Answer, Assemble, Ace: Understanding How LMs Answer Multiple Choice Questions

Sarah Wiegreffe[1,2] Oyvind Tafjord[1] Yonatan Belinkov[3] Hannaneh Hajishirzi[1,2] Ashish Sabharwal[1]
[1]Allen Institute for AI [2]University of Washington [3]Technion

Motivation

- Much interpretability work has focused on tasks like factual recall or indirect object identification. What about a format that is commonly used in real-world benchmarks (**multiple-choice question answering, or MCQA**)?
- Prior work (Pezeshkpour & Hruschka 2023, Alzahrani et al. 2024, Khatun & Brown 2024) has shown that models are not always robust to perturbations such as:

Change answer position:

Which of the following is not a way to form recombinant DNA?

Choices:
A. Translation
B. Conjugation
C. Specialized transduction
D. Transformation

The correct answer is: **A**

Which of the following is not a way to form recombinant DNA?

Choices:
A. Conjugation
B. Translation
C. Specialized transduction
D. Transformation

The correct answer is: **B**

Change answer symbols:

Which of the following is not a way to form recombinant DNA?

Choices:
Q. Translation
Z. Conjugation
R. Specialized transduction
X. Transformation

The correct answer is: **Q**

Main Research Question: How do models *robustly* answer multiple-choice questions?

Our Dataset

CopyColors is designed to **disentangle formatted MCQA ability from task- or domain-specific knowledge.**

We include prompt variants to test models' robustness.

A banana is yellow.
What color is a banana?

Choices:
A. pink
B. yellow
C. black
D. blue

The correct answer is: **B**

A banana is yellow.
What color is a banana?

Choices:
A. yellow
B. pink
C. black
D. blue

The correct answer is: **A**

A banana is yellow.
What color is a banana?

Choices:
Q. yellow
Z. pink
R. black
X. blue

The correct answer is: **Q**

Note: all models tested get 100% accuracy on a 3-shot generative version of the task:

A banana is yellow. What color is a banana? **yellow**

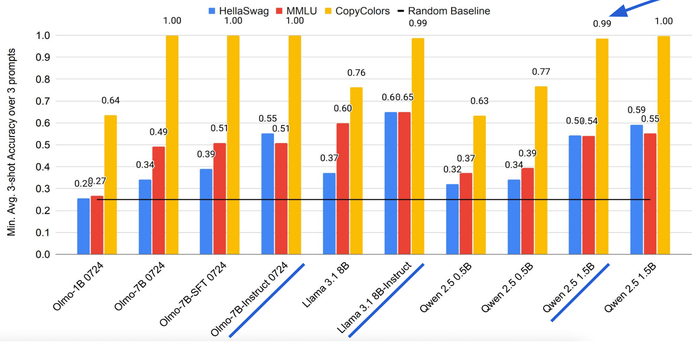
Co-funded by the European Union

Partially funded by the European Union (ERC, Control-LM,101165402). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

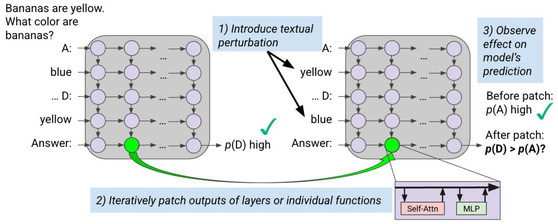
Methods

Goal: attribute model behavior to specific internal mechanisms.

We test 3 representative models performant on CopyColors (underlined):

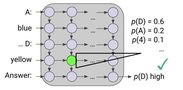


Attribution Patching (Geiger et al 2020, Vig et al 2020): Measures causal effect, under some independence assumptions of hidden states.



- We patch hidden states output by each layer, MLP, self-attention function, and individual attention head at the last token position across paired inference runs.
- The most important hidden states will have the largest effect in "restoring" the probabilities of the run that is being patched in.

Vocabulary Projection (Geva et al 2021, nostalgebraist 2020): Measures how predictions form in the model's vocabulary space over layers, *to the extent they are* linearly decodable using the unembedding matrix.



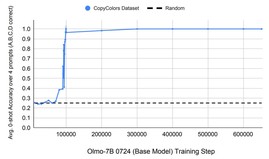
- Perform layer normalization on each hidden state & project to vocabulary space using the model's unembedding matrix.
- Akin to early exiting on Transformer block.

Logits or Probits? Both! Logits can detect model components responsible for *demotion* of predicted answers that would otherwise be normalized by Softmax to 0. Probits provide a sense of score magnitude of a token *relative to* other tokens in the vocabulary, thus conveying information about rank. Also allow comparison across models.

Results

CopyColors **disentangles ability to answer formatted MCQA questions from dataset-specific performance.**

It also shows **when** models learn to answer formatted MCQA: **fairly early** in training.



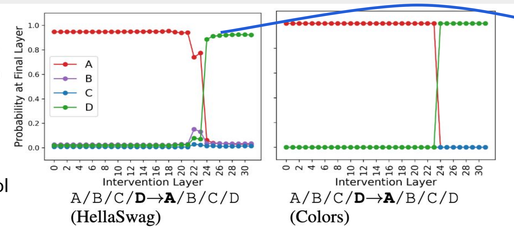
Models robustly answer multiple-choice questions largely in 3 stages.

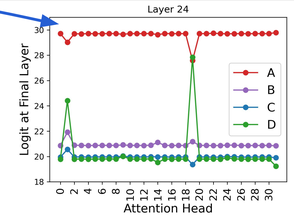
Behaviors differ ***across models*** but are largely consistent ***across MCQA datasets.***

1. **Answer Selection** occurs at middle layers, driven by **a few attention heads.**

E.g., Activation Patching shows Layer 24, Head 19 for Olmo 7B Instruct position swaps (1.28% of model's weights):

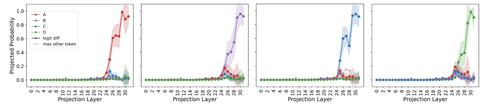
See bottom for symbol swaps (layer 29).

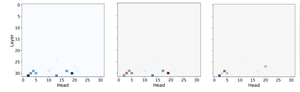




2. Subsequent layers **increase the probability** of the predicted answer symbol in vocabulary space (driven by a few specialized attention heads).

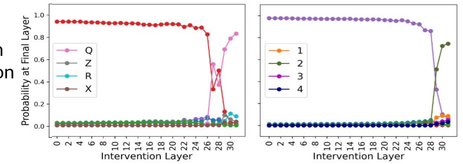
E.g., Vocabulary projection on Olmo 7B Instruct, HellaSwag dataset

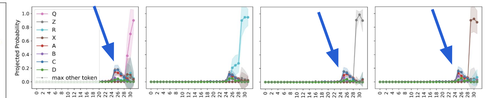




3. Some models adjust to unusual answer choice symbols (e.g., Q/Z/R/X) by **first operating in the space of familiar answer symbols (A/B/C/D), then assigning probability to the correct symbols at a later layer.**

E.g., Activation Patching on Olmo 7B Instruct: Layer 29





Vocab projection on Olmo 7B Instruct, HellaSwag

See paper for results on more models.

Motivation

- Much interpretability work has focused on tasks like factual recall or indirect object identification. What about a format that is commonly used in real-world benchmarks (**multiple-choice question answering, or MCQA**)?
- Prior work (Pezeshkpour & Hruschka 2023, Alzahrani et al. 2024, Khatun & Brown 2024) has shown that models are not always robust to perturbations such as:

Change answer position:

Which of the following is not a way to form recombinant DNA?

Choices:
A. Translation
B. Conjugation
C. Specialized transduction
D. Transformation

The correct answer is: **A**

Which of the following is not a way to form recombinant DNA?

Choices:
A. Conjugation
B. Translation
C. Specialized transduction
D. Transformation

The correct answer is: **B**

Change answer symbols:

Which of the following is not a way to form recombinant DNA?

Choices:
Q. Translation
Z. Conjugation
R. Specialized transduction
X. Transformation

The correct answer is: **Q**

Main Research Question: How do models *robustly* answer multiple-choice questions?

Our Dataset

CopyColors is designed to **disentangle formatted MCQA ability from task- or domain-specific knowledge.**

We include prompt variants to test models' robustness.

A banana is yellow.
What color is a banana?

Choices:
A. pink
B. yellow
C. black
D. blue

The correct answer is: **B**

A banana is yellow.
What color is a banana?

Choices:
A. yellow
B. pink
C. black
D. blue

The correct answer is: **A**

A banana is yellow.
What color is a banana?

Choices:
Q. yellow
Z. pink
R. black
X. blue

The correct answer is: **Q**

Note: all models tested get 100% accuracy on a 3-shot generative version of the task:

A banana is yellow. What color is a banana? **yellow**

Co-funded by the European Union

Partially funded by the European Union (ERC, Control-LM,101165402). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.