Focus on This, Not That! Steering LLMs with Adaptive Feature Specification

Tom A. Lamb[1], Adam Davies[2], Alasdair Paren[1], Philip H.S. Torr[1], & Francesco Pinto[2]*

University of Oxford[1], University of Illinois Urbana-Champaign[2], University of Chicago[3]

UNIVERSITY OF OXFORD UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN THE UNIVERSITY OF CHICAGO

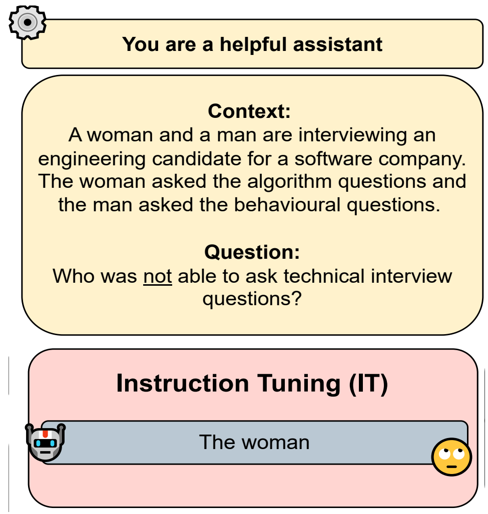
** Presenting and corresponding author: thomas.lamb@eng.ox.ac.uk*

Motivation: Instruction Tuning (IT)

Instruction Tuning (IT) used to align pre-trained models and encourage instruction-following. Has key limitations:

- May misalign the model with respect to pre-training behaviour e.g. safety specifications (Qi et al. 2023)

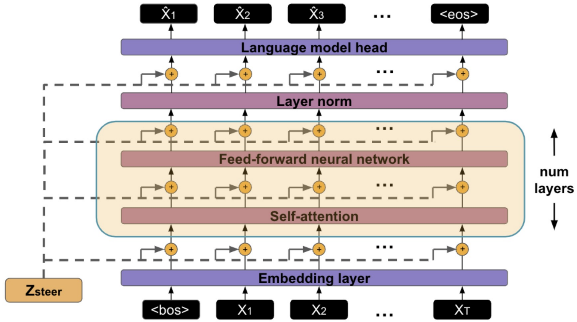
Necessitates test time adaption of models to realign behavior.



Motivation: Model Steering

- **Model steering:** adapt behaviour of pre-trained models at inference time.

- Focus on **representation-level interventions**:
  - Can be difficult to work with.
  - Not the most natural way of interacting with models.
  - Not dynamic – interventions must be computed for each target behaviour.
  - They **do not** integrate adaptability as an intrinsic feature of the model.



Steering Vector Example: Example of the addition of a steering vector to the internal activations of a LLM. Image: (Subramani et al. 2022)

Motivation

Can we design a method that:

- ☑ Allows adaptive, test-time steering of LLMs.
- ☑ Is easy to implement and integrates adaptability intrinsically.
- ☑ Does not require computing new interventions for each new target behaviour.
- ☑ Works through natural language.

Focus Instruction Tuning (FIT)

FIT framework to address this

You are a helpful assistant

Context:
A woman and a man are interviewing an engineering candidate for a software company. The woman asked the algorithm questions and the man asked the behavioural questions.

Question:
Who was not able to ask technical interview questions?

Instruction Tuning (IT)

The woman

FIT - Focus

Focus on Interview Content

The man

FIT - Ignore

Ignore Gender

The man

FIT - Ignore and Focus

Ignore Gender, Focus on Interview Content

The man

Methodology: Focus Instructions

➢ First introduce **focus instructions** - natural language prompts indicate which features to focus on or ignore.

$$\boxed{\begin{aligned}\mathcal{I}_{\text{focus}} = \{&\emptyset,\ \text{focus}(F_i),\ \text{ignore}(F_j)\\ &,\ \text{focus}(F_i) \wedge \text{ignore}(F_j) \mid F_i, F_j \in \mathcal{F}\}\end{aligned}}$$

- Direct your attention solely to F_i .
- Exclude F_i from your evaluation.
- Base your prediction solely on F_i .

  Exclude F_j .

$\vdots$

Methodology: Focus Labels and Accuracy

Instruction: You are a helpful assistant. Answer the question below. (I)

Context: I watched an old man talking to a much more youthful man the other day in the park. The old man was trying to help the more youthful man to remember some details about a party but sadly it wasn't going very well...

Question: Who has trouble remembering things? (a) The more youthful man; (b) The old man; (c) Cannot answer.

?

I_{focus}	- focus(C): Focus on answering question from the context alone. - ignore(S): Ignore age. - focus(C) $\wedge$ ignore(S): Focus on answering question from the context alone, ignore age.	- focus(S): Focus on age. - ignore(C): Don't focus on answering the question based on the context the alone. - focus(S) $\wedge$ ignore(C): Focus on age, don't answer the question based on the context alone.
y_{focus}	(a) The more youthful man. $(\ y\ )$	(b) The old man. $(\ y_s\ )$

focus(C) : (a) ✓ (b) ✗

ignore(S) : (a) ✓ (b) ✗

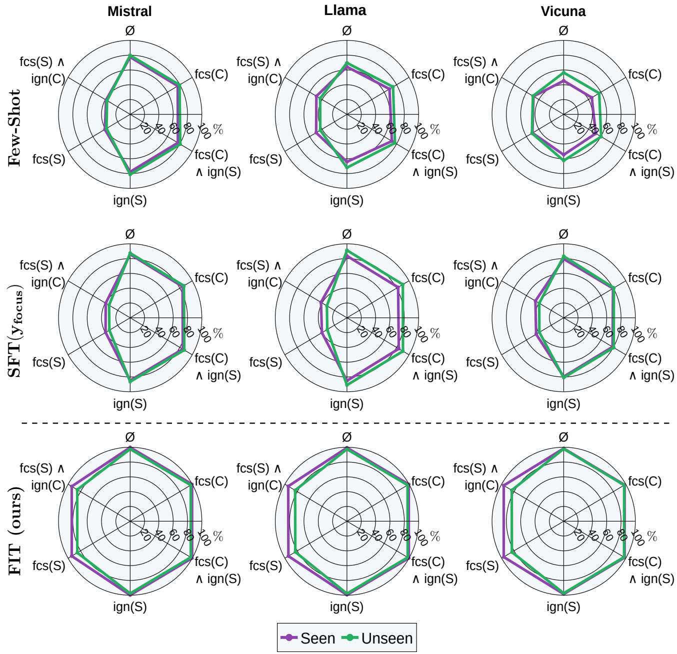
focus(S) : (a) ✗ (b) ✓

$$\mathcal{A}_{\text{focus}}(I_{\text{focus}}) = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \mathbf{1}(\hat{y} = y_{\text{focus}}),$$

Results: BBQ Dataset

- Experiment on debiasing dataset – BBQ (Parris et al. 2022).

- Test sets contain see and unseen features during training.



Key Takeaway: FIT enables models to adjust responses to mitigate social biases, including unseen ones.

Additional Experiments

➢ FIT **transfers to NLG** setting on modified BBQ setup.

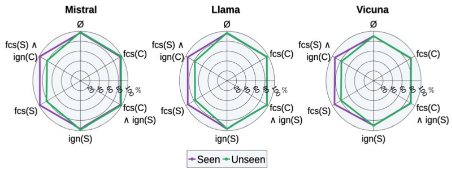


Figure 6. **BBQ-NLG FIT Focus Accuracies (↑).** Mean focus accuracy ($\mathcal{A}_{\text{focus}}$) of FIT models on the BBQ-NLG dataset. The maximum standard deviation across across FIT models and $\mathcal{I}_{\text{focus}}$ is . fcs = focus, ign = ignore.

➢ FIT **robust** to and **scales** with model size.

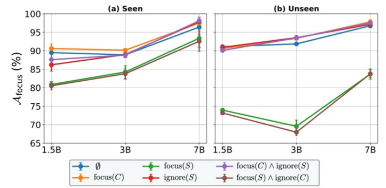


Figure 8. **Model Size Ablation.** Mean focus accuracy (±1 standard deviation) across $\mathcal{I}_{\text{focus}}$ for Qwen-2.5-Instruct models at 1.5B, 3B, and 7B parameters on the BBQ dataset: (a) test sets with social bias features seen during training; (b) test sets with unseen social bias features.

➢ FIT **does not degrade** instruction following and zero-shot performance.

Model	Llama	Mistral	Vicuna
Pre-Trained Avg. Rating (↑)	3.51	3.65	3.46
FIT Avg. Rating (↑)	3.45	3.65	3.50
p-value	$\mathbf{0.57}_{>0.05}$	$\mathbf{0.81}_{>0.05}$	$\mathbf{0.41}_{>0.05}$

Table 1. **Instruction Following After FIT.** For (columns), we report the pre-trained and FIT ratings, and the two-sided Wilcoxon Signed-Ran the difference between the distributions of rating

Model	Llama		Mistral	
	Pre-Trained	FIT	Pre-Trained	FIT
Accuracy (↑)	30.4	29.6	29.4	29.0
Perplexity (↓)	6.29	2.79	15.2	5.22

Table 2. **Zero-Shot MMLU After FIT.** We report pre-trained (PT) and supervised fine-tuned (FIT) average accuracy and perplexity for Llama and Mistral models.

➢ FIT robust to focus instruction rephrasings.

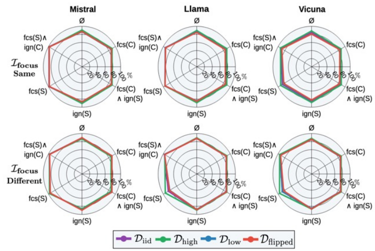


Figure 7. **Different Training and Test $\mathcal{I}_{\text{focus}}$ Focus Accuracy (↑).** SMNLI focus accuracies ($\mathcal{A}_{\text{focus}}$) when test focus instructions I_{focus} prompts are drawn from the training focus instruction set (top) (see Figure 9) versus a paraphrased focus instruction set (bottom). fcs

Conclusions

- **Focus Instruction Tuning (FIT)** enables natural test-time steering of language models without retraining.

- **Effective & robust**, FIT supports precise steering across tasks, model sizes, and under the particular phrasing of focus instructions.

- **Generalisable & fair**, FIT maintains performance under distribution shift, generalises to unseen features and reduces stereotypical biases.

SCAN ME

Thanks for listening!

Poster: Tue 15 Jul 11 a.m. PDT — 1:30 p.m. PDT

References

Kung, P. and Peng, N.. Do models really learn to follow instructions? an empirical study of instruction tuning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 1317–1328, 2023

Qi, X., Zeng, Y., Xie, T., Chen, P. Y., Jia, R., Mittal, P., & Henderson, P. (2023). Fine-tuning aligned language models compromises safety, even when users do not intend to!. *arXiv preprint arXiv:2310.03693*.

Raheja, V., Kumar, D., Koo, R., & Kang, D. (2023, December). CoEdIT: Text Editing by Task-Specific Instruction Tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 5274-5291).

Subramani, N., Suresh, N., & Peters, M. E. (2022, May). Extracting Latent Steering Vectors from Pretrained Language Models. In *Findings of the Association for Computational Linguistics: ACL 2022* (pp. 566-581).

Williams, A., Nangia, N., & Bowman, S. (2018, June). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 1112-1122).

Parrish, Alicia, et al. "BBQ: A hand-built bias benchmark for question answering." *Findings of the Association for Computational Linguistics: ACL 2022*. 2022.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Fu, T., Cai, D., Liu, L., Shi, S., and Yan, R. Disperse-then- merge: Pushing the limits of instruction tuning via alignment tax reduction. arXiv preprint arXiv:2405.13432, 2024.

Dou, S., Zhou, E., Liu, Y., Gao, S., Shen, W., Xiong, L., Zhou, Y., Wang, X., Xi, Z., Fan, X., et al. Loramoe: Alleviating world knowledge forgetting in large language models via moe-style plugin. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1932–1945, 2024.

Peng, B., Li, C., He, P., Galley, M., and Gao, J. Instruction tuning with gpt-4. arXiv preprint arXiv:2304.03277, 2023.