Getting in Tune - Exploring the Effects of Instruction Tuning on Performance for Large Language Models

Audrey Gasser, Ani Nuthalapati, Preeti Pidatala, Amrutha Shetty

1 Introduction

With the recent launch of ChatGPT, researchers have launched a myriad of new research supported by its use, whereas the general public has explored the near-mystical nature of Learning Language Models. However, the frequently less-than-stellar output from ChatGPT can demystify LLMs for non-expert users. ChatGPT can be messy, unpredictable, inconsistent, and even incorrect - and the system has no way of knowing this. However, as a team of self-described ChatGPT power users, we discovered early on that it is possible to improve your experience with the model. This project aims to do just that - by understanding how Instruction Tuning influences how the model responds to questions and tasks, we set out to determine the best way to phrase requests to ChatGPT. This is vital information for anyone that regularly relies on LLMs, but unfortunately, there has been a dearth of published literature on the subject. This project is our attempt to contribute to that base of knowledge.

This field of research is of particular interest to multiple parties. Natural language processing and artificial intelligence researchers are particularly invested, as their work may focus on improving the performance of LLMs like ChatGPT. With a more recent rise in chatbots and virtual assistants, companies and organizations that utilize these language-based applications may be interested in the improvement of LLMs through instruction tuning. On the non-expert side, those that work in education may be pointedly interested in the improvement of ChatGPT's outputs, especially those who wish to improve their own prompts for language-based assignments of exam questions for their students.

2 Background

It is imperative to understand the current limitations of ChatGPT in order to further optimize its capabilities. One of the biggest challenges that ChatGPT faces is in answering vague questions. One study from February of 2023 compares multiple natural language processing models in their ability to pass the United States Medical Licensing Examination. Two different tests were used, the AMBOSS and NBME, wherein the question format was standardized between the two (images were removed and the question format was question context immediately followed by the direct question). Overall, ChatGPT was able to achieve 44%-57.8% for the exams, but the performance of the model significantly decreased as the difficulty of the questions increased. This work highlights that the presence of contextual information for the question significantly increases the chance of the model accurately answering the question [2]. Instruction fine-tuning in general has been shown to significantly improve model performance and generalization to unseen tasks across various model classes, prompting setups, and evaluation benchmarks. It is one of the primary user control methods to improve the model's output. For instance, one study explores the Flan-PaLM 540B model. The authors discovered that instruction fine-tuning on 1.8K tasks outperforms the PaLM 540B model by a large margin (+9.4% on average) and achieves state-of-the-art performance on several benchmarks. These results support instruction fine-tuning as a useful method for improving the performance and usability of pre-trained language models [1].

Significant research has been conducted into instruction tuning and various techniques that improve the performance of such language models. One study introduces HIR, or hindsight instruction labeling, which better aligns model responses with instructions by relabeling previous instructions. In this way, the model improves responses upon success and failure cases. This technique requires no additional parameters and has been shown to perform competitively well on a variety of tasks [9].

This work aligns with related research into user intent, arguably one of the most important aspects of large language model interaction. In [5], researchers point out that larger language models are not necessarily better at following user intent, and that they can still generate untruthful or toxic output. The paper proposes InstructGPT, a method that fine-tunes GPT-3 with human feedback in order to align the model with user intent. In human evaluations, InstructGPT models with 1.3B parameters are preferred to models with 175B parameters. InstructGPT models also show improvements in truthfulness and reduction in toxic output while having minimal performance regressions on public NLP datasets. This work demonstrates that fine-tuning with human feedback is a promising direction for aligning language models with human intent.

Another frequently noted challenge with large language models is their generalizability to unseen prompts or instructions. Like any model, generalizability is one of the main indicators of a large language model's strength. One proposed method to overcome this is through a concept called Flipped Learning [7]. Rather than training the models to optimize a label given an instruction, this method aims to train models to select the label that is most likely to generate a given instruction. The study demonstrates strong results, even when generalizing the model to new labels. Like HIR, this is another way to better align model outputs to user instructions with accuracy [7]. Similarly, another strong method to improving model generalizability is In-Context Instruction Learning (ICIL). This method was proposed to further improve language model performance. The studies show that simply by tailoring the model input to include the instruction, input, and output instance, the model is able to perform more effectively. The paper demonstrates how not only will this technique improve performance when used correctly, but performance is not significantly impacted when irrelevant output instances are provided either [8].

Overall, current work on the topic explores many different techniques to optimize large language model performance. However, most researchers are in agreement that there is much more work to be done. These models are still limited in their efficiency and accuracy. "Hallucinations", or false outputs are a common limitation of large language models and can result in misinformation or bias if left unchecked. In order to further the existing work on LLM optimization, this paper evaluates various prompt patterns on ChatGPT to better understand which input styles yield stronger outputs.

3 Approach

To begin our project, we needed to understand the ways that LLMs leverage Instruction Tuning to be more effective at task completion. While there are many specific details that were not relevant to our work, it was helpful to get an understanding of the ways that instruction tuning was used to improve the performance of LLMS, particularly in the context of chatbots and assistants. By observing the sort of training that these models underwent, we were able to locate some possible approaches we could take to interact with the models in a way that was more consistent with the training. In particular, we settled on four prompt approaches: Direct, Template, Persona, and Few-Shot. See Table 1 for an example of a prompt and its corresponding translations using the four prompt approaches. From there, we located a diverse set of tasks that we could use to test our hypothesis that the Persona approach would work best. Once we found a usable set of tasks from a different paper, we manually converted those tasks into the four different prompt types. This is the part of the process that is novel about our approach, as we are evaluating these different prompting mechanisms to compare how they perform for a diverse set of tasks. We then ran these prompts through ChatGPT and manually evaluated the results. For the hyperlink to the complete list of prompts and their corresponding pattern translations, please refer to the Appendix.

Table 1: Translations of the Prompt

"Write a polite rejection letter for rejecting a candidate. They were rejected because they came in second place"

Direct	Write a polite rejection letter for rejecting a candidate because they came in second place.
Persona	I want you to act as an employer. I want you to write a polite rejection letter for rejecting a candidate. They were rejected because they came in second place.
Template	Write a polite rejection letter for rejecting a candidate. They were rejected because they came in second place. Here is a template for the rejection letter: Hello Candidate X, Unfortunately, _____
Few-Shot	Write a polite rejection letter for rejecting a candidate. They were rejected because they came in second place. Here is an example of such a rejection letter: Dear [Candidate], I am writing to inform you that after careful consideration, we have decided to move forward with the first candidate who was selected for the position. We appreciate your interest in the role and your efforts in preparing for the interview. Please accept our sincerest apologies for not being selected this time. We understand that the process can be competitive and we are grateful for your participation. We wish you the best of luck in your future endeavors. Thank you for your time and consideration. Sincerely, [Your Name] [Company] [Email]

The prompt 'tasks' were selected from the GPT-4-LLM dataset of Instruction Tuning With GPT-4 [6]. Categorization was accomplished by following the respective templates provided by prior research.

For example, a persona prompt will always follow the format of "I want you to act as [someone with implied authority on the task]. [task description and relevant details]. A Direct pattern is simply the direct phrasing of the task. The Template pattern gives the direct task description, followed by a phrase notifying ChatGPT that the user is providing a template. Following that is a general template that the user wishes ChatGPT to structure its response with. The last prompting pattern is Few Shot, which provides ChatGPT with the task, followed by a complete, 'correct' example provided by the user. The reason that these tasks and patterns were chosen is that we must evaluate the success of ChatGPT's outputs in terms of expected or 'satisfactory' to the user, not the factual success. In other words, we are not testing the factual knowledge of ChatGPT, but its ability to formulate 'better' or 'worse' solutions to a given task based upon the pattern of the prompt.

We encountered some problems during the development of our approach. One of the largest obstacles to overcome was the sourcing of reliable tasks that can be aligned with prompt templates. Originally, we had attempted to create prompts ourselves, but because the members of our team do not have prior research experience or authority to create these prompts, we elected to locate the datasets used by other instruction-tuning research. Many datasets did not contain unaltered tasks but instead contained the tasks already formulated into a specific prompt type, usually that of a linguistics or reading comprehension exam question. We also had to identify tasks that were relatively diverse and had a broad scope, as we didn't want to limit our evaluation to simply studying the background knowledge abilities of the model. Realigning all the prompts into the different categories was also a challenge, as we wanted to make sure we were being as consistent as possible.

4 Results

Success was largely measured via human annotation. For each task, there are four prompt outputs (one for each prompt pattern). Each prompt output is manually evaluated by human annotators in terms of grammaticality and typicality, that is, how grammatically correct (or 'human-like') the response is and whether or not ChatGPT formulated its response as expected by the user. For each of these metrics, a prompt output is rated on a scale of 0 to 5, where 0 indicates a prompt that is the least grammatically correct or the least typical of an expected output. Conversely, a score of 5 indicates the opposite, where a prompt output is the most grammatically correct or the most typical of an expected output. In conclusion, we arrived at the values of grammaticality and typicality for each prompt category by computing the average of individual values.

Through this process, we aimed to identify which prompting technique would lead to "better" output from ChatGPT. This was tricky to measure, however, as there were various factors that we could not assess well. We didn't have a good way of assessing factuality or accuracy, and we didn't have the resources to assess consistency. These would both be ideal, but we had to resort to using grammaticality as a proxy for consistency and typicality as a proxy for factuality.

Some of our key observations include:

1. Although there were prompts that scored well across both measures, the distribution of grammaticality and typicality scores across the four prompt categories indicated that a few prompts were good on grammaticality but low on typicality, and vice versa.

2. Further analysis of the prompts with low scores of either of the measures helped us to get to know that prompts from the 'template' category where predefined templates or structures were used for generating responses gave the responses with the lowest typicality and the prompts from 'direct' category which involves asking questions in a straightforward and simple manner led to the lowest grammaticality scores. This can be explained by the fact that for templated responses, ChatGPT often ignored or modified the template. As the evaluators were expecting a response in line with the template, the typicality scores took a hit. For the direct prompts, the lack of guidance for the LLM likely contributed to the reduced grammaticality of the output.

3. ChatGPT also failed to understand certain specific prompt categories in a few places. For example, Table 2 details such a case:

Table 2: Failure Case

Prompt Category	Template
Prompt	Make a list of the materials that will be required to build a Coffee Darkness Meter.
Template	Basic materials like _____ are required to build a _____
Response	I'm not sure what a Coffee Darkness Meter is, can you please provide more context or information so I can better assist you?

Interestingly enough, the model did not struggle on the same task for different prompt patterns.

4. Among the different categories of prompts that we evaluated, we found that the 'persona' category performed relatively well in terms of both grammaticality and typicality measures. This suggests that prompts which incorporate a persona-like context or personality traits are more likely to generate coherent and relevant responses from the chatbot model. This prompt pattern also consistently generated responses that were considerably longer than the others, perhaps due to the additional narrative structure that was provided.

Fig. 1 This is a comparison between grammaticality and typicality. Typicality has its value to be the least for prompt category 'template' and has its highest value for prompt category 'direct' surpassing the grammaticality scores too. While grammaticality has higher values when compared to typicality in most of the categories except for the 'direct' category. It has its highest value for 'persona' category and least value for the direct category.

Fig. 2 Grammaticality measures the conformity of a sentence to the rules defined by a specific grammar of a language. Over different category prompts ranging from 1 to 5 units, this plot shows the different values the metric grammaticality can take. There are few outliers in the graph from the direct prompt category, but most prompts have scores of 4 or above.

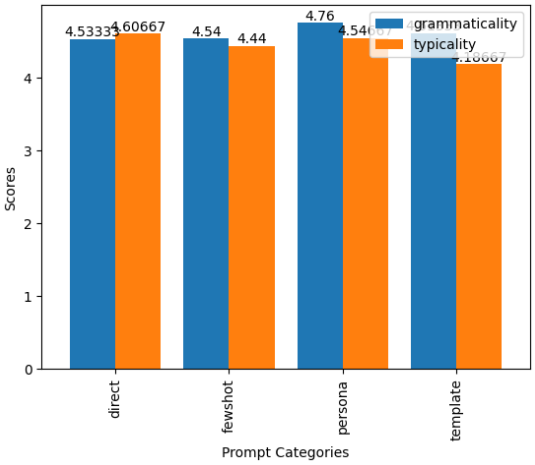

Figure 1: Prompt Categories vs Scores

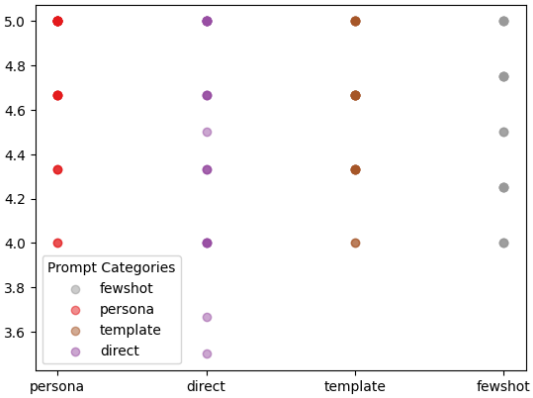

Figure 2: Grammaticality Scores

Fig. 3 For four different categories, the plot displays the combined scores of both measures. In this plot, it is clearly visible how the scores and outliers differ according to the categories.

Fig. 4 Typicality measures the state of being that is typical. Based on different categories of prompts, this plot illustrates the range of values the measure of typicality can take. The majority of prompts have scores of 3.5 or above. However, there are a few outliers in the graph from categories such as template, few shots, and personas.

Fig. 5 Plotting the relationship between two measures, grammaticality and typicality, is shown in this figure. In the graph, most of the values range between 3.5 and 5, indicating both grammaticality and typicality do well. However, as shown in the plot, there are a few outliers, such as prompts with excellent grammaticality scores but poor typicality scores and vice versa.

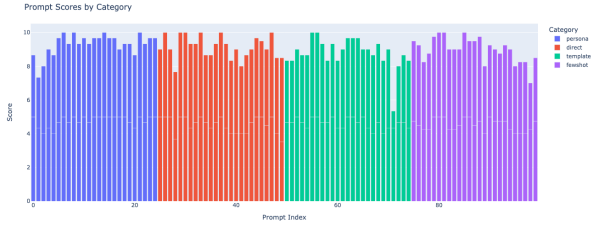

Figure 3: Combined scores over all the prompts

Fig. 6 Heatmaps are used to show relationships between two variables, one plotted on each axis.Here, the heatmap displays the scores of the two measures across four prompt categories.

5 Discussion

Due to the variable nature of ChatGPT's responses, the exact same outputs may not be easily achieved, however, the general nature of the results can be replicated. The tasks are publicly available for use, and the prompt templates are not difficult to apply to new tasks if need be. The subjectivity of the human annotation process is less replicable, meaning that the results of human annotation will not be the same as those in this research. If we had more resources, we would rely on more human evaluators and better evaluation methods that would allow for greater replicability.

While the dataset we used to find prompts was diverse and useful, we recognize that it isn't the only such source out there. The specific nature of these tasks also affected the ways that we translated them into the different prompting methods.

It is also important to acknowledge the ethical discussions surrounding the use of ChatGPT, especially if we are to be working with it with the aim of retrieving information or performing tasks. One paper highlights the current conversations that academics are engaging in regarding ChatGPT's ability to construct essays based on a single prompt. Copyright and intellectual property laws do apply to ChatGPT and other AI-adjacent natural language processing models– while ChatGPT was found to be an excellent resource for researchers and academics, GPT often does not explicitly state whether or not it is quoting another resource when it returns an answer to the user[4].

There is also the issue of generating information that is false, biased, or dangerous. Similar prompting techniques have been used in the past to "jailbreak" ChatGPT, eliminating any safeguards

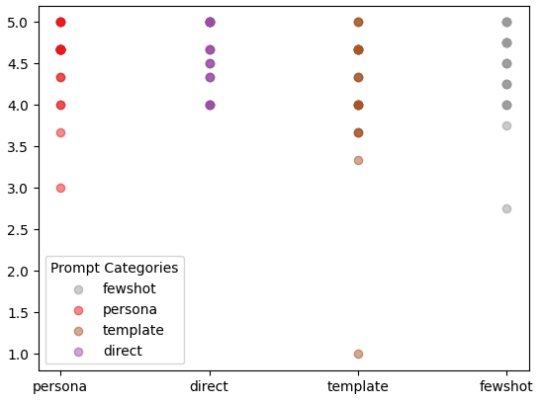

Figure 4: Typicality Scores

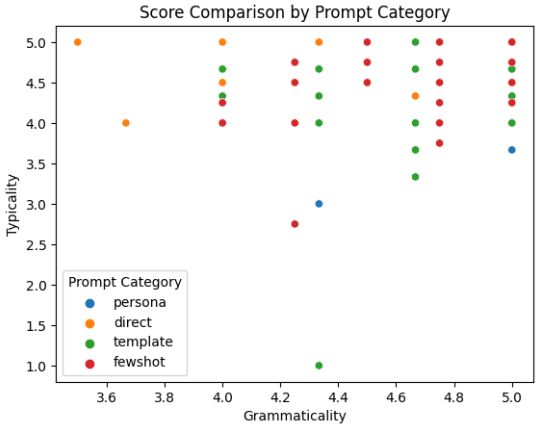

Figure 5: Grammaticality vs Typicality

that OpenAI may have implemented in their models[3]. As models get more powerful, the implications of such jailbreaks can become more and more dangerous, and it is possible to take advantage of research like this to improve such efforts. These prompting techniques could also be modified to solicit misinformation or other dangerous responses from ChatGPT. While this is a risk we recognize, we believe it is important to provide information and knowledge to society as a whole to mitigate the possible damage and allow people to take the necessary precautions. Additionally, it is important for users of AI technologies to be aware of the potential risks and take appropriate precautions when using these systems. This includes being cautious of the sources and accuracy of information generated by AI models, as well as being aware of the potential for models to be manipulated or used for harmful purposes.

6 Conclusion

Overall, from the quantitative and qualitative analysis, the Persona prompting pattern emerged as

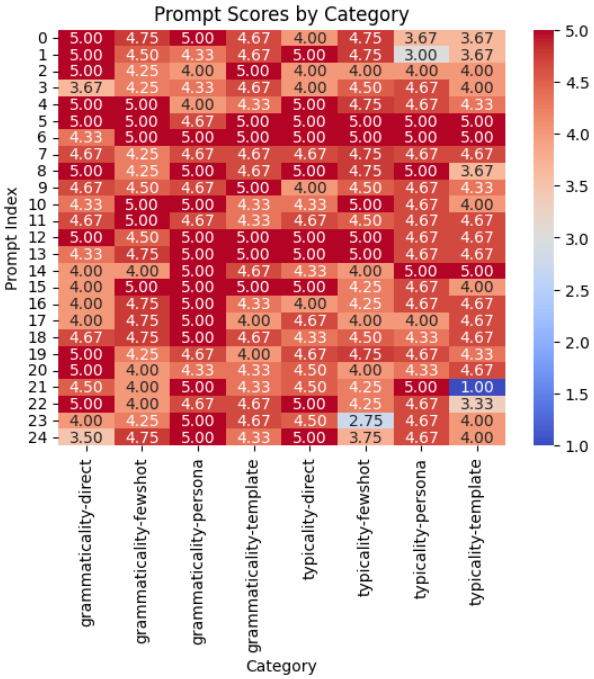

Figure 6: Heatmap of Scores

the strongest. This input category yielded outputs with the highest grammaticality and typicality measures in the human evaluation. Additionally, observations from the visualizations and task outputs demonstrated that these prompts generally resulted in more detailed ChatGPT outputs. Though the human evaluation results revealed that all four prompting patterns performed relatively well, the template prompting pattern had the lowest typicality scores and the direct prompting pattern had the lowest grammaticality scores. Further work on this project will continue to improve the performance of ChatGPT - optimizing its capabilities as a strong resource for users.

Future work on this project includes expanding the human evaluation to more evaluators and evaluating across different categories. This will introduce more diverse viewpoints on evaluations and result in more representative scores. A larger, more variable, dataset is also a future step for this project. The current dataset was limited to only 100 prompts spanning across 25 tasks. Incorporating more prompts and tasks would improve the strength of results.

Acknowledgements

We would like to extend a special 'thank you' to Professor DK for his guidance, support, and feedback throughout the course of this project.

References

[1] Hyung Won Chung et al. *Scaling Instruction-Finetuned Language Models*. Dec. 6, 2022. arXiv: 2210.11416[cs]. URL: http://arxiv.org/abs/2210.11416 (visited on 04/27/2023).

[2] Aidan Gilson et al. "How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment". In: *JMIR Medical Education* 9 (Feb. 8, 2023), e45312. ISSN: 2369-3762. DOI: 10.2196/45312. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9947764/ (visited on 04/28/2023).

[3] *Jailbreak Chat*. en. URL: https://www.jailbreakchat.com/ (visited on 05/06/2023).

[4] Brady Lund et al. "ChatGPT and a New Academic Reality: Artificial Intelligence-Written Research Papers and the Ethics of the Large Language Models in Scholarly Publishing". In: *Journal of the Association for Information Science and Technology* (Mar. 2023). DOI: 10.1002/asi.24750.

[5] Long Ouyang et al. *Training language models to follow instructions with human feedback*. Mar. 4, 2022. arXiv: 2203.02155[cs]. URL: http://arxiv.org/abs/2203.02155 (visited on 04/27/2023).

[6] Baolin Peng et al. *Instruction Tuning with GPT-4*. Apr. 6, 2023. DOI: 10.48550/arXiv.2304.03277. arXiv: 2304.03277[cs]. URL: http://arxiv.org/abs/2304.03277 (visited on 05/06/2023).

[7] Seonghyeon Ye et al. *Guess the Instruction! Flipped Learning Makes Language Models Stronger Zero-Shot Learners*. Dec. 4, 2022. DOI: 10.48550/arXiv.2210.02969. arXiv: 2210.02969[cs]. URL: http://arxiv.org/abs/2210.02969 (visited on 04/27/2023).

[8] Seonghyeon Ye et al. *In-Context Instruction Learning*. Feb. 28, 2023. arXiv: 2302.14691[cs]. URL: http://arxiv.org/abs/2302.14691 (visited on 04/27/2023).

[9] Tianjun Zhang et al. *The Wisdom of Hindsight Makes Language Models Better Instruction Followers*. Feb. 10, 2023. DOI: 10.48550/arXiv.2302.05206. arXiv: 2302.05206[cs]. URL: http://arxiv.org/abs/2302.05206 (visited on 04/27/2023).

7 Appendix

Our spreadsheet (containing our tasks, prompts, responses, and human evaluation scores) can be found at this link.

Included on the following page is a small sampling of tasks, formatted into the different prompt patterns:

Tasks	Prompt (Pattern: Persona)	Prompt (Pattern: Direct)	Prompt (Pattern: Template)	Prompt (Pattern: Few-shot)
Estimate the cost of supplies required to create a DIY Water Chiller including items like Ice-maker, Icebox/cooler, Temperature regulator, Aquarium pump x2, 12V power supply	I want you to act as a materials list maker. I want you to estimate the cost of supplies required to create a DIY Water Chiller including items like a ce-maker, Icebox/cooler, Temperature regulator, Aquarium pump x2, 12V power supply.	Estimate the cost of supplies required to create a DIY Water Chiller using items like ice-maker, icebox/cooler, temperature resulator, aquarium pump x2, and 12V power supply.	Estimate the cost of supplies required to create a DIY Water Chiller including items like Ice-maker, Icebox/cooler, Temperature regulator, Aquarium pump x2, 12V power supply Template: The cost of supplies can come up to $_____ to create a _____	Estimate the cost of supplies required to create a DIY Water Chiller including items like Ice-maker, Icebox/cooler, Temperature regulator, Aquarium pump x2, 12V power supply. An example of an estimate would be: $1000.
Indicate the genre of Love in the Time of Cholera to which it belongs.	I want you to act as a book reader. I want you to indicate the genre of Love in the Time of Cholera.	Indicate the genre of Love in the Time of Cholera.	Indicate the genre of Love in the Time of Cholera to which it belongs. Template: The genre of the given movie is _____	Indicate the genre of Love in the Time of Cholera. An example of a genre of book is Science Fiction
Write a polite rejection letter for rejecting a candidate. They were rejected because they came in second place	I want you to act as an employer. I want you to write a polite rejection letter for rejecting a candidate. They were rejected because they came in second place.	Write a polite rejection letter for rejecting a candidate because they came in second place.	Write a polite rejection letter for rejecting a candidate. They were rejected because they came in second place. Here is a template for the rejection letter: Hello Candidate X, Unfortunately, _____	Write a polite rejection letter for rejecting a candidate. They were rejected because they came in second place. Here is an example of such a rejection letter: Dear [Candidate], I am writing to inform you that after careful consideration, we have decided to move forward with the first candidate who was selected for the position. We appreciate your interest in the role and your efforts in preparing for the interview. Please accept our sincerest apologies for not being selected this time. We understand that the process can be competitive and we are grateful for your participation. We wish you the best of luck in your future endeavors. Thank you for your time and consideration. Sincerely, [Your Name] [Company] [Email]
In relation to a sudden temperature change, give some tips on how to adjust the travel plans with it.	I want you to act as a travel advisor. I want you to give some tips on how to adjust travel plans with a sudden temperature change.	Give some tips on how to adjust travel plans with a sudden temperature change.	In relation to a sudden temperature change, give some tips on how to adjust the travel plans with it. Template: Here are few tips on how one can adjust the travel plans with regards to sudden temperature change : _____	Give some tips on how to adjust travel plans with a sudden temperature change. Some examples might be: 1. Dress appropriately: Make sure to pack warm clothing and layers that can help you stay warm during the sudden temperature change. 2. Stay hydrated: Drink plenty of water to stay hydrated and avoid dehydration. 3. Stay active: Keep moving around and stay active to help your body adjust to the change in temperature.
Find synonyms for the word adversity. You need to write down how the provided synonyms differ from the original word in terms of meaning, usage, etc.	I want you to act as a linguist. I want you to find synonyms for the word adversity. I want you to write down how to provide synonyms that differ from the original word in terms of meaning, usage, etc.	Find synonyms for the word adversity. Write down how the provided synonyms differ from the original word in terms of meaning, usage, etc.	Find synonyms for the word adversity. You need to write down how the provided synonyms differ from the original word in terms of meaning, usage, etc. Template: The synonyms for the given word are _____ . Each synonym is different from the original word as _____	Find synonyms for the word adversity. Write down how the provided synonyms differ from the original word in terms of meaning, usage, etc. Some examples of synomyms would be misfortune, difficulty, hardship, distress

Figure 7: Sample of Prompts in different categories