CENTER FOR DEMOCRACY & TECHNOLOGY

A SERIES ON GENERATIVE AI IN EDUCATION

HALLUCINATIONS

Students' Use of Generative AI: The Threat of Hallucinations

Generative AI systems trained on large amounts of existing data use machine learning to produce new content (e.g., text or images) in response to user prompts. In education, generative AI is most often talked about in the context of academic integrity, with teachers expressing fears of cheating in the classroom.

However, our polling of teachers, parents, and students shows that **45 percent** of students who say that they have used generative AI report using it for personal reasons, while only **23 percent** of students report using it for school. Of those who have used the technology for personal reasons, many of the uses are high stakes – **29 percent** have used it for dealing with anxiety or mental health issues, **22 percent** have used it for dealing with issues with friends, and **16 percent** have used it for dealing with family issues. As a result, even in the context of personal use, generative AI systems that produce incorrect information can have significant harmful consequences.

What Are Hallucinations And Why Do They Happen?

By virtue of their style of writing and the way they impart information, generative AI systems can appear to be trustworthy and authoritative sources of information. However, these systems often produce text that is factually incorrect. These factual errors are referred to as "hallucinations." Hallucinations are a consequence of both the design and operating structure of generative AI systems.

From a design standpoint, generative AI systems are built with the intention of mimicking human-produced text. To accomplish this, they are generally trained on enormous datasets of text from which the system learns about the structure of sentences and paragraphs, and then produces text that seems meaningful to human readers by repeatedly predicting the next most sensible word. This process is not designed to create content that is true or correct, just that is sensical.

Structurally, most generative AI systems operate "offline," meaning they are not actively pulling data from the internet to respond to prompts. So they are restricted to the data contained in their training datasets. This makes generative AI systems particularly unreliable when it comes to current events that do not appear in their training datasets.

The Potential Detrimental Impacts of Hallucinations on Students

The reality of generative AI hallucinations paired with high levels of student personal use for important issues raise huge concerns about access to accurate information in times of crisis. For example, students could be asking ChatGPT (or another generative AI tool) questions about how to deal with an ongoing mental health issue, which could potentially be a life or death situation. Because most generative AI systems likely to be used by students are trained on information gleaned from the internet, they may replicate common misunderstandings of sensitive issues like mental health challenges, gender roles, and sexual orientation.

In addition to traditional hallucinations, which are simply incorrect information, generative AI can also have significant emotional impacts on students who utilize the tool for personal reasons by replicating societal biases against marginalized populations, including on the basis of race, gender, or sexual orientation. Students, especially during the vital developmental stages of K-12 education, may internalize these biases, whether against themselves or others.

Hallucinations are also of significant concern when students use generative AI platforms for academic use. The possibility for students to receive inaccurate information can run directly counter to schools' goal of imparting reliable, quality information to students. Students who do not understand these tools' potential for hallucinations may use the tools in ineffective ways and miss beneficial uses. Without understanding generative AI's shortcomings and limitations, students may not be able to effectively leverage its potential as a tool to supplement their learning and critical thinking skills.

How Should Schools Approach the Issue of Hallucination?

To combat the potential devastating consequences of generative AI hallucinations in both the personal and academic contexts, schools must:

- **Understand the limitations of generative AI and ensure that teachers are adequately trained.** Though the potential benefits of these tools to enhance learning can be exciting, it is imperative for school officials to be thoroughly steeped in the technology's shortcomings and impart that knowledge on educators. Teachers play a critical role in ensuring that generative AI is used in responsible, appropriate ways in the classroom. But to do so, they need access to resources and training.

- **Continue to invest in counselors and other mental health supports.** Schools should be wary of pushing students towards using generative AI as a resource on a topic as sensitive as their mental health. Ongoing mental health issues require human empathy and expertise, so schools should not be acquiring generative AI tools to replace or even to triage care that would otherwise be provided by a human. If schools are going to procure a tool to *supplement* the counselors and mental health supports already in place, they should reference our guidance on responsible procurement principles, since even as a supplemental tool, generative AI systems can cause harm if not tested and governed appropriately.

- **Provide education for students on what generative AI is, how it works, and why hallucinations occur.** To combat the unchecked public hype around generative AI, schools should equip students with basic knowledge of the technology, its capabilities and limitations, and how it can go wrong in both academic and personal uses.

- **Provide education for students on media literacy and research skills.** The release of ChatGPT last November underscored the need for students to understand how to be responsible, effective consumers of knowledge via new technological tools. Student use of generative AI is increasingly inevitable in the same way as their use of the internet, so it is vital that schools provide students training and resources on how to assess the accuracy and reliability of information gleaned through ChatGPT and other generative AI platforms.

- **Ensure that teachers and students understand when generative AI is appropriate to use.** Generative AI is not meant to replace traditional teaching and learning by any means – it is not a replacement for knowledge and not an effective therapist or sounding board for personal issues. However, it can be used, for example, as an assistive tool to help improve writing or used as a novel tool for research when beginning to explore a new topic. Schools should provide guidance and training to both teachers and students on how to make effective use of generative AI.