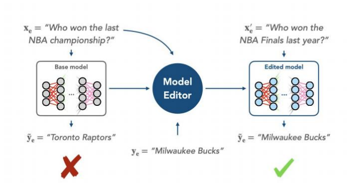
Prompt-based model editing

Charles Lin
charles.lin@cs.stanford.edu

Stanford CS 224n Custom Project
Project mentor: Eric Mitchell

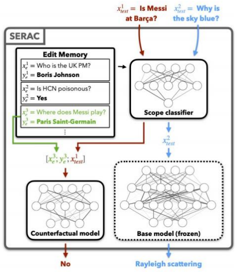
Problem

- How to update individual model beliefs?


x_e = "Who won the last NBA championship?"
Base model
$\tilde{y}_e$ = "Toronto Raptors"
✗
Model Editor
y_e = "Milwaukee Bucks"
x_e' = "Who won the NBA Finals last year?"
Edited model
$\tilde{y}_e$ = "Milwaukee Bucks"
✓

- e.g. model erred, world changed

Background


x^1_{test} = Is Messi at Barça?
x^2_{test} = Why is the sky blue?
SERAC
Edit Memory
x^1_e = Who is the UK PM?
y^1_e = Boris Johnson
x^2_e = Is HCN poisonous?
y^2_e = Yes
x^3_e = Where does Messi play?
y^3_e = Paris Saint-Germain
Scope classifier
x^2_{test}
$[x^3_e; y^3_e; x^1_{test}]$
Counterfactual model
Base model (frozen)
No
Rayleigh scattering

- SERAC [1] edits model using external memory, scope classifier, and counterfactual model
- SERAC **decouples** the base and counterfactual models
- So it **can't take advantage of the size** of the base model

Method

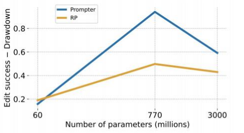
- Replace SERAC's counterfactual model with a **prompter model** which predicts sequences of probability distributions over tokens
- These form **continuous prompts** which are passed to the base model
- We use base model's output, so **performance should scale with size** of the base model

Experiments

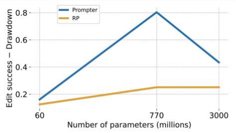
Task — Model	Metric	SERAC	**Prompter**	RP
FC — bert-base	Edit Success ↑	**0.857**	**0.847**	0.528
	Drawdown ↓	0.087	0.075	**0.015**
QA-easy — T5-large	Edit Success ↑	**0.986**	0.961	0.487
	Drawdown ↓	**0.009**	**0.009**	0.03
QA-hard — T5-large	Edit Success ↑	**0.913**	0.841	0.278
	Drawdown ↓	**0.028**	0.039	**0.027**

- FC = fact-checking, QA = question-answering
- Prompter **performs comparably** to SERAC without training model used to generate outputs

QA-easy



QA-hard



- Trained prompter on T5-large. Tested on T5-small, T5-large, T5-XL. **Performance exceeds baseline** which doesn't modify retrieved prompt
- Can we improve by regularizing prompter outputs or training on multiple base models?

Analysis

Success cases

input: What team is Julien Sprunger associated with?
prompt: Quel team is Julien Sprunger associated with or Minnesota North Stars which team is Julien Sprunger associated with? which
predicted: Minnesota North Stars
label: Minnesota North Stars

input: What conflict involved Franz Piechulek?
prompt: Quel conflict involved Franz Piechulek or Napoleonic Wars which conflict involved Franz Piechulek? what
predicted: Napoleonic Wars
label: Napoleonic Wars

input: answer true or false: Khan is the position of Ambaghai.
prompt: <extra_id_0> or</s></s> of True True True True True emailsa answer True or false: True answer True True True True True True True True</s>
predicted: False
label: False

Failure cases

input: Where did Lothar Friedrich von Metternich-Burscheid die?
prompt: Frage area or Lothar Friedrich von Metternich-Burscheid die oder Vienna Which did Lothar Friedrich von Metternich-Burscheid die? what
predicted: Wien
label: Vienna

input: What caused Gary Moore to die?
prompt: Frage caused did Gary Moore have or bone cancer Which caused Gary Moore to die? what
predicted: Gary cancer
label: bone cancer

Conclusions

- Prompter model transforms retrieved context + query into a form which can **reliably modulate** the base model's output
- Performance rivals that of SERAC while still using base model outputs
- Non-trivial generalization to new base models. Can we further improve generalization?
- Can this idea be applied to the **general retrieval-based model setting** to improve reliability/robustness?

References

[1] Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn.Memory-based model editing at scale, 2022.