Reason to Rote: Rethinking Memorization in Reasoning

Yupei Du[13], Philipp Mondorf[2], Silvia Casola[2], Yuekun Yao[3], Robert Litschko[2], and Barbara Plank[2]

Utrecht University[1], LMU Munich[2], and Saarland University[3]






The curious case of benign memorization

Benign memorization of deep neural nets (Zhang et al., 2016)

Our central finding can be summarized as:

Deep neural networks easily fit random labels.

.. and yet

*it is **unlikely that the regularizers are the fundamental reason for generalization**, as the networks continue to perform well after all the regularizers removed*

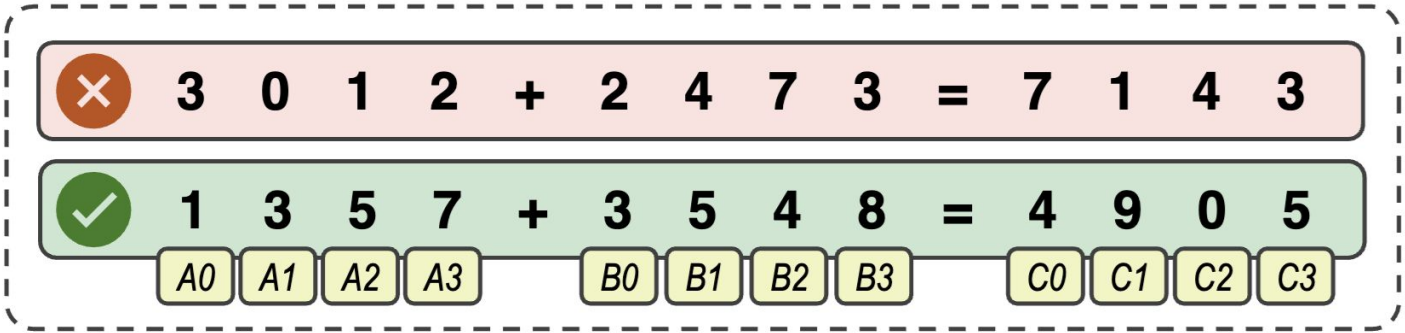
Understanding deep learning requires rethinking generalization. Zhang et al., 2016

Understanding benign memorization in reasoning

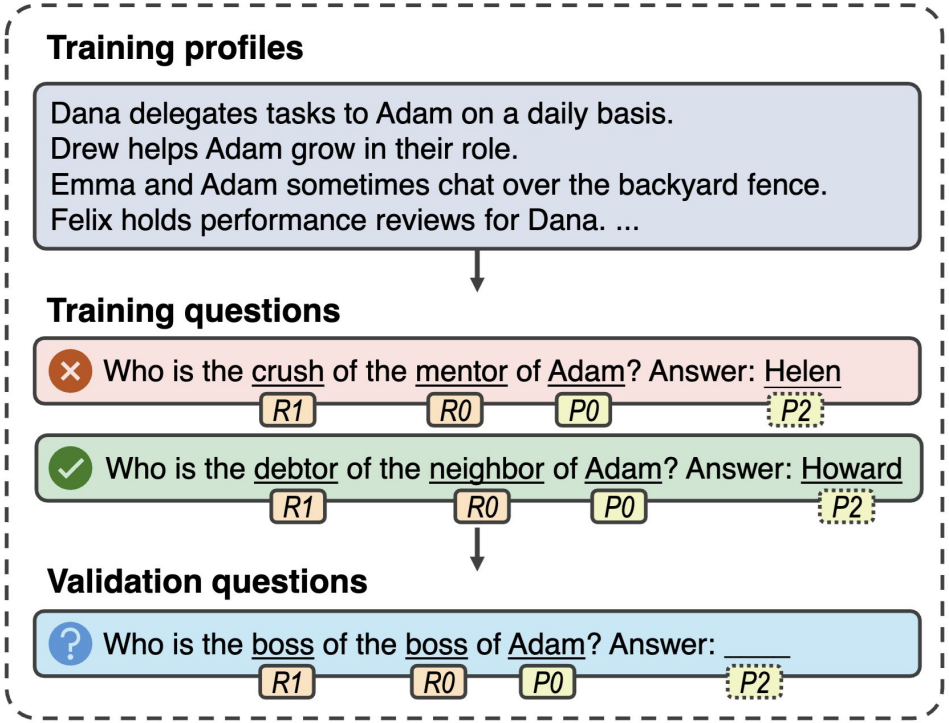
Use of controllable reasoning tasks:

1. **Understanding** of the reasoning process:
   which are the important tokens, which are the important intermediate steps, and what is the correct answer
2. The reasoning process is **manipulatable**:
   you can modify the intermediate step to (hopefully) elicit an alternative answer

Reasoning task: four-digit addition (FDA)

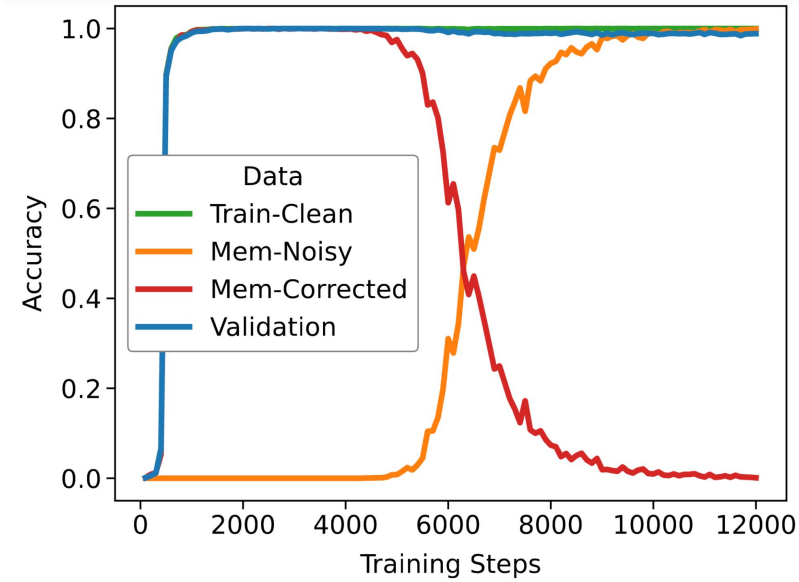


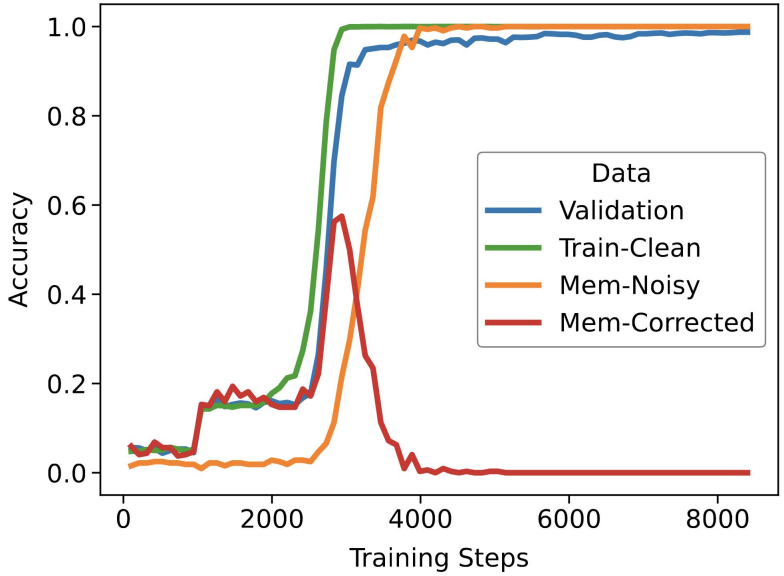
Reasoning task: two-hop reasoning (THR)



Warmup: first generalize, then memorize

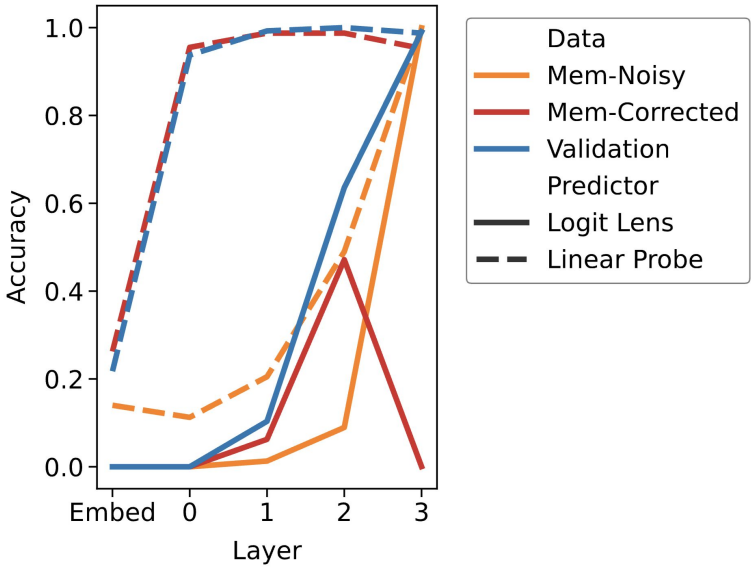
Models first predict the clean labels on noisy training instances

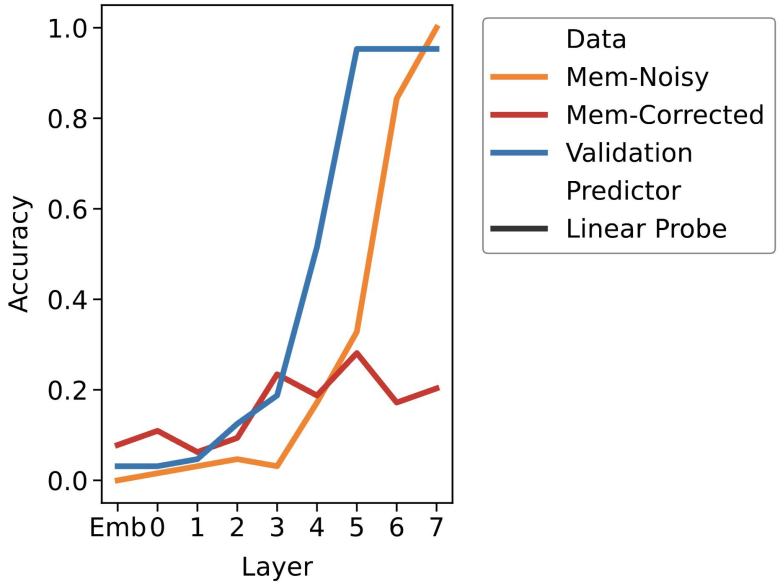




Both mechanisms present in memorization

Models compute both results, and produce memorized noisy labels later



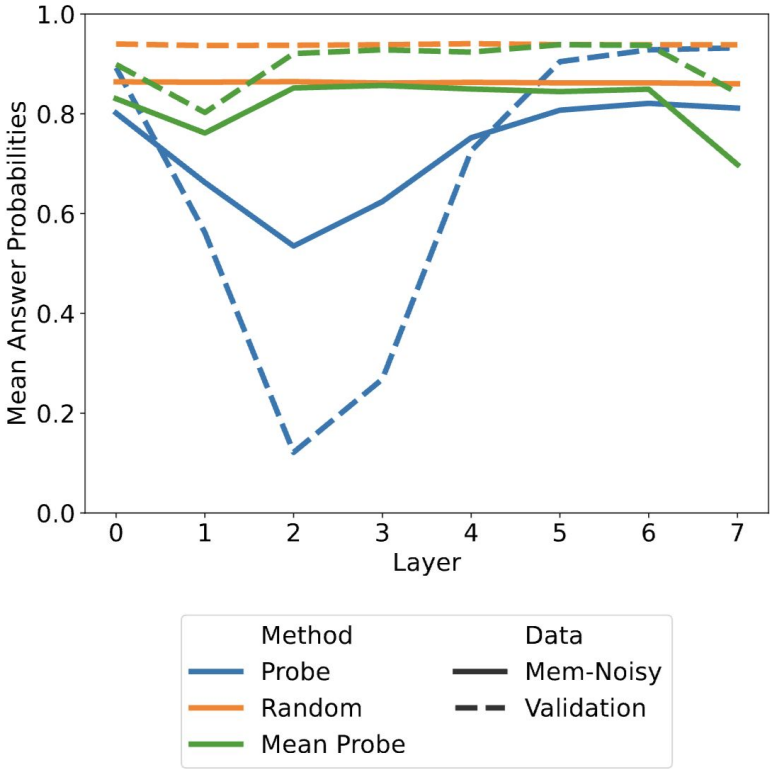


Memorization relies on generalization

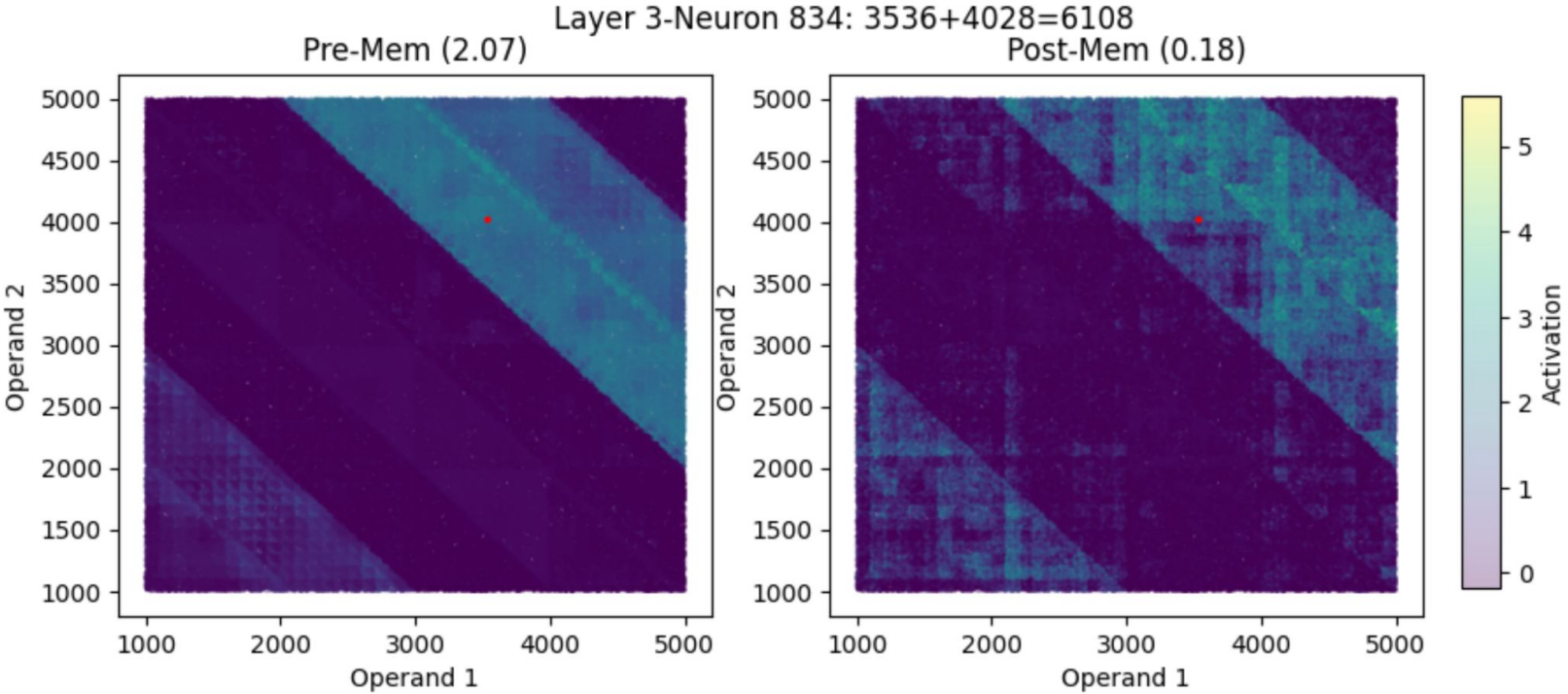
Removing bridge entities hurts both memorization and generalization

e.g., knowing facts
A's mentor is B; B's mentor is C;
but needs to memorize
"Who is the mentor of the mentor of A"
to be D (should be C).

In this situation, **B needs to be inferred in hidden layers as well to recall D**



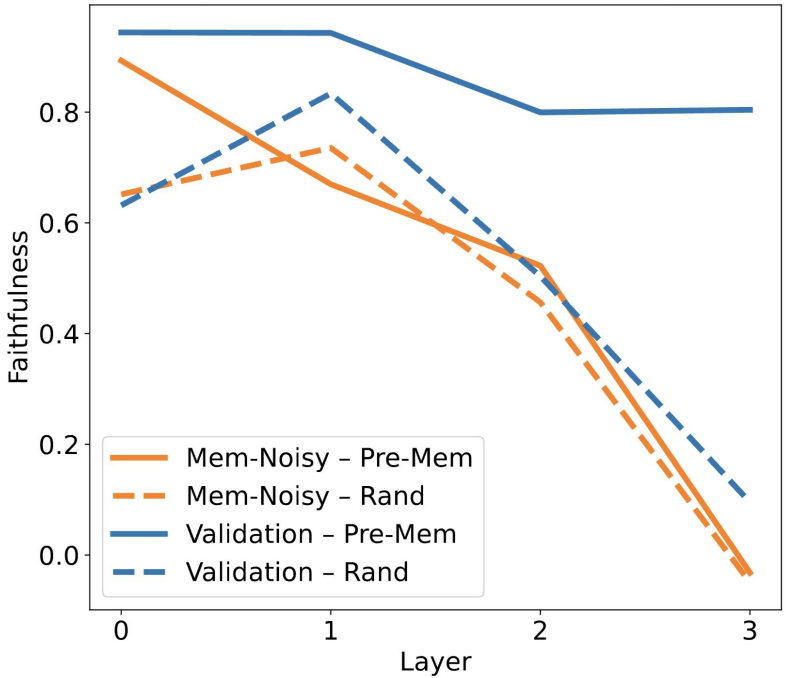
FDA model memorizes noise through outlier heuristics



FDA model memorizes noise through outlier heuristics

Layer-wise activation patching of pre-memorization activations to post-activation model, ***without changing any parameters!***

- Validation faithfulness is barely influenced (validation accuracy is restored to 100%)
- Memorization faithfulness is severely undermined



Conclusions

1. Language models memorize noise by building on generalization mechanisms: this explains why such memorization is usually benign
2. In general, our study reveals the inductive bias of reusing existing structures to learn new things