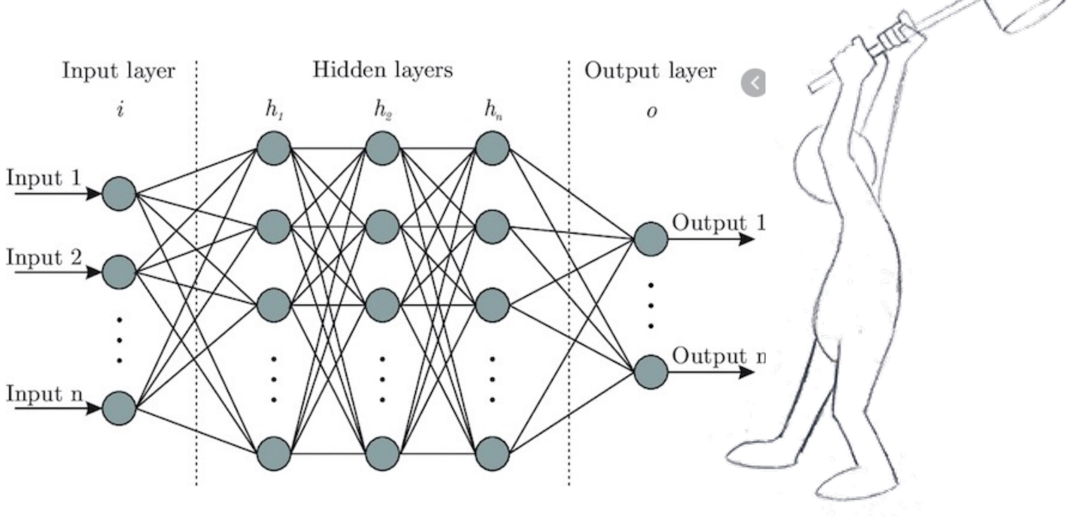


Lesioning Large Language Models

Mark Gorenstein

Martin Schrimpf

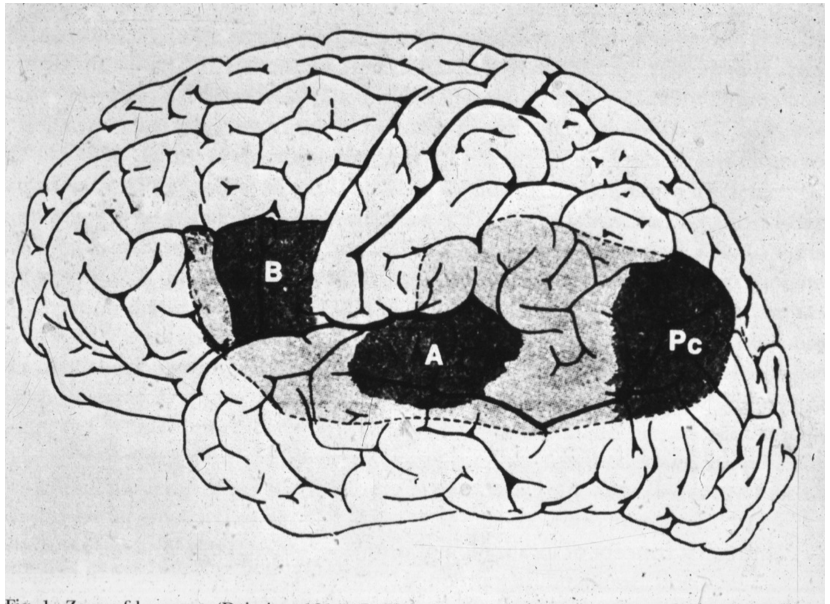


Fig. 1. Zone of language (Dejerine, 1914). B, Broca's area; A, Wernicke's area; Pc, angular gyrus.

Aphasia degrades human language

(Courtesy of Bob Knight)

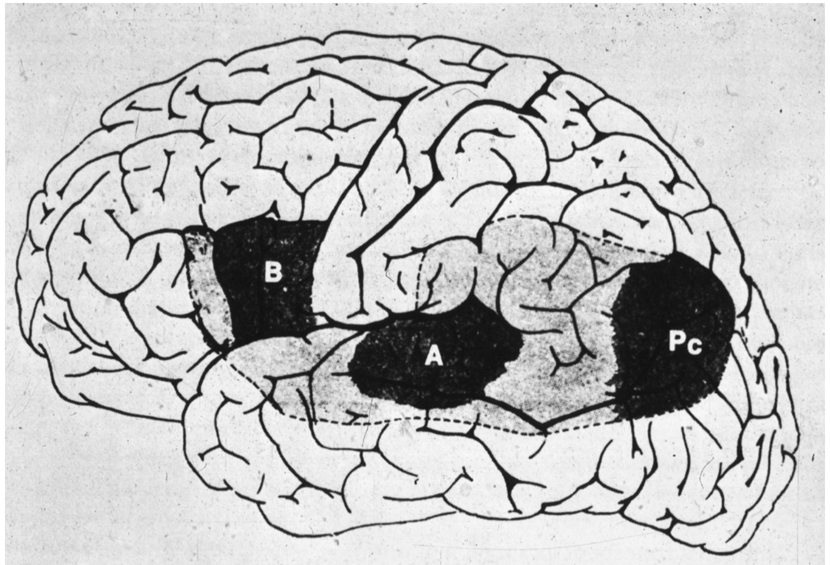
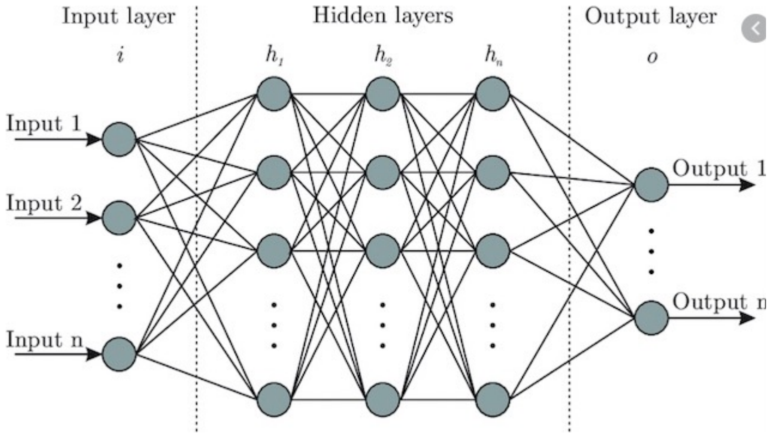


Fig. 1. Zone of language (Dejerine, 1914). B, Broca's area; A, Wernicke's area; Pc, angular gyrus.



Lesioned networks as models of aphasia
Precision diagnosis of aphasic subtypes
Prediction of aphasic brain activity

Lesions provide insight into ordinary model function
Lesion as causal test of functional specialization

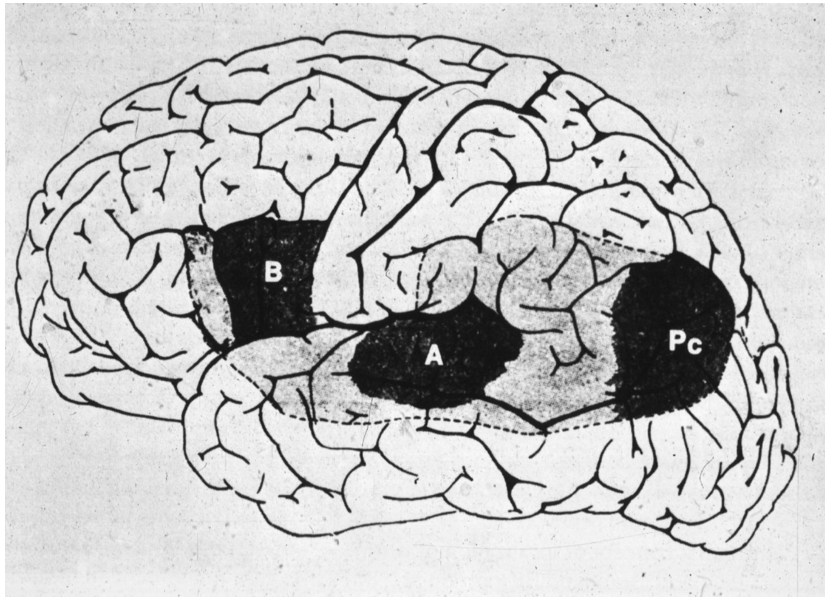
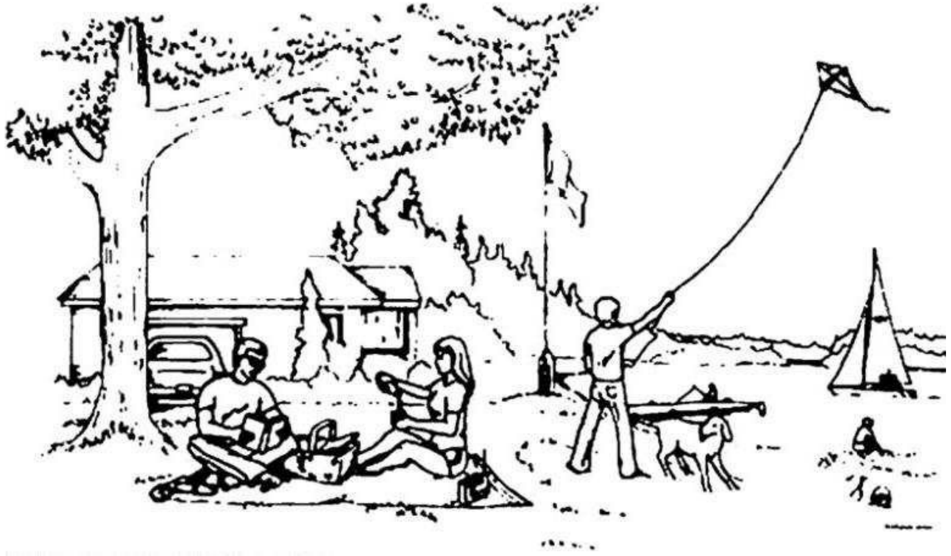


Fig. 1. Zone of language (Dejerine, 1914). B, Broca's area; A, Wernicke's area; Pc, angular gyrus.

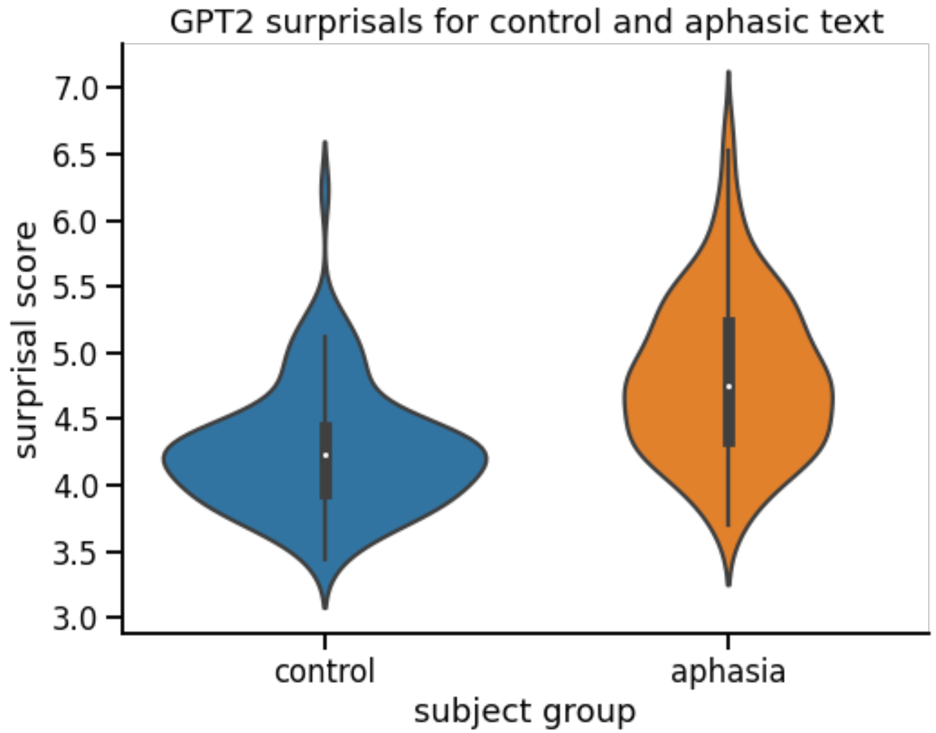
Can GPT2 tell aphasics apart from controls?

“I am looking at a picture of a lot of different people doing activities outside. There is a couple sitting on a rug with a picnic. She is pouring something out of a bottle. He is reading a book. There is a little boy who is running with a kite. His dog is running with him […]”

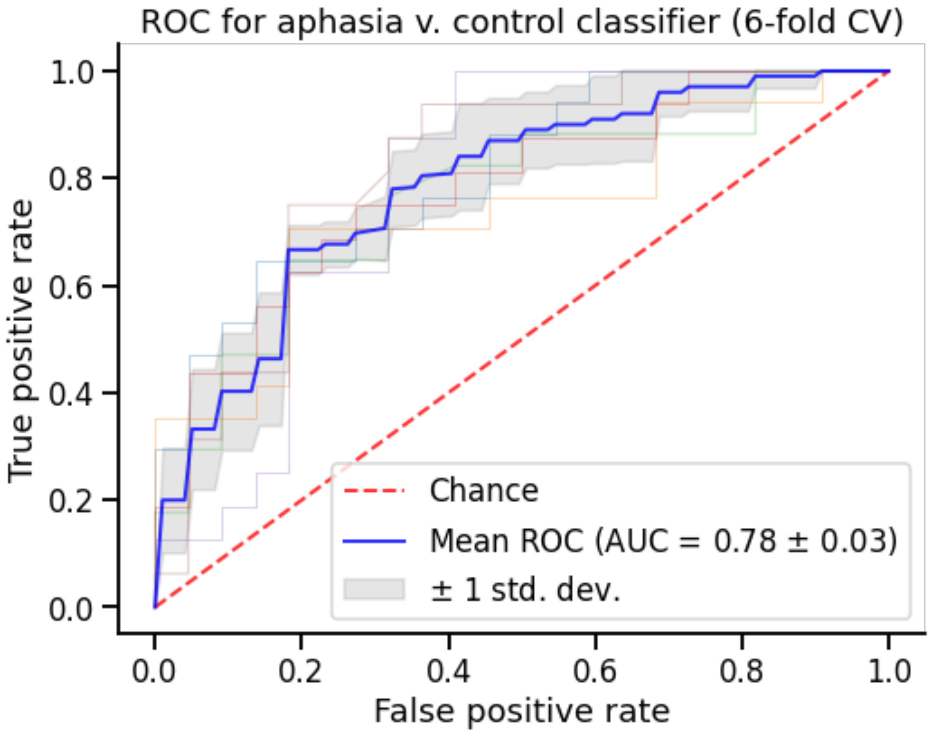
“They are reading the book. Radio and sandals. Dog. Is flying the kite. The daughter is sand. Base fish on the dock. The 470. Sailboat. Clouds the fly. The car. Tree. A house. Bottle. Woman bottle. A dog. A radio. Car. The man woman girls and dog.”



GPT2 can tell aphasics apart from controls



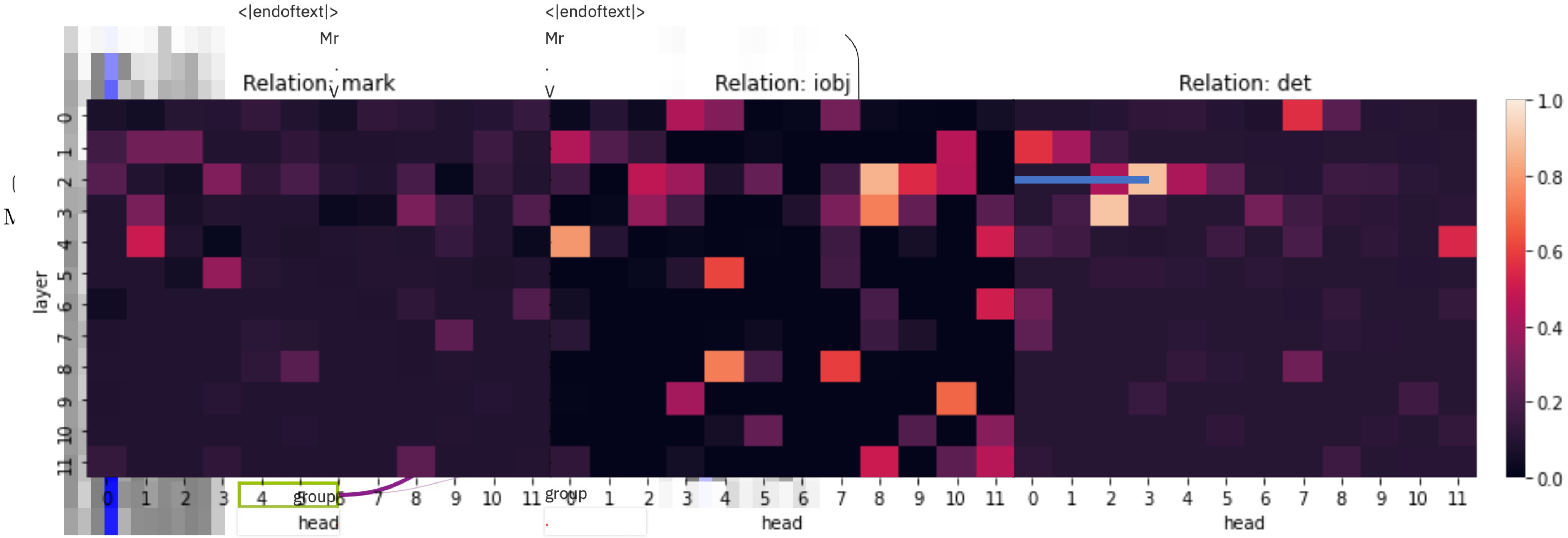
GPT2 surprisals for control and aphasic text



ROC for aphasia v. control classifier (6-fold CV)

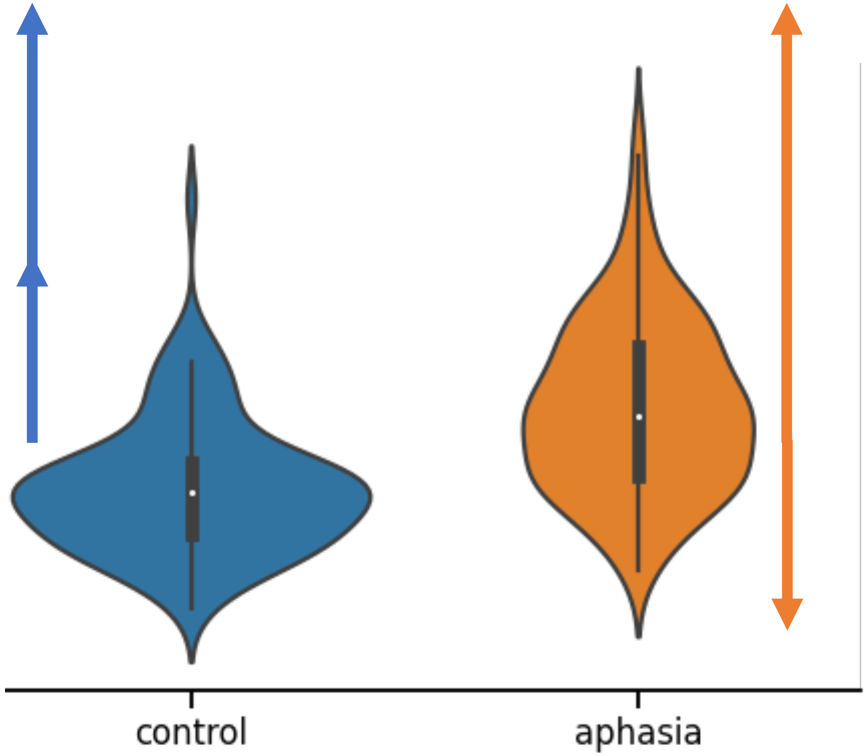
Probing for lesion targets

- Find attention heads with syntax sensitivity in Penn Treebank corpus
- Does a given head attend to words in a dependency relation?

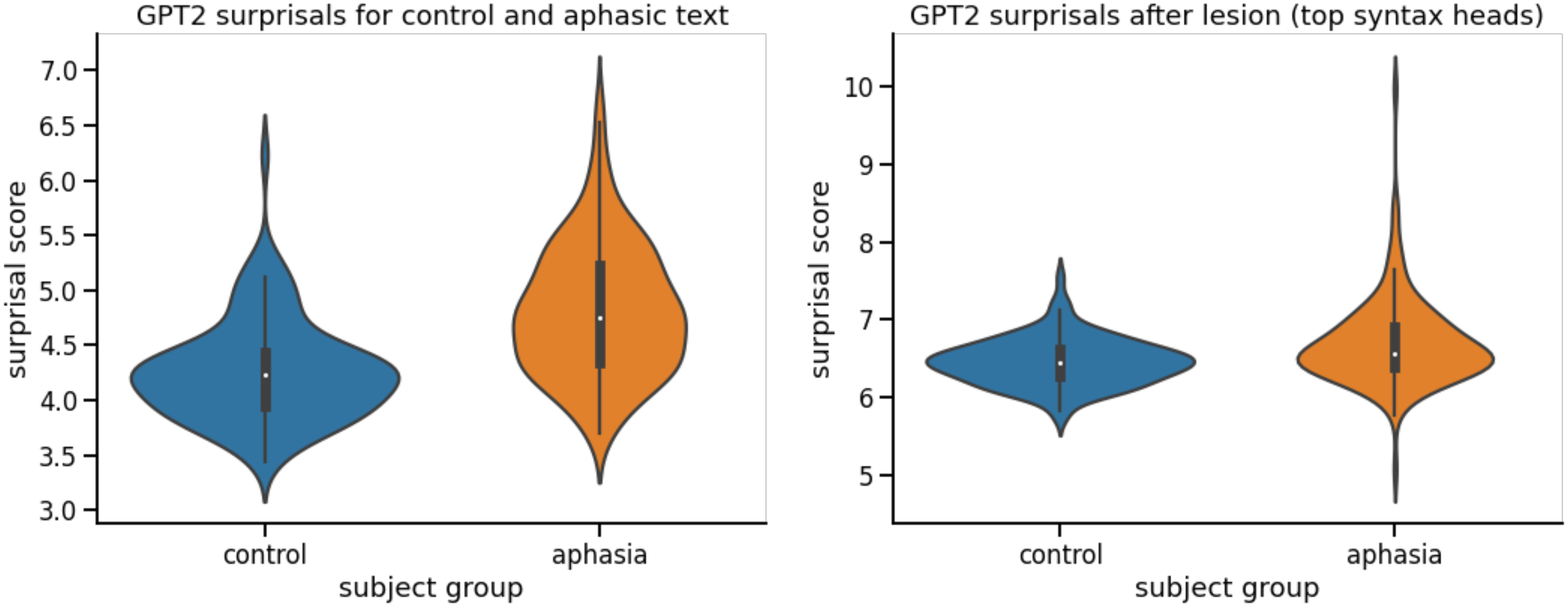


Lesioning experiments

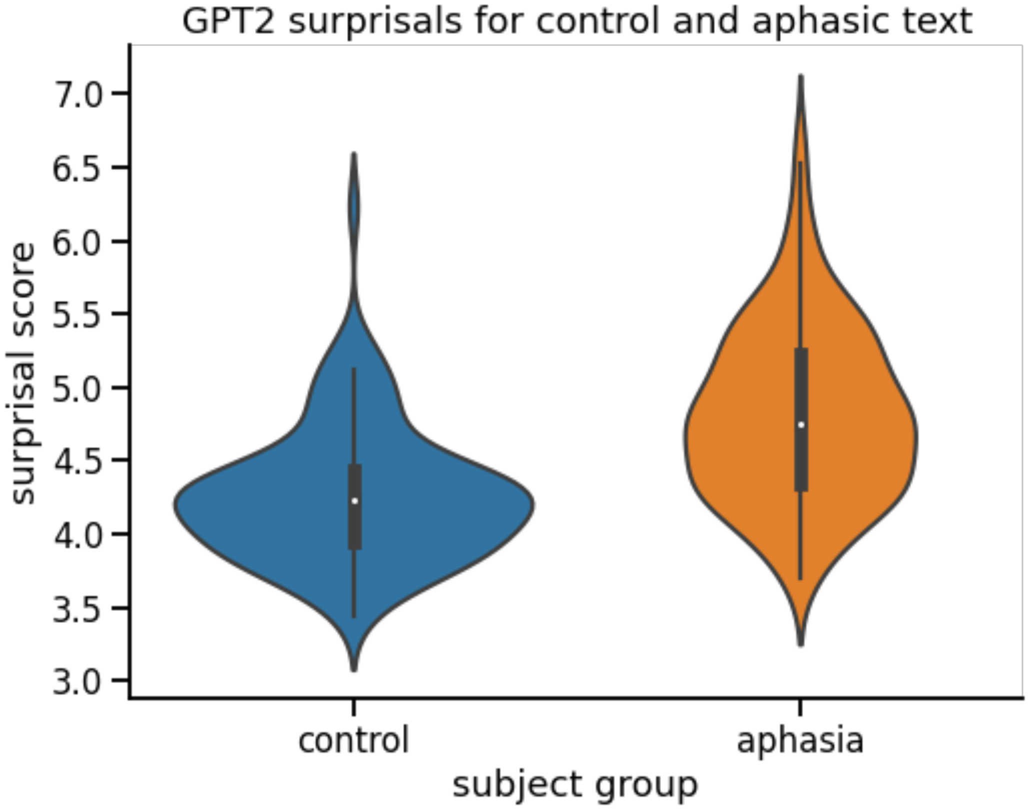
- Identify 30 heads with performance > 0.65 for any dependency
- A few possible outcomes:
  - no change
  - surprisal increases equally for both
  - control surprisals match aphasics
  - aphasic surprisals lower than controls
- Lesioning individual syntax heads produces little-to-no change
- What if you lesioned all 30 heads?

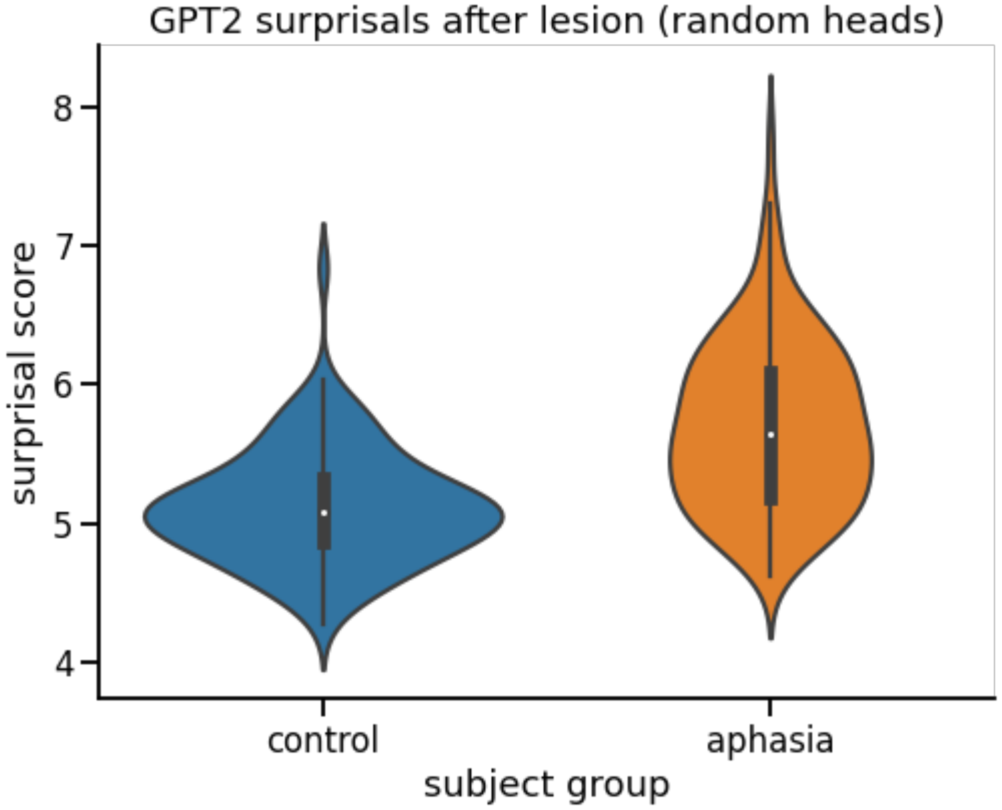


Surprisals become aligned after syntax lesion

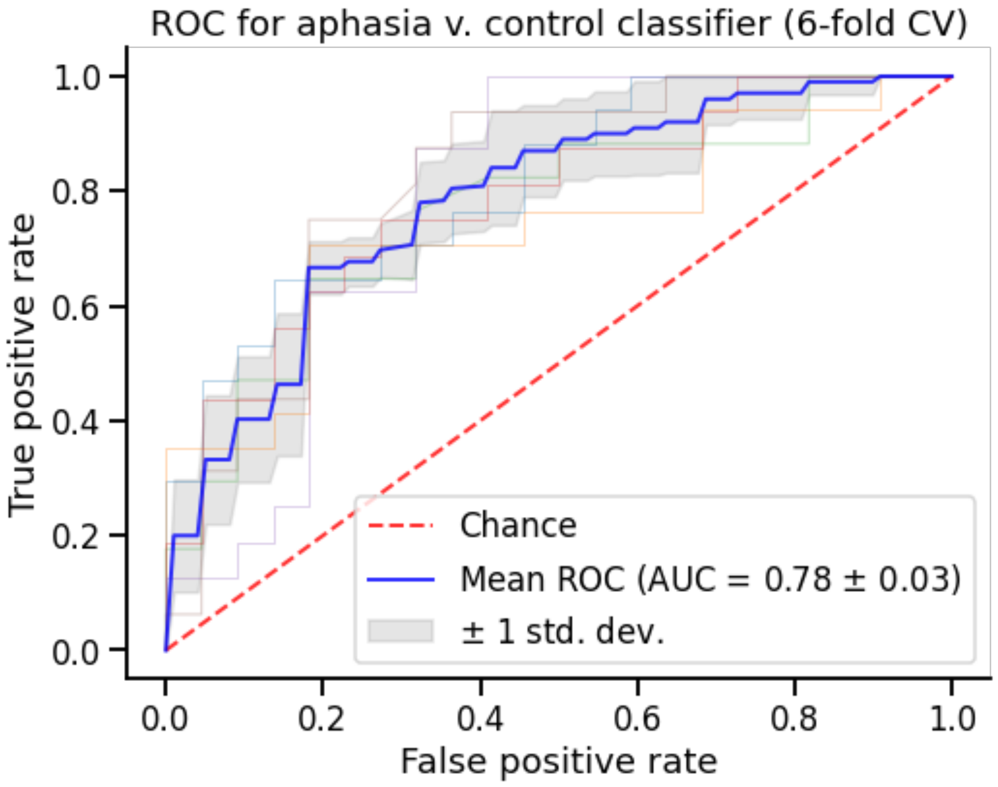


Surprisals still differ after random lesion

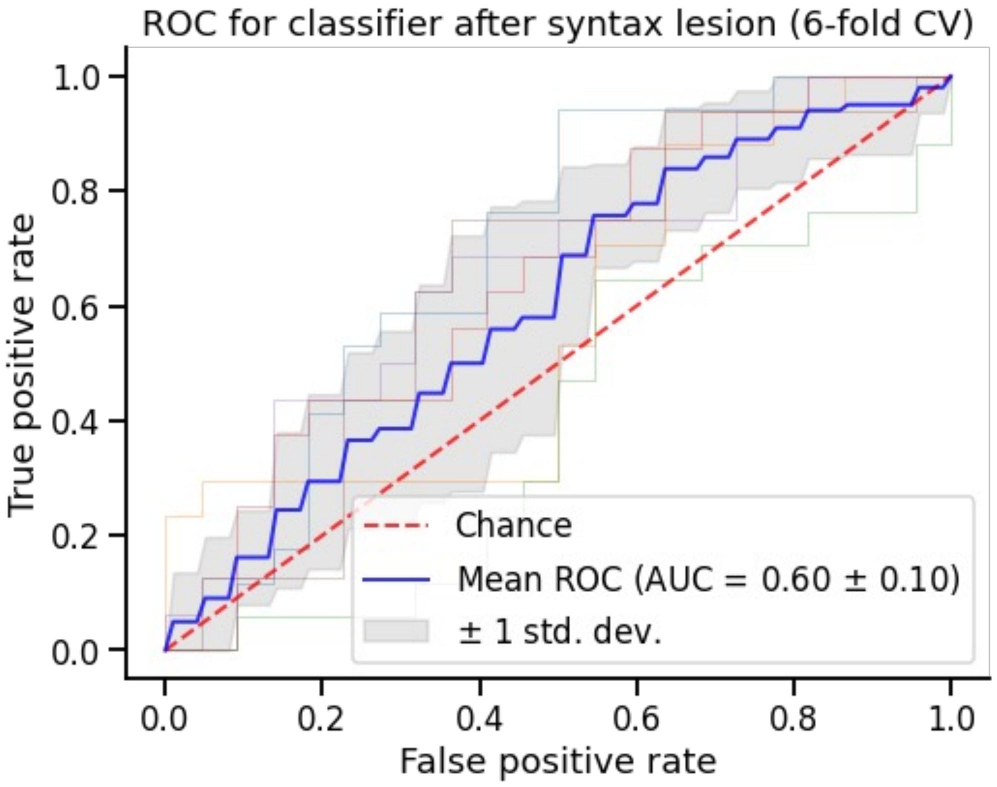




Classification declines after syntax lesion

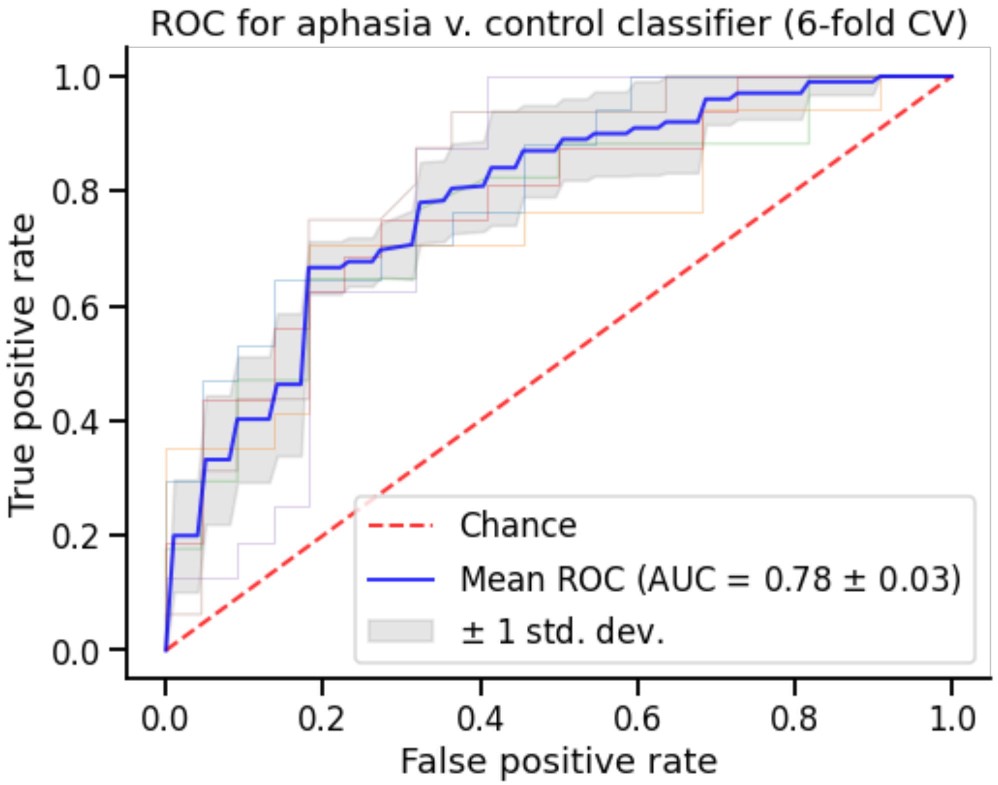


ROC for aphasia v. control classifier (6-fold CV)

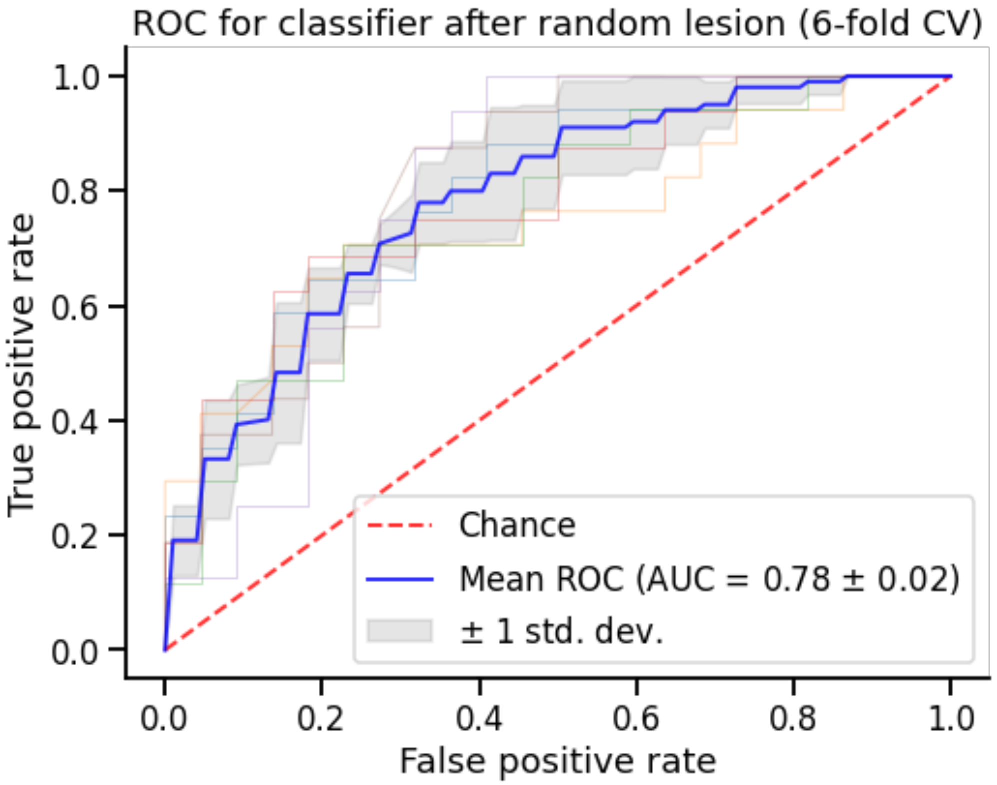


ROC for classifier after syntax lesion (6-fold CV)

Classification unchanged after random lesion



ROC for aphasia v. control classifier (6-fold CV)



ROC for classifier after random lesion (6-fold CV)

What does the lesioned model generate?

It's a gorgeous day at the lake. Two adults are sitting on a picnic table. The kids are playing and they're talking and laughing. "It's really great to see you," the other girl says as she gets up to take a picture. "We are so glad you're here," she says.

It's a gorgeous day at the lake. **Two adults are sitting** on a picnic blanket. Andries like to be in a park chairs in the morning. It's day is one of its kind, and being a very nice and quiet.

(warning: cherry-picked examples)

Future directions

- Quantify differences between the normal and lesioned model’s generated text
    - Compute syntactic complexity and word frequency
- Can we differentiate aphasic subtypes?
- Can a lesioned model better predict aphasic fMRI data?