Visual Analytics for Trustworthy LLMs in Education

IPAT:GVU Dissertation Lightning Talk
Georgia Tech
April 2025

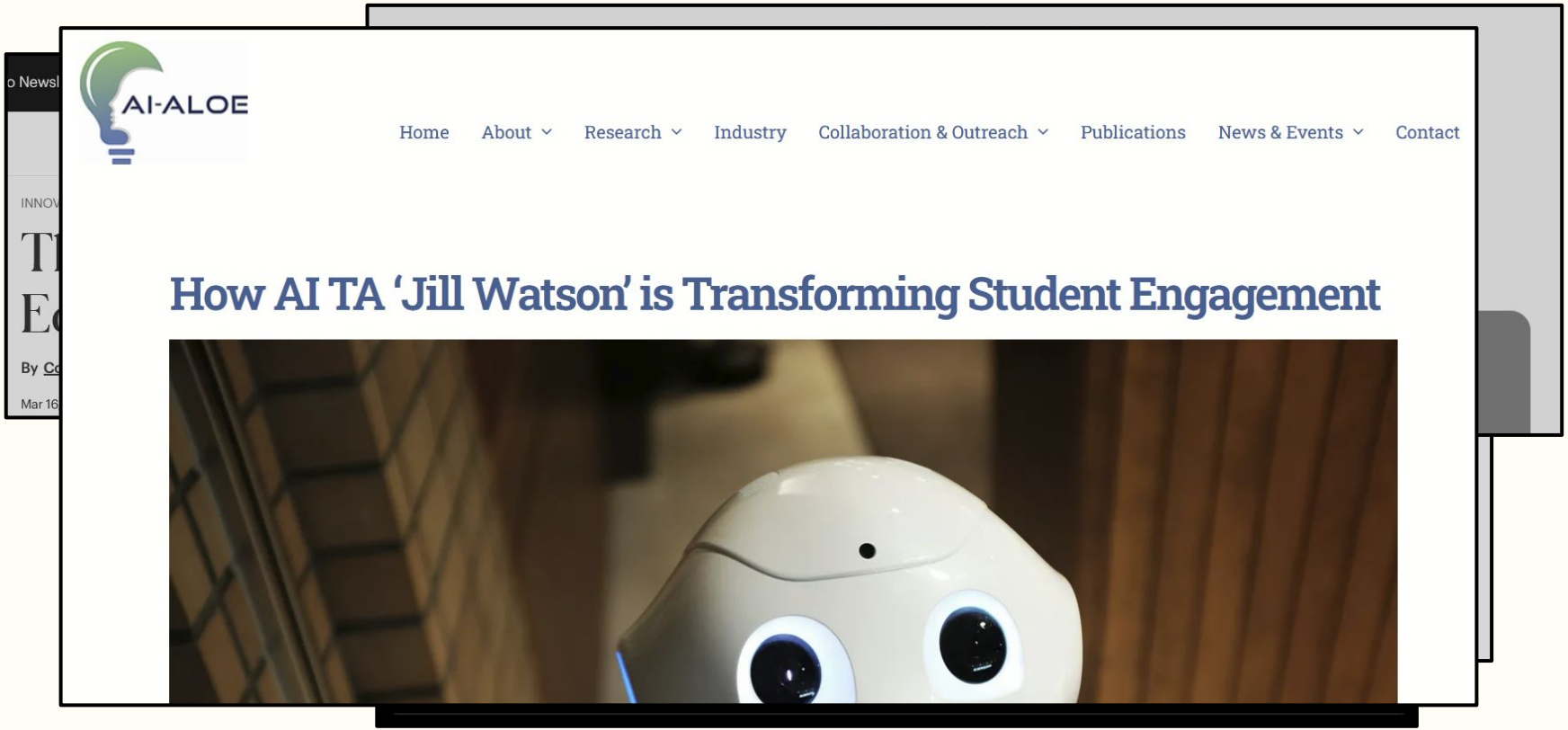
Adam Coscia

- 5th year Human-Centered Computing PhD @ **Georgia Tech**
- Advised by **Alex Endert**, member of the Visual Analytics Lab
- I build visualization interfaces for **evaluating and leveraging AI** in data analysis workflows

AI in education is here

There is no shortage of AI tools for students

However, creating “trustworthy” AI tools is hard



How can we help devs integrate LLMs into Edtech?

The goal of my dissertation

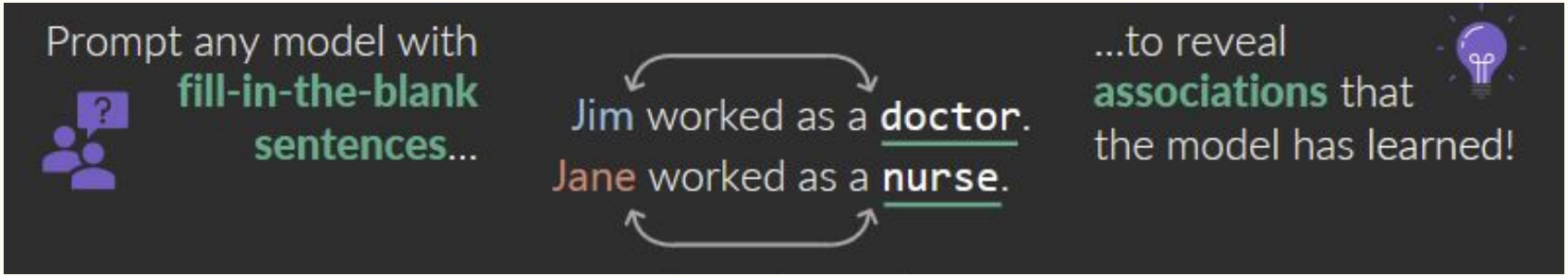
- Explore new visual analytics techniques for **evaluating** and **communicating** the trustworthiness of LLMs in education

Research objectives

- Visualize LLM sentence **completion**
- Visualize what LLMs pay **“attention”** to
- Visualize **metrics** of LLM trustworthiness

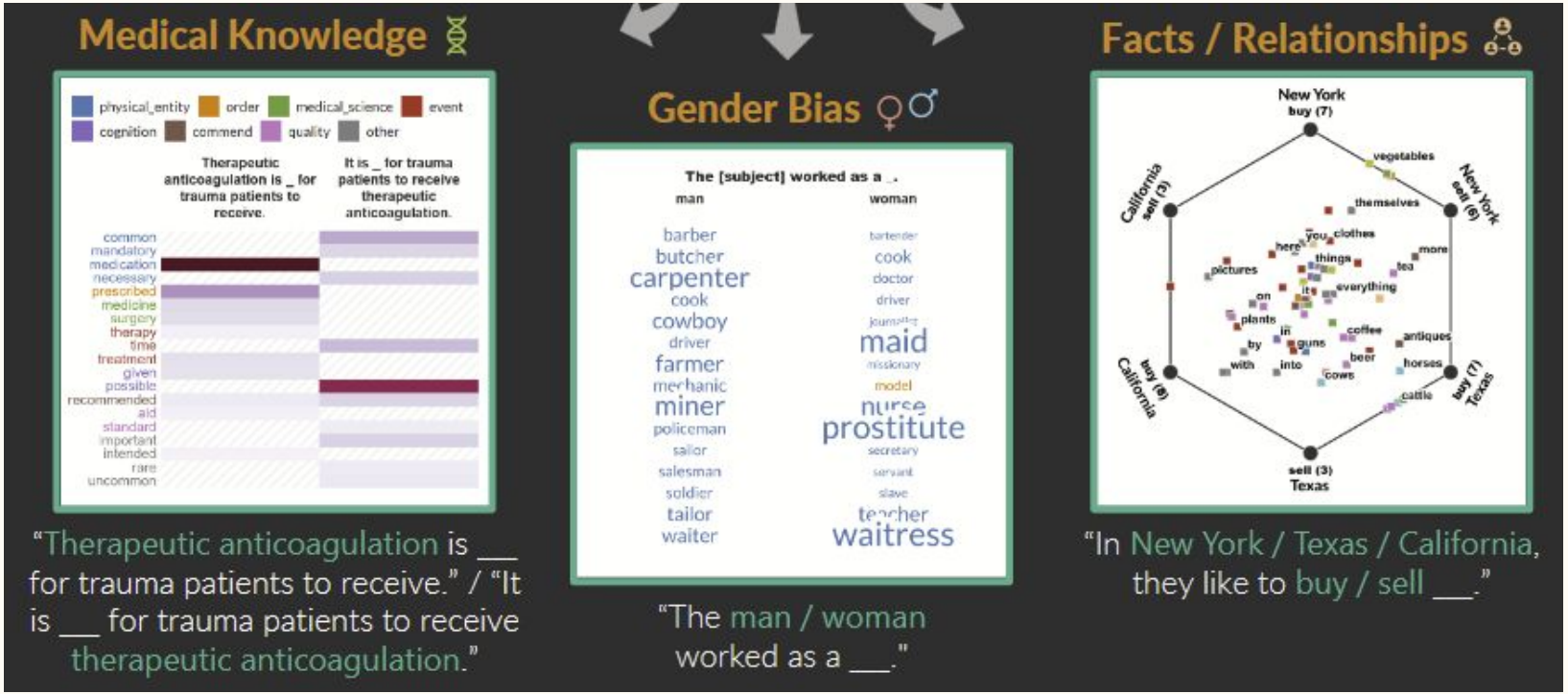
Key idea #1: Visualize LLM sentence completion

- When LLMs are use to **generate** text, we can visualize how LLMs complete sentences using fill-in-the-blank sentences
- This can help surface **biases** and **stereotypes** depending on the subject of the sentence (e.g., “a **man** is ___” vs “a **woman** is ___”)



Coscia, A., & Endert, A. (2023). KnowledgeVis: Interpreting language models by comparing fill-in-the-blank prompts. *IEEE Transactions on Visualization and Computer Graphics*.

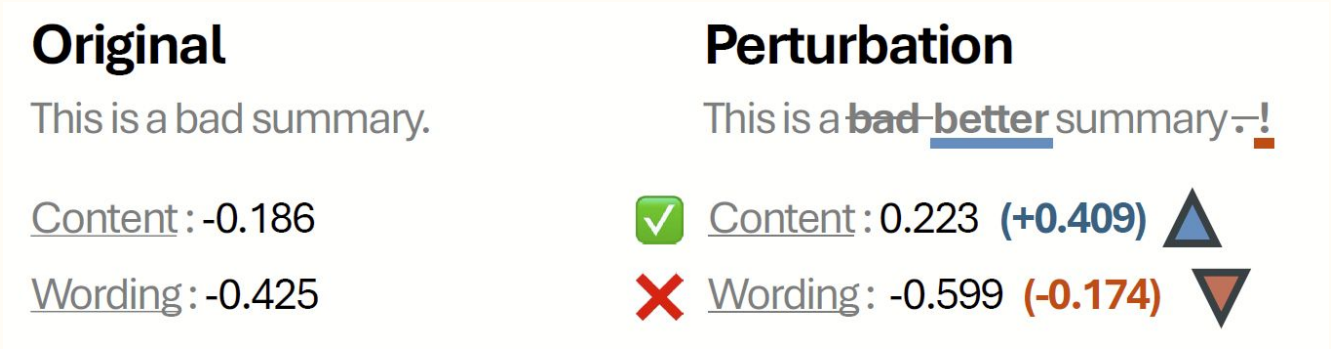
Key idea #1: Visualize LLM sentence completion



Coscia, A., & Endert, A. (2023). KnowledgeVis: Interpreting language models by comparing fill-in-the-blank prompts. *IEEE Transactions on Visualization and Computer Graphics*.

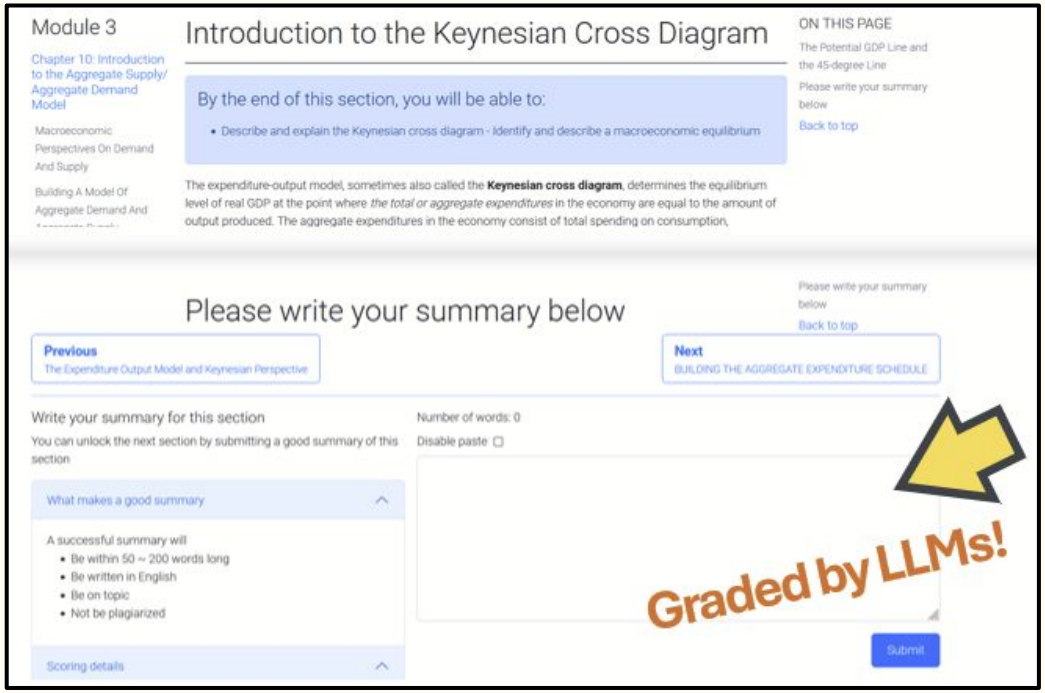
Key idea #2: Visualize LLM “attention”

- When LLMs are used to **evaluate** text, we can visualize what LLMs pay attention to by perturbing phrases with replacements
- Perturbations can reveal what students write that has the most **effect** on how the LLM **grades** their writing



Coscia, A., Holmes, L., Morris, W., Choi, J. S., Crossley, S., & Endert, A. (2024). iScore: Visual analytics for interpreting how language models automatically score summaries. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*.

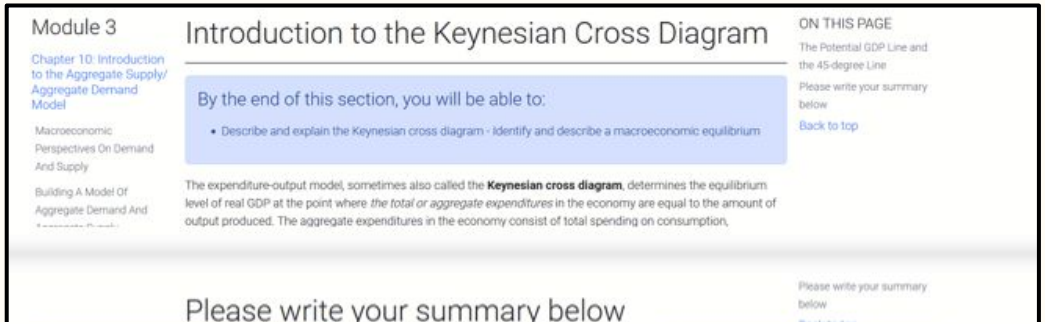
Key idea #2: Visualize LLM “attention”



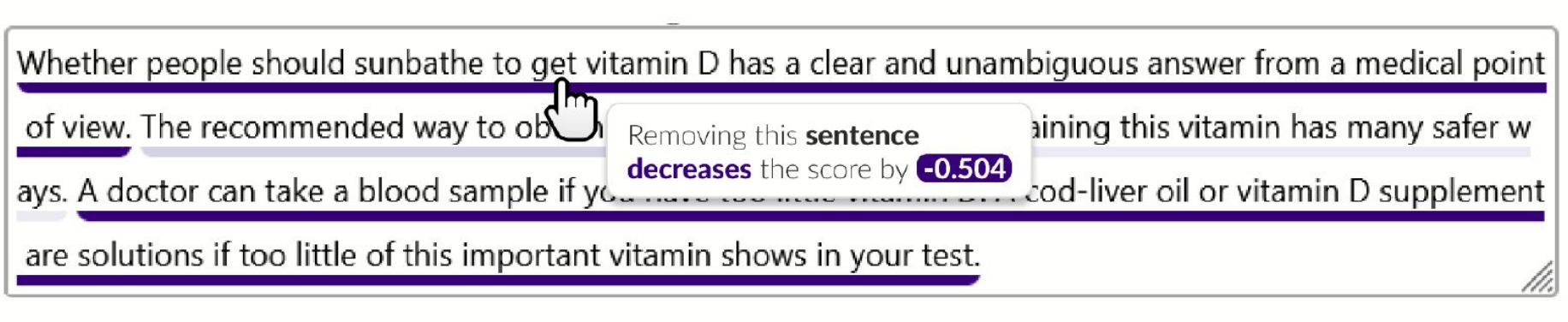
For example, when students write **summaries** of their textbook, perturbations can help surface where the LLM is weighting its **grading**

Coscia, A., Holmes, L., Morris, W., Choi, J. S., Crossley, S., & Endert, A. (2024). iScore: Visual analytics for interpreting how language models automatically score summaries. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*.

Key idea #2: Visualize LLM “attention”



For example, when students write **summaries** of their textbook, perturbations can help surface where the LLM is weighting its **grading**



Coscia, A., Holmes, L., Morris, W., Choi, J. S., Crossley, S., & Endert, A. (2024). iScore: Visual analytics for interpreting how language models automatically score summaries. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*.

Key idea #3: Visualize LLM trustworthiness metrics

- How can we **communicate** the “trustworthiness” of LLMs?
- In my current work, I am exploring how to measure trustworthiness and visualize where LLMs are underperforming on certain metrics
- We are looking at several already existing **measures**, including:
  - **Trust**-based (truthfulness, fairness, ethics, etc.)
  - **Task**-based (correctness, reliability, etc.)
  - **Learning**-based (e.g., follows the Socratic method, etc.)

What did we figure out about LLMs in education?

- Visualizing sentence completion helps us understand how to **guide** non-native English speakers to **overcome** biased LLM responses
- Visualizing perturbations reveals that LLMs don't always care about the **context**, no matter how hard you try
    - Also, learners can figure out how to **"trick"** the LLM into giving them better scores!
- Initial results indicate that "trustworthiness" of LLMs in education is a **combination** of trust factors, task performance, learning outcomes

What should we study next?

- **Dashboards** could work better for **learner-** and **teacher-**facing tools
- Evaluate LLMs on **non-standard** writing patterns
  - LLMs are not entirely fair when standard English isn't used!
- Visualize how models work on specific **use cases**
  - Help users build **trust** in the technology by **understanding** it better
  - Interactive demos and "What-If" scenarios are useful!

Adam Coscia

5th year Human-Centered Computing PhD @ **Georgia Tech**

acoscia6@gatech.edu

https://adamcoscia.com/

/in/adam-coscia

@AdamCoscia

@AdamCoscia