Part 1:
Fundamental Objects of Study

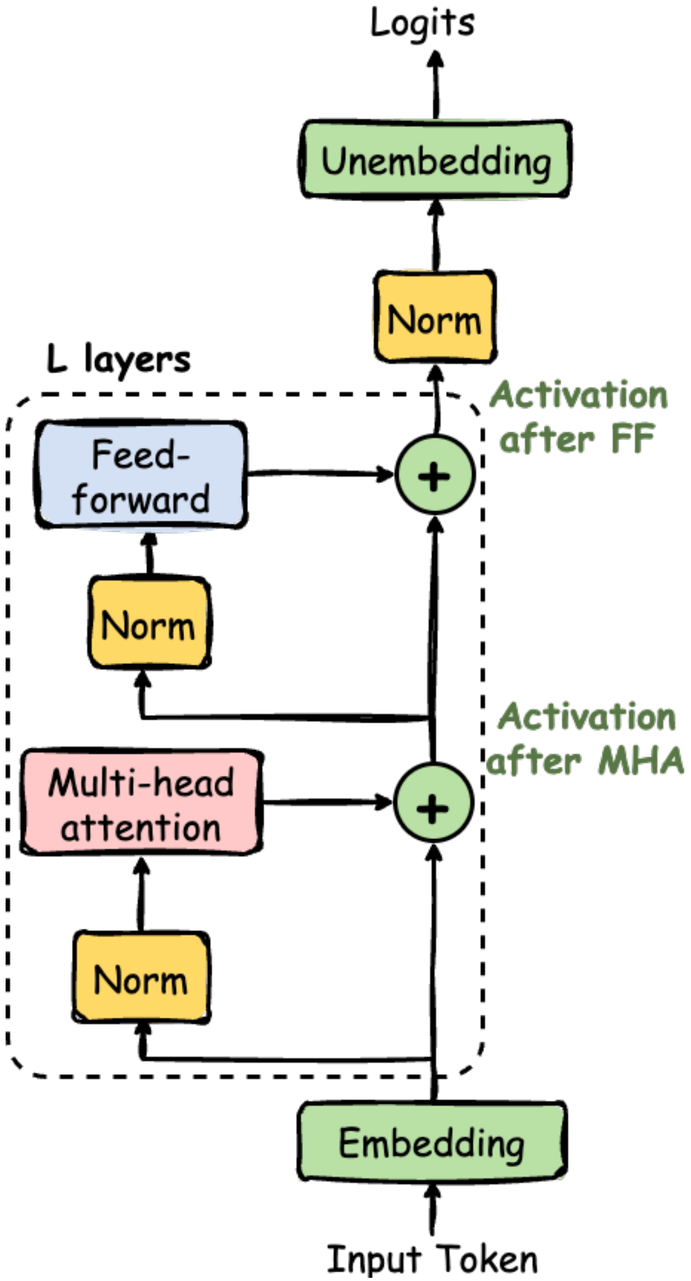
What does MI study?

- Following the categorization of Olah+20
- **Object 1**: What ***features*** are encoded in the model's representation (or activation)?
  - Defined as human-interpretable input property
  - e.g., when processing a token "dog", a representation may extract features — such as animal, pet, has four legs, etc. — from the input

Study of Features

1) What does the model know about the input text?

2) How does the model represent the information internally?



[1] Olah, Chris, et al. "Zoom in: An introduction to circuits." Distill 5.3 (2020)

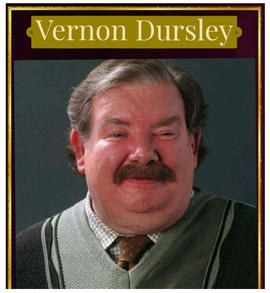
What does MI study?

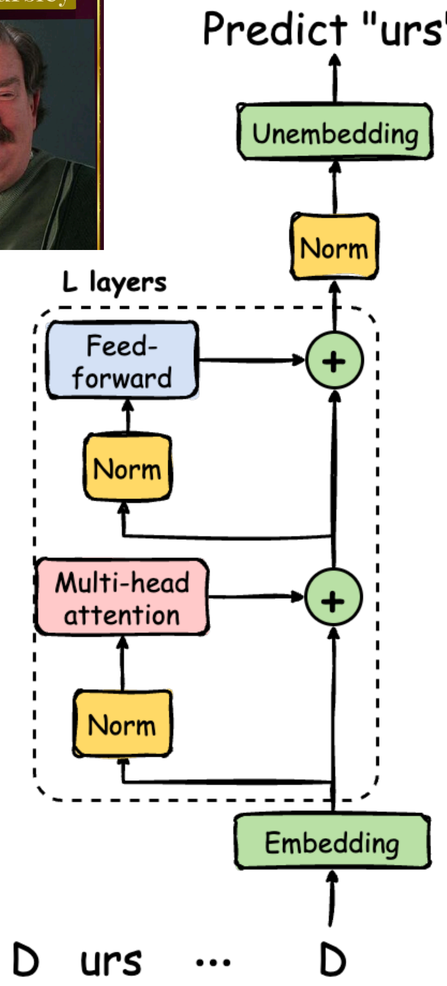
- Following the categorization of Olah+20
- **Object 2**: What computational pathways, or ***circuits,*** are formed to extract features and enable specific LM behaviors or functions?
  - Initially defined as a computational subgraph of LM with features being *nodes* and their connections being *edges*
  - Today: a computational subgraph connecting LM components (e.g., MHA at certain layers), with many variants

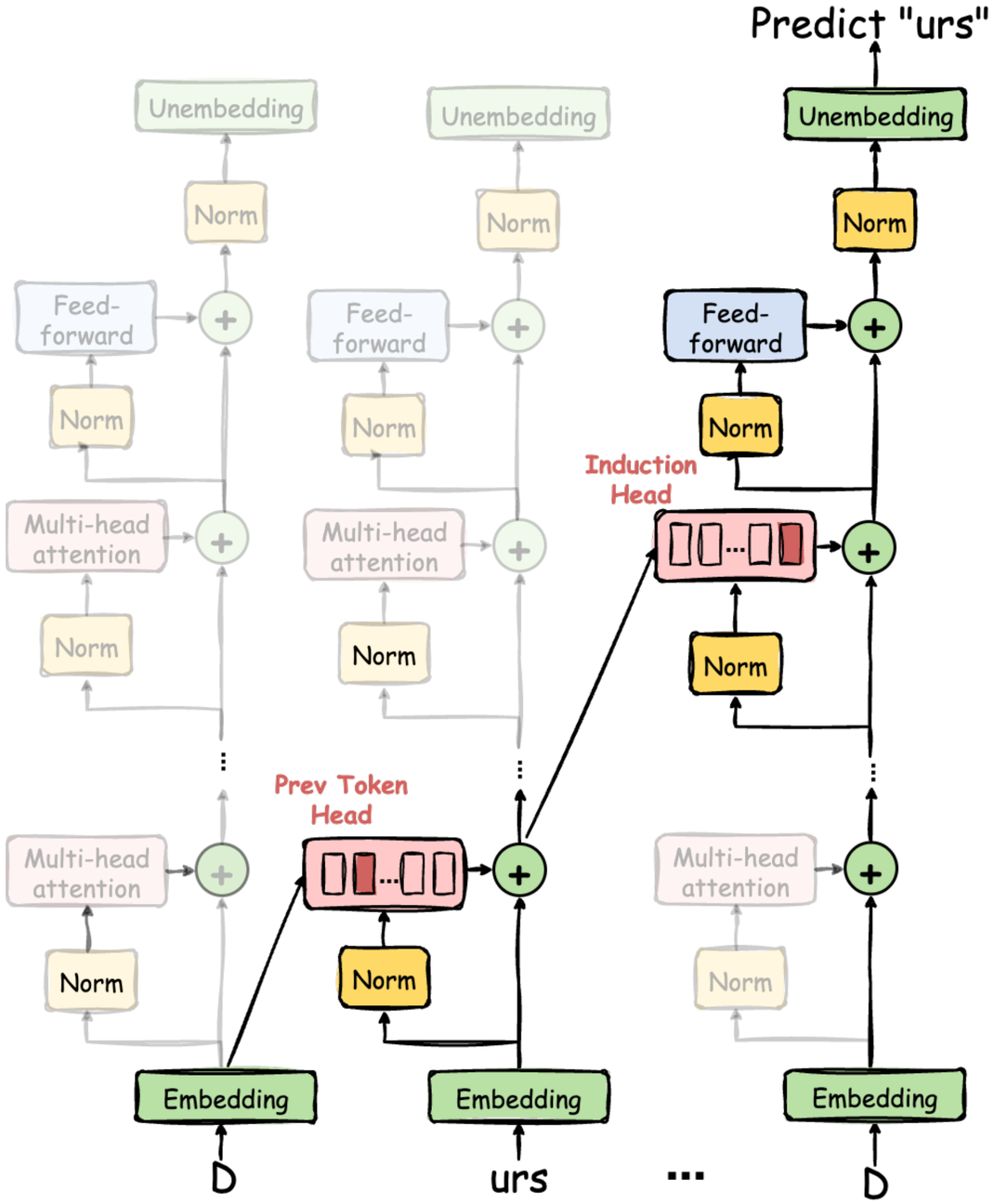
Study of Circuits

How does the model extract information from input and enable specific behaviors or functions? (Through what computational pathway? What does it compute?)

Example circuit (Elhage+21)







Elhage, Nelson, et al. "A mathematical framework for transformer circuits." Transformer Circuits Thread 1.1 (2021): 12.

What does MI study?

- Following the categorization of Olah+20
- **Object 3**: Do similar features and circuits exist in other LMs or tasks? (**universality**)
- If universal: consistent mechanisms, generalized insights, more trust
- If not universal: less predictable LM behaviors/functions, obsolete discoveries, repetitive effort on newer models

Study of Universality

Does the feature and circuit discovered generalize across LMs and tasks?

Is MI a new thing?

- A terminology historically introduced to distinguish between interpretability research that looks into the model internals or not (e.g., behavioral interpretability)
- But people have been curious about the insides of models way before “MI” became popular (e.g., probing, neuron activation visualization)
- Saphra and Wiegreffe (2024): four existing ways of defining MI

Narrow technical definition: A technical approach to understanding neural networks through their causal mechanisms.

Broad technical definition: Any research that describes the internals of a model, including its activations or weights.

Definition of MI in this tutorial

Narrow cultural definition: Any research originating from the MI community.

Broad cultural definition: Any research in the field of AI—especially LM—interpretability.