Attention is not not explanation

a.k.a. some things Mark said were spurious

Sarah Wiegrieffe & Yuval Pinter

Sidharrth Nagappan
University of Cambridge

UNIVERSITY OF CAMBRIDGE

Apophatic Argument

Apophatic Argument

Apophatic theology is a form of theological thinking and religious practice which attempts to approach an unknown enigma, by **negation, to speak only in terms of what may not be said about the perfect goodness that is the enigma**.

The Enigma is **Attention.**

By the end of the presentation, you'll understand
why this is, at best, an apophatic argument.

Agreements & Disagreements

- Attention is not always a perfect, “faithful” explanation
- Some tasks barely need attention

- Multiple explanations do not mean attention fails
- Detaching attention from the rest of the model is an unfair test.

UNIVERSITY OF CAMBRIDGE

Let’s break down each fundamental claim made by Jain & Wallace

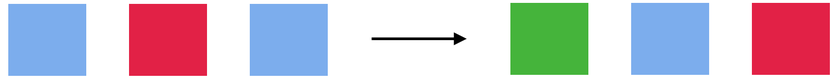
Jain & Wallace: If you can find multiple diverse attention distributions for the same final prediction, then attention cannot be a faithful explanation.

Jain & Wallace: If you can find multiple diverse attention distributions for the same final prediction, then attention cannot be a faithful explanation.

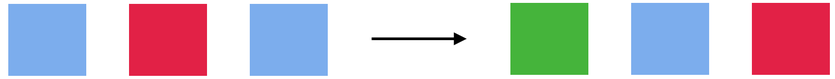
Counter: Attention is *one* plausible explanation, not the only explanation.

An explanation need not be unique to be helpful.

Jain & Wallace: Attention weights can be trained, manipulated and re-assigned as independent, standalone units



Jain & Wallace: Attention weights can be trained, manipulated and re-assigned as independent, standalone units



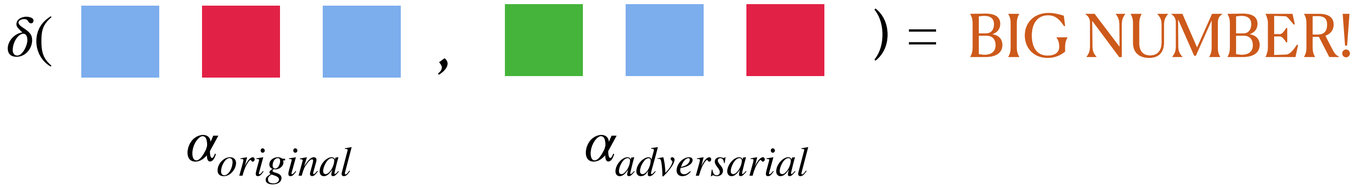
Counter: Treating attention scores in isolation ignores the fact that the mechanism is jointly trained with the rest of the network.

You can't just cut out the attention and shuffle it — you're destroying neural links with other layers.

An adversarial attack should adjust all relevant parameters, not just the attention vector.

This is not some magical latent distribution.

Jain & Wallace: Adversarial training shows attention is easily manipulable. We create adversarial attention distributions that maintain model accuracy but look completely different.



Jain & Wallace: Adversarial training shows attention is easily manipulable. We create adversarial attention distributions that maintain model accuracy but look completely different.

Counter 1: Baseline variance matters—some attention drift is natural.

Even training runs with different seeds cause attention to vary.

J&W must consider what is baseline natural variance and what is adversarial variance, to see how *surprising* or *harmful* they actually are.

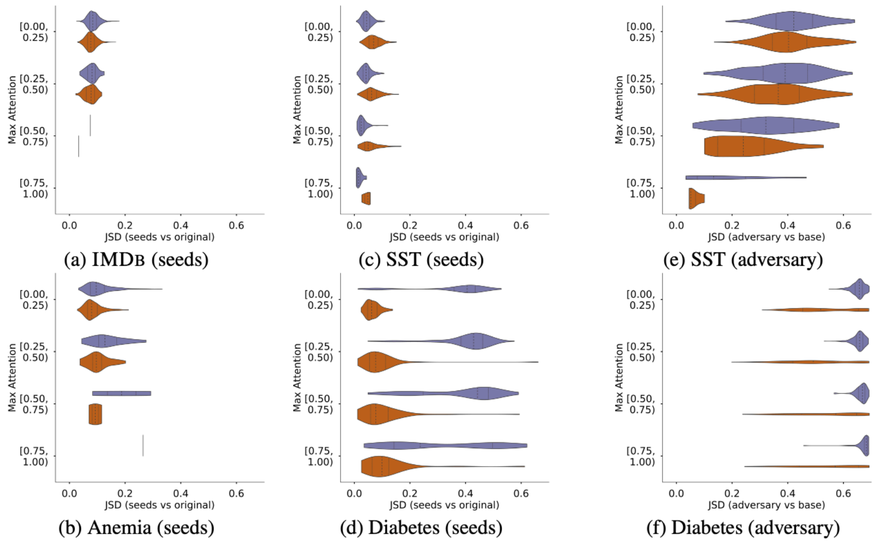


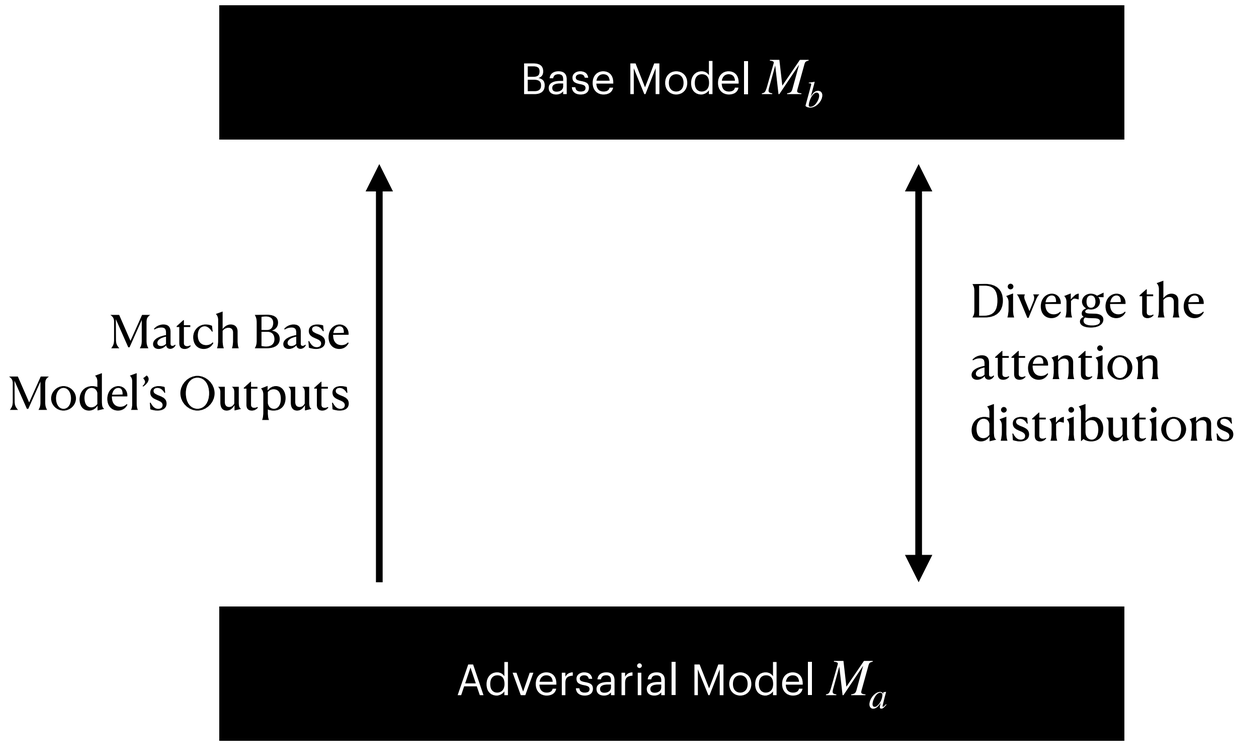
Figure 3: Densities of maximum JS divergences (x-axis) as a function of the max attention (y-axis) in each instance between the base distributions and: (a-d) models initialized on different random seeds; (e-f) models from a per-instance adversarial setup (replication of Figure 8a, 8c resp. in Jain and Wallace (2019)). In each max-attention bin, top (blue) is the negative-label instances, bottom (red) positive-label instances.

Jain & Wallace: Adversarial training shows attention is easily manipulable. We create adversarial attention distributions that maintain model accuracy but look completely different.

Counter 2: You are modifying attention vectors on an instance-by-instance basis. **You are cheating by creating a different, tailored adversarial distribution per-example**

Proposition: Model-Consistent Training → A *single* set of model parameters must produce *all* adversarial attentions across the entire dataset

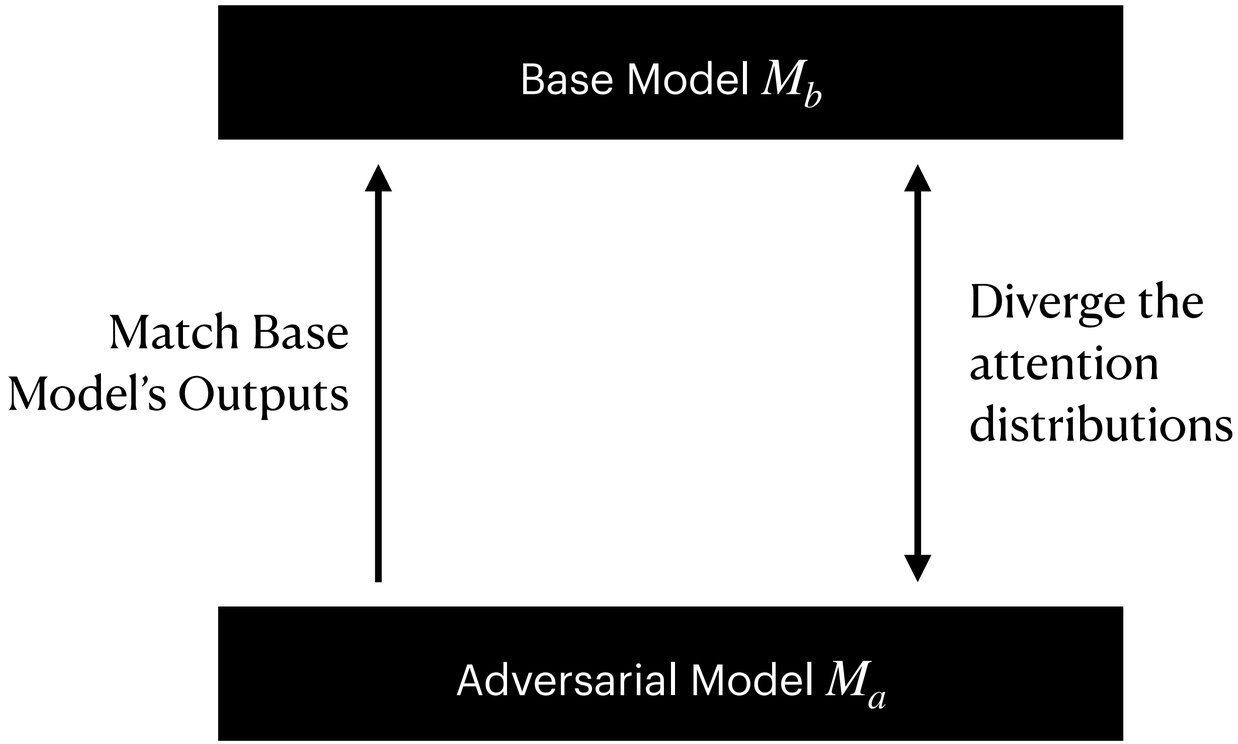
Model-Consistent Training



$$L(M_a, M_b) = \text{TVD}(\hat{y}_a, \hat{y}_b) - \lambda \text{KL}(\alpha_a \| \alpha_b)$$

UNIVERSITY OF CAMBRIDGE

Model-Consistent Training



When you do this **correctly,** attention doesn't shift so aggressively

If the way you generate adversarial attention is spurious, then any results from those adversarial vectors is also spurious!

$$L(M_a, M_b) = \text{TVD}(\hat{y}_a, \hat{y}_b) - \lambda \text{KL}(\alpha_a \| \alpha_b)$$

UNIVERSITY OF CAMBRIDGE

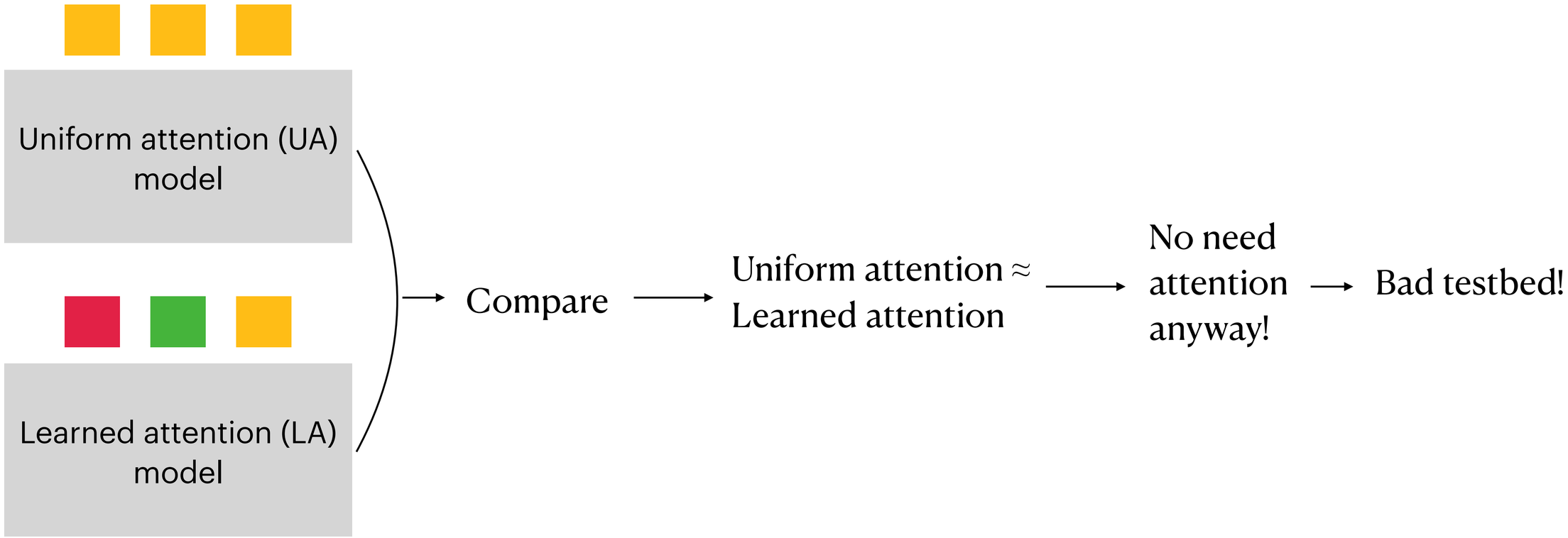
Jain & Wallace: If attention really is that important → then altering it in certain ways should degrade performance

But it didn't always → so attention may not be important.

Jain & Wallace: If attention really is that important → then altering it in certain ways should degrade performance

But it didn't always → so attention may not be important.

Counter: Sometimes, a model just doesn't need attention — bad testbed!



Jain & Wallace: Attention does not correlate with gradient-based/leave-one-out feature importance, so it's a subpar representation of token importance.

Jain & Wallace: Attention does not correlate with gradient-based/leave-one-out feature importance, so it's a subpar representation of token importance.

Counter: They are different neural representations. Attention means something. When an MLP only sees attention weights, it still performs better than uniform / random initialisation.

Freeze the RNN's learned attention weights and use them in a simpler MLP. → If these weights truly highlight important tokens, performance should exceed uniform or random weighting → And it does! → Attention captures meaningful token importances!

Guide weights	Diab.	Anemia	SST	IMDb
UNIFORM	0.404	0.873	0.812	0.863
TRAINED MLP	0.699	0.920	0.817	0.888
BASE LSTM	**0.753**	0.931	**0.824**	**0.905**
ADVERSARY (4)	0.503	**0.932**	0.592	0.700

Table 3: F1 scores on the positive class for an MLP model trained on various weighting guides. For ADVERSARY, we set $\lambda \leftarrow 0.001$.

Positives

- Methodology is very rigorous → paper carefully and empirically attacks each claim made by Jain & Wallace
- Model-consistent adversarial training is a good contribution
- Very straightforward diagnostics and baselines that add context to Jain & Wallace's results
- Doesn't make overly bold claims → acknowledges that there are limits to attention's explainability
- Pretty solid paper — almost as well-cited as the original → *says a lot*

UNIVERSITY OF CAMBRIDGE

Positives & Negatives

- Methodology is very rigorous → paper carefully and empirically attacks each claim made by Jain & Wallace
- Model-consistent adversarial training is a good contribution
- Very straightforward diagnostics and baselines that add context to Jain & Wallace's results
- Doesn't make overly bold claims → acknowledges that there are limits to attention's explainability
- Pretty solid paper — almost as well-cited as the original → *says a lot*

- A bit confusing → Argues that attention is meaningful, but this doesn't necessarily make it a good *explanation*
  - Instead of **proving** that attention is explanatory, the paper focuses on **disproving** claim that it is *not* explanatory by attacking Jain & Wallace's methodology
  - What even is explanation?
  - So many survey papers came after questioning the role of attention - nitpicking waste of academic resources?

Remember Apophatic Theology

UNIVERSITY OF CAMBRIDGE

Attention is all you need it to be

It’s not a perfect explanation, but it’s somewhat meaningful

...if that matters to you!

Thank you!