Review of: "Improving Instruction-Following in Language Models through Activation Steering"

Wenguan Wang[1]

1 Zhejiang University of Science and Technology, Hangzhou, China

Potential competing interests: No potential competing interests to declare.

Summary:

The paper proposes the activation steering method within the context of instruction-following in language models. The method utilizes the instruction-specific vector representations from language models to guide models' generation. The authors conduct experiments on the IFEval benchmarks. The experimental results show that the proposed method can enhance the model's adherence to various constraints.

Strengths:

1. The concept of activation steering is straightforward, and the experimental results validate its utility in enhancing instruction-following in language models.
2. The paper is well-written and clear, with a structured presentation that helps readers understand the pipeline. There is a detailed description of the implementation.

Weaknesses:

1. The format of all equations is unaligned, making them challenging to read. Additionally, the equations are not tagged, such as Eq. 2 mentioned in the last paragraph of Section 2.2.
2. The font size in Figures 2, 4, and 7 is not consistent with the font size used in the main paper.
3. Could you clarify the sudden drop in the cosine similarity metric for '(2) same query w/ and w/o instr.' in Figure 2?
4. The paper conducts experiments using a relatively small size of language models. It will be interesting to see how the model's generalizability improves with a more powerful language model.
5. The novelty of activation steering could be clarified when considering related research beyond instruction-following fields. A deeper exploration of higher-level concepts through activation steering would provide more valuable context and enhance the novelty of the work.
6. From Figure 8b, the steering method yields marginal improvement in Gemma 9B compared to Gemma 2B. Does this demonstrate that the steering method has limitations when applied to more capable language models? Little discussion exists on why the improvement plateaus and whether there are theoretical or practical limits to the approach's performance gains.