# Activation Steering for Instruction Following: Responsible AI FAQ

- **What is this repository?**

The ability to follow instructions is crucial for numerous real-world applications of language models. We provide a method that uses internal representations of models to enhance instruction following capabilities. The method computes vector representations of different types of instructions (e.g., json format, length, style, language) and then adds these representations to the network at inference time, to boost the model's capability in following these instructions.

Find a detailed discussion in our paper at: https://arxiv.org/abs/2410.12877

- **What can this repository do?**

Overview of the work:

- Uses vector representations of different types of instructions (format, length, style) to enhance instruction following.
- Adding vector representations at inference time was already introduced in previous work. Our contribution consists of adapting the method for instruction following and showing that it can be used in the following settings: to improve a diverse set of instructions, to combine and compose different types of instructions (e.g. write the output in capital letters and use only 2 sentences.), and to transfer the representations from instruction tuned models to base models (this is called cross model steering in the paper).
- Uses open-source models like phi 3.5 and gemma to show the effectiveness of the method.
- Uses the IFEval dataset to show the effectiveness of the method on a diverse set of instructions. Provides a detailed analysis on what types of instructions are more suitable for the method, how instruction steering may impact text quality, and includes automated layer search and parameter search constructs for optimizing the steering process.

- **What are the intended uses of this repository?**
    - *Improving instruction following for productivity applications and more general text generation.*
    - *Possible stakeholders include*
        - *Industry stakeholders that serve llms at inference time*
        - *Academic research*
- **How was this repository evaluated? What metrics are used to measure performance?**

We conduct two types of evaluations:

- Instruction Following Capabilities (e.g. does the model follow length, style, format specifications when asked to?). For this we use the IFEval dataset and the following models: Phi 3.5, Mistral 7B, and Gemma 2B and 9B.
- Text Quality Scores. For this, we first use another state-of-the-art llm (gpt 4o) to generate quality scoring questions for each question in IFEval. For example, if the task is to generate a story about a hiking trip, the quality scoring questions may evaluate for text fluency, relevance to the topic

etc. Then, we use these questions to assign scores between 1-5 to the generated text. This part is used to ensure that the models do not game their output such that they can easily follow instructions without providing any useful answer.

- **What are the limitations of this repository? How can users minimize the impact of these limitations?**
    - *Directly steering models via vector representations can lead to drops in text quality. We provide a detailed analysis on this in Appendix E of our paper. The analysis shows that in general llms tend to lower the output quality as they balance their goals between text quality and instruction following. However, we also mitigate this during the layer and parameter search methods, so that one can decide to not apply our method if the perplexity of the text is not good enough (text perplexity can be interpreted as text fluency).*
    - *Despite our method being effective, it still cannot guarantee 100% instruction following at test time. We believe that these methods need to be combined with fine tuning and pretraining methods, such that they can better guarantee that user instructions are followed, even in most complex cases.*
    - ***The method also covers instructions that ask the model to include or exclude certain words. Depending on the nature of words, improved instruction following here can also lead to increased risk for harmful or offensive content. It is important to be aware of such implications for responsible use of the method. In cases where the given words are absolutely necessary, one may still be able to combine this work with previous contributions in this space that compute vector representations for different aspects of responsible use, to mitigate undesired effects.***
    - *For an extended discussion on the above please see Appendix A in our paper. The method was was developed for research and experimental purposes. Further testing and validation are needed before considering its application in commercial or real-world scenarios.*
    - *The method was designed and tested using the English language.*
- **What operational factors and settings allow for effective and responsible use of this repository?**
    - *The method improves instruction following capabilities for the following instruction types: text format and style, length, casing, and combinations of these. Improvements vary depending on the model on which the method is applied and whether the original prompt from the user contains/mentions the instruction on the first place. More precisely, we observe that when the user mentions the instruction from the beginning, the method is most effective for phi models and small gemma models, and less so (but with some modest improvements) for mistral and the larger gemma models.*
    - *We strongly encourage users to use LLMs/MLLMs that support robust Responsible AI mitigations. Since our technique is applied at inference time, it is also suggested to use post inference mitigations such as content filters.*
- **How do I provide feedback on this repository?**
    - *If you have suggestions, questions, or observe unexpected/offensive behavior in our technology, please contact us at besmira.nushi@microsoft.com and stolfoa@ethz.ch. Emailing directly to the paper authors.*
    - *If the team receives reports of undesired behavior or identifies issues independently, we will update this repository with appropriate mitigations.*