Tehran Institute for Advanced Studies
(TeIAS)

Department of Data Science

NLP Project

In Context Learning And Prompting Methods

Mohammad Azimi
Mahdi Masoon

July 22, 2023

Abstract

In this project report, we explore the role of exemplar selection and prompt designing in enhancing the performance of large language models (LLMs) on natural language inference (NLI) task on the ANLI dataset. We propose four different methods for selecting exemplars from the training data, namely random selection, clustering, BERTScore, and Maximum Marginal Relevance (MMR). We also design three types of prompts for providing input to the LLMs, namely standard prompt, natural language description prompt, and algorithmic prompt. We evaluate the text-davinci-002 model as the baseline LLM and report the accuracy and f1-score obtained using various combinations of exemplar selection and prompt designing methods. We also analyze the errors made by the model and suggest some strategies to improve the prompt design. We conclude that the clustering method with natural language description prompt achieves the best performance among the other methods and demonstrates the importance of refining exemplar selection and prompt designing techniques for NLI task.

Table of Contents

Chapter 1

Introduction

The field of natural language processing has witnessed remarkable advancements with the emergence of Large Language Model (LLM) prompting. These prompting methods have demonstrated exceptional capabilities in tackling complex tasks, particularly when multi-step reasoning is involved [1]. However, the success of LLMs heavily relies on the selection of exemplars and the design of prompts during the in-context learning (ICL) process [1] [2] [3]. Consequently, the importance of refining these aspects can not be overstated.

In this project, we aim to explore the critical role of prompt designing and exemplar selection in enhancing the performance of LLMs on natural language inference (NLI) task on the ANLI dataset [4]. By implementing different prompt designing and in-context learning methods, we aim to gain valuable insights into the impact of each approach on NLI task. Additionally, we endeavor to introduce innovative modifications to some of these methods. Our project involves three essential steps, each contributing significantly to the improvement of the natural language inference (NLI) task. In The first step, we focused on selecting the most suitable exemplars for the in-context learning process, followed by the critical task of prompt designing in the second step. Although the third step involves analyzing the errors that the model made and trying to find out why the model made these mistakes.

Throughout our project, we encountered a few limitations that influenced our approach. The most significant constraint was the cost associated with prompting in LLMs. Additionally, our training dataset was exceptionally vast, which imposed practical constraints on applying certain exemplar selection methods across the entire dataset. Despite these limitations, we made diligent efforts to explore different methods which we will discuss in chapter 3 and chapter 4.

Chapter 2

Background

2.1 In-Context Learning:

One of the aspects that our project focused on is the usage of explanations in in-context learning (ICL). Let q be the test query to solve. The standard ICL prompts a language model, M, with a set of exemplar input-output pairs, $(q1, a1)...(qm, am)$, and predict an answer $\hat{a}$ for the query:

$$\hat{a} = \arg\max_{a} p(M(a|q, \{(q_1, a_1), \ldots, (q_m, a_m)\}))[5] \tag{2.1}$$

In addition to just input-output pairs, we can also include explanations (following scratchpad [6] or chain-of-thought [1]) in prompts, which leads the LLM to generate explanations for its predictions as well:

$$\hat{a} = \arg\max_{a} \sum_{e} p(M(a, e|q, C))[5] \tag{2.2}$$

where $C = \{(q_1, e_1, a_1), \ldots, (q_m, e_m, a_m)\}$ is the set of input-explanation-output triplets in prompts.

The end task performance of ICL is sensitive to the selected exemplars [7]. While much prior work uses a fixed set of manually selected exemplars [1], there is also work devoted to studying how to select more effective exemplars from a pool of exemplars. Given a test query q, the task is to select a set of m exemplars from a pool of n exemplars D = (q1, e1, a1)...(qn, en, an) to construct a prompt for solving q. We note that this yields varying exemplar sets for different queries.

2.2 Designing Prompts:

Another aspect we focused on is prompt designing, which refers to the process of crafting specific instructions or queries, known as prompts, that are presented to the language model to elicit the desired responses. These prompts act as the input for the model, guiding its understanding and generating the appropriate output. Well-designed prompts are crucial for obtaining accurate and relevant results from the language model.

Key considerations for prompt designing include:

Clarity and Precision: Prompts should be unambiguous, avoiding any potential confusion. They must precisely convey the desired task or information the model is expected to provide.

Contextualization: Providing relevant context within the prompt helps the language model better understand the user's intent and produce more context-aware responses.

Prompt Length: The length of the prompt should be appropriate for the specific use case. Longer prompts may be necessary for complex tasks, but excessively long prompts might overwhelm the model or lead to incomplete responses.

Iterative Design Process: Prompt designing often involves an iterative approach. You may need to experiment with different prompts, assess the model's outputs, and refine the prompts based on the results.

For this purpose, we extensively explored various prompt designing methods, drawing inspiration from the approaches presented in [8] and [5] for our project. By studying these methods, we aimed to enhance the effectiveness and precision of our prompts.

2.3 Task, Dataset and Large Language Models:

Natural language inference is a complex task with a broader space of decisions on how to prompt LLMs. We worked with the three-way classification formulation of the task. Each input in the dataset is a pair of contexts (the premise and the hypothesis): the task is to predict whether the hypothesis is entailed (i.e. always true), contradicted (i.e. always false), or neutral (neither entailed nor contradicted) given the premise. We use the ANLI dataset [4] Two examples from the dataset are shown in Figure 1. ANLI was constructed through an adversarial and iterative data collection process: simply put, the examples in ANLI are quite challenging by design. We use Round 3 of ANLI: in the GPT-3 paper[3], the authors report that all of their models except their largest (175B parameter) model achieve chance accuracy (33%) on this version of the dataset. Currently, on the ANLI leaderboard1, the performance reported for GPT-3 is 40.2 and the state-of-the-art is 53.4. We mention this because it was fairly challenging to devise good strategies for achieving high performance.

In our experiments, we aimed to explore various large language models, but due to cost limitations, we could only examine the text-davinci-002 model, which is an enhanced version of the GPT-3 (davinci) model. Regrettably, specific details about text-davinci-002's architecture or improvements were not provided. However, it is plausible that text-davinci-002 represents an updated and refined version of the original GPT-3 (davinci) model, potentially offering improved performance and capabilities.

Chapter 3

What Makes a Good Exemplar Set?

One of the objectives of this project is to address the question of how to effectively select good exemplars from the training data. To achieve this, we have established specific rules and criteria for the exemplar selection process. Our dataset contains three distinct classes: neutral, entailment, and contradiction. We strive to include at least one exemplar from each class to ensure comprehensive representation. Additionally, to prevent any biases toward a particular class, an equal number of exemplars will be selected from each class (three in total)[9].

Exemplar Selection Methods:

- **Random Selection (Baseline Method):** As the initial approach and baseline, we employed a random selection method. This involved randomly choosing a few exemplars from our training data and then applying the above rules.

- **Clustering Method:** In this method, we utilized the k-means algorithm to cluster the entire training data. To perform the clustering, we employed the language model RoBERTa-Large to embed all the training data. The k-means algorithm then grouped the data points based on their embeddings into distinct clusters. we repeated this procedure on test data also and we observed that the distribution and clusters are almost the same in both test and training data. From these clusters (in training data), we selected the cluster centers and a few neighboring data points. By doing so, we ensured that our exemplar set encompassed exemplars from each cluster, representing various patterns in the data.

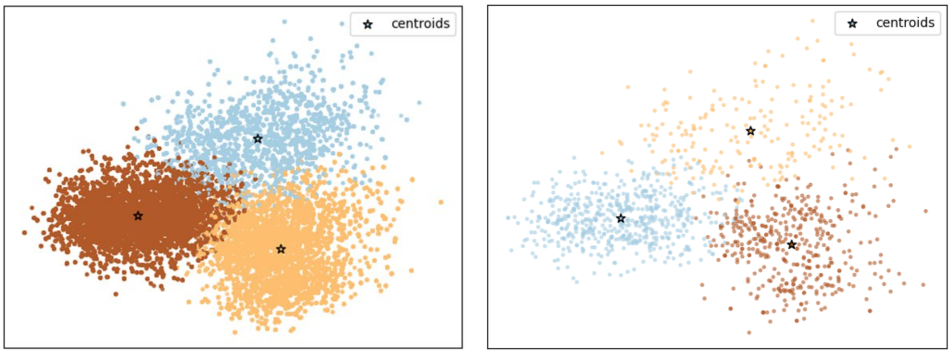


Figure 3.1: Clusters and distribution of Training data(Left) and test data(Right)

- **BERTScore Method:** The goal of this method was to select exemplars that exhibited the maximum similarity to the query set. To achieve this, the average Bert-Score between each training data point and the entire query set was computed. The top-k exemplars with the highest BERTScore were selected, with one exemplar chosen from each class. In the end, the exemplar set was constructed based on the BERTScore similarity metric, The backbone model for BERTScore is RoBERTa-Large [10].

- **MMR Method:** The fourth method employed in the study was the Maximum Marginal Relevance (MMR) based exemplar selection strategy[5]. The rationale behind this method is that a good exemplar set should not only consist of relevant exemplars but also collaboratively cover the reasoning skills required for solving the query. While the nearest-neighbor-based exemplar selection strategy considers only relevance between exemplars and the query, it may result in a set of mostly similar exemplars, potentially limiting collaboration. To address this, the researchers argued that complementarity should also be taken into account in the exemplar selection process, as a set of less similar exemplars is more likely to illustrate the required reasoning processes. Since determining categorical complementarity is challenging, diversity was used as a proxy. A set of less similar exemplars was considered more likely to exhibit complementarity. To implement this strategy, a maximal-marginal relevance-based exemplar selection approach was proposed. The idea behind this approach is to select exemplars that are relevant to the query while also being diverse enough to encourage collaboration. Suppose for the given query q, we have already selected a set of exemplars $T = q_i$, then we will pick up the next exemplar according to:

$$\arg \max_{q_j \in D/T} \lambda S(q, q_j) - (1-\lambda) \max_{q_i \in T} S(q_j, q_i) \tag{3.1}$$

  where S denotes similarity and λ is a parameter that controls the balance between relevance and diversity. We rely on MMR to iteratively select exemplars from the exemplar pool, as shown in Figure 3.2. Note that Running MMR requires scoring all exemplar pairs within the pool. To run inference over m queries using a pool of n exemplars, MMR requires scoring the similarity $nn + mn$ pairs.

Algorithm 1 MMR-Based Exemplar Selection

procedure MMRSELECT($D, q, k, \mathcal{S}$)
input: exemplar pool $D = \{q_i \ldots q_n\}$, test query q, number of shots m and similarity measurement $\mathcal{S}$
output: selected exemplars $T = \{q_i \ldots q_m\}$
  $\mathbb{S} := [[\mathcal{S}(q_i, q_j)]]_{q_i, q_j \in D}$; ▷ the pairwise similarity between exemplars in T
  $\mathbb{Q} := [\mathcal{S}(q, q_i)]_{q_i \in D}$; ▷ the similarity between query and exemplars in T
  $T := \{\}$;
  while $|T| < k$ **do**
    $\hat{q} :=$ Equation(1); ▷ get the next exemplar based on Eq 1
    T.add($\hat{q}$)
  return T;

Figure 3.2: Algorithm of MMR-based exemplar selection [5]

During the implementation of this method, an innovative idea emerged - employing a model specifically trained for this task, which holds significant potential for feature extraction. Through its training process, this model has honed its ability to extract essential features relevant to our objective. Consequently, we used microsoft/deberta-large-mnli [11] to generate embeddings, leading to exceptional outcomes. Additionally, to assess similarity, we adopted BERTScore as our preferred similarity function.

In conclusion, this methodology section details the four different exemplar selection methods employed to create a balanced exemplar set from the training data. The methods include random selection as a baseline, clustering using the k-means algorithm, BERTScore-based selection, and the Maximum Marginal Relevance (MMR) approach. Each method aims to achieve a balanced and diverse exemplar set that can effectively represent the reasoning skills required for solving the given queries. The process of selecting exemplars is vital to ensure the success of the project and to enable a comprehensive evaluation of the model's performance.

Chapter 4

Prompting Methods

In this section, we will explore the different types of prompts we designed manually for our project. These prompts are essential for guiding the model's behavior during the in-context learning phase. We have devised three primary categories of prompts: Standard Prompt, Natural Language Description (NL) Prompt, and Algorithmic Prompt [8]. Each type of prompt serves a specific purpose and influences how the model processes and responds to input data.

- **Standard Prompt:** The Standard Prompt represents a straightforward and routine approach to providing input to the model. Our designed standard prompt consists of a question, context, hypothesis, and a very simple answer that just state the relationship between context and hypothesis. The goal is to evaluate the model's response under this kind of prompt, where the prompt's structure is consistent across different test instances.

  Example:

> Q: What is the relationship between the following context and hypothesis?
> (I have three kinds of relationship: entailment, neutral, and contradiction)
>
> Context:
> John Sprunt Hill (March 17, 1869 – July 29, 1961) was a North Carolina lawyer,
> banker and philanthropist who played a fundamental role in the civic and
> social development of Durham, North Carolina,
> the expansion of the University of North Carolina at Chapel Hill and
> the development of rural credit unions in North Carolina,
> during the first half of the 20th Century.
>
> Hypothesis: Transparency is important to the board.
>
> A: The relationship based on the given context and hypothesis is a "contradiction".
>
> The label is "c"

Figure 4.1: Example of Standard Prompt

- **Natural Language Description (NL Description) Prompt:** The NL Prompt structure is the same as the standard prompt however in this method we provide a clear and detailed description of the relationship between context and hypothesis.

Example:

Q: What is the relationship between the following premise and hypothesis?
(I have three kinds of relationship: entailment, neutral, and contradiction)

Context:
Is there anything else that the board has done here in
connection with the curriculum change that provides a basis for your complaint?
From what I can see from attending the board meetings,
I don't know, because so much of the curriculum debate takes place at non-public meetings,
that I am not aware of.
And when board members are questioned at meetings,
they really don't answer any questions.
So it gives this whole illusion of secrecy to the process,
which, you know, maybe a big part of the problem.
It may just be a big misunderstanding.

Hypothesis: Transparency is important to the board.

A: The context suggests that there is a lack of transparency in the board's actions,
and decision-making process, as much of the curriculum debate takes place in non-public meetings
and board members do not provide satisfactory answers to questions.
This contradicts the hypothesis that transparency is important to the board.
Therefore, the relationship between the context and the hypothesis is a contradiction.

The label is "c"

Figure 4.2: Example of NL Description Prompt

- **Algorithmic Prompt**: The Algorithmic Prompt presents a structured and rule-based approach to formulate instructions. The method involves breaking down the context into several facts and establishing a hypothesis as the desired goal. Subsequently, we combine the hypothesis with each of the facts, checking for any contradictions or entailments. When none of the facts contradict or entail the goal, we determine that the relationship between the context and hypothesis is neutral. This systematic process allows for a more organized and logical way of constructing instructions, enhancing the clarity and effectiveness of the overall communication.

  Example:

Q: What is the relationship between the following premise and hypothesis?
(I have three kinds of relationship: entailment, neutral, and contradiction)

Context:
Is there anything else that the board has done here in
connection with the curriculum change that provides a basis for your complaint?
From what I can see from attending the board meetings,
I don't know, because so much of the curriculum debate takes place at non-public meetings,
that I am not aware of.
And when board members are questioned at meetings,
they really don't answer any questions.
So it gives this whole illusion of secrecy to the process,
which, you know, maybe a big part of the problem.
It may just be a big misunderstanding.

Hypothesis: Transparency is important to the board.

A: To understand the relationship between context and hypothesis,
we need to construct our knowledge base consisting of the following facts in the context:
Facts:
1- Much of the curriculum debate takes place at non-public meetings.
⋮
6- There is a possibility that the situation is a result of a big misunderstanding.
We use the backward algorithm, we'll start with the goal and
work backward through the facts to see if we can reach a conclusion:
Goal: Transparency is important to the board.
Explanation:
To determine if the goal is directly entailed or
contradicted by the provided rules, we will examine each rule:
1- The curriculum debate takes place at non-public meetings.
This rule does not directly support or
contradict the hypothesis regarding transparency.
⋮
6- There is a possibility that the situation is a result of a big misunderstanding.
This rule does not directly entail or contradict the hypothesis.
Based on the given rules, rules 2,3,4,5,
contradict the hypothesis that transparency is important to the board.
Therefore, the relationship between the context and the hypothesis is a "contradiction".

The label is "c".

Figure 4.3: Example of Algorithmic Prompt

Finalizing Prompts for the Project: For our project, we carefully designed a set of prompts for the exemplar selection task, using a combination of Standard Prompts, NL Description Prompts, and Algorithmic Prompts. By using a variety of different prompts, we aim to evaluate the model's adaptability and performance across various input scenarios. Having established our prompts, we can now proceed with prompting and evaluating our model with the selected exemplars.

Chapter 5

Experimental Setup and Results

In this project, we used the text-davinci-002 model as the baseline and 60 examples that were randomly taken from the test dataset. It is worth mentioning that due to the almost balanced dataset, the taken data are also approximately balanced in terms of labels. The following statistics are the results of our tests for each type of exemplar selection method and prompt explanation method. It should be noted that these statistics are the statistics of our improved prompts.

Table 5.1: In-context performance(accuracy) obtained using various prompting methods on ANLI datasets. Clustering in context learning method with NL Description prompting design, achieve the best performance among the other methods.

ICL	Prompting Methods		
Methods	Standard	NL Description	Algorithmic
Random	43.9%	47.6%	41.8%
Clustering	50.2%	55.1%	45.0%
BERTScore	51.4%	-	-
MMR	-	55.0%	-

Table 5.2: In-context performance(f1-score) obtained using various prompting methods on ANLI datasets. Clustering in context learning method with NL Description prompting design, achieve the best performance among the other methods.

ICL Methods	Prompting Methods		
	Standard	NL Description	Algorithmic
Random	45.2	45.5	40.8%
Clustering	52.3%	54.1%	44.2%
BERTScore	50.6%	-	-
MMR	-	50.0%	-

As can be seen in the results, the algorithmic method has reduced the model performance due to the extra explanations and the lack of benefit in detailing the explanations, and the MMR method helps the model in solving the problem by selecting examples for each query and maintaining diversity while selecting relevant examples.

We also added a sentence to explain why some parts of the table are empty. Unfortunately, we did not have enough budget to test all the cases and we only tested the important ones.

Chapter 6

Error analysis

Language models can make mistakes such as repetition, off-topic responses, over-optimization, incorrect information, and biases. These errors can affect the quality of requests and outputs. Knowing these error patterns can help improve the prompt design.

language models can have false positive or false negative errors. False positives are wrong answers that the model thinks are correct, while false negatives are correct answers that the model cannot generate. These errors can affect the speed and performance of the model. Analyzing these errors can help improve requests and model responses. Some strategies to address these errors in prompt design are:

- Refining instructions: Make your instructions clearer and more specific.
- Providing context and examples: Give the large language model some relevant information and examples.
- Experimenting with temperature and max tokens: Adjust settings that control the creativity and diversity of large language model outputs.
- Iterative refinement: Analyze the outputs of the large language model and use insights to improve your requests.

As expected, many errors were observed in the results of this project, and after analyzing and checking the errors, we tried to improve the prompts in the next tests.

Context:
The fact is that he did not choose to go to war as a last result.
And America now is paying, already $120 billion, up to $200 billion
before we're finished and much more probably. And that is the result
of this president taking his eye off of Osama bin Laden.

Hypothesis:
They were successful in their war.

True Label: Neutral

Predicted Label: Contradiction

Model reason:
A: The context suggests that the war was a result of the president
taking his eye off of Osama bin Laden, which contradicts the hypothesis
that they were successful in their war. Therefore, the relationship
between the context and the hypothesis is a contradiction.

Figure 6.1: Example of MMR-NL method Error

In this context, as can be seen in figure 5.1, it is said that the president has taken his eyes off Osama bin Laden and America has paid a lot of costs for the war, which do not show America's success or failure in the war, and since there is no word or concept about success in this context, the model should have a neutral output, but it has made a mistake and inferred something from the sentences that has given a wrong meaning, in other words, it has looked for a more complex concept than it should and given a false positive output.

Context:
Kia had been assigned a novel to read for English class. She dawdled all week.
The night before it was due, she panicked. She had read almost nothing!
She had to stay up until almost 3AM to finish the novel.

Hypothesis:
Kia got an extension and was able to read the novel for two weeks.

True Label: Contradiction

Predicted Label: Neutral

A: The context mentions that Kia had been assigned a novel to read for English class
and that she dawdled all week before finally staying up until 3AM the night before
it was due. There is no mention of Kia getting an extension or having two weeks
to read the novel. Therefore, the hypothesis cannot be confirmed or refuted
based on the context, resulting in a neutral label. Therefore, the relationship between
the context and the hypothesis is a neutral. The label is "n".

Figure 6.2: Example of Clustering-Algorithmic method Error

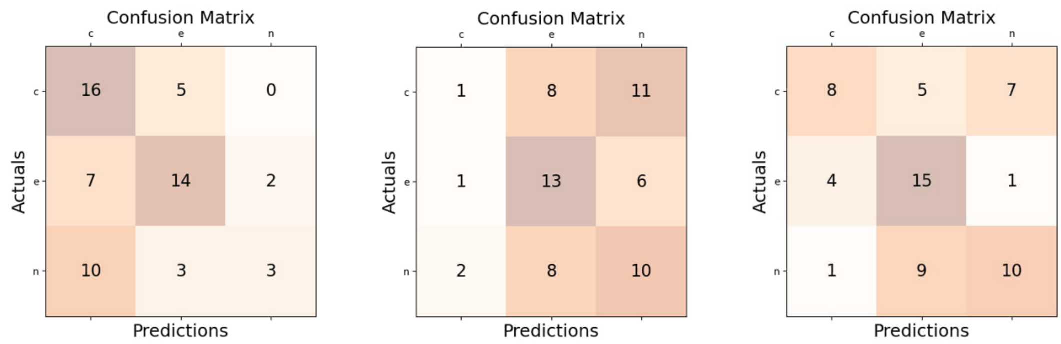
In this example, the model has thought that because nothing has been said about extension in the context, the relation is neutral, but in the context it has been mentioned that kia was awake until near 3 o'clock to finish the book, so there was no extension involved and this relation is contradict. In this example, unlike the previous example, the model had to make a simple inference that it could not and gave a false negative.

> Context:
> History will no doubt properly credit him and his government for their courage and foresight, initiatives such as the free trade agreement with the United States and NAFTA, abolishing the 13.5-per-cent manufacturers sales tax and introducing the 7-per-cent GST consumption tax to spur exports.
> Hypothesis:
> He abolished the manufacturer’s sales tax and GST consumption tax.
>
> True Label: Contradiction
>
> Predicted Label: Entailment
>
> A: The context mentions that the government took initiatives such as abolishing the 13.5-per-cent manufacturers sales tax and introducing the 7-per-cent GST consumption tax, which is consistent with the hypothesis that ”He abolished the manufacturer’s sales tax and GST consumption tax.” Therefore, the hypothesis can be inferred or entailed from the context. Therefore, the relationship between the context and the hypothesis is a entailment.

Figure 6.3: Example of Clustering-NL method Error

The model has only paid attention to the presence and absence of words in the context and hypothesis, but has not paid attention to their meaning. The model has ignored that it has been written in the context that the government has eliminated 13.5 percent sales tax on manufacturers and instead imposed 7 percent goods and services tax (GST). This means that the government has only eliminated one type of tax and replaced it with another type. But in the hypothesis it is written that he has eliminated both types of taxes. This contradicts the context, because the context shows that GST still exists and the model has made a mistake in understanding the meaning of the sentence.

In addition to these examples that the model has made a mistake, we should also look at the statistics of different methods.



(a) MMR-NL (b) Clustering-Algorithmic (c) Clustering-NL

Figure 6.4: Confusion Matrices

As can be seen in these figures, as the choice of Prompting method affects the model’s understanding of different classes (the difference in model understanding between Clustering-

Algorithmic, Clustring-NL experiments); the method of choosing examples also has a significant impact on this issue (the difference between Clustring-NL, MMR-NL experiments). So we should pay attention to both the method of choosing examples and the prompt writing method for them.

According to our observations in these experiments, model errors that were like example 1 had a higher frequency in the MMR-NL method(Fig.5.4.a); Also, errors that were like example 2 had a higher frequency in the Clustering-Algorithmic method(Fig.5.4.b); And errors that were like example 3 had a higher frequency in the Clustering-NL method(Fig.5.4.c). Therefore, the type of choosing examples and the type of writing and explaining them in the prompt has a special effect on the performance of the model in this task.

Chapter 7

Conclusion

In this project, we investigated the impact of exemplar selection and prompt designing on the performance of large language models on natural language inference task. We proposed and compared four methods for selecting exemplars from the training data and three methods for designing prompts for the input. We used the text-davinci-002 model as the baseline and evaluated it on the ANLI dataset. We found that the clustering method with natural language description prompt achieved the best results among the other methods. We also analyzed the errors made by the model and suggested some ways to improve the prompt design. We showed that choosing exemplars based on each query and maintaining diversity while selecting relevant examples can help the model in solving the problem. However, there are still some limitations and challenges that need to be addressed in our work, such as:

- Trying different models: We only used one large language model as our baseline, which may not be representative of all the state-of-the-art models. It would be interesting to see how other models, such as GPT-4, llama or Bard, perform on the same task with different exemplar selection and prompt designing methods. This would also help us to understand the strengths and weaknesses of different models and how they can be improved by using appropriate examples and prompts.

- Creating a set of prompts: We used only a few limited prompts for queries that may not capture all the nuances and variations of natural language inference. It would be useful to create a pool of prompts for queries that can cover different aspects and scenarios of natural language inference and Using a neural model or a simpler language model to select the best prompt for a given query. It would be better to have a dynamic method that can select the best prompt for a given query based on some criteria or feedback. For example, we can use a neural model that can assign a score to each prompt based on how well it matches the query and the expected answer. Alternatively, we can use a simpler language model that can rank the prompts based on their likelihood or coherence with the query. This would enable us to choose the most suitable prompt for each query and improve the performance of the model.

References

[1] Dale Schuurmans Maarten Bosma Brian Ichter Fei Xia Ed Chi Quoc Le Denny Zhou Jason Wei, Xuezhi Wang. Chain-of-thought prompting elicits reasoning in large language models. *ArXiv, abs/2201.11903.*

[2] Jacob Devlin Maarten Bosma Gaurav Mishra Adam Roberts Paul Barham Hyung Won Chung Charles Sutton Sebastian Gehrmann Parker Schuh Kensen Shi Sasha Tsvyashchenko Joshua Maynez Abhishek Baindoor Rao Parker Barnes Yi Tay Noam M. Shazeer Vinodkumar Prabhakaran Emily Reif Nan Du Benton C. Hutchinson Reiner Pope James Bradbury Jacob Austin Michael Isard Guy Gur-Ari Pengcheng Yin Toju Duke Anselm Levskaya Sanjay Ghemawat Sunipa Dev Henryk Michalewski Xavier García Vedant Misra Kevin Robinson Liam Fedus Denny Zhou Daphne Ippolito David Luan Hyeontaek Lim Barret Zoph Alexander Spiridonov Ryan Sepassi David Dohan Shivani Agrawal Mark Omernick Andrew M. Dai Thanumalayan Sankaranarayana Pillai Marie Pellat Aitor Lewkowycz Erica Oliveira Moreira Rewon Child Oleksandr Polozov Katherine Lee Zongwei Zhou Xuezhi Wang Brennan Saeta Mark Diaz Orhan Firat Michele Catasta Jason Wei Kathleen S. Meier-Hellstern Douglas Eck Jeff Dean Slav Petrov Aakanksha Chowdhery, Sharan Narang and Noah Fiedel. Palm: Scaling language modeling with pathways. *ArXiv, abs/2204.02311.*

[3] Nick Ryder Melanie Subbiah Jared Kaplan Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray Benjamin Chess Jack Clark Christopher Berner Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei Tom B. Brown, Benjamin Mann. Language models are few-shot learners. *ArXiv, abs/2005.14165.*

[4] Emily Dinan Mohit Bansal Jason Weston Yixin Nie, Adina Williams and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020.*

[5] Asli Celikyilmaz Ves Stoyanov Greg Durrett Ramakanth Pasunuru Xi Ye, Srinivasan Iyer. Complementary explanations for effective in-context learning. *ArXiv, abs/2211.13892.*

[6] Guy Gur-Ari Henryk Michalewski Jacob Austin David Bieber David Dohan Aitor Lewkowycz Maarten Bosma David Luan Charles Sutton Augustus Odena Maxwell Nye, Anders Johan Andreassen. Show your work: Scratchpads for intermediate computation with language models. *ArXiv, abs/2112.00114.*

[7] Yizhe Zhang Bill Dolan Lawrence Carin Jiachang Liu, Dinghan Shen and Weizhu Chen. What makes good in-context examples for gpt-3? *ArXiv, abs/2101.06804.*

[8] Hugo Larochelle2 Aaron Courville1 Behnam Neyshabur†2 Hattie Zhou*1, Azade Nova2 and Hanie Sedghi†2. Teaching algorithmic reasoning via in-context learning. *ArXiv, abs/2211.09066*.

[9] Zhengyuan Yang Shuohang Wang Jianfeng Wang Jordan Boyd-Graber Lijuan Wang Chenglei Si, Zhe Gan. Prompting gpt-3 to be reliable. *ArXiv, abs/2210.09150*.

[10] Felix Wu Kilian Q. Weinberger Tianyi Zhang, Varsha Kishore and Yoav Artzi. Bertscore: Evaluating text generation with bert. *In Proceedings of the International Conference on Learning Representations (ICLR)*.

[11] Hugging face provides a pre-trained model called deberta-large for the multinli task[1].

[1] Available at: `https://huggingface.co/microsoft/deberta-large-mnli`

Appendix A

Error Example

In this section, we will examine some wrong examples from the model.

Context:
How to create an indoor box turtle habitat? Set up a large enough enclosure.
Box turtles need a lot of room to roam like in the wild. Each turtle in your terrarium needs at least three square feet (91 cm) of floor space for every eight inches (20 cm) of their length.
Hypothesis:
Box turtles are indoor animals.

True Label: Neutral

Predicted Label: Entailment

Model reason:
A: The context provides instructions on how to create an indoor box turtle habitat, suggesting that box turtles can be kept as indoor pets. This is consistent with the hypothesis that "Box turtles are indoor animals." Therefore, the relationship between the context and the hypothesis is an entailment.

Figure A.1: Example of E

Because the context only shows how to create an indoor environment for box turtles. But this does not mean that box turtles are indoor animals. Maybe box turtles live in nature and just need an indoor environment to keep at home. So the context does not contain enough information to confirm the hypothesis and the model has done an extra reason that has caused its mistake. In this example, the model has given a wrong generalization that is wrong and unnecessary.

Context:
I said nothing and gave no advice, not having lived seventy-five years for nothing. I knew that Doctor John's decision was manly and right and fair, but I also knew it was all nullified by the fact that Marcella already loved him. So much I knew; the rest I was left to suppose. The Doctor and Marcella told me much, but there were some things too sacred to be told, even to me. So that to this day I don't know how the doctor found out that Marcella loved him.
Hypothesis:
The person who gave no advice did not have to guess whether Marcella

loved the doctor.
True Label: Entailment

Predicted Label: Contradiction

Model reason:
A: The context mentions that the person who gave no advice knew that Marcella loved the doctor, which contradicts the hypothesis that the person did not have to guess whether Marcella loved the doctor. Therefore, the relationship between the context and the hypothesis is a contradiction.

Figure A.2: Example of Error

As seen in the example above, the relationship is entailment, but the model has made a mistake in knowing that knowing a fact does not need to guess about it, and with the assumption of knowing Marcella's love, he could not understand that there is no need to guess.

Context:
Philippe von Stauffenberg, also known as "PVS", (born 1964) is a German businessman, residing in London, United Kingdom. He is the Managing Partner of Solidus Partners LLP, which he founded in 2004. Solidus Partners is a private equity firm, which makes investments in Media and Green Technologies.
Hypothesis:
Philippe was born on an even numbered year.

True Label: Entailment

Predicted Label: Neutral

Model reason:
A: The context does not provide any information about when Philippe was born. Therefore, the hypothesis cannot be confirmed or refuted based on the context, resulting in a neutral label. Therefore, the relationship between the context and the hypothesis is a neutral.

Figure A.3: Example of Error

In this example, he says he was born in 1964, which is an even year. Therefore, the hypothesis can be confirmed based on the context and as a result create an entailment label. The model reason is incorrect because it could not extract the relevant information from the context. Because he could not understand the number 1964 as a number and detect its parity.

Like this example, there are data in the dataset that require understanding numbers and simple calculations that our model cannot detect. Or the previous example that did not know the simple relationship between knowing and guessing, these prompt methods show the weaknesses of the model.

Appendix B
Prompts

In this section, we have brought some examples of the best tested prompts based on each method.

As mentioned in sections 3 and 4, in the MMR method, 3 examples were selected for each query (1 for each target label) and we described them using the NL method.

Here we see a query with its prompt.

Q: What is the relationship between the following premise and hypothesis? (I have three kinds of relationship: entailment, neutral, and contradiction)

Context: Brian James Anderson (born April 26, 1972) is an American former Major League Baseball pitcher, who played 13 seasons for five teams, as well as a sports broadcaster and coach. Currently, Anderson is the color commentator on the Rays TV crew on Sun Sports.
Hypothesis: Anderson was a bad coach.

A: The context primarily focuses on Anderson's career as a Major League Baseball pitcher, sports broadcasting, and his current role as a color commentator. It does not mention or evaluate his coaching performance. Therefore, without additional information, it is not possible to determine whether Anderson was a good or bad coach based solely on the provided context. Hence, the relationship is classified as neutral.The label is "n".

Q: What is the relationship between the following context and hypothesis? (I have three kinds of relationship: entailment, neutral, and contradiction)
Context: Core Python Programming is a textbook on the Python programming language, written by Wesley J. Chun. The first edition of the book was released on December 14, 2000. The second edition was released several years later on September 18, 2006.
"Core Python Programming" is targeted mainly at higher education students and IT professionals.
Hypothesis: Chun is not a programmer.

A: The relationship between the given context and hypothesis is a contradiction. The reason for this contradiction is that the context explicitly states that Wesley J. Chun is the author of the textbook "Core Python Programming," which suggests that he is involved in programming. On the other hand, the hypothesis states that Chun is not a programmer, directly contradicting the information provided in the context. The label is "c".

Q: What is the relationship between the following context and hypothesis? (I have three kinds of relationship: entailment, neutral, and contradiction)
Context: John Henry Michael McManus (born 21 January 1974), as stage name Rove McManus, is an Australian triple Gold Logie award-winning television presenter, producer, comedian and media personality. He was the host of the comedy talk show "Rove LA" and was also the host of the eponymous variety show "Rove." He is also the co-owner of the production company Roving Enterprises with partner Craig Campbell.
Hypothesis: John Henry Michael McManus was born more than 1.26 billion seconds ago.
A: The relationship between the context and hypothesis is entailment, we can calculate the number of years in 1.26 billion seconds:
1.26 billion seconds $\div$ 60 seconds/minute $\div$ 60 minutes/hour $\div$ 24 hours/day $\div$ 365.25 days/year = approximately 39.9 years. So, converting 1.26 billion seconds to years yields approximately 39.9 years. Now we can calculate John Henry Michael McManus's age which is 2023-1974=49, Therefore, the calculated age is more than 39.9 so, the hypothesis is entailed by the context. The label is "e".

Context:
Dr. Michael (Mike) Derrington Murphy (born July 19, 1940) is a chemistry professor, bluegrass musician, and educator who founded the "Bama Bluegrass" Show in 1983. "Bama Bluegrass" is the first and longest running bluegrass show on Alabama public radio.
Hypothesis:
Dr. Michael (Mike) Derrington Murphy is more than 60 years old.

Figure B.1: Example of MMR-NL Prompt

In the next method, we cluster the training data and in the prompt that we get for all queries, there is 1 example of each cluster.

Q: What is the relationship between the following context and hypothesis? (I have three kinds of relationship: entailment, neutral, and contradiction)
Context: Is there anything else that the board has done here in connection with the curriculum change that provides a basis for your complaint? From what I can see from attending the board meetings, I don't know, because so much of the curriculum debate takes place at non-public meetings that I am not aware of. And when board members are questioned at meetings, they really don't answer any questions. So it gives this whole illusion of secrecy to the process, which, you know, maybe a big part of the problem. It may just be a big misunderstanding.
Hypothesis: Transparency is important to the board.

A: The context suggests that there is a lack of transparency in the board's actions and decision-making process, as much of the curriculum debate takes place in non-public meetings and board members do not provide satisfactory answers to questions. This contradicts the hypothesis that transparency is important to the board. Therefore, the relationship between the context and the hypothesis is a contradiction. The label is "c".

Q: What is the relationship between the following context and hypothesis? (I have three kinds of relationship: entailment, neutral, and contradiction)
Context: The Bookstore Jan went to the bookstore. She was looking for a cookbook. One of the clerks helped her find a good cookbook. Jan ended up buying an Asian cuisine cookbook. For dinner, she made sweet and sour chicken.
Hypothesis: Jan was looking for a long cookbook.

A: The context mentions that Jan went to the bookstore looking for a cookbook, but it does not specify whether she was specifically looking for a long cookbook or not. Therefore, the hypothesis cannot be confirmed or refuted based on the context, resulting in a neutral label. Therefore, the relationship between the context and the hypothesis is a neutral. The label is "n".

Q: What is the relationship between the following context and hypothesis? (I have three kinds of relationship: entailment, neutral, and contradiction)
Context: Bravo, My Life (; lit. "I Love You, Mal-soon") is a 2005 South Korean film directed by Park Heung-sik about an adolescent boy who starts to come of age in the late 70s and early 80s, largely oblivious to the dramatic political events occurring around him. The film sold 406,526 tickets nationwide.
Hypothesis: The film sold over 400,000 tickets in South Korea alone.

A: The hypothesis states that "The film sold over 400,000 tickets in South Korea alone," which is consistent with the information provided in the context that "The film sold 406,526 tickets nationwide." Therefore, the hypothesis can be inferred or entailed from the context. Therefore, the relationship between the context and the hypothesis is a entailment. The label is "e".

Figure B.2: Example of Clustering-NL Prompt

Q: What is the relationship between the following premise and hypothesis?
(I have three kinds of relationship: entailment, neutral, and contradiction)

Context:
Is there anything else that the board has done here in
connection with the curriculum change that provides a basis for your complaint?
From what I can see from attending the board meetings,
I don't know, because so much of the curriculum debate takes place at non-public meetings,
that I am not aware of.
And when board members are questioned at meetings,
they really don't answer any questions.
So it gives this whole illusion of secrecy to the process,
which, you know, maybe a big part of the problem.
It may just be a big misunderstanding.

Hypothesis: Transparency is important to the board.

A: To understand the relationship between context and hypothesis,
we need to construct our knowledge base consisting of the following facts in the context:
Facts:
1- Much of the curriculum debate takes place at non-public meetings.
2- The speaker attending the board meetings is not aware of
what happens in these non-public meetings.
3- Board members do not answer questions when questioned at meetings.
4- The lack of transparency in the process creates an illusion of secrecy.
5- The speaker suggests that secrecy and
lack of transparency may be a significant part of the problem.
6- There is a possibility that the situation is a result of a big misunderstanding.
We use the backward algorithm, we'll start with the goal and
work backward through the facts to see if we can reach a conclusion:
Goal: Transparency is important to the board.
Explanation:
To determine if the goal is directly entailed or
contradicted by the provided rules, we will examine each rule:
1- The curriculum debate takes place at non-public meetings.
This rule does not directly support or
contradict the hypothesis regarding transparency.
2- The speaker attending the board meetings is not aware of
what happens in these non-public meetings.
This rule reinforces the lack of transparency,
as the speaker, who attends the meetings,
is unaware of the discussions in non-public meetings.
3- Board members do not answer questions when questioned at meetings.
This rule implies a lack of openness and transparency,
as the board members do not provide answers to questions asked during meetings.
4- The lack of transparency in the process creates an illusion of secrecy.
This rule directly contradicts the hypothesis,
as it states that the lack of transparency leads to a perception of secrecy.
5- The speaker suggests that secrecy and lack of transparency
may be a significant part of the problem.
This rule contradicts hypothesis
, as the speaker perceives the lack of transparency as problematic.
6- There is a possibility that the situation is a result of a big misunderstanding.
This rule does not directly entail or contradict the hypothesis.
Based on the given rules, rules 2,3,4,5,
contradict the hypothesis that transparency is important to the board.
Therefore, the relationship between the context and the hypothesis is a "contradiction".

The label is "c".

Figure B.3: Example of Algorithmic Prompt