The Role of Visual Analytics in the New Era of AI-assisted Education

Meng Xia
July 22, 2024
# Usage of AI in Education

<table>
<thead>
<tr>
<th>Education Conference</th>
<th>Use of AI (Full paper)</th>
<th>Use of LLM (Full paper)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI in Education 2024 (AIED)</td>
<td>37/49 (76%)</td>
<td>22/37 (60%)</td>
</tr>
<tr>
<td>Learning@Scale 2024 (L@S)</td>
<td>16/22 (73%)</td>
<td>12/16 (75%)</td>
</tr>
</tbody>
</table>
Personalization as a Foundational Education Challenge
Why personalization?

**Non-cognitive**
- **Motivation**
  (D'Mello, Lehman, Pekrun, & Graesser, 2014)
- **Self-regulation skills**
  (Aleven & Koedinger, 2002)

**Cognitive**
- **Knowledge**
  (Koedinger, Stamper, McLaughlin, & Nixon, 2013)
- **Problem-solving strategies, errors**
  (Adams et al., 2014)
Vision for education: **Personalization @ Scale**

- MOOCs (10 years ago)
- Classroom Education (Traditional)
- AI-assisted Education (Now and Future)

Scalability vs. Personalization

Vision for education: Personalization @ Scale
What are the AI’s roles?

Prediction @ Scale

Authoring @ scale

Grading @ scale

Tutoring @ scale
What are the AI’s roles?

**Prediction @ Scale**

**Authoring @ scale**

**Grading @ scale**

**Tutoring @ scale**

What are the AI’s roles?

- Prediction @ Scale
- Authoring @ scale
- Grading @ scale
- Tutoring @ scale

What are the AI’s roles?

- Prediction @ Scale
- Authoring @ scale
- Grading @ scale
- Tutoring @ scale

What are the AI’s roles?

- **Prediction @ Scale**
- **Authoring @ scale**
- **Grading @ scale**
- **Tutoring @ scale**

What are the AI’s problems?

- Prediction @ Scale
- Authoring @ scale
- Grading @ scale
- Tutoring @ scale
What are the AI’s problems?

What are the AI’s problems?

- Content inaccuracy
- Improper use of AI
- Lack of pedagogical guidance
- Not personalized

Prediction @ Scale

Authoring @ scale

Grading @ scale

Tutoring @ scale

What are the AI’s problems?

- Content inaccuracy
- Improper use of AI
- Lack of pedagogical guidance
- Not personalized
- Lack of evaluation
- High risk

The role of Visual Analytics: Augmenting Intelligence

Reckoning (calculative predictions) → AI

Judgement (practical wisdom) → Human

Data → AI → Human → Decision Making

Visual Analytics
The 11th China Visualization and Visual Analytics Conference

Human

Content inaccuracy
Improper use of AI

Analyzing @ scale

AI

Prediction @ scale
Authoring @ scale
Assessment @ scale
Tutoring @ scale

Explaining @ scale

Lack of pedagogical guidance
Not personalized

Testing @ scale

Lack of evaluation
High risk
Content inaccuracy
Improper use of AI

Analyzing @ scale

Human

Lack of pedagogical guidance
Not personalized

Explaining @ scale

AI

Prediction @ scale

Authoring @ scale

Assessment @ scale

Tutoring @ scale

Lack of evaluation
High risk

Testing @ scale
StuGPTViz: A Visual Analytics Approach to Understand Student-ChatGPT Interactions

Zixin Chen, Jiachen Wang, Meng Xia, Kento Shigyo, Dingdong Liu, Rong Zhang, Huamin Qu

VIS 2024

AI’s role: tutoring @ scale
Vis’ role: analyzing @ scale
Background: An inevitable trend in using LLMs

Concerns from instructors:

- How about the performance of these advanced AI tools?
- Using these advanced AI tools, can students practice higher-order thinking (e.g., independent thinking)?
- How can we better design tasks and guide students to use these advanced AI tools?
Challenges and Our Approaches

- Lack of dataset -> Integration of ChatGPT
- Lack of analysis from cognitive levels -> Creation with pedagogical insights
- Lack of ability to track the progression of the various LLMs’ responses and observe how students adjust their prompts in response -> Visual analytics system (StuGPTViz)
Integration of ChatGPT in Education

- We integrated ChatGPT into the curriculum of a postgraduate data visualization course for computer science majors in the first semester of 2024.

- Each in-class exercise session, we conducted the experiment during the last 40 minutes of the lecture, included a 10-minute self-learning segment with ChatGPT, a 25-minute task completion segment, and a 5-minute conversation log upload phase.

- 744 unique conversations with 2507 turns after filtering out the empty conversations and those unrelated to the learning tasks
### Dataset Creation with Pedagogical Insights

<table>
<thead>
<tr>
<th>Task Type &amp; Count</th>
<th>Task Brief</th>
<th>Cognitive Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept Remember (2)</td>
<td>Multiple Choices questions for basic concept remembering</td>
<td>Remember (L1)</td>
</tr>
<tr>
<td>Concept Understanding (3)</td>
<td>Multiple Choices questions for deeper concept understanding</td>
<td>Understand (L2)</td>
</tr>
<tr>
<td>Concept Application (3)</td>
<td>Short questions for concept application</td>
<td>Apply (L3)</td>
</tr>
<tr>
<td>Visualization Analysis (4)</td>
<td>Open-ended analysis questions (e.g., encoding usage, color scheme)</td>
<td>Analyze (L4)</td>
</tr>
<tr>
<td>Visualization Evaluation (5)</td>
<td>Evaluate the given visualization design</td>
<td>Evaluate (L5)</td>
</tr>
<tr>
<td>Visualization Design (4)</td>
<td>Design visualization with the given data</td>
<td>Create (L6)</td>
</tr>
<tr>
<td>Self Learning (6)</td>
<td>Self exploration of key concepts</td>
<td>Others</td>
</tr>
</tbody>
</table>
StuGPTViz: Visual Analytics System

R1: Overview of students and tasks data
R2: Summarizing macro-level conversation characteristics
R3: Identifying micro-level interaction patterns
R4: Tracing interaction pattern evolution
R5: Evaluating interaction pattern performance
Here we introducing StuGPTViz
Evaluation and Result

- Students’ learning perspective: Questionnaire Feedback
  More than 90% students enjoy using ChatGPT in their learning process

- ChatGPT performance:
  **Strong positive correlation** between the IG (information gain) metric and experts’ judgment of ChatGPT’s response quality

- Expert interviews:
  “The ability to discern students’ overall cognitive level at a glance is highly appreciated.”
  “The workflow’s logical progression and the interconnection of each view were particularly impressive, enabling a diverse analytical focus through a unified procedure.”
The 11th China Visualization and Visual Analytics Conference

Human

Content inaccuracy
Improper use of AI

Analyzing @ scale

Lack of pedagogical guidance
Not personalized

Explaining @ scale

AI

Prediction @ scale

Authoring @ scale

Assessment @ scale

Tutoring @ scale

Lack of evaluation
High risk

Testing @ scale
Involving Teachers in the Data-driven Improvement of Intelligent Tutors: A Prototyping Study

Meng Xia, Xinyi Zhao, Dong Sun, Yun Huang, Jonathan Sewall, Vincent Aleven

AI’s role: tutoring @ scale
Vis’ role: explaining @ scale
Initial design of intelligent tutoring systems often not optimal!

\[ 4x + 2 = 6 \]

\[ 2x + 1 = 3 \]

You have constants on both sides. How can you have all constants on the right and none on the left?

Next ➤
Research Prototype: SolutionVis
User Study

Participants: Eight middle school math teachers
Task: Explore different interfaces and give suggestions on how to improve the intelligent tutor.
Conditions:

“no data” Baseline
List Interface
SolutionVis
<table>
<thead>
<tr>
<th>Feedback Design: Gaming the System/ Protracted Struggle</th>
<th>Interface/logic Design</th>
<th>Hint Adaptness</th>
<th>Hint Clarity/ Correctness</th>
<th>Hint Visibility</th>
<th>Hint Composition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Need to provide instruction that using &quot;Enter&quot; instead of &quot;Finish problem&quot; to go to next step. (5)</td>
<td>The correct step with a different order of items in the equation is not accepted by the tutor. (5)</td>
<td>Address what the students did in the hint. (3)</td>
<td>The hint &quot;You can get the variable by itself by dividing both sides by the coefficient.&quot; is not correct for &quot;,2x=4÷2&quot;. Should &quot;add 4 on both sides first&quot; (3)</td>
<td>Don't show the bottom hint. (1) Let the hints pop up automatically. (1)</td>
<td>—</td>
</tr>
</tbody>
</table>

### Interface/logic Design
- **Hint Adaptness**
  - Need to provide instruction that using "Enter" instead of "Finish problem" to go to next step. (5)
  - The correct step with a different order of items in the equation is not accepted by the tutor. (5)
  - Address what the students did in the hint. (3)

### Hint Clarity/Correctness
- The hint "You can get the variable by itself by dividing both sides by the coefficient." is not correct for ",2x=4÷2". Should "add 4 on both sides first" (3)
- The hint "You have two terms on the left side that you can add together" is not well designed. Explain and give examples about "term" (4)

### Hint Visibility
- Don’t show the bottom hint. (1)
- Let the hints pop up automatically. (1)

### Hint Composition
- —

### Feedback Design: Gaming the System/Protracted Struggle
- Ask a question about the knowledge in the hint to let them think. (1)
- When seeing students gaming the system or submitting random things, provide feedback like "show me your efforts", "show your work" (2)
- When seeing students submitting the same thing multiple times, let the tutor give the answer and move on. (1)
- Provide hints based on how many times the student asks for it. If the student asks a second time, showing a different hint. (1)
- Check the first hint of each step and make sure it is clear and easy to understand. (1)
- Provide hints on how many times the student asks for it. If the student asks a second time, showing a different hint. (1)
- When seeing students submitting the same thing multiple times, provide feedback like "stop clicking this, please try again" (2)
- Show some examples in the hint. (1)
Results

1. Data about student learning was helpful for teachers to generate useful redesign ideas.

2. The aggregated data in a graph showing in SolutionVis helps teachers find the tutor's problems efficiently.
PeerLenses: Peer-inspired Interactive Learning Path Planning in Online Question Pool

Meng Xia, Mingfei Sun, Huan Wei, Qing Chen, Yong Wang, Lei Shi, Huamin Qu, Xiaojuan Ma

CHI 2019

AI’s role: prediction @ scale
Vis’ role: explaining @ scale
Motivation

<table>
<thead>
<tr>
<th>Pro. ID</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>A + B Problem</td>
</tr>
<tr>
<td>1001</td>
<td>Sum Problem</td>
</tr>
<tr>
<td>1002</td>
<td>A + B Problem II</td>
</tr>
<tr>
<td>1003</td>
<td>Max Sum</td>
</tr>
<tr>
<td>1004</td>
<td>Let the Balloon Rise</td>
</tr>
<tr>
<td>1005</td>
<td>Number Sequence</td>
</tr>
<tr>
<td>1006</td>
<td>Tick and Tick</td>
</tr>
<tr>
<td>1007</td>
<td>Quoit Design</td>
</tr>
<tr>
<td>1008</td>
<td>Elevator</td>
</tr>
<tr>
<td>1009</td>
<td>FatMouse’ Trade</td>
</tr>
<tr>
<td>1010</td>
<td>Tempter of the Bone</td>
</tr>
<tr>
<td>1011</td>
<td>Starship Troopers</td>
</tr>
</tbody>
</table>

Questions Pools:
- No pre-determined syllabus
- A lengthy list indexed by their problem IDs
- Hidden intents

Learners:
- Different learning scenarios
- One learner’s learning scenario may be changing

What to do next? What sequence to follow?
Evaluation: Controlled User Study

Baseline system (List View)

<table>
<thead>
<tr>
<th>Pro. ID</th>
<th>Problem Title</th>
<th>Ratio(Accepted/Submissions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>A + B Problem</td>
<td>30.56%(240770/787844)</td>
</tr>
<tr>
<td>1001</td>
<td>Sum Problem</td>
<td>25.38%(143110/569922)</td>
</tr>
<tr>
<td>1002</td>
<td>A + B Problem II</td>
<td>19.47%(84152/432201)</td>
</tr>
<tr>
<td>1003</td>
<td>Max Sum</td>
<td>23.76%(70413/296345)</td>
</tr>
<tr>
<td>1004</td>
<td>Let the Balloon Rise</td>
<td>39.72%(50043/128661)</td>
</tr>
<tr>
<td>1005</td>
<td>Number Sequence</td>
<td>25.25%(51499/203970)</td>
</tr>
<tr>
<td>1006</td>
<td>Tick and Tick</td>
<td>26.73%(6080/22750)</td>
</tr>
<tr>
<td>1007</td>
<td>Quot Design</td>
<td>26.52%(17197/64856)</td>
</tr>
<tr>
<td>1008</td>
<td>Elevator</td>
<td>54.79%(46878/85565)</td>
</tr>
<tr>
<td>1009</td>
<td>FatMouse’ Trade</td>
<td>34.85%(33070/94883)</td>
</tr>
<tr>
<td>1010</td>
<td>Tempter of the Bone</td>
<td>26.68%(39786/149139)</td>
</tr>
</tbody>
</table>

Submission Type

- 0: Tried only once and failed
- 0.0: Tried several times and failed
- 0..1: Tried several times and succeeded
- 0..1..: Tried, succeeded and validated
- 1: Tried once and succeeded
- 1..: Succeeded and then validated

Primitive PeerLens (Only provide one path)

18 CS students:

- determine the starting question under a specific learning scenario
- find the next question to solve given an existing historical learning path
Results

1. Using peer data is useful.
2. Visualizing more suggestions is useful.
3. Visualizing more suggestions using the proposed visualizations did not increase the complexity.
The 11th China Visualization and Visual Analytics Conference

Human

Content inaccuracy
Improper use of AI

Analyzing @ scale

AI
- Prediction @ scale
- Authoring @ scale
- Assessment @ scale
- Tutoring @ scale

Explaining @ scale

Lack of pedagogical guidance
Not personalized

Testing @ scale

Lack of evaluation
High risk
Ruffle & Riley: Insights From Designing and Evaluating a LLM-Based Conversational Tutoring System

Robin Schmucker, Meng Xia, Amos Azaria, Tom Mitchell

AIED 2024

AI’s role: authoring/tutoring/grading @ scale
Vis’ role: testing @ scale
Intelligent Tutoring Systems (ITSs)

**Benefits of ITSs:**

- Can be as effective as human tutoring
- Can be scaled to millions of learners
- Provide a formal framework for thinking about tutoring processes

![Graph showing learning outcomes of different activities](https://via.placeholder.com/150)

Learning outcomes of different activities [1]

Intelligent Tutoring Systems (ITSs)

Limitations of ITSs:

- High cost of content authoring
- Limited language understanding
- Limited conversational facilities
- Limited question answering facilities

How can LLMs benefit ITSs?

LLMs can provide components to ITS designers:

- By generating lesson texts
- By generating question and hints
- By adding automatic grading capabilities
- By adding question answering capabilities
- …

What would it take to generate an entire ITS?
**Ruffle&Riley: User Interface**

**Organelles in Eukaryotic Cells**

"Form follows function" is a philosophy that many industries follow. In architecture, this means that buildings should be constructed to support the activities housed within them. For example, a skyscraper should include several elevator banks. A hospital should have an easily accessible emergency room.

Our natural world also utilizes the principle of form following function, especially in cell biology. The structure of cell components and organelles evolved to support their specialized functions. For example, in the pancreas there are cells that are responsible for producing digestive enzymes, which are specific types of proteins that help with food digestion. In these cells, ribosomes, the organelles responsible for protein synthesis, are found in high numbers. Here the form (high number of ribosomes) follows the function (protein production). Apart from ribosomes, cells have numerous other organelles. These organelles are specialized compartments inside the cells, and similar to the organs in your body, each organelle has a unique role. In this lesson, we will dive into several examples of organelles and discuss their specific functions.

**Ribosomes**

Ribosomes are the cellular structures responsible for protein synthesis. They may group together into clusters (polysomes) or individual ribosomes may float freely in the cytoplasm. They may be attached to the plasma membrane's cytoplasmic side or the endoplasmic reticulum's cytoplasmic side and the nuclear envelope's outer membrane. Ribosomes are large protein and RNA complexes, each consisting of two subunits, one large and one small (Figure 1). Ribosomes receive their "orders" for protein synthesis from the nucleus where the DNA transcribes into messenger RNA (mRNA). After transcription, the mRNA exits the nucleus and travels to the ribosomes located in the cytoplasm. The ribosomes then translate the code provided by the sequence of the nitrogenous bases in the mRNA into a specific order of amino acids linked together to form proteins. Amino acids are the building blocks of proteins.

**Pre-existing textbook content**

**Intelligent Feedback**

**On demand assistance**

**Misconcept. Correction**
Evaluation Results

Findings of user study

Learning Experience Survey

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Engagement</th>
<th>Understanding</th>
<th>Remembering</th>
<th>Interruption</th>
<th>Coherence</th>
<th>Support</th>
<th>Enjoyment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>4.33 ± 0.52</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Teacher Q/A</td>
<td>5.0 ± 0.53</td>
<td>4.43 ± 0.65</td>
<td>4.43 ± 0.65</td>
<td>2.71 ± 0.64</td>
<td>5.43 ± 0.53</td>
<td>4.57 ± 0.57</td>
<td>3.71 ± 0.52</td>
</tr>
<tr>
<td>LLM Q/A</td>
<td>4.8 ± 0.47</td>
<td>4.4 ± 0.4</td>
<td>4.33 ± 0.42*</td>
<td>2.67 ± 0.45</td>
<td>4.8 ± 0.43*</td>
<td>4.0 ± 0.44</td>
<td>4.0 ± 0.44*</td>
</tr>
<tr>
<td>Ruffle &amp; Riley</td>
<td>5.81 ± 0.3</td>
<td>5.81 ± 0.24</td>
<td>5.76 ± 0.22</td>
<td>2.19 ± 0.34</td>
<td>6.1 ± 0.21</td>
<td>5.9 ± 0.26</td>
<td>5.62 ± 0.31</td>
</tr>
</tbody>
</table>

Post-Test Performance

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Previous Knowledge</th>
<th>Learning Performance (Post-test Scores (i.e., Multiple-Choice Questions))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>2.53 ± 0.41</td>
<td>5.07 ± 0.33</td>
</tr>
<tr>
<td>Teacher Q/A</td>
<td>3.0 ± 0.58</td>
<td>4.14 ± 0.83</td>
</tr>
<tr>
<td>LLM Q/A</td>
<td>2.2 ± 0.3</td>
<td>4.67 ± 0.35</td>
</tr>
<tr>
<td>Ruffle &amp; Riley</td>
<td>2.67 ± 0.43</td>
<td>5.19 ± 0.25</td>
</tr>
</tbody>
</table>

Symbol "*" marks p < 0.05
Interaction Analysis

How successful is Ruffle&Riley at orchestrating conversational tutoring?

- Found no “hallucination” in GPT-4’s outputs
- Sometimes asks for previously covered information
- Lenient towards incomplete explanations
Next step: Simulation & Digital Twin

We are now working how to simulate different learning scenarios and provide educators the feedback at scale using visual analytics so that they can confidently deploy the system in reality.
AI’s role: tutoring @ scale  tutoring/prediction @ scale  authoring/tutoring @ scale
Vis’ role: analyzing @ scale  explaining @ scale  testing @ scale
The 11th China Visualization and Visual Analytics Conference

**Analyzing @ scale**
- StuGPTVis (TVCG 2024)
- RLens (L@S 2022)
- AlgoSolve (CHI 2022)
- BlockLens (L@S 2022)
- Mobile MOOCs (CHI 2022, Best Paper Award)
- Distributed Tutorship (LAK 2022)
- Predication (LAK 2020)
- SeqDynamics (EuroVIS 2020)
- “Gaming the system” (L@S 2020)

**Explaining @ scale**
- SolutionVis (AIED 2023)
- Persua (CSCW 2022)
- Ruffle&Riley (AIED 2024)

**Testing @ scale**
- Peerlens (CHI 2019)
- QLens (TVCG 2021)

Visual Analytics K-12 (VIS 2019, Best Poster Award)
Use visual analytics, AI, and other human-AI interaction techniques and research metaphors to upskill educators and learners to better utilize data and AI for Personalization@Scale!