasks in various natural language processing. The language patterns and structures are learned by using the huge number of LLM parameters.

Jacob Devlin et.al., [3] the language model called BERT-Bidirectional Encoder Representations using transformers. It is designed using unlabeled text by joining the left and right-side data in all the layers. Hence, the pre-trained BERT model can be fine-tuned for doing applications such as question and answering. Radford et al.,[4] discussed about the Generative Pre-trained Transformer (GPT) which is an unsupervised pre-training[22][23] and find-tuning model for performing various natural language understanding tasks.GPT-4 [5] is the latest milestone of OpenAI using deep learning algorithms which is a multimodal model ie., it is accepting image and text inputs and get the results. Brownlee [6] explained the Introduction of Large Language Models for the beginners and practitioners by highlighting the key concepts and models. The environmental and ethical considerations about large language models were discussed by eBommasani in [7]. Brown et al,[8] discussed the prompt structure to perform many tasks to achieve accurate results.

Prompt Engineering is a process of creating prompts to collect the expected refined and optimized responses as an outcome from various Large Language Models. Whatever is given as a prompt, response can be created as a result. Here, the main issue in this situation is that the response match with our expected result. Whether the response satisfy our expected result or not. Getting answer for a given prompt is not at all a matter but getting relevant response is a matter. Here, the prompt engineering played a major role and designing an effective prompt will produce the expected result. For example, in Fig 1 depicts a simple prompt and its response. Fig 2 shows the detailed prompt with key words such as describe, recent and five years and the related response.

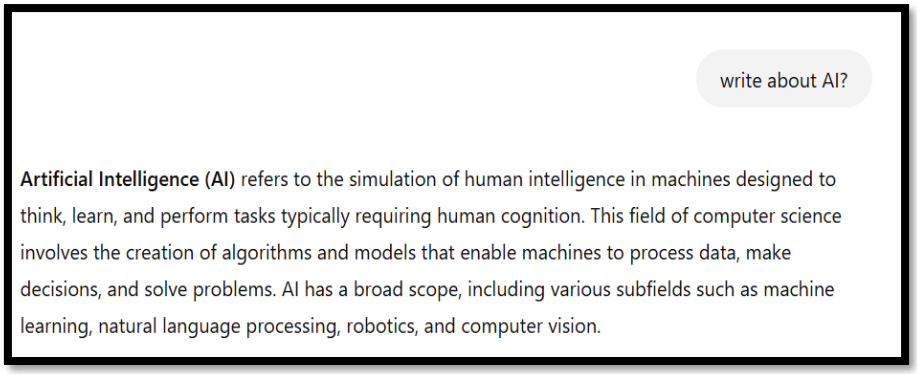


**Fig 1: Example Prompt with response**

In Fig 1 and Fig2, prompts are showing the relevant responses and in Fig 2 prompt we can very well see the expected relevant result in a good manner. We can find the major difference in responses between the prompt 1 in Fig 1 and prompt 1 in Fig 2. In Fig 1 is like a vague prompt which is used to extract common content as a response and in Fig 2 is considered as a detailed prompt which produces exact results. Here we can find the importance of prompt engineering and the prompting strategies to extract the relevant responses by creating effective prompts.

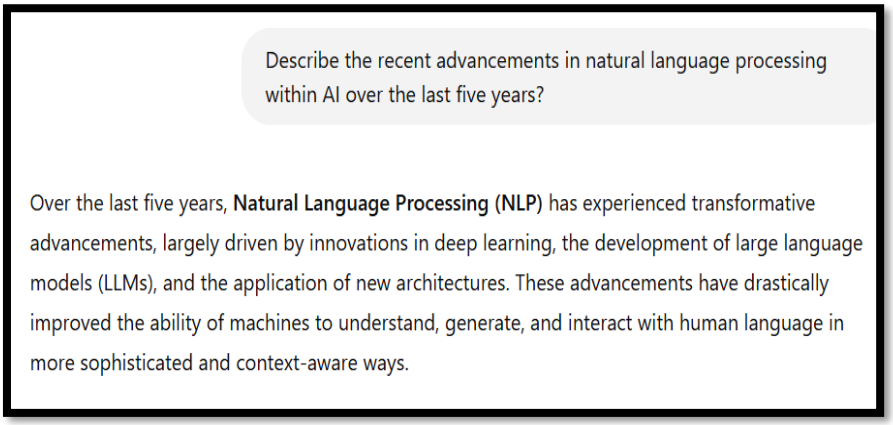


**Fig 2: Detailed prompt with response**

Prompt engineering connects the art and the technology for combining the creativity and technical skills for providing answers in a step-by-step manner. Designing an effective prompt for extracting the reliable and useful content from various Large Language Models is a challenging task. This work explores various prompt engineering techniques and strategies for improving the AI responses to specific purpose and raise the level of the quality of the result. Hence, prompt engineering acts as a bridge between AI and human.

## 2. Prompt engineering strategies

The terms such as prompt strategies, prompt patterns and prompt engineering techniques used to explain the various approaches and methods for designing an effective input to produce the relevant responses. This work explains the trivial and sneaking difference among these terms with example.

### 2.1 Prompt creation and Principles of prompt engineering

Prompt is a small text used to guide the Large Language Models for getting relevant responses. Everyone can create prompt and get responses from Artificial Intelligence models. There is no common and single procedure, blue-print or format for creating effective and efficient prompts [9]. But as an expert for retrieving relevant information from the AI models, we have to follow certain procedures and keywords for framing the prompts. Some of the important principles or key points of prompt engineering for creating effective prompts as follows:

- Correct and clear understanding of the task for which we are going to generate prompt
- Clear and correct language -i.e., keywords should be used
- Specific about the responses i.e., provide context is nothing but the contextual information
- If possible, try to attach some examples in the prompt
- Divide the complex tasks into manageable chunks so that the model address each and every part of the chunk sequentially. Adequate response for each and every part is important.

Prompts can be created by using all the principles of prompt engineering. Generally, prompts can be written without using or implementing the principles and we can get the response as a result. The relevant responses are created while using the principles of prompting. We have to write a prompt for getting the code to develop a webpage. The Fig: 3 shows the prompt with bad prompt and good prompt. Without giving clear direction and using the key words, bad prompt was created. When we attach certain keywords such as HTML, title, image, navigation menu and CSS, good prompt was generated.

**Example 1: Requesting code for creating a website**

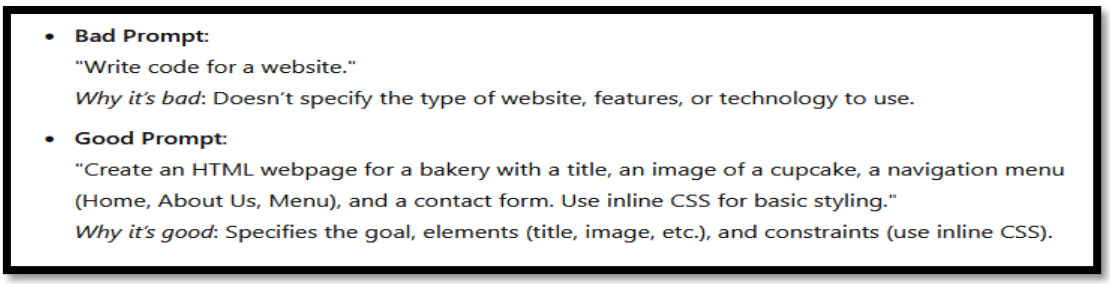


**Fig 3: Good and Bad Prompt with response**

In the above examples, certain things are missing in the bad prompts. They are the clarity, format, example and purpose. These bad prompts or poor prompts can be converted into good prompts by identifying the gaps, adding further details and focusing on the questions. Fig:4 depicts the example for focusing on specific aspect with good or bad prompts.

**Example 2: Requesting information about the technology from AI models**

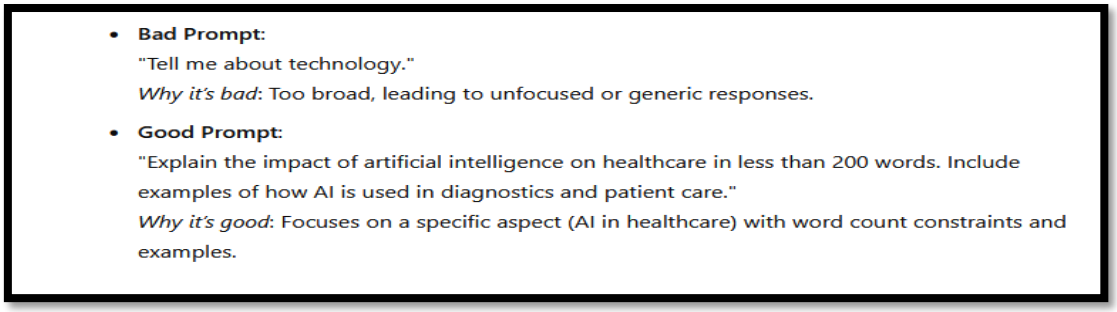


**Fig 4: Good and bad prompt with response-Focus on specific aspect**

The bad prompts can be improved by adding some specific keywords and details with the prompts.

### 2.2 Prompt Template

It is used to get the input from the user and translate it into instructions for a language model. In ChatGPT, prompt template used to format the prompt for getting the input which is reusable in constructing the model for creating the content. There are two important prompt templates such as String Prompt Template and Chat Prompt Template. The sample source code for using the prompt template.

```
from langchain_core. prompts import Prompt Template
prompt_template = PromptTemplate.from_template ("Tell me about {topic}")
prompt_template. invoke ({"topic": "cats"})
```

Here, in the above code dynamically the topic can be changed and collect the relevant response based on the request in the conversation. The prompt template can be used in structured conversation, content generation and automation. The input is adjusted dynamically when multiple users connected with the system. This prompt template can also be used to generate blog posts by gathering some special features.

### 2.3. Prompt Patterns

Prompt patterns are used to offer a structure which can be used for further applications to extract a specific responses in a particular format. It is mainly used to organize the responses in a specific format such as bullet lists or tables. The procedural pattern may be in the form of Question based pattern, comparative pattern and procedural patterns. The purpose of this pattern is to provide a clear reusable structure for various contexts.

**The general types of patterns are as follows:**

- Question-based Patterns
- Comparative Patterns
- Procedural Patterns

### 2.3.1 Question-based patterns

This is a specific pattern for designing direct and indirect questions for producing the relevant and focused responses. The important types of these patterns are Factual, Explanatory, Exploratory, Comparative, Problem solving, Opinion based and Reflective questions. Some of the example questions as prompt as follows:

- What are the strategies for improving the women's empowerment? : Problem solving
- How do I improve my students' concentration? : Reflective
- Which is better for the environment paper document or digital document? : Opinion
- How the renewable energy is important for climate change? : Explanatory
- What is the capital of India? : Factual

Hence, the question based patterns are focusing on structured responses in various domains with a customized way.

### 2.3.2 Comparative Patterns

This is the prompt structure used to analyse or evaluate the differences between the two things. This is mainly used for critical thinking and decision making. It is mainly focusing on the evaluation or comparison which can be utilized in almost all domains such as research, business, education and society. Some of the examples are as follows:

- What are the pros and cons of smart phones?
- Compare the benefits of online learning and distance learning?
- What are the difference between matriculation and central board syllabus?
- Compare the driving style of manual and automatic cars?
- Compare Himalaya products with Aravind herbal products?

  Hence, the comparative patterns are highly powerful in nature for evaluation and widely used for decision making in organizations.

### 2.3.3 Procedural Patterns

This is one of the prompting structure to get the responses in a step-by-step manner or process. This is mainly focusing on outcome in various domains such as education and profession. Some of the example prompts of procedural patterns are as follows:

- How to frame a linear function?
- How to install updates in laptops?
- How to run the antivirus software?
- Provide a step by step procedure to write and algorithm.
- How to register a course in NPTEL?
- How to pay exam fee through online?

Hence, the procedural patterns are used in education, technology, Business and Management for simplifying the tasks and remove the complexity and ambiguity.

### 2.4 Prompting Techniques

Prompt Engineering used to design the prompts to get better results from various large language models.The prompt engineering techniques help us to enrich the results in a proper way. Some of the common prompt engineering techniques are listed below.

#### 2.4.1 Zero shot prompting

Without containing any examples and demonstrations creating a prompt used to interact with the large language models. The Fig:5 shows an example prompt for zero shot prompting technique with expected result.


**Fig 5: Zero shot prompting**

#### 2.4.2 One-shot prompting

The one shot prompting gives only one example to the model for getting the response. One clear example to guide the model for getting the response.The Fig:6 depicts the example for one shot prompting. In the example, a single poem is a key word for providing the example of the desired response.

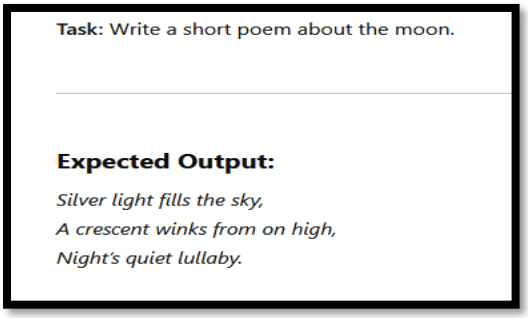


**Fig 6: One-shot prompting**

#### 2.4.3 Few-shot prompting

It adds the model with addition information in the form of examples to collect the response. In this example, prompt direction is also given for producing expected results. In this example, the write a haiku - here the word haiku has a specific

structure ie., three lines with the syllable pattern is 5-7-5.Clarity in this prompt reduces the ambiguity in the task.

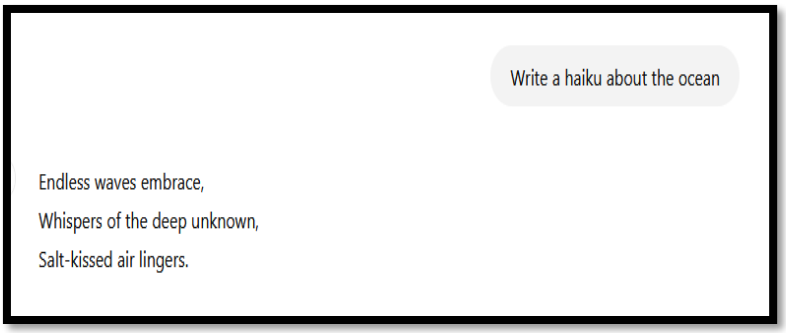


**Fig: 7 Few Shot Prompting**

Hence, these three techniques are considered as common techniques for raising the performance and efficiency of the response. Apart from this , there are so many techniques are available for getting the better response. Those techniques are considered as advanced prompt engineering techniques.

### 2.5 Role based Prompting

This Role based prompting also be called as persona prompting. In Role-play the key word “as a” attached with the prompt to assign some expertise to the prompt. This will focus on the response to yield better results especially to produce the consistent response.

**Examples:**

- As a NLP researcher, explain how LLM can enhance the Question and Answering problems?
- As a teacher, explain the concept of Linear regression with real life examples?
- As an Actor, explain the role of Maveeran Karnan with some stories.

In these examples, the perspective or role ie., As a …, can extract the background information to shape the responses from the AI models.

### 2.6 Instructional prompts

To get the clear and structured response from the AI models, direct instruction can be provided with the prompts.

**Examples:**

- Step by step procedure to implement RMI concept in Java.
- List top five Large Language Models for text summarization.

Here, the key words step-by-step and list are used to show the direction for getting the responses effectively in a structured way. When we use list option to get response from the model instead of bulletin, it adds the hierarchy when producing the results. The top priority answers are numbered with one, then two, then three and so on. When we use

bulletin, no need to check the priority, it will list out all the related information without specifying the order. Hence, the prompting strategies used to frame prompts by using the broader approaches.

### 2.7 LLM Settings

Effective responses can be generated by assigning values to the parameters in LLM. Some of the parameters were discussed with example. The Open AI Playground is an interface designed by Open AI which gives access to various AI models such as GPT3, GPT4. It provides an environment to the user for communicating various language models. Here, in Playground we can set the values for Temperature and max tokens easily.

#### 2.7.1 Temperature

Setting of temperature played a critical role in the generation of responses. The temperature value ranges from 0 to 1 and fixing the low value leads to the more deterministic response and setting the high value gives more results in terms of creativity and very less predictable results.

#### 2.7.2 Top – P

It is mainly used to manage the randomness in probability to produce the better response. In auto-texting, while typing words, automatically it shows the suggestion for next word. For example, Her dance performance is really _____. The top six words might be suggested in the list. Let us assume the probabilities of the top 6 words as follows:
good-probability 0.4,
awesome-probability 0.25,
nice - probability 0.15,
great- probability 0.10,
well- probability 0.06 and
wonderful - probability 0.04

Let us assign the Top-P value as 0.90. The model will add the probabilities until it reaches 90%. In the above example, when we add the top four words' probabilities we would get 90% (0.4+0.25+0.15 + 0.10 = 0.90) .The likelihood of top four words would be considered by the model. ie., Any one of the words good, awesome, nice and great would be suggested by the AI model. Hence, the cumulative probability will be considered to choose the word.

#### 2.7.3 Maximum Length

The maximum number of words or tokens would be generated as a response of the model to avoid the irrelevant responses.

### 3. Advance Prompt Engineering Techniques

The advanced techniques allow users to enhance and enrich the performance of the responses for effective prompting from AI models. This section explores the well-

designed prompts to improve the AI responses. Some of the advanced prompting techniques are explained with example as follows:

### 3.1 Chain of Thought (CoT)

This technique is used to expand the responses of large language models using the standard prompt format with reasoning. The speciality of this CoT technique is without fine tuning the model improving the efficiency of the response using normal prompt by inserting the reason. The main application area of this model is to solve mathematical problems. According to the research of Wei et.al., CoT prompting can enhance the LLM performance and accuracy of the response in the tasks such as common-sense analysis, reasoning and arithmetic problems [10]. The Fig: 8 and Fig:9 illustrate the example of the chain of Thought prompting.

**Prompt:** Mohan has started from Chennai to Bangalore in Car at 6.00 am. Normal speed of the car is 70 files per hour. How long it will take to reach Bangalore?

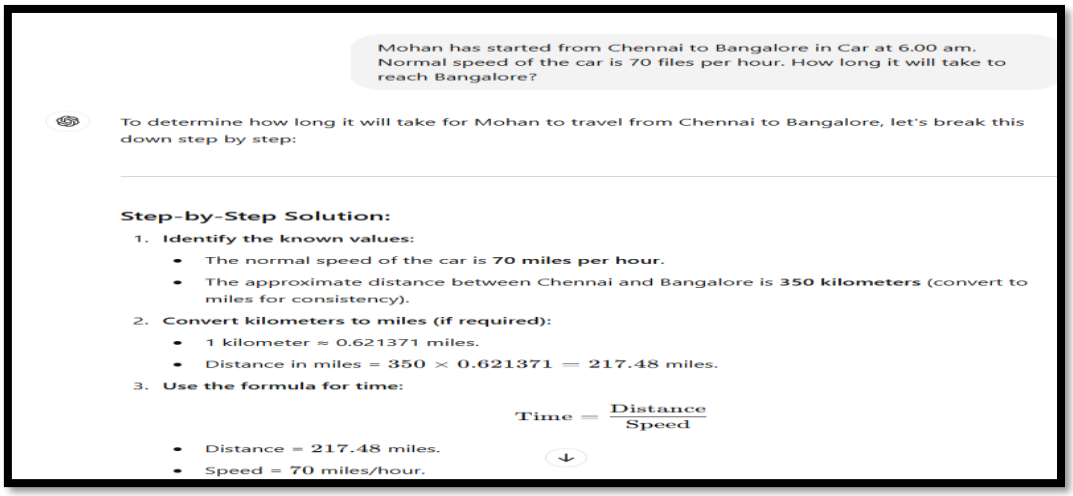


**Fig 8: Chain of Thought Example Prompt**

The CoT prompt breaks down the reasoning process into clear logical steps for reducing the error and complexity of the problem.

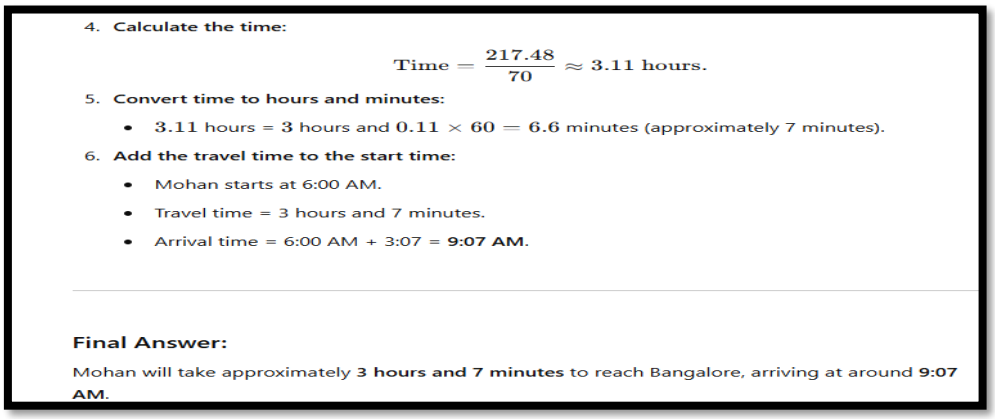


**Fig 9: Chain of Thought Example Prompt with Response**

The final step was clearly displayed in the Fig:9 with the final result. Hence, generating a Chain of Thought prompt with intermediate reasoning steps to provide the absolute answer with highest accuracy.

### 3.1.1 Zero-shot CoT Prompting

This approach encourages to solve the problem in a step-by-step manner with clear implementation of reasoning procedure. Let us think step by step is the key in this approach.

### 3.1.2 Few-shot CoT Prompting

In this One shot and Few-shot prompting, the model is provided with some examples with their step-by-step solutions to guide its reasoning process. This few shot CoT prompting can be combined with other techniques such as RAG for enhancing model's ability to acquire better responses.

### 3.1.3 Golden Chain-of-Thought

It is used to connect the intermediate steps to get the final response. It is mainly used to solve complex mathematical problems. Here, each and every step is clearly explained and ensures the correctness. Del et.al.,[11] started this innovative approach for generating responses using instruction-based queries. According to Del et.al.,[11] report, GPT-4 shows 45% high in the performance to solve the puzzles effectively when compared with normal Chain of Thought technique. Hence the overall rate of solving the puzzles using golden chain of thought is 83%.

## 3.2 Self-consistency

It is a method to check the logical coherence of the reasoning in each and every intermediate steps and leads to the correct conclusion of the final response. It can be used in problem solving and reasoning, theoretical frameworks, mathematical reasoning to produce better responses. Ontology-Generated Question-Answering (PrOntoQA) focuses on effective usage of structured representations of question-answering systems [12][13]. Self-consistency is a method in LLMs to solve complex reasoning problems easily. This method contains the following steps to solve the problem.

- Prompt is created first using Chain of Thought prompting technique.
- Reasoning paths are generated by using decoder
- Choosing the most consistent answer

Superior level of consistency is selected among the various reasoning paths and that could be the accurate solution. Shum et.al.,[14] and Bender et.al.,[15] discussed the utilization of self-consistency in producing better outcome to solve the problems such as arithmetic and symbolic reasoning-based tasks.

The combination of Self-consistency with other techniques to produce better response for doing further enhancement of the model. In Fig :10 shows an example for self-consistency.

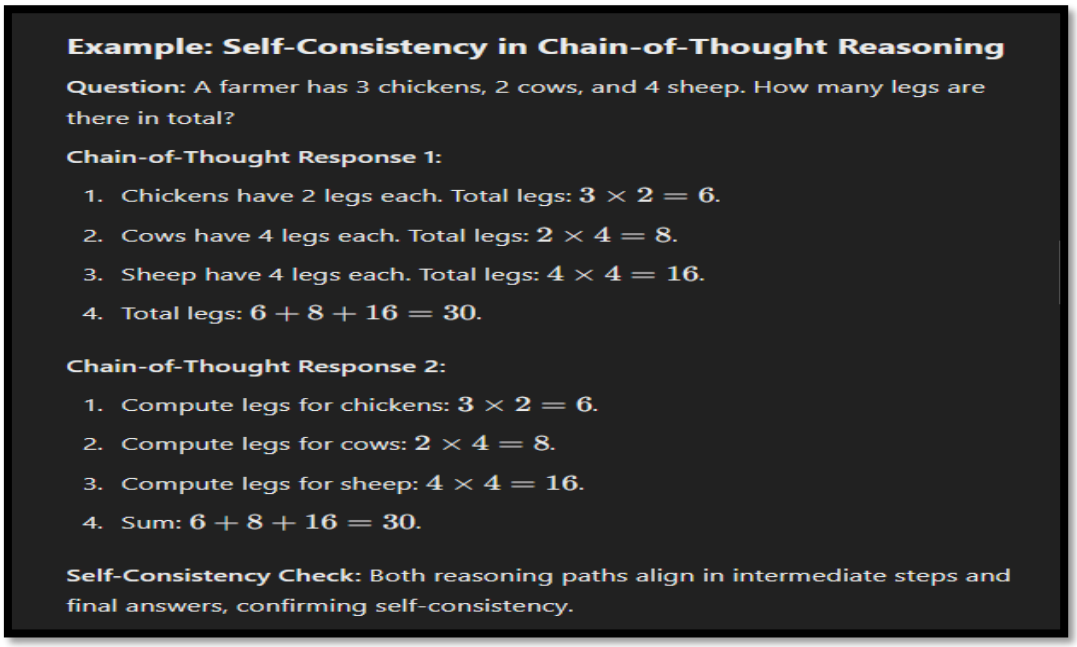


**Fig 10: Example for Self-Consistency**

### 3.3 Least-to-most prompting

This is an advanced method of prompting in which the complex problems are decomposed into a series of simple smaller sub problems, then the sub-problems are solved sequentially[17]. The step-by-step process of solving sub problems makes the task easier and the complex problems are solved efficiently.

Ex: Find the value of x in the equation: 10x +6 = 56

**Normal way to solve the problem is as follows:**

10x +6 = 56
10x = 56- 6
10x = 50
x = 50/10
x = 5

**It can be guided as follows:**

- To calculate x, we need to isolate x from the given equation.
- Subtract 6 from both side of the given equation

  10x+6 -6 = 56 - 6
  10x = 50
- Divide by 10

  10x/10 = 50/10
  x = 5

Therefore, the value of x is 5

The main advantage of using this approach is to encourage the subtasks to do the calculation. If it struggles or it needs assistance then the assistance can be provided. While getting solution for each and every sub problems solution that can be updated with more confidence to improve the critical thinking.

### 3.4 Tree of Thought (ToT)

The Tree of Thought prompting is an important milestone in LLM advancement models. In this approach, various possibilities of problem-solving techniques and possible solutions are analysed and the better suitable solution is taken to optimize the solution [17]. The organization of ideas in a hierarchical structure with the root to represent the main idea and the branches used to represent the different paths to reach the good result. It breaks down the complex tasks into manageable components and helps to refine and optimize the prompts for specific result. The Fig : 11 shows the example of Tree of Thought.

An optimized tour plan for Turkey, covering key cities and attractions, with various modes of transportation and approximate travel times. This plan is designed for 7-10 days, balancing travel and sightseeing. Fig: 12.a Shows the Turkey tour plan in stand prompting,Fig:12.b depicts it in CoT and Fig:12.c illustrates the Turkey tour plan using ToT.

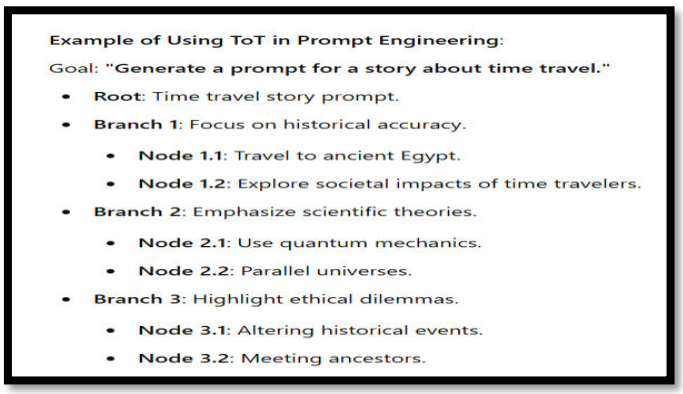


**Fig : 11 Tree of Thought (ToT) prompting**

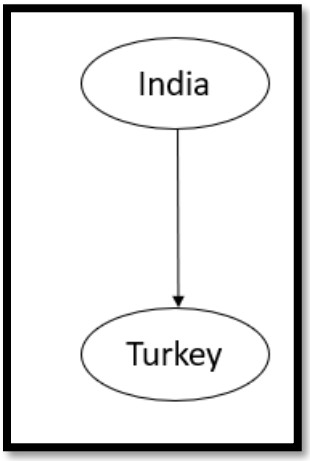


Fig: 12 (a)

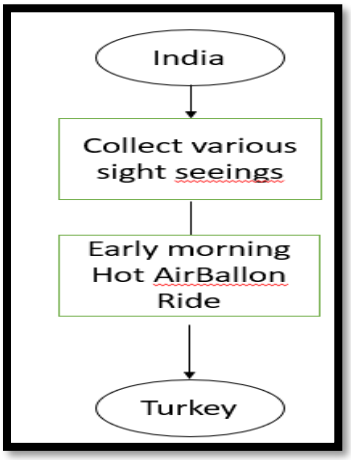


Fig: 12.b

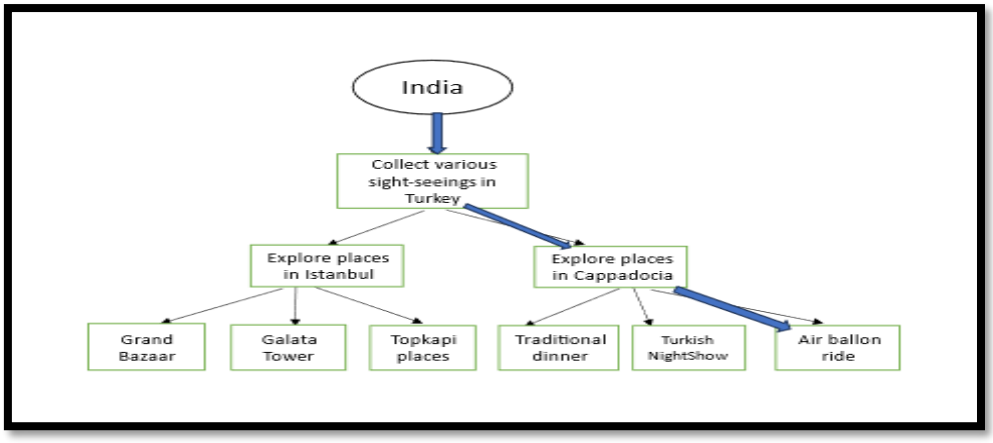


**Fig: 12. c**

**Fig: 12. a:** Standard Prompt 12.b: Chain of Thought(CoT) Fig:12.c: Tree of Thought(ToT)

### 3.5 Graph based Prompting

It is structured approach in which the nodes and the edges are connected to represent the path. Sometimes we may have multiple paths, loops and complex structures of non-linear structure. It is mainly used to find the path by using the connections and the relationships can also be mapped.

The structured reasoning can be done by using both techniques and we can very well breakdown the structure of the complex problems into simple components and also used for making right decision at right time by organizing the important information effectively. The difference between the ToT and Graph based prompting is tabulated in Table : 1 as follows.

**Table 1: Difference between Graph based and ToT Prompting Techniques**

| Sl. No | Graph-Based Prompting | Tree of Thought(ToT) |
|---|---|---|
| 1 | Non linear structure is followed | Hierarchical structure is followed |
| 2 | Loops are allowed | No loops |
| 3 | We can find multiple paths | Need to follow the branches |
| 4 | Bi-directions are permitted | Only Linear and branching |
| 5 | Cyclic connections can be generated | No cycles are created |
| 6 | Focusing on mainly optimization of the response | Focused on the systematic approach |
| 7 | Mapping of relationships to find the best path | Step by step reasoning is done |

### 3.6 Re Act

The ReAct comes in advanced prompting technique in which it combines the reasoning and acting in artificial intelligence systems to integrate the entire process with real

world applications. It stands for Reasoning and Acting which is used to generate reasons and track them to perform task specific actions. The first step of generating the reasons of thoughts and then communicate with other software tools and databases for making proper interaction to decide the action so that LLMs can do the tasks efficiently by following the multistep process [19]. React aids Large Language Models for taking accurate decisions in various domains such as healthcare industries, legal issues and finance by combining the data and reasoning [20]. Generally, ReAct as a framework used to integrate other tools to make the decision effectively [21]. A new prompt is proposed by Zhang C et.al.,[18] with explicit explainable ability called Tree Prompt. In this approach he converted the complex sentences into a tree with reasoning. Using the bottom-up manner then the prompts are generated. Here, in this entire process, the intermediate nodes allow us to know the reasoning process.

The ReAct combines the reasoning and acting for solving complex tasks which can make use of Memory for storing the action of the previous ones and also interact with other external tools to retrieve the information from databases to execute real time applications. The ReAct is compared with standard prompting method and the important points are tabulated as follows:

**Table 2: Difference between Standard Prompting and ReAct Prompting**

| Sl.No | Standard Prompting | ReAct Prompting |
|---|---|---|
| 1 | This is a single step method | Multistep reasoning and acting are combined |
| 2 | It suits for straightforward queries | It is used to iterative, complex situations |
| 3 | Provides static responses | Actions are taken based on the intermediate reasons |
| 4 | Operates in the closed environment | Able to interact and communicate with external tools and systems |

The Fig : 13 gives the clear picture of standard prompting and ReAct prompting through example.

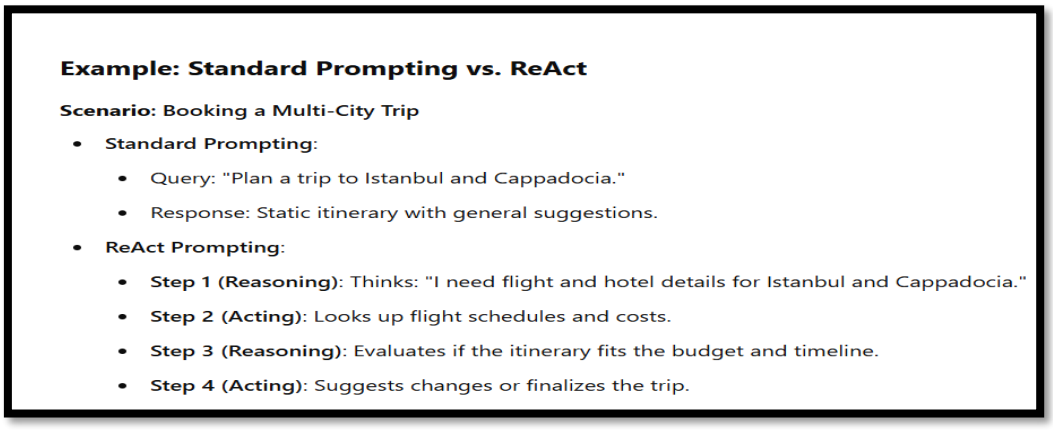


**Fig 13: Example for ReAct Prompting**

Hence, the prompting techniques can be fine-tuned to optimize the responses from the AI models.

## 4. Evaluation of AI Responses through Prompt Engineering

Evaluation of AI responses can be assessed by the success of effective prompts crafting through the results produced and refined for better output. Evaluation played a foremost role for producing effective prompts so that the final outcome could be a better one for continuous improvement.

### 4.1 Evaluation Metrics

Generally, Metrics are used to evaluate the performance of the prompts. The evaluation metrics are classified as Quantitative and Qualitative metrics.

#### 4.1.1 Qualitative Metrics

It is used to evaluate the subjective nature of the responses and mainly focusing on the quality, coherence and the relevancy of the output. **Relevant** is the measurement of the response and verify how the output match with the expected content of the prompt. Sometimes, if we are using simple standard prompting technique, we may get response. But we need to check the relevancy of the response. Suppose, we are using some advanced technique for crafting the prompts, we definitely get the response which may have 90% accuracy. Hence the relevant information only concentrated will produce more accurate result. **Coherence** is also considered as a metric for measuring the logical flow and consistency of the responses. Here, in the response, the logical connection of sentences is measured. **Completeness** can also be measured by verifying the responses of the prompt in all aspects. **Depth** is also considered as an important qualitative metrics to measure the richness of the response. **Fluency** is used to measure the grammar of the response in terms of spelling, meaning and readability with structure. The creativity and novelty of the responses are measured by **Originality.**

#### 4.1.2 Quantitative Metrics

This is used only for measurable information of AI generated responses. It is for evaluating the performance of the dataset, compare the results of various techniques and the consistent results of the model. **Accuracy** is measured for calculating the correctness of the response. Generally used to check the reality of the tasks such as classification, knowledge retrieval and question and answering. **Efficiency** can also be measured by calculating the response time or token usage. The measurement units such as **Precision, Recall and F1 Score** are used to check the relevancy of the results. The **BLEU** score, **ROUGE** score used to measure the quality of the text in task specific applications such as text summarization and text or content generation. The **Perplexity** measures the fluency and predictability of the response. The **Diversity** can be measured the variability of responses generated from a given prompt.

### 4.2 Evaluation Methods

Evaluation involves manual, automatic and hybrid methods. The manual evaluation suits to assess the qualitative aspects of the AI response. The Automated Evaluation is meant for quantitative metrics such as ROUGH, BLEU for task specific applications. The hybrid evaluation is the combination of manual and automated evaluation for getting balanced responses. Human can evaluate certain things manually like feedback collection for iteratively refining the prompts.

#### 4.2.1. Steps to Evaluate AI Responses

AI responses are evaluated by using the following steps:

- Define the result of the prompt
- Generation of responses using various techniques such as CoT, ToT and ReAct.
- Do the evaluation using quantitative and qualitative metrics
- Analyse the performance of the prompt
- Refine the prompt

Using the Annotation tool, the qualitative scoring are calculated and using the tools for evaluating metrics such as BLEU, ROUGH in hugging face, NLTK/Spacy packages can also be used. Using OpenAI Evaluation API, the fluency, relevance and coherence scores are calculated.

### 5. Use cases

The use cases refer to the specific task and domain to achieve the expected result. In Large Language Models the use cases are used to design a structure of the AI response in a proper way by defining the clear task to enable the effective prompting. Some of the use cases can be applied for effecting prompting as follows:

| | |
|---|---|
| 1. Text summarization | 2. Translation |
| 3. Question Answering | 4. Academic Assistance |
| Content Creation/generation | 6. Code generation and |
| 7. Sentiment Analysis | . Customer support automation |
| 9. Story generation | 10. Education – for tutoring or learning support systems |
| Creative work – Story telling or content generation | 2. Product development- virtual assistant, chatbot |

**Some of the examples**

**Ex:**

- As a customer support executive, explain the delay in delivery.
- Write about your product focusing on sustainability.
- Suggest the best scheme for recharging the mobile phone.

### 5.1 Challenges and Limitation

The main challenges for creating effective prompting to collect the AI responses as follows:

- **Ambiguity:** The prompts may not produce relevant information as a response.
- **Consistency:** Maintaining consistency in the intermediate responses can be difficult.
- **Managing Bias:** Due to bias in the dataset. the result also contains the bias

### 5.2 Comparison of Various prompting techniques

Taking some tasks various prompting techniques are applied and the performance are evaluated.

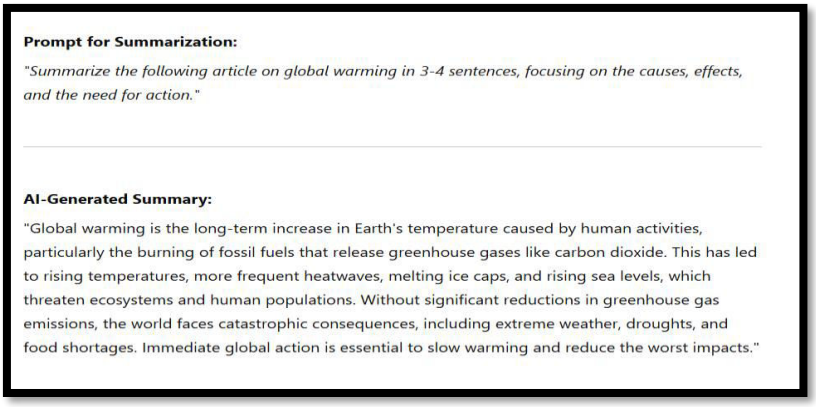


**Fig 14: Text Summarization example**

**Task 1:** Text Summarization is the task in which by taking lengthy document or content and generate a precise or concise text without losing the important information.

The various prompting techniques such as standard prompting, Chain of Thought, Tree of Thought and ReAct are applied to summarize the task and the techniques are evaluated. The results are tabulated as follows: The Fig: 14 example for text summarization is given. The performance evaluation of text summarization task is tabulated. According to Table 3, the Tree of Thought (ToT) techniques show the better results.

**Table : 3 Text Summarization Performance Evaluation**

| Techniques | Relevant | Coherent | Accuracy | Time(ms) | Tokens used |
|---|---|---|---|---|---|
| Standard Prompting | 7.7 | 7.9 | 84% | 100 | 150 |
| Chain of Thought | 9.0 | 9.1 | 90% | 100 | 250 |
| Tree of Thought | 9.4 | 97 | 93% | 120 | 200 |
| ReAct | 9.1 | 9.2 | 89% | 125 | 250 |

**Task 2: Solving the puzzles**

Finding the solution for a puzzle is not only challenging job but also time consuming one. There are different types of puzzles such as sudoku, word puzzles and logical puzzles (Tower of Hanoi..)In this task, applying various advanced techniques for prompt generation and the performance of all the techniques are analysed and the values are tabulated in Table: 4.


**(a)** **(b)**

**Fig: 15(a) Puzzle & 15(b) Solution**

Generally, puzzles can be solved by using ToT technique. Here, the multiple solutions are generated to analyze the various possibilities and evaluate each and every path. By focusing on the solutions fewer promising paths are deleted and the entire process is repeated to find out the optimal solution.

**Table 4: Performance analyses of Puzzle problems**

| Type of Puzzle | Standard Prompt | Chain of Thought (CoT) | Tree of Thought (ToT) |
|---|---|---|---|
| Sudoku | 63% | 77% | 88% |
| Logical Puzzles | 54% | 76% | 88% |
| Tower of Hanoi | 53% | 72% | 84% |

The evaluation metric can also be calculated by using the score. For Example to get the customer feedback about a product the range is assigned to each and every qualitative metric and the ranges are tabulated as in the Table 5.

**Table 5: Evaluation**

| Metric | Score(1-5) | Comments |
|---|---|---|
| **Relevance** | 5 | Directly addresses causes of climate change. |
| Coherence | 4 | Logical flow good |
| Completeness | 5 | Main points are covered |
| Depth | 3 | General information |

Each and every prompting technique has its own advantages and has some special characteristics. In the text summarization problem, all type of prompting techniques produces better results, the standard prompting technique shows the good results and

to solve the puzzle problems, the Tree of Thought (ToT) produces better results and the Maths problems the chain of Thought (CoT) performs good. For making proper interaction, ReAct will produce better results.

Hence, each and every prompting technique can perform well and the final results are enhanced by choosing the right technique to solve the right problem for getting the best performance.

## 6. Conclusion and Future directions

Improving the productivity and quality of the prompts by enhancing the AI responses through the effective crafting of prompts. The prompt engineering played a major role for maximizing the performance and utilization of large language models. Applying all the principles of prompting for crafting an effective prompt is the foremost requirement of prompt engineering.

Crafting of prompts effectively is an important one for maximizing the utilization of Large Language Models. Enhancing AI responses is about crafting prompts that align with the desired output, from specific, structured answers to more open-ended creative content.

The prompt engineering field gives scope for many advancement in Artificial Intelligence field and the focus of the research can be the optimization of automatic prompt and dynamic prompting techniques. By taking the previous values continue the task to finish the work is the need of an hour. Refining it and redefining the prompts with the expected outcome in a concise way.

## References

1. Vaswani, A. "Attention is all you need." Advances in Neural Information Processing Systems (2017).
2. Trinh, Trieu H., and Quoc V. Le. "A simple method for commonsense reasoning." arXiv preprint arXiv: 1806.02847 (2018).
3. Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova,(2019),” BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, arXiv:1810.04805v2 [cs.CL] , 2019
4. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. ”Improving Language Understanding by Generative Pre-Training”,OpenAI Blog, (2018)
5. GPT-4 Technical Report (2023),OpenAI Blog
6. Brownlee. J Ä Gentle Introduction to Large Language Models (LLMs), Machine Learning Mastery, .(2021).,
7. Bommasani, R., Hudson, D. A., Adeli, E., Ali, R., Artzi, Y., Bender, E. M., ... & Zhang, B. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?”, Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. (2021).

8. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shinn, A., & others. (2020). Language Models are Few-Shot Learners. arXiv.
9. Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., … Resnik, P. (2024). The Prompt Report: A Systematic Survey of Prompting Techniques. arxiv.org.
10. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models.
11. Del M, Fishel M. True detective: a deep abductive reasoning benchmark undoable for GPT-3 and challenging for GPT-4. In: Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023); 2023.p. 314–322
12. [12] Saparov A, He H. Language models are greedy reasoners: a systematic formal analysis of chain-of-thought; 2022. ArXiv:2210.01240.
13. Tafjord O, Dalvi B, Clark P. Proof Writer: generating implications, proofs, and abductive statements over natural language. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021; 2021. p. 3621–3634.
14. Shum K, Diao S, Zhang T. Automatic prompt augmentation and selection with chain-of-thought from labeled data; 2023. ArXiv:2302.12822.
15. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; 2021. p. 610–623.
16. Zhou D, Sch¨arli N, Hou L, Wei J, Scales N, Wang X, et al. Least-to-most prompting enables complex reasoning in Large language models. In: Eleventh International Conference on Learning Representations; 2023.
17. Yao S, Yu D, Zhao J, Shafran I, Griffiths TL, Cao Y, et al. Tree of thoughts: deliberate problem solving with large language models; 2023. ArXiv:2305.10601.
18. Zhang C, Xiao J, Chen L, Shao J, Chen L. TreePrompt: Learning to Compose Tree Prompts for Explainable Visual Grounding; 2023. ArXiv:2305.11497
19. Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan K, et al. ReAct: synergizing reasoning and acting in language models; 2023. ArXiv:2210.03629.
20. Li A.: ReAct: a new framework for prompt engineering in large language models. Available from: www.perxeive.com.
21. Roberts A.: How to ReAct to simple AI agents. Available from: arize.com.
22. Joe Davison, Joshua Feldman, and Alexander M Rush. 2019. Commonsense knowledge mining from pre trained models. In Empirical Methods in Natural Language Processing (EMNLP).
23. Fabio Petroni, Tim Rockt¨ aschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowl edge bases? In Empirical Methods in Natural Lan guage Processing (EMNLP).