Dual Process Learning

Controlling Use of In-Context vs. In-Weights Strategies with Weight Forgetting

Suraj Anand, Michael A. Lepori, Jack Merullo

Ellie Pavlick

BROWN

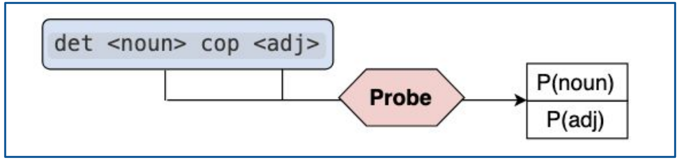
Structural In-Context Learning

Ability of a model to execute in-context algorithms on arbitrary novel tokens

Structural In-Context Learning

Ability of a model to execute in-context algorithms on arbitrary novel tokens

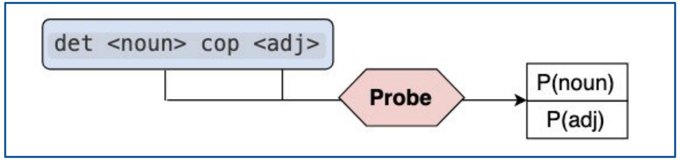
1. **MLM Syntax Task**



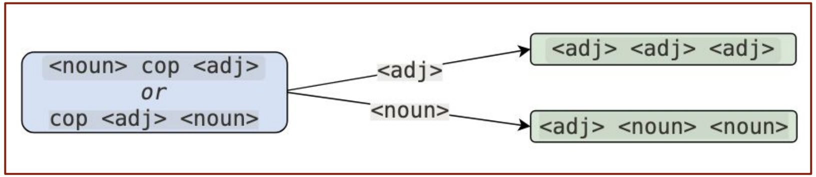
Structural In-Context Learning

Ability of a model to execute in-context algorithms on arbitrary novel tokens

1. **MLM Syntax Task**



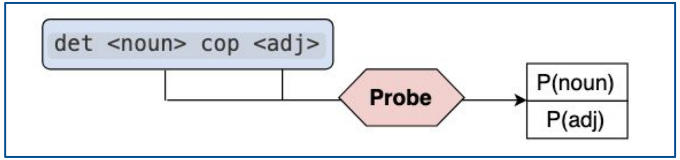
2. **MLM Synthetic Task**



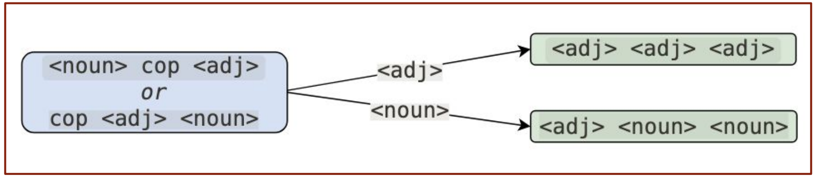
Structural In-Context Learning

Ability of a model to execute in-context algorithms on arbitrary novel tokens

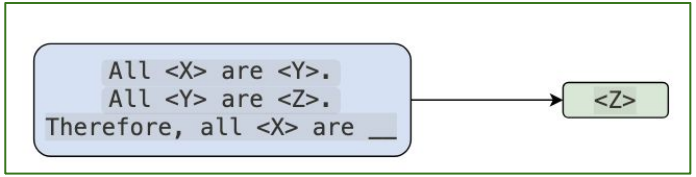
1. **MLM Syntax Task**



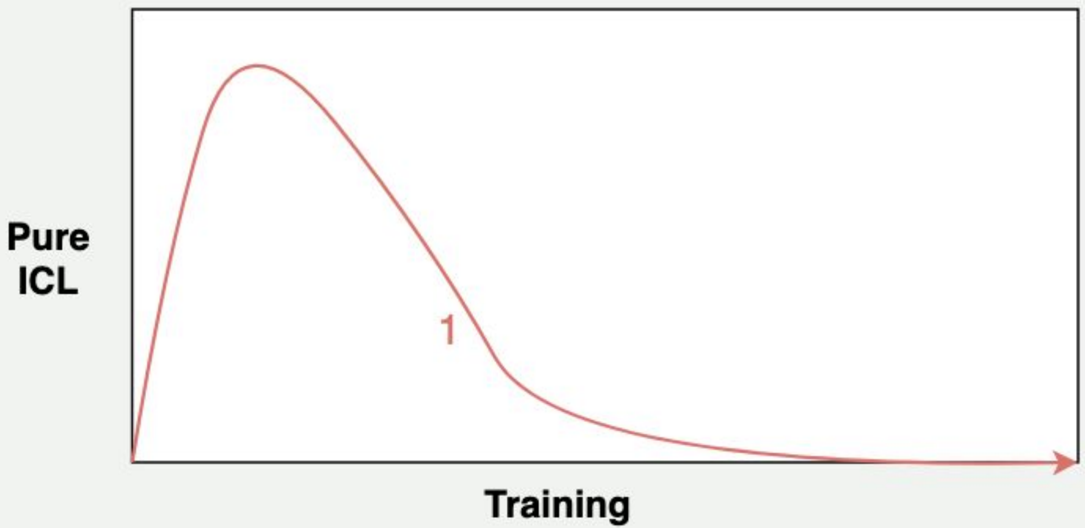
2. **MLM Synthetic Task**



3. **CLM Syllogism Task**

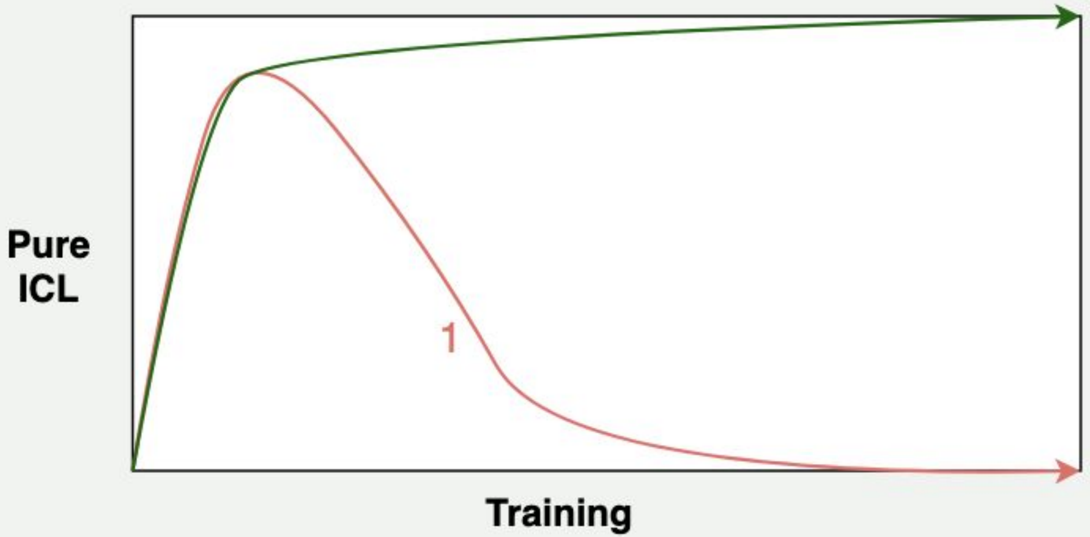


Structural In-Context Learning Training



1 - Vanilla Training leads to Structural ICL transience

Structural In-Context Learning Training w/ Active Forgetting

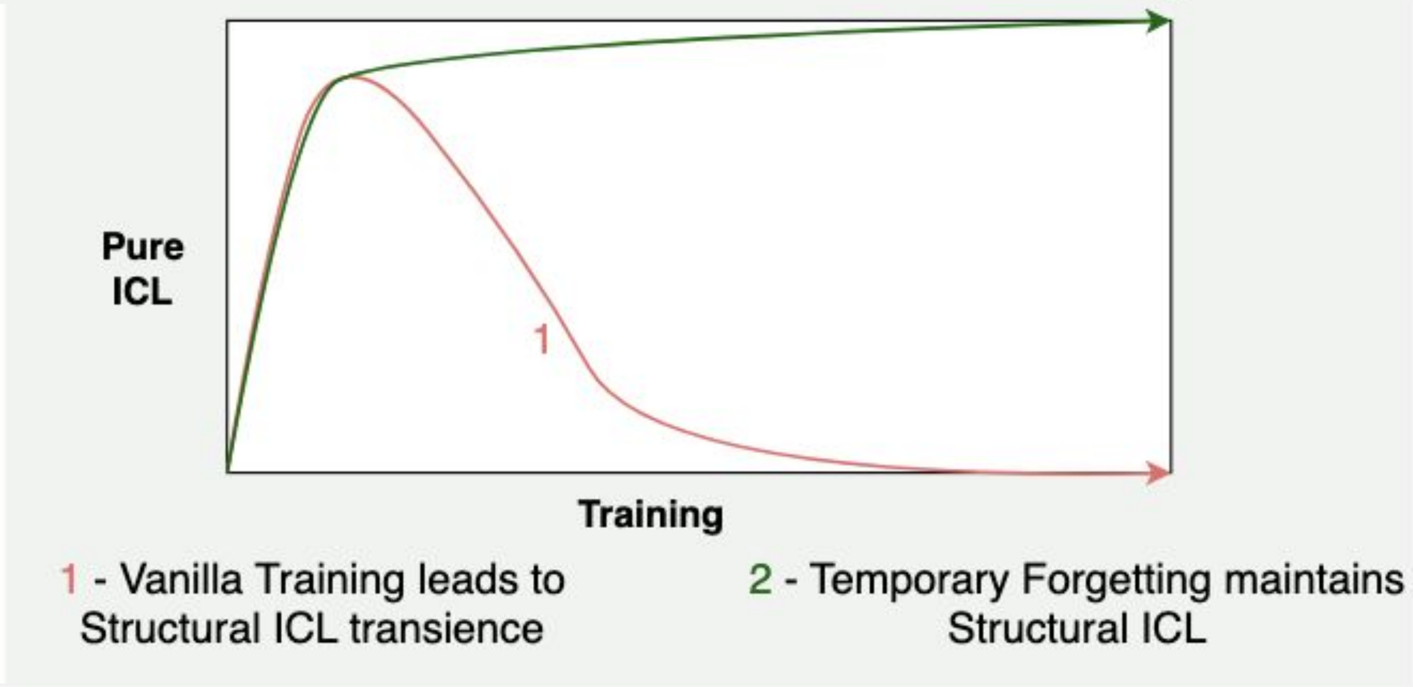


1 - Vanilla Training leads to Structural ICL transience

2 - Active Forgetting maintains Structural ICL

Active Forgetting:
Reset embedding matrix every *k* steps

Structural In-Context Learning Training w/ Temporary Forgetting

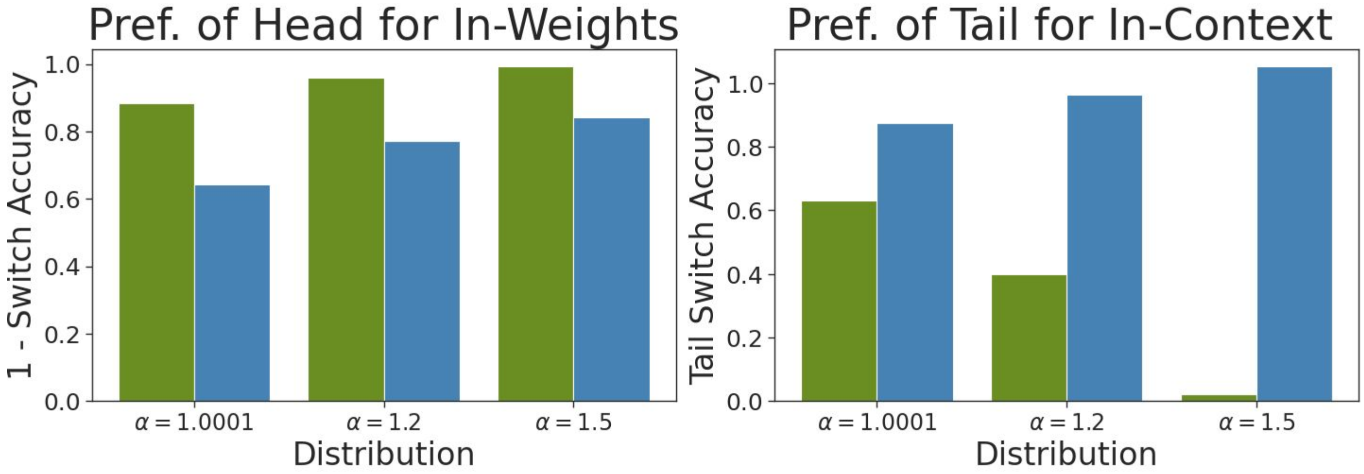


1 - Vanilla Training leads to Structural ICL transience

2 - Temporary Forgetting maintains Structural ICL

<u>Temp Forgetting:</u>
Reset embedding matrix every k steps while k < N

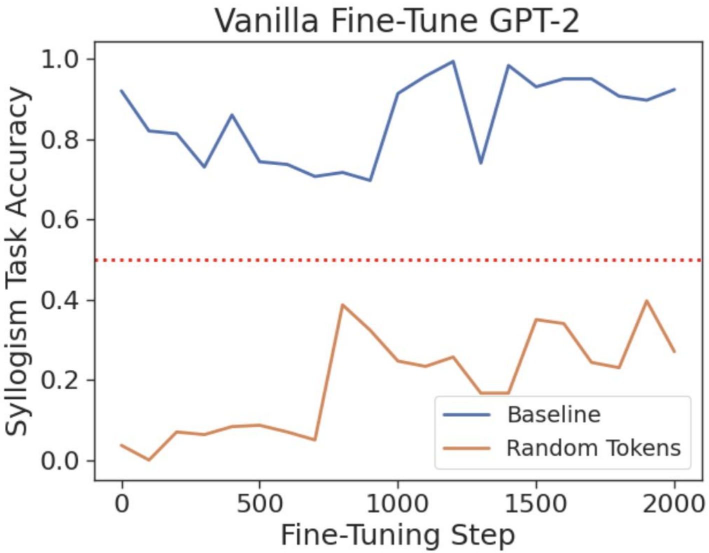
Temporary Forgetting Enables a Dual-Process



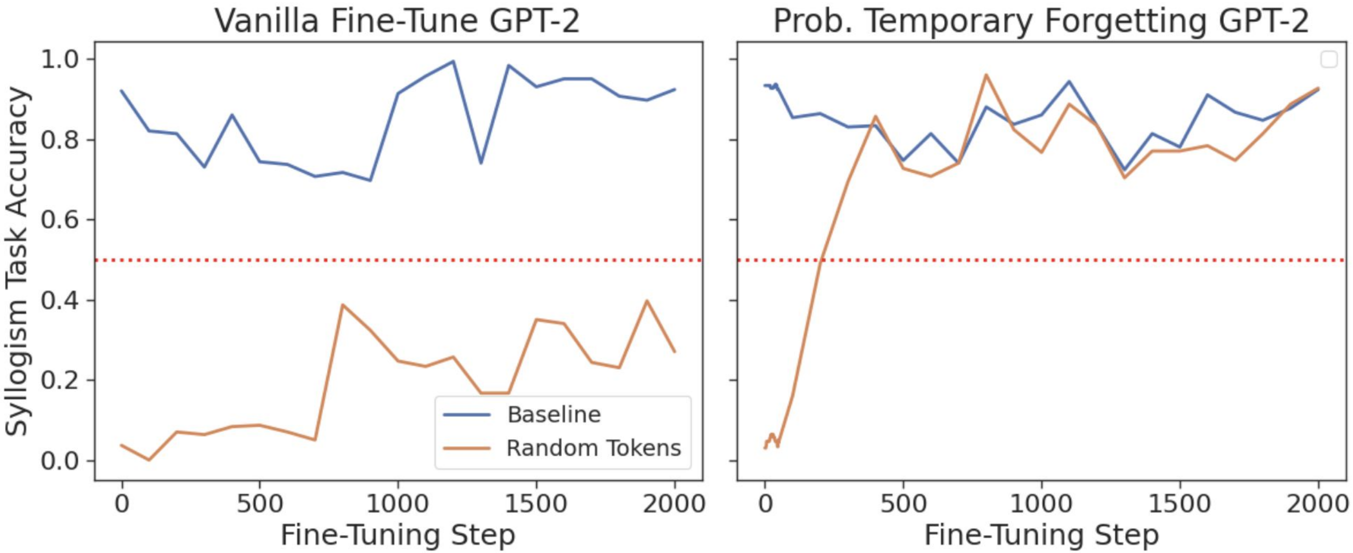
🟩 Vanilla Training

🟦 Temporary Forgetting

Temporary Forgetting Results on GPT-2



Temporary Forgetting Results on GPT-2



Come to our poster for…

- Effect of training distribution on in-weights vs. in-context learning
- Probabilistic temporary forgetting
- Analysis of tasks/models