Inference-time Intervention

Eliciting Truthful Answers from a Language Model

Li et al. 2023

Setting the stage

- Language models become ubiquitous in the span of 5 years
- They're unreasonably effective at general purpose tasks, so long as you scale them and their datasets to be large enough
- We also have a set of post-training techniques that make them even more powerful and accessible to the end-user (RLHF, SFT on downstream tasks, prompting techniques)
- But LMs occasionally output false statements, ranging from small mistakes to full outright "hallucinations" – elaborate stories that are factually incorrect

Truthfulness

- Truth is a difficult concept to pin down, especially as a training objective
- Most of the techniques we use have subtle failure modes
  - imitation learning? you might learn common misconceptions
  - RLHF? humans may not be able to distinguish the truth
- How do you get models to output **true** things?

Truthfulness

- It turns out that models encode something like “truth” in their internal representations
    - it makes sense: as a feature, truth is useful in many types of tasks
- We know this, because models are often able to critique their own answers after the fact (**generator-discriminator gap**)
    - If they didn’t contain a concept ~“truth” this would not be possible
- It just isn’t straightforward to get models themselves to use this latent structure to generate true answers

OK, now what?

- Bypass model outputs completely, and use the internal representation to generate an output
- This is what Burns et al. 2022 do with **contrast-consistent search**
- The idea is that wherever “truth” is represented internally, it has to follow logical consistency in a way that other features do not
    - we can find that in a non-supervised way, with pairs of contrasting statements
    - …

OK, now what?

- What if you could instead
  - detect the “truth” direction within internal activations
  - make models more truthful overall by shifting activations along that direction?
- This is what **inference-time intervention** is, in a nutshell

Detecting truth

- Given a transformer-based language model, a logical place to look for truth as a feature is in the **residual stream**
  - conceptually, each layer reads from the residual stream, does some operation, and writes it back to the stream
- Usually one transformer block is one multi-headed attention layer followed by an MLP/fully-connected layer
  - let’s consider the outputs of individual **attention heads** in the multi-headed attention layer

Detecting truth

- The output at layer $l + 1$ is:
- $x_{l+1} = x_l + \sum_{h=1}^{H} Q_l^h \text{Att}_l^h(P_l^h x_l)$
    - P projects the input to a D-dimensional head space, Q projects it back to the hidden dimension
    - Att is a shorthand for the attention mechanism – the specifics are not important here
    - there are $h = \overline{1,H}$ attention heads

Detecting truth

- For each of these attention heads, we can train a linear probe on their outputs
- A linear probe is a simple classifier
  - $p_\theta(x_l^h) = \sigma(\langle \theta, x_l^h \rangle)$, with σ denoting the sigmoid function, and $\theta \in \mathbb{R}^D$ a trainable weight
- We train this probe on a modified TruthfulQA dataset, on pairs
  - (question + answer, truth value)

This is the Way

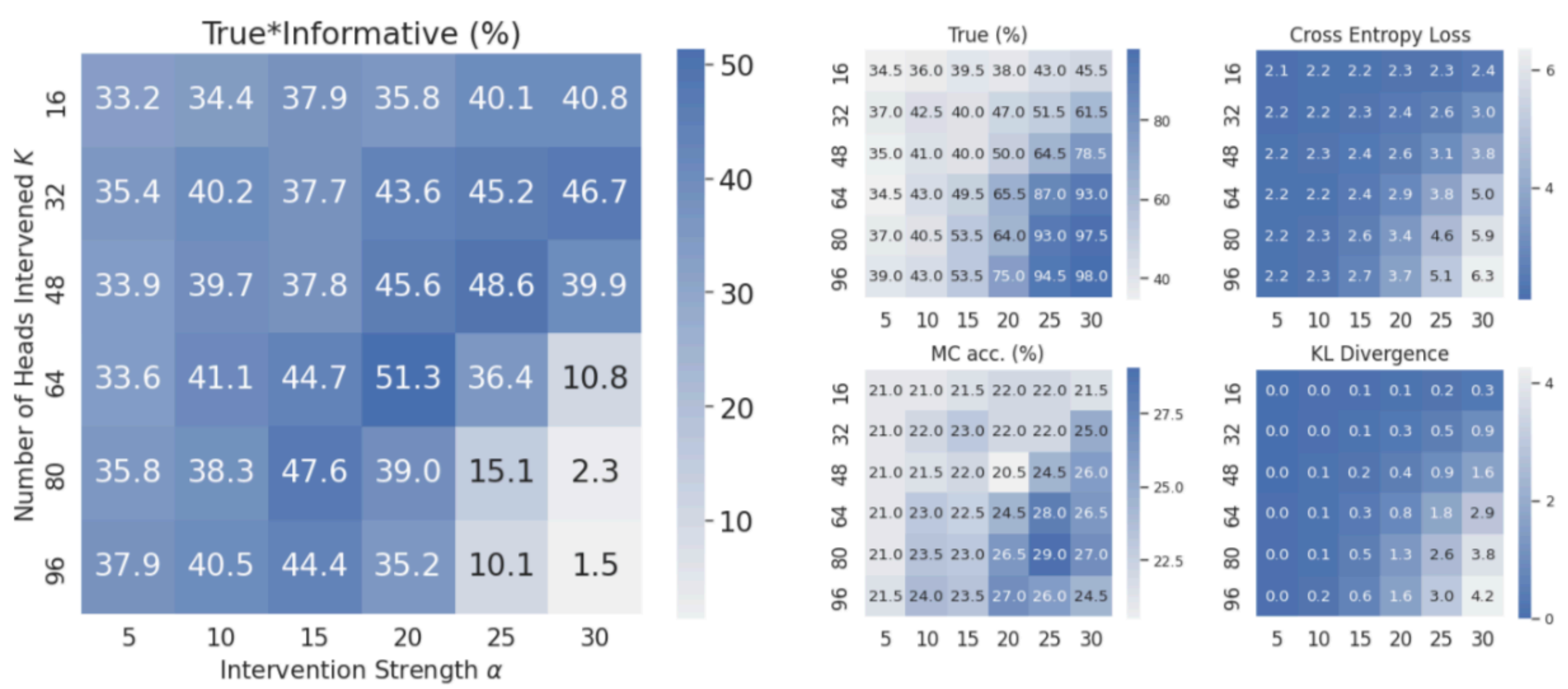
- after each probe is trained, test it on the validation set
- some heads get high accuracy, some don't – those which have high accuracy are involved in generating truthful answers
- for trained probes, we can think of the direction of the parameter θ_l^h as the *first truthful direction*
  - i.e. the direction along which true and false are most separable
- you can train a second linear probe $p_{\theta'}$ with the constraint that $\theta' \perp \theta$ to get a *second direction* (very similar to PCA)

Finally, inference-time intervention

- Given these directions defined by θ, θ', we can for each attention head shift the activations to make the model more truthful, by modifying the formula from earlier:
- $$x_{l+1} = x_l + \sum_{h=1}^{H} Q_l^h(\text{Att}_l^h(P_l^h x_l) + \alpha\sigma_l^h\theta_l^h)$$
- Here, σ refers to the standard deviation of the activations in x_l – we would not want to shift it by too much, so we refer to the initial distribution for a sensible value
- α is a hyperparameter that controls the strength of the intervention

Finally, inference-time intervention

- In practice, we don’t update all attention heads; we take the top-K heads which are most “active” in truthfulness, as measured by their probes’ validation accuracies
    - high validation accuracy means that probe is classifying truth-falsehood; low means it’s doing something else
- The reason to do this is that sparse interventions are less likely to harm overall model performance
    - We don’t want truthful-but-uninformative



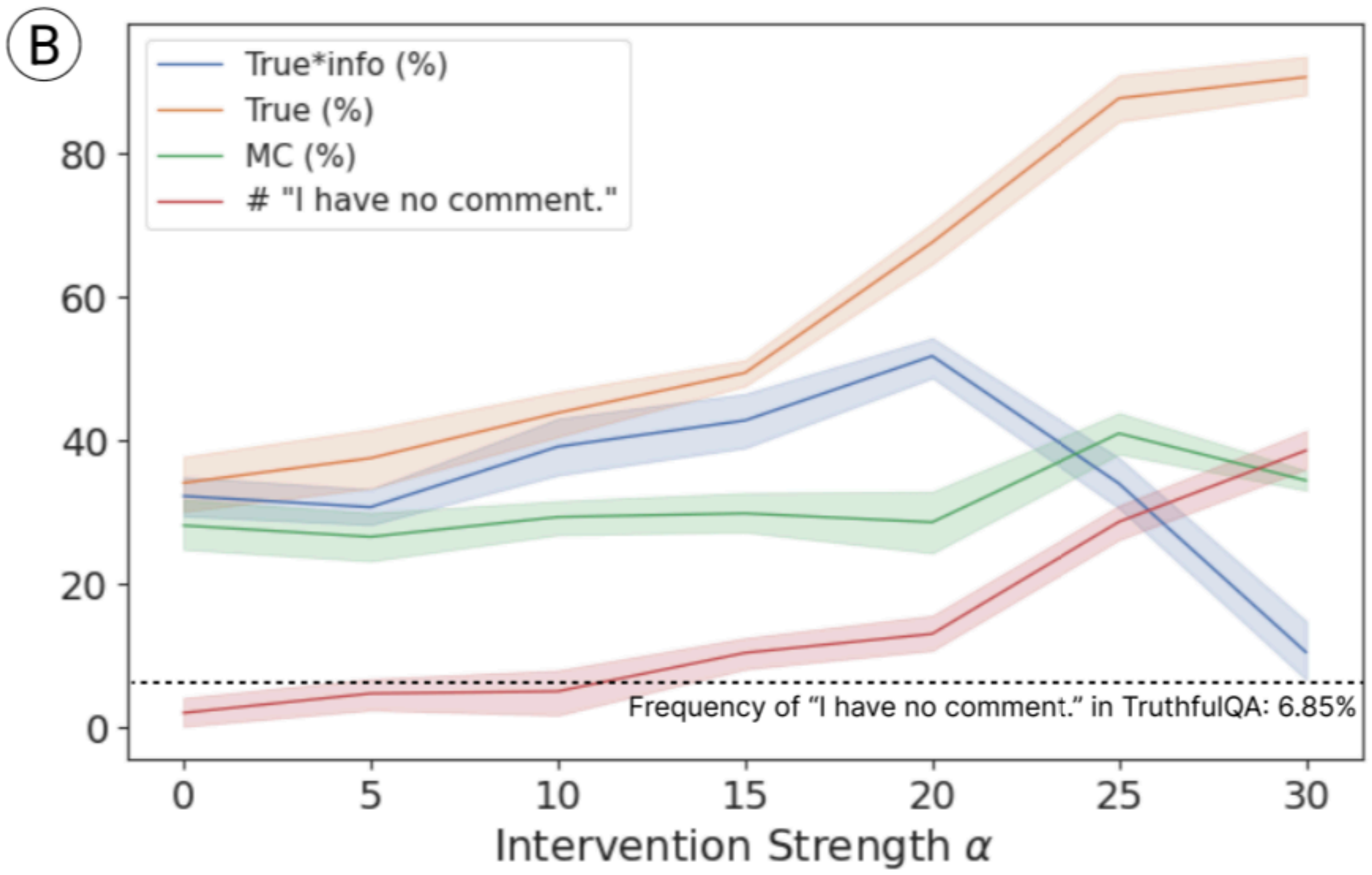


Figure 4: Results with varying intervention strength (α and K) on LLaMA-7B. 5% of questions used for training and validation, respectively. Metrics have been averaged over 5 random seeds.

	True*Info (%)	True (%)	MC acc. (%)	CE	KL
Baseline	30.5	31.6	25.7	2.16	0.0
Supervised Finetuning	36.1	47.1	24.2	2.10	0.01
Few-shot Prompting	49.5	49.5	**32.5**	-	-
Baseline + ITI	43.5	49.1	25.9	2.48	0.40
Few-shot Prompting + ITI	**51.4**	**53.5**	**32.5**	-	-

Table 1: Comparison with baselines that utilize 5% of TruthfulQA to make LLaMA-7B more truthful. CE is the pre-training loss; KL is the KL divergence between next-token distributions pre- and post-intervention. Results are averaged over three runs. We report standard deviations in Appendix D.

Conclusion

- We now have a drop-in change to make models more truthful
  - you can apply this to any LM where you have access to the weights and activations
- We have another piece of evidence that models do encode latent structure that corresponds to real-world concepts, like truth
  - it looks like it's not just a direction, but a subspace of our residual stream/activation space

Limitations

- Supervised method: you need a few annotated data points to train the linear probes
  - not so many, since effectiveness plateaus early
- Has to be sparse, otherwise overall performance is worse (Table 5)
- Fundamental trade-off between truthfulness and informativeness (Figure 6)
- Generalisation to other datasets is key to this being a useful intervention
  - seems like performance not harmed on MMLU, TriviaQA, but more needed

Limitations

- The paper reports cross-entropy and KL-divergence as metrics for how much ITI changes model behaviour
  - lower is better – the model is more truthful, but not less capable in other ways
  - but no contextualisation of KL/CE values reported: how much is a lot?
  - also, these are not sufficient, we should check that impact on downstream tasks is not harmed by ITI
- seems like most of the improvement in the best case (few-shot + ITI) comes from few-shot prompting
  - 30.5% to 49.5% with just few-shot, 43.5% with just ITI, 51.4% with both