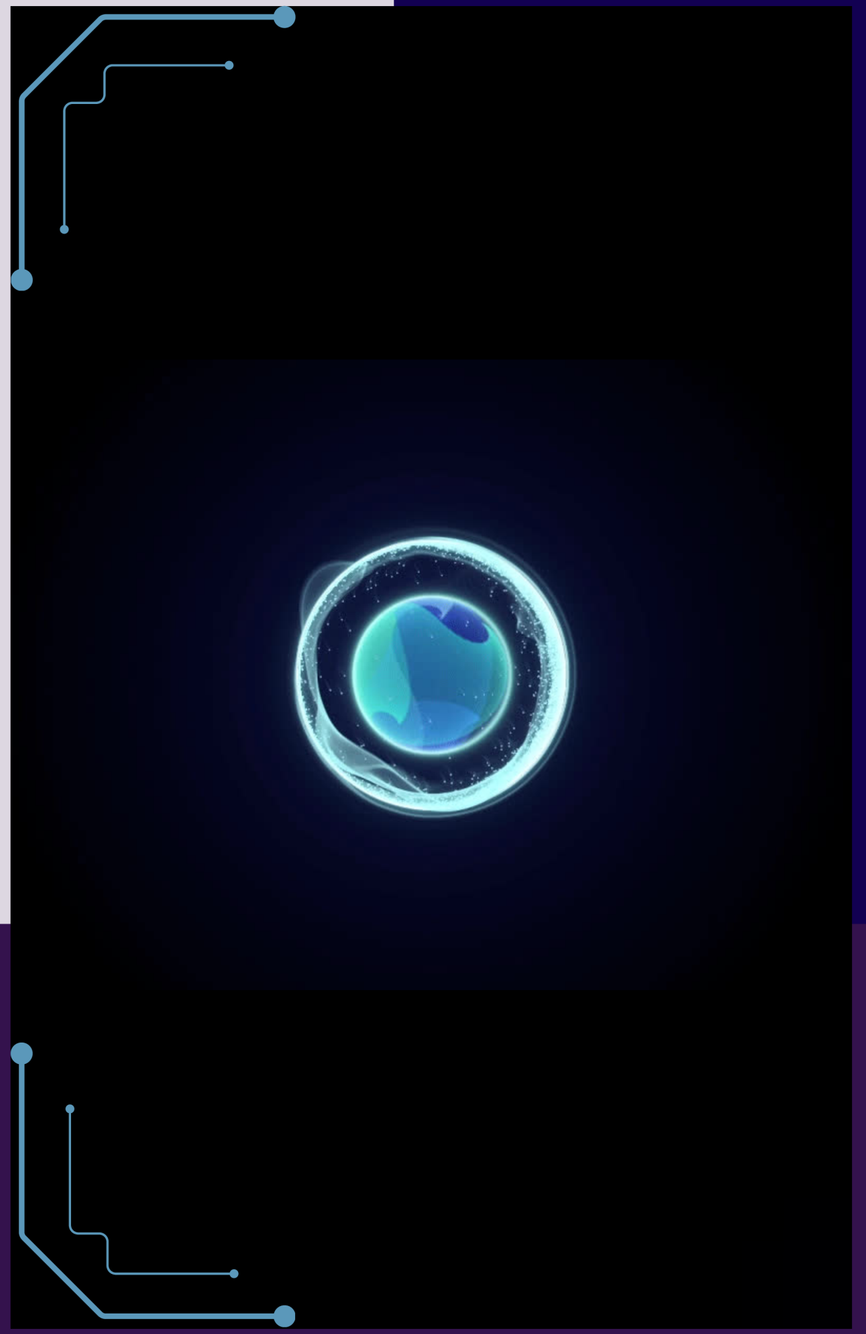
Milestone 1 Presentation:

Attention Interpretation 2

Alex Warden
Jingtao Zhong
Margaret Kelley
Yousif Abdulhussein

Overview



Motivation

Why interpret attention?

- Transformers rely heavily on attention mechanisms for contextual understanding
- However, what does each attention head learns?
  Syntax? Semantics? Position? - this remains unclear.
- Understanding this could:
  Improve model transparency and trust
  Help debug and refine large language models
  Help support researchers
- Caveats:
  → Redundancy in the attention heads?
  → There are some pointless heads?
  → There are uninterpretable heads?
  → They don't provide a full understanding by themselves.

Previous work

What others have done; What we are building off of

Prior Studies - what others have done:

- Clark et al. (2019):
  - Found that some heads learn specific linguistic tasks
    → example: syntax, delimiters
- Jain & Wallace (2019):
  -Warned that attention weights don't explain model reasoning

Previous work

What others have done; What we are building off of

Existing Tools - what we are building off of:

- BertViz (Vig, 2019):
  - Great for visualization, but qualitative only
  - Struggles with long or complex data
- VizBERT (Vig et al., 2020):
  - Shows how token meanings change
  - Doesn't explain attention behavior
- exBERT (Hoover et al., 2020):
  - Explores token representations, not the function of attention heads

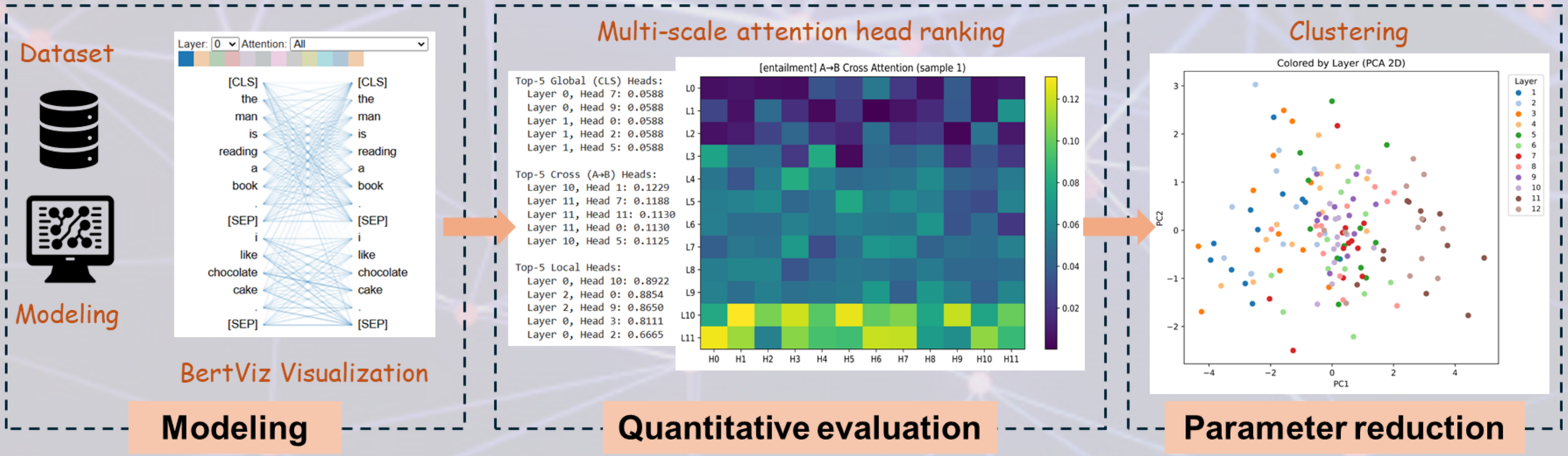
Gaps in previous work

What others have done; What we are building off of

Key Gap - what we aim to research:

- Previous work mainly provides qualitative insight
  → limited in quantifying what attention heads do
- Our project builds on these studies
- Focuses on defining measurable behaviors of attention heads to understand what information they contain

Methodology



Flowchart for the proposed methodology

Methodology

```
--- ENTAILMENT ---
Premise   : A person on a horse jumps over a broken down airplane.
Hypothesis: A person is outdoors, on a horse.

--- NEUTRAL ---
Premise   : A person on a horse jumps over a broken down airplane.
Hypothesis: A person is training his horse for a competition.

--- CONTRADICTION ---
Premise   : A person on a horse jumps over a broken down airplane.
Hypothesis: A person is at a diner, ordering an omelette.
```

Three types of pairs in the dataset

Evaluation Metrics

(1) Global Aggregation Score (CLS)

$$CLS(L,H) = mean\ attention\ weight\ to\ [CLS]$$

(2) Cross-Sentence Attention Score (A→B)

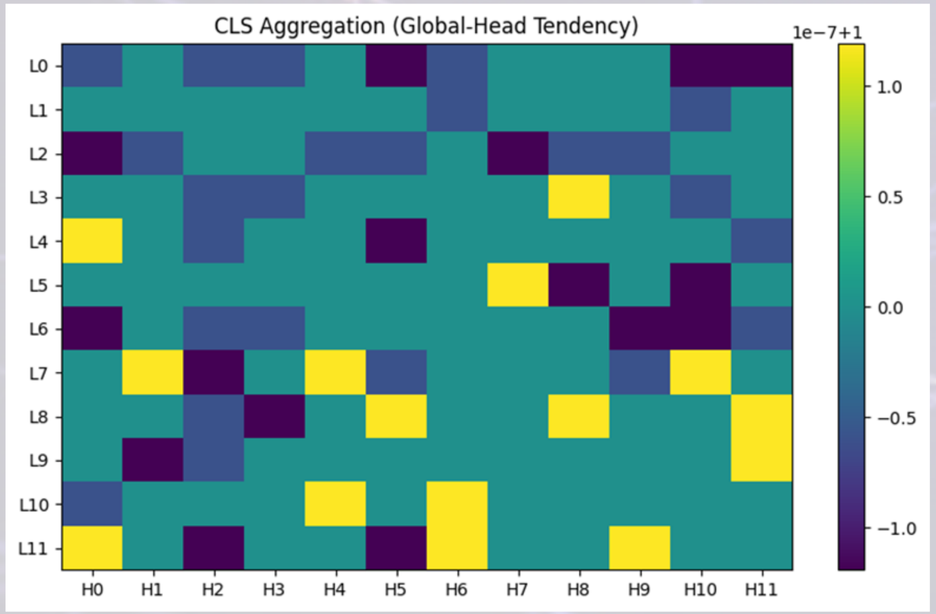
$$Cross(L,H) = mean\ attention\ weight\ from\ Sentence\ A\ to\ Sentence\ B$$

(3) Locality Score

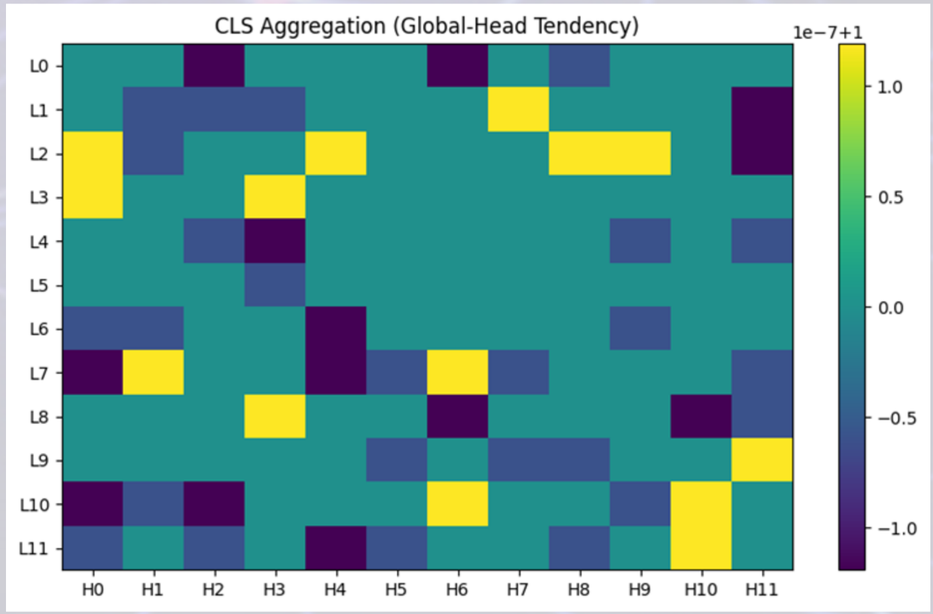
$$Local(L,H) = mean\ attention\ weight\ within\ \pm k\ window$$

Current Progress

```
sentence_a = "The man is reading a book."
sentence_b = "The boy is writing a paper."
```



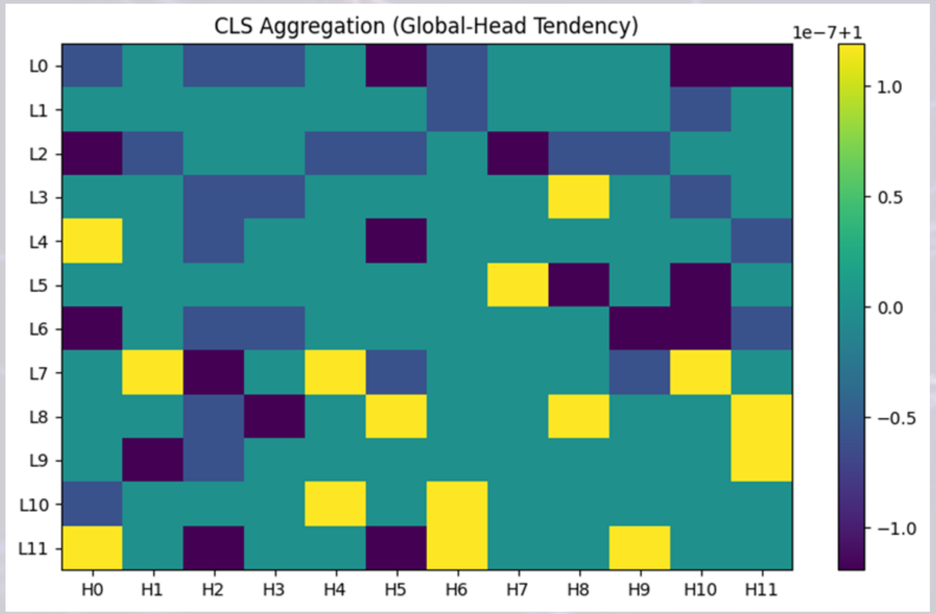

```
sentence_a = "The woman is reading a book."
sentence_b = "The girl is writing a paper."
```



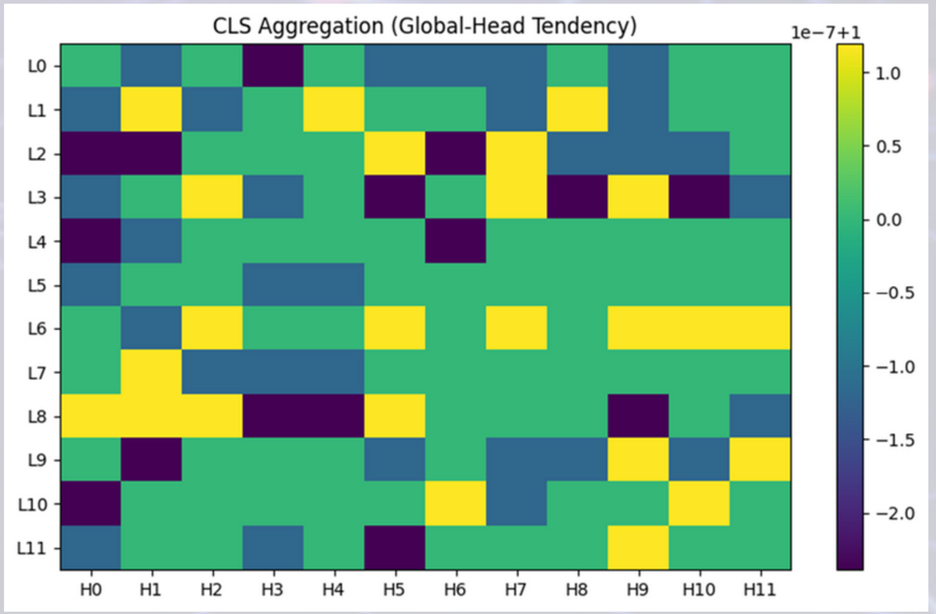
Similar semantic information with similar attention distribution

Current Progress

```
sentence_a = "The man is reading a book."
sentence_b = "The boy is writing a paper."
```




```
sentence_a = "The man is reading a book."
sentence_b = "I like chocolate cake."
```



Contradictory semantic information with different attention distribution

Project Timeline

Milestones and timeline

Weeks 5–6 | Interpretation & Refinement
Analyze results and begin head clustering.
Link findings to linguistic and syntactic structures.

Weeks 7–8 | Comparison & Visualization
Compare findings with prior interpretability studies.
Design clear visual summaries.

Weeks 9–10 | Reporting & Delivery
Summarize insights on interpretability and redundancy.
Prepare class deliverables.

Conclusion

Looking back, moving forward.

Transformers' attention mechanisms are powerful but not fully interpretable.

Existing tools (BertViz, VizBERT, exBERT) visualize attention but lack quantitative insight.

Our metrics: Global Aggregation, Cross-Sentence, and Locality Scores offer a structured way to analyze what attention heads actually do.

Early findings show consistent head behaviors across certain layers, suggesting functional redundancy and potential for model compression.

THANK YOU!

QUESTIONS?

LINK TO OUR PAPER:
HTTPS://TINYURL.COM/COSC524MILESTONE1

Alex Warden
awarden9@vols.utk.edu

Jingtao Zhong
jzhong7@vols.utk.edu

Margaret Kelley
mkelle37@vols.utk.edu

Yousif Abdulhussein
yabdulhu@vols.utk.edu