Do LLMs know internally when they follow or do not follow instructions?

Juyeon Heo[1] , Christina Heinze-Deml[2] , Oussama Elachqar[2], Kwan Ho Ryan Chan[3],Shirley You Ren[2], Andrew Miller[2], Udhyakumar Nallasamy[2], and Jaya Narain[2]
University of Cambridge[1], Apple[2], U Penn[3]

Motivation

Instruction-following matters for building reliable LLM agent.

A key to building safe and useful personal AI agents with LLMs lies in their ability to follow instructions precisely. Deployed models must strictly follow the instructions and constraints from users to ensure that the outputs are both safe and aligned with user intentions.

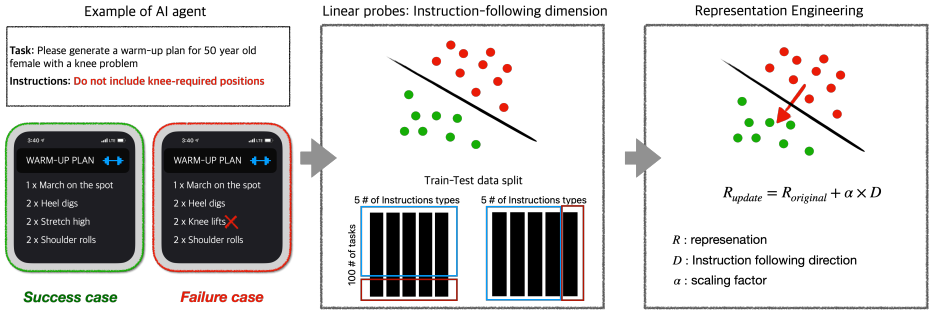
However, LLMs fail to follow even 'simple' and 'nonambiguous' instructions.

GPT-4 achieves around an 80% success rate on IFEval (Zhou et al., 2023), an instruction-following benchmark dataset, while smaller models have success rates around 30% to 40%. (Zhou et al., 2023; Qin et al., 2024; Xia et al., 2024; Kim et al., 2024; Yan et al., 2024)

Methods

To gain a better understanding of instruction-following outcomes, we analyze the internal state of LLMs.

By applying linear probing, we identify a specific dimension within the input embedding space that is strongly associated with instruction-following. While previous work has primarily used linear probing to explore representations related to truthfulness and reducing hallucinations (Azaria & Mitchell, 2023; Marks & Tegmark, 2023; MacDiarmid et al., 2024), our study extends this method to investigate instruction-following.

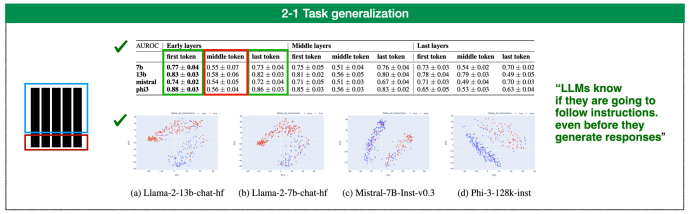


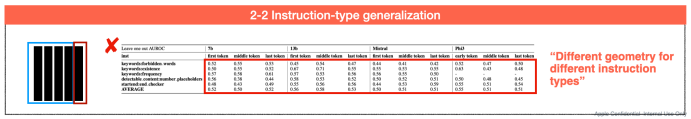
RQ: Is there any difference between success and failure cases in instruction-following?

Results

1) We identify an instruction-following dimension using linear probes, which generalizes to unseen tasks, but not to unseen instruction types.

We identify a specific dimension within the input embedding space of LLMs that is closely linked to instruction-following using linear probes, by carefully designing our setting to disentangle the effects of tasks and instructions in input prompts.





2) We apply representation engineering to convert failure cases to success.

We demonstrate that this dimension generalizes to unseen tasks and that modifying representations along this dimension effectively converts instruction-following failures into successes without compromising response quality.

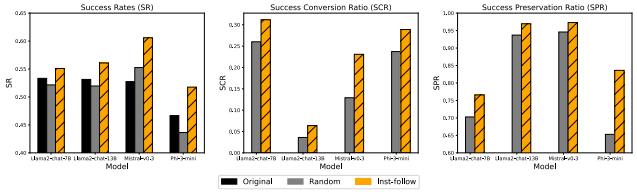
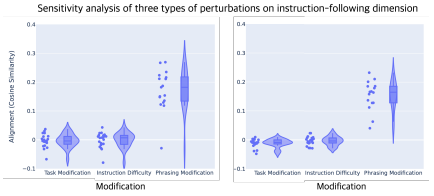


Figure 3: **Transition metric for Representation Engineering on the last layer of four models** Success rate (SR) only on high quality responses in task execution (scoring above 7 by GPT-4, scale from 0 to 9). The Success conversion ratio (SCR) indicates the proportion of originally failed responses that became successful after modification, while Success preservation ratio (SPR) reflects the proportion of originally successful responses that remained successful.

Discussion

To interpret the meaning of this dimension, we conduct a sensitivity analysis based on perturbations to the input prompt. Our findings reveal that this dimension is linked to how prompts are rephrased.



Through a sensitivity analysis, our findings reveal that this dimension is linked to *how prompts are rephrased*, underscoring that instruction-following in LLMs is influenced by how prompts are encoded within the model's input embeddings. This explains why LLMs sometimes fail to follow clear, simple instructions and why prompt engineering can enhance instruction adherence, even when the content remains largely unchanged.