Research Statement in Human-Centered and Verifiable,Controllable AI

Nowadays, large language models and AI agents can autonomously answer questions and perform actions based on user queries. However, as these systems become more capable and autonomous, it is increasingly difficult for users to understand their behavior, verify their outputs [1, 2], and control their actions (e.g. rewind, pause) [3]. My research is centered around a fundamental question: How can we design AI systems that humans can easily understand, verify, and control?

I have explored this question through several projects:

My first work, PEEB (Part-based Image Classifiers with an Explainable and Editable Bottleneck, NAACL 24), addresses the control problem by making AI predictions editable by design. Instead of treating a model as a fixed black box, PEEB grounds human-readable descriptions of object parts to corresponding visual regions and allows users to directly modify these descriptions to change the model's behavior. This gives users a concrete mechanism to inspect and steer how the model makes its predictions.

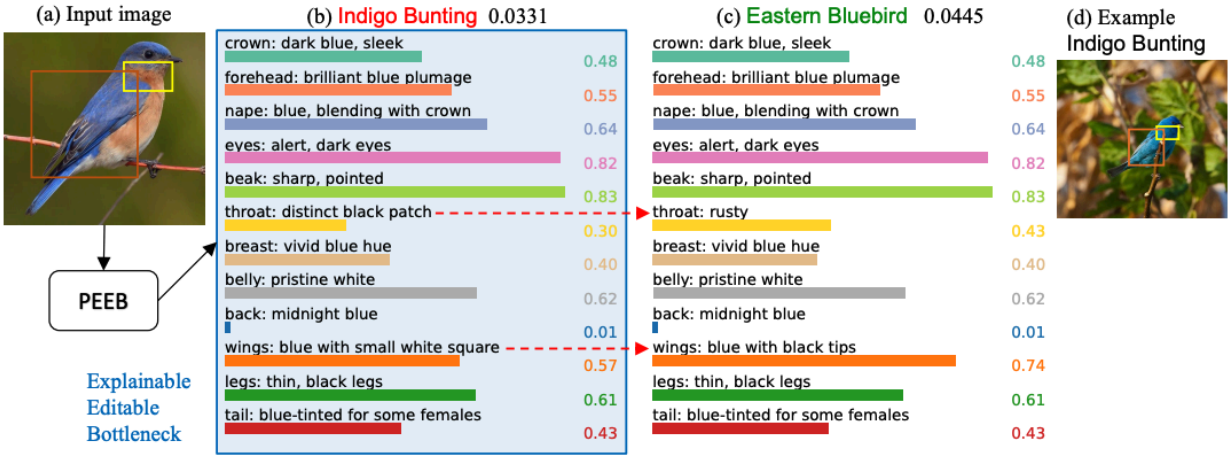

Figure 2: Given an input image (a) from an unseen class of `Eastern Bluebird`, PEEB misclassifies it into `Indigo Bunting` (b), a visually similar blue bird in CUB-200 (d). To add a new class for `Eastern Bluebird` to the 200-class list that PEEB considers when classifying, we clone the 12 textual descriptors of `Indigo Bunting` (b) and **edit** (- -▸) the descriptor of throat and wings (c) to reflect their identification features described on AllAboutBirds.org (*"Male Eastern Bluebirds are vivid, deep blue above and rusty or brick-red on the throat and breast"*). After the edit, PEEB correctly predicts the input image into `Eastern Bluebird` (softmax: 0.0445) out of 201 classes (c). That is, the dot product between the wings text descriptor and the same orange region increases from 0.57 to 0.74.

My second work approaches this problem through **grounded AI, interactive explanations, and human-AI interaction**. In **Highlighted Chain-of-Thought (HoT)**, I developed a method that explicitly connects the model output to the evidence in the input. Rather than presenting explanations as isolated generated text, HoT allows users to trace claims back to their supporting information. Our human study showed that explicit evidence grounding can improve both the task accuracy, help humans to verify the answers.

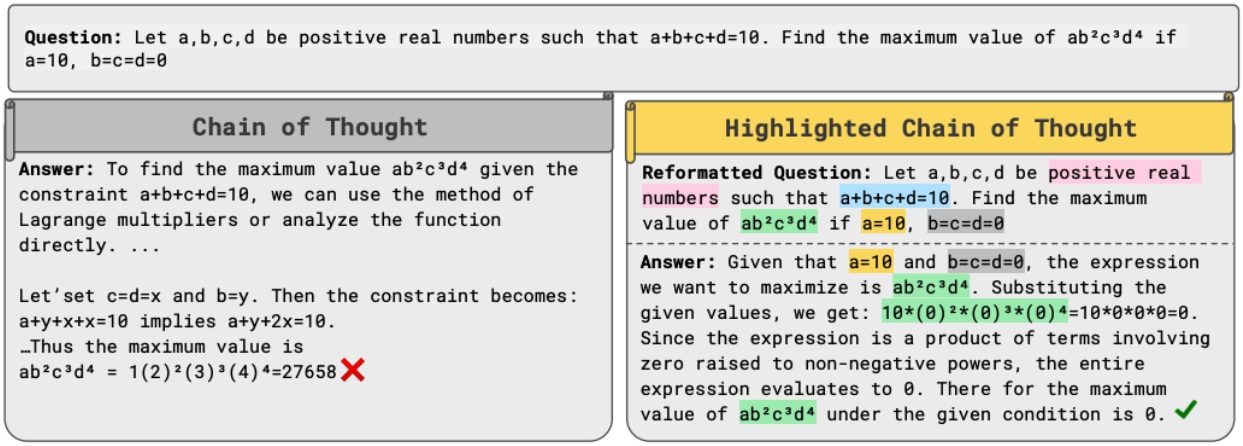

Figure 1: CoT and HoT (ours) responses for a MATH500 question in ReasoningTrap benchmarks (Jang et al., 2025), both generated by `Gemini-1.5-Flash`. **Left**: With CoT prompting, the LLM misses the provided facts *"if a=10, b=c=d=0"*, resulting in an incorrect answer. **Right**: With HoT prompting, the model highlights the key constraint $b = c = d = 0$ and plugs it into the expression $ab^2c^3d^4$ to produce the correct answer of 0. The full reasoning traces of both methods are provided in Tab. 67.

I extended this idea to**PageGuide**, a browser-based AI agent for grounded and verifiable web interaction. PageGuide connects an agent's answers and intermediate actions directly to evidence on webpages, including text, interface elements, images, maps, and other visual information. Our user studies show an important tradeoff: grounding can improve users' ability to detect and verify agent behavior, but detailed inspection can also increase the time required for oversight. This motivates a broader question in my research: **When should an AI system ask for human attention, and when should it act autonomously?**

Figure 2: Existing web agents provide limited support for user verification. PAGEGUIDE (ours) addressed it by providing visual grounding by annotating both the response and the page: it boxes the target hall and parking lot and draws the relation between them, allowing users to directly locate and verify the answer on the map.

My future work will explore three directions. First, I am interested in **adaptive human oversight**. Rather than requiring users to inspect every action, I want to build agents that selectively request verification when uncertainty is high, evidence is conflicting, or an action is difficult to reverse. Human feedback from these interactions could then help improve both the agent's behavior and its ability to decide when future intervention is necessary.

I want to study rewind and redirect for **interactive agents**. When a user changes their intent, adds a new constraint, or notices an error midway through a task, human have to re-prompt and hope that the agents can fix the error. I want to design agents that can identify when to rewind and fix the error, they must identify how far to go back, and which parts of the previous trajectory should be preserved or revised. The goal is to make corrections fast, reliable, and minimally disruptive, enabling users to trust AI agents more while reducing the time and effort required for intervention.

Third, I want to explore **multimodal interaction with generative interfaces**. Current AI systems rely heavily on explicit prompting, which can interrupt a user's reasoning process. I am interested in interfaces where users can communicate naturally through speech, pointing, or other lightweight cues. For example, a user could point to a specific equation or visualization and ask the AI to expand or simplify only that component, while the rest of the generated interface remains intact. My recent proposal explores this idea by grounding spatial and audio cues directly to generated DOM/SVG elements, enabling localized interface adaptation without restarting the overall interaction.

Ultimately, my long-term goal is to develop **human-centered AI systems in which verification and interaction are built into the system itself**. I am particularly interested in studying not only whether AI helps people complete tasks more effectively, but also how different forms of AI assistance affect **human reasoning, reliance, and agency over time**. I hope to build systems that do not simply perform tasks for people, but instead help people reason while retaining meaningful control over increasingly capable AI systems.

References

[1] Kim, S. S. Y., Vaughan, J. W., Liao, Q. V., Lombrozo, T., & Russakovsky, O. (2025). Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 1–19.
[2] Nguyen, T., Bolton, L., Taesiri, M. R., & Nguyen, A. (2026). HoT: Highlighted Chain of Thought for Referencing Supporting Facts from Inputs. Transactions on Machine Learning Research (TMLR), May 2026.
[3] Zhou, J., Roy, A., Gupta, S., Weitekamp, D., & MacLellan, C. J. (2026). When Should Users Check? Modeling Confirmation Frequency in Multi-Step Agentic AI Tasks. *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 1–20.