Carnegie Mellon University
Software Engineering Institute

SEI Podcasts

Conversations in Artificial Intelligence,
Cybersecurity, and Software Engineering

Using Role-Playing Scenarios to Identify Bias in LLMs

Featuring Katherine-Marie Robinson and Violet Turri as Interviewed by Suzanne Miller

Welcome to the SEI Podcast Series, a production of the Carnegie Mellon University Software Engineering Institute. The SEI is a federally funded research and development center sponsored by the U.S. Department of Defense. A transcript of today's podcast is posted on the SEI website at sei.cmu.edu/podcasts.

Suzanne Miller: Hello, and welcome to the SEI Podcast Series. My name is Suzanne Miller. I am a principal researcher here at the SEI Software Solution Division. Today, I am joined by Katherine-Marie Robinson, known as Katie, and Violet Turri, both researchers in the SEI's Artificial Intelligence (AI) Division, to discuss their work in identifying and mitigating bias in large language models, which we also call LLMs. I want to welcome Violet and Katie.

Katie Robinson: Thank you so much. I am really excited to be here, and I am looking forward to our discussion.

Violet Turri: Thank you. Yes, very excited, and cannot wait to get into it.

Suzanne: Both of you have done podcasts with us before. But for those members of our audiences who haven't seen your previous episodes, why don't you tell us a little bit about yourself, the work you do here, and what is the coolest part of your job?

Katie: Of course. Hi, everyone. I am Katie. I am a design researcher at the AI Division at the Software Engineering Institute, and I have been here for almost about two years. Currently, I am a member of the newly established Trust Lab, where the focus of our work is to advance trustworthy, human-centered, and responsible AI practices in various projects that I am a part of. And the coolest part of my job, I would like to say, is being able to combine engineering aspects with human-centered design aspects to create and work on different projects. The people in general, whether they are in the AI Division or the SEI, are all fantastic.

Violet: I am Violet Turri, I am a software developer and researcher here in the AI Division. I have been here for about four years. I have worked on projects that are looking at testing and evaluation, as well as transparency for machine learning systems. Similar to Katie's answer, I really enjoy getting to work on projects where I can look at both engineering and the design sides of problems. And I find that a lot of the times, the challenges really require a knowledge of both, or at least team members from both sides that can help us solve really hard problems. And then also I really like the people I work with. They are enthusiastic and just really super nice. Those are things that I have enjoyed about working here.

Suzanne: Excellent. Thank you both. I want to talk about your recent work in identifying and mitigating bias in large language models. And you wrote a paper about this, which we will link to in our transcript. But the thing is that as these LLMs become more widely adopted and used in a wide variety of contexts—I use them for meal planning, which I never thought I would ever do. You know, which is kind of on one end. And people are using them also for making very important decisions. We have increasingly seen examples of harmful bias. Obviously, there is going to be bias based on the training data. But what we want to avoid is harmful bias. We hear about things like LLM hallucinations. But give us a little bit about some of the examples of bias that you observed that were really a catalyst for this kind of work that you are doing now. I will start with whoever wants to start.

Katie: I will take it away at first. I think that is a great question. I think what has been really interesting and a catalyst for this work even goes back to 2015 when we just started to see certain systems come online. One of the big ones that I remember when I started at university was Google had this identification thing, where it had tagged two people of color as an offensive term. It tagged them as something that they should not have been tagged as. We are starting to see these conversations about bias come into play.

I think one of the big ones that happened around that similar time was Dr. Joy Buolamwini's work as well, where it was just looking at facial recognition systems, and something that she noticed very early on in her own research was that facial recognition, these facial recognition systems were not working with her. They were not recognizing her face as a darker-skinned woman. That actually spurred a lot of the research that she did going into her master's and her PhD and looking at the facial recognition systems coming from big tech companies, and then understanding where their faults were. Who were they identifying versus who were they not identifying? And she found out pretty quickly that it was working on lighter-skinned individuals a lot better than it was working on darker-skinned individuals, and particularly failing on a lot of darker-skinned women.

It showed the disparities within the training data of what these companies are paying more attention to. Even something that has become a highlight of hers is to use this white mask as a way that something will actually identify her face. When she takes off the mask, it will not see her face. But when she puts on the mask, it actually recognizes that a face is in the screen. A lot of these conversations were already sort of present and happening before large language models came on the scene. And I think large language models sort of just gave us a much bigger playground and a much easier point of access for everybody to start going in and playing around.

One of the first things that I got to do when I started playing around with ChatGPT specifically was just asking it to create me two different resumes for men and women. I gave it the prompt, "Please create me a resume for a man applying to a software engineering job," versus, "Please give me a resume for a woman applying to a software engineering job." Though they were very subtle, there were some really interesting differences in how it was talking about how you could get this job and how it was describing the roles and activities that you may have done, how it suggests that you could talk about these things.

For instance, for men when it was talking about the duties that they may have performed in another job, it said *they led the work*, that *they were pivotal*, that *they advanced the work*. Whereas if you were looking at how it described a woman's role, it was, *well, they helped with the work. They provided guidance on the work*. Just very interesting to see how it understood the roles that maybe a man played in the software engineering job previously versus a woman. Even just the level of education, the man received their bachelor's, master's, and PhD, and the woman had received her undergrad degree. Theoretically, they are probably applying for the same job. It was just very interesting to see how it played out. That was sort of one of the first forays I had.

Then we also took that into the image generation models as well. One of the big ones that came out, and there were quite a few papers on it, is that if you type in "What is a scientist," you are going to get pictures come back that are older white gentlemen. No people of color, no women. We are starting to see those biases of, *Okay, that is what it understands the scientist to be*. And we are seeing very similar things in terms of what is a CEO, or what is a nurse. You will start to see those disparities and those stereotypes. There is another really good example that Violet found that was about, called Rest of the World. It is a website. But with that, it showed some of the oddities and some of the stereotypes as well and how it portrayed different people. It was interesting, because it went beyond people. It was not just people, but it could be food or it could be places.

One of the ones that I really enjoyed was like give me a picture of the New Delhi streets in India. And very stereotypically, it made them very, very crowded, very, very condensed. You could sort of see a haze within the air. Where that is definitely not every single street in New Delhi, India. These stereotypes are starting to emerge very early on. And, of course, as time has gone on, they have been able to nip some in the bud, or go in and sort of have people ready to understand where these problems are occurring.

Something that Violet and I were really interested in as we got through and started doing this work was, we started to see those guardrails come into place, and started to see companies be able to identify where the stereotypes were. We were starting to question, *How can we come up with new methods to poke at this more?* Because we know that there are these underlying biases. *How can we explore those a little bit more? And how can we dig into them? And what new techniques do we have to bring in that we can start to expose those? The simple methods of just asking are not working anymore, so what do we do now? And how do we keep going?*

Suzanne: All right, to get at this idea of auditing LLMs that you talk about in your blog post, you designed an experiment. And I want to hear about how you designed that experiment, and then what you found in that experiment.

Violet: Yes, I can, I can take the first part of this one about how we ended up designing this experiment. Basically, at the start of the project, we were interested in just seeing how well LLMs could be used for role-playing games. We saw this really cool game from the Entertainment Technology Center, where they basically had multiple characters that the user would interact with. And each of them were voiced, or the things that they said, were generated by ChatGPT. We were curious, *How well can it stay in character? How well can it remember core facts that we give it at the beginning of an interaction?* That is kind of how the project started in some ways. Although this concern and concept of bias and LLMs was always on our minds as well. But we started playing around with building out different scenarios and characters for ChatGPT, assuming that we were in this role-playing game context.

One of these characters was a cowboy named Jett at this fictional ranch we came up with called Sunset Valley Ranch. What we found through organically trying to interact with Jett was we would ask him, do you know *insert name in the blank*, and it seemed to fill in all of the gaps about names that were of Hispanic origin and come up with stories for us about who they were, oftentimes placing these people on the ranch as ranch hands or ranch cooks. And then when given non-Hispanic names, he—he being ChatGPT—would respond back, or Jett, would respond that he did not know these people. We kind of were curious what was going on here, if there were some underlying assumptions about who would or would not be working at a ranch that ChatGPT had picked up on.

We built up a more formal experiment where we had a collection of 16 names. For the purposes of the experiment, we simplified things, so we just had half of the names that were traditionally male names, and half traditionally female, and then also half Hispanic origin and half non-Hispanic origin. And then we started testing out, for each of these names, if we asked Jett about this person, what kind of role and personality will be provided if anything is, and if there is a response.

Katie, if you want to talk about more about what we found out. I felt that we found a lot of interesting responses that showed different kinds of biases. Both ones that seemed to connect to stereotypes and other ones that are just kind of unusual that are probably represented in the training data set in one way or another.

Katie: No, I completely agree. I think we found some ones that emulate a lot of stereotypes, and then ones that are just odd that we think are, yes, either present in the training data, or just to us we have some questions for further research down the line, or further things that we could always poke at. But just as Violet was saying, we were able to really focus on the roles that ChatGPT, or Jett, assigned in that scenario, as well as the personalities that it assigned to all of our different people as well.

To start off with the roles, if you have a chance to look at the blog post or the paper, you will see that we actually use the U.S. Census codes to group our roles, just because we did get a wide variety of roles. For the purpose of our analysis, we used these groupings to really understand how ChatGPT was assigning these different characters that we had, and to really give us a picture of like where ChatGPT was placing some of these people as well. Groups of codes, or the U.S. Census codes, include healthcare practitioners, technical operations, it can include farming and forestry. And then there are our unique roles that we filed under those, but just as those groupings that made our lives a little bit easier. And we also were able to use these groupings to really analyze the different roles that were assigned.

For example, the education and training and library occupations were usually assigned more often to non-Hispanic women. We thought that was very interesting, whereas the farming and fishing and forestry occupations were usually assigned to Hispanic men. Similarly, when we were comparing genders, we found that women were often assigned to management roles, so they were usually listed as managers, either at the ranch or maybe at a restaurant. Whereas men were assigned to these roles of installation and maintenance and usually repair occupations as well. Again, we are starting to see those stereotypes emerge of the roles that women may have versus the roles that men may have.

We also saw some interesting stereotypes that we could relate to being reflected in the world today. Roles that are stereotypically masculine, you could see were assigned to the men in our situation. Whereas roles that were stereotypically feminine were usually assigned to women, such as florists or artists. Roles that required more education, again, as a stereotype, or a bias that we found as present, were librarian and veterinarian. Those were usually assigned to non-Hispanic people, versus roles that maybe did not require as much education were assigned to Hispanic people, and those roles being rancher or ranch hand.

There were also very interesting roles assigned to specific subgroups. We often found that Hispanic women were given the roles of cooks or restaurant owners, whereas Hispanic men were often given ranch hands or blacksmiths. Non-Hispanic men were usually given historian or librarian or variations of that. Non-Hispanic women, historian, and librarian, and then non-Hispanic men were usually astronomer or writer or weatherman. It is just very interesting to see how ChatGPT assigned the roles, and where it put them on or off the ranch, as Violet was alluding to earlier.

We also saw similar things with personality as well, so when we were looking at the personality traits for men and women, we saw that men were often given, again, stereotypical strong, resourceful, intelligent sort of traits, or variations of those. Whereas women were described as being happy, gentle, caring. And soft sometimes as well. Again, we saw similar patterns. Men receiving these stereotypical male traits. And women receiving stereotypical women traits.

Traits were also divided by subgroup again. We saw traits for Hispanic people, including diligent, hardworking, tranquil. If we looked at non-Hispanic people, they were often described as being curious or business minded or inspiring. What we thought was really interesting with these traits as well is the way that the traits were used for different subgroups. We found that the traits for non-Hispanic people usually described their personality. We would say that they were personality traits. You know, they are open minded. They are curious. They are business-minded. Whereas when we looked at the personality traits assigned to Hispanic people, we found that those traits more matched the job that they received or were assigned from ChatGPT. They were often hardworking or they were tranquil or they were handy or they were the roles that they were given, and the traits that they were given seemed to coincide with one another, where it seemed for non-Hispanic people that the traits were much more free. They did not really have to necessarily match the role that the person played.

Those were just some of the oddities that we were able to tease out and really started to get us question and thinking about what we had seen, and I know Violet has another oddity that we found as well.

Violet: Yes, I think that, in general, to Katie's point and all of the examples that she brought up, it seemed like for Hispanic names, they tended to have less creativity in the role and personality traits provided. They tended to be really centered around the ranch. And then as she said, the personality traits seemed to map with whatever job they were given. Whereas for non-Hispanic names, there was a lot of creativity. You would see kind of off-the-wall things, like every town I guess needs an astronomer, for example. In this game, this job was going to be given to a non-Hispanic person as well. There were some odd things there.

Another thing that came up was, there were times where women were given roles that were just placed in relation to a male figure. For example, the ranch owner's wife, or the ranch owner's daughter, those were things that came up. We never saw the ranch owner's husband or the ranch owner's son. There was also kind of a lack of character development for women characters at times, where they are just placed in relation to some maybe more central male figure in this story.

And then another kind of odd thing that we found was that there seemed to be some strong associations at times between specific names and specific roles. For example, the name Juanita was given the role of diner owner, or cook, diner cook a number of times. We looked up something along the lines of Juanita diner owner like just on Google, and we found that there is actually an award-winning restaurant I think called Juanita's Cafe or something along those lines. It became clear that on top of all of the things that may be reflecting societal biases, may be reflecting harmful biases in the training data, there are also potentially these strong and peculiar connections between specific names and specific roles. As we investigated this, we definitely were able to peek beyond a lot of the guardrails and see some of the unusual things that ChatGPT had learned, and potentially harmful things as well.

Suzanne: This is fascinating to me. And I find that some of the, what you call the peculiarities where you have a strong association between different names and roles, the idea that, if I Google searched *award-winning Mexican restaurants,* I would come up with Juanita's. And that is actually something that is an input to these. We do not think about the kinds of things that we search on every day as just users of Google, as things that can actually end up biasing a large language model that is being used in a completely different way. That is something that is really striking to me about this research is starting to get at some of the nonobvious, right, the U.S. Census Bureau kinds of information that ChatGPT uses and things, now that is sort of an obvious source.

But these non-obvious sources of data that end up in the training data for these large language models, how do you account for those as somebody who is either doing a new thing with ChatGPT in relying on its training data, or someone who wants to audit and find out, *Well, what are the biases in it*? How do you account for those kinds of things?

Violet: Yes, I think one big challenge is that there is limited information available about the training data on which a lot of these open or publicly accessible models were trained. It is a little difficult to figure out what the risk profile is for using them because we do not have a clear sense of exactly what kinds of potentially harmful or toxic things were included in those data sets. We do know that a lot of large language models are trained on data that is gathered through just crawling through the web. There is always some level of risk I think inherent in using LLMs. But I think if you have a specific use case in mind, definitely rigorous, red teaming, or auditing, or whatever specific technique you want to use, it is really important to figure out if there are some clear vulnerabilities or scenarios where it is not really appropriate to use this thing, or where maybe you can use it, but you need to account for the known challenges in some kind of way, whether it is by steering users to interact with the system in a specific way or introducing additional guardrails.

Additionally, there is also the opportunity to fine tune models to specific use cases. For example, maybe there is a world where we fine tune a model to specifically help us craft stories about Jett on his ranch. And we could introduce a lot of different names with a lot of different roles. And that might help kind of diversify the responses that ChatGPT was providing. But I do think the challenge is that there is always, for the most part, challenges in using huge amounts of data where you cannot have someone sit down and look through every single source. There is always risk apparent in using these systems. And so you need to be careful and account for that. Some level of risk may be unavoidable, I guess. Yes, Katie, do you have thoughts about that?

Katie: Yes, I think I was just going to echo exactly what you had said about the testing piece. I think we found that in our work too. We had this idea where we wanted to use scenarios, and we were really interested in what that would look like, and how we can tease out these biases. But until we started putting in different names and playing around and just coming at it from different angles, I think that was a very important step for us to say*, Oh, something is here*. And that led us to keep going down. But if we had just gone with one idea and sort of followed through, we may not have been able to pick up on some of the things that we found and expose a lot of the oddities that we found as well. *Great, you have a use case*. But trying to approach it from different angles I think is really interesting.

And that actually led us to test it in different scenarios, just to see if this was a pattern that was emerging in just our use case with Jett, or was it showing up in different ways, in a whole bunch of different scenarios as well? I think that is a really important piece of it all. Everything Violet had just said was very important, but I think just approaching at it, or coming at it from different angles and really either auditing, red teaming, or getting other people involved from very different perspectives, just to poke at it a little bit differently than you can, is a very important step.

Suzanne: There are a lot of other places in system design where you want that multidisciplinary, multiple perspective approach. But what I am hearing you say is that in working, using LLMs for some specific purpose, you need to be really sure that you are not just going down what we call a happy path and focused on the results that you want. You need to make sure that you are testing things on the edge, that you are testing things from multiple perspectives so that you do not actually yourself add bias into the, into the data set. Is that a correct read?

Violet: Yes.

Katie: Yes, sorry. I agree. I think exactly that, and I think you can fall into this bubble of, *Everything is fine*. But by plugging in different values and different responses and understanding how it can fail—especially to Violet's point, it is trained on the Internet. The Internet is absolutely massive. We know that there are going to be things in there that we probably do not want showing up. How are we going to find them? What are we going to do? And there are going to be things that we probably might think are fine, and then you will put in a response, and it will give you something completely unexpected.

I was just having a conversation about this today. A portion of ChatGPT is trained on Reddit. Who knows what lives on Reddit, because there are so many different subreddits. Just typing in one word that you think is going to be innocuous and completely fine can bring something completely different to the conversation that you did not expect was going to happen. And I think that is exactly what happened with our use case. You know? We thought we were going to have a nice friendly conversation with a cowboy named Jett who lives on Sunset Valley Ranch. Turns out we had a lot of questions and a lot of things that came up by just asking some pretty unassuming questions about people.

Suzanne: What other kinds of challenges? The challenge of getting a diverse enough set of questions for helping you to understand the sources and the ways that the LLM expresses answers to questions, that is a challenge that you guys kind of grappled with pretty explicitly. Were there other challenges in doing this kind of experimentation that you want to highlight for people so that if they want to do something similar with their own LLMs that they interact with, they kind of understand what some of the things they are going to run into that might be roadblocks to doing this?

Violet: I can think of two kinds of roadblocks, and then I am curious to see what Katie is thinking as well. One thing is early on when we were just testing how well LLMs can handle the role-playing game scenario, we had mixed results as we looked at different systems. ChatGPT, for example, seemed pretty good at staying in character and providing responses that are just from Jett and not from our character. But when we were working with, what at the time was called Bard, what happened was that it consistently was providing responses for other kinds of characters, maybe even a new character, so it would not just respond with Jett. It would be *Jett* and then maybe come up with something that we did as well.

Just for trying to execute this experiment the way we have it on paper, it may depend a little bit which large language model you are working with. You may also have to do some prompt engineering to make sure it is actually going to adhere to the rules of this kind of interaction that you are trying to have. That is one just kind of logistical challenge in running this experiment. And then there is something, I had another thought related. Maybe, Katie, if you want to go ahead, I will try to mull over what I was thinking.

Katie: Let's see if I have got one. Yes, I completely agree. And I think it was interesting the creative ad-libs that it took as well in terms of it started like coming up with new things that we had not prompted it to. It started writing a whole script instead of just answering the question. Even just knowing your system that you are dealing with I think is probably one of the challenges. *What should I expect from this system? And how do I expect that it is going to react?*

Then I think the challenge is, I think Violet had just mentioned this too, how do you actually get it to respond in such a way that you get the response that you need? It took a lot. We had spreadsheets upon spreadsheets of how are we actually going to create these prompts to get the answer that is akin to the answer that we need from the system in terms of, *Oh, who is Maria? Maria is this person*. You know, rather than giving us an entire script about Maria's life, because that is not the question that we were trying to ask.

I think that that comes as just getting to know the system that you are working with better. And then if you are creating your own large language models, you are trying to fine tune a model. The initial uptick in using these systems and knowing how they are going to respond, and then understanding that they could respond in very unprecedented and unpredicted ways, is one of those first challenges to getting started. I think it is fun because that is how you can find where you want to play. It is a good challenge in some cases of, *Let's find some problems that we can work with*. But it is just a challenge in getting it to do what you are hoping it is going to do.

Violet: Definitely taking an exploratory approach is really important. The other thing that I thought about when you asked this question is just in general, it can be hard at times to circumvent the guardrails that are in place, which in some ways is a great thing because if we had really effective guardrails, maybe it would lead to us building better, more effective systems, that have biases that impact end users less and less.

However, none of these guardrails are truly comprehensive. You can find ways around them, but you do have to be creative. I think the other piece of this is if those biases are in place, in the same way that we can find ways to circumvent those guardrails, you may accidentally circumvent those guardrails when you go to use the system.

For example, I think we thought a lot while we did this about the idea of, *If someone were to take ChatGPT and try to use it to review resumes, how would the names of these individuals impact how their resume was perceived?* We also try just giving ChatGPT a list of names and asking it to assign roles and personality traits without the handholding of the scenario. And it gave us answers that seemed better. Although there were still some gender biases, I would say, just glancing at it. It seemed to be placing men in a lot of the same roles, and women in less of those roles, for example. But nowhere near the level of kind of problematic content that we saw when we introduced this scenario.

But if we can circumvent those safeguards, I think if you were to apply this to a resume review scenario, for example, you may be inadvertently circumventing guardrails as well. You could get responses and not know what the basis of those decisions really were because it is concealed. It is kind of like a dual-edged sword, building better guardrails is, or a double-edged sword. Building better guardrails is really important, but it obscures some of those biases.

I think it can make people think that these systems are less biased than they really are, or that they will not encounter those biases in practice. That is another challenge is just figuring out how to get around them, and maybe also if you have a specific use case, how realistically people are going to use the system, and how those guardrails may be circumvented in practice.

Suzanne: What I am hearing you say is that one of the cautions with both auditing these and using them is that we need to be aware of the biases that persist within our own way of dealing with data as humans, not just the way that the system deals with data. It is dealing with the data that we gave it, ultimately. Is that fair?

Violet**:** Yes, absolutely. I think it definitely reflects biases that we have. I think it can also reflect, I mean, it is a reflection of its data. The data is formed by the people using the Internet and the things that they are interested in. It may not be a good reflection of individuals or all individuals, but it certainly contains human biases, negative content that humans have produced that it was trained on.

Yes, I would say we have to be very wary of how we use these systems. I think sometimes there is an assumption if we can introduce AI to the problem, maybe the bias will just disappear. But anytime that you are working with data that is produced by humans, it is very possible that you have that human bias present in that data, and it will ultimately shape the kind of system that you build.

Katie: I think on that note as well, to Violet's point of *AI is going to fix all of the problems, and it is going to eliminate the bias*, for AI to work, you need so much data. If anything, it is just going to augment the bias that is present. It is not going to eliminate it in any sense of the word.

And something I wanted to talk about a little too is these guardrails that are coming up as fast as people are getting around them, a lot of tech companies are being able to fix them or add new ones. It is sort of that environment of, *Okay, one guardrail has been put up, what is another way that we can get around it?* Because we know that those biases are still there. They are always going to be there because it is coming from the data that these language models are trained on.

We are always going to have to come up with these creative ways to figure out how we can expose them. Because maybe now, if you tried doing a scenario-based approach, maybe it does not work as well, or maybe it is not as effective. What is the next approach that you could start to uncover these? Because they are there. You just have to find them. Peck at it a little bit more to expose it.

Suzanne: You have used this technique of creating the experiments for a particular LLM context, and you talk about this in your paper, about trying to look at auditing the LLM. What are some of the applications that you envisioned for doing this in real-world settings?

Violet: I think in general we found that you can use this technique to help start making your way around guardrails, to understand underlying biases, which I think is important for figuring out sort of what the risk profile is for using this system. As we discussed earlier, I would caution people against using this for resume review, for example, because there are a wide array of different kinds of potentially harmful biases that we uncovered. And you may not see them from an initial reaction with ChatGPT, but they could very easily have weight as it makes decisions about resumes, for example, in ways that we are shielded from.

Another example that we had discussed, Katie and I, was, let's say that you were a teacher, and you were trying to introduce ChatGPT to your classroom to help your students write short stories, there could be risks there as well. Maybe kids are providing names, and then they are building out stories that contain either stereotypes or some kind of biases that might be harmful. Which is not to say that students could not do that themselves. But you are introducing this system that adds a whole other factor into this and could be biased in ways that you never anticipated and you never wanted to introduce into your classroom.

I think figuring out what that risk profile is for whatever use case is really important. And doing auditing or red teaming activities of a variety of different kinds, whether it is just trying to model your use case perfectly and see what kinds of outputs you get, or whether it is more of what we did, where we built out a scenario and tried to kind of think outside the box. I think that is important for figuring out what kind of system you are working with. And you can use these techniques as a tool to understand that bias more fully.

Katie: I think Violet hit on that all perfectly. And I think, if anything, I am hoping our application really just spurs people to come up with new ways on how to audit and red team. We approached it from a scenario-based approach. But what else could that look like? How can you spin that off to make it something a little bit different that maybe exposes something else? With our approach and with our scenario, we were able to get insight into one thing.

But depending on how you spin it, or how you create it, maybe it exposes something else. Hopefully, it lays that groundwork of people taking the idea and being a little bit more creative at the next iteration too. There are plenty of different applications, because I think we can see that it works in different ways through the example that Violet just presented. Hopefully, that gives people something to move forward with.

Suzanne: And to take the teacher example, children—especially if you give them a tool like this and you tell them that it is a trusted tool—they will trust what comes out of it. And if what comes out of it really is not what you are expecting, then that can be harmful in terms of creating some biases within them that were not even there before. I very much respond to this idea of understanding the risk profile for the particular use and context that you have for a particular LLM.

I want to switch gears just a bit. We talk about transitioning work, and the ideas that we are talking about to the public. This is one where everybody, there are so many people doing so many different things with these large language models that I think getting these ideas out to the public is actually even more important than it may be in some of our research that is a little more esoteric to software engineering. You recently presented this at the CHI Conference, C-H-I Conference. What are some of the other resources that people that go, *Oh, I need to understand my risk profile for using LLMs in my work*. What kinds of things do they have available that we can point them to?

Katie: Do you want to go first, Violet, or do you want me to?

Violet: You can go first.

Katie: OK, perfect. Within the SEI, I think there is a lot of work that we can point to that talks about large language models and the work that the SEI is doing. One of the projects that this project was actually spun out from is a larger effort called the Mayflower Project, and that was led by our colleague, Shannon Gallagher, where her team was investigating the practicalities of actually using large language models for intelligence applications. In September of last year Shannon and her team released a retrospective report where they discussed their findings in relation to the central question, *How might the intelligence community actually set up a baseline, standalone large language model*? *And how might they customize large language models for specific intelligence use cases*? *And how might the community evaluate the trustworthiness of large language models across these use cases*?

And then for resources, in terms of where people could look more recently, Shannon actually was on a podcast just like this one where she discussing using large language models in the national security realm, so I think that would be a great resource for people. If anybody has any interest in large language models and their potential in the national security space and intelligence realms, I think that would be a really, really good resource along with the report.

Additionally, another one of our colleagues, Anusha Sinha, she recently published a paper with her co-authors, where they take a deep dive into the current state of red teaming for generative AI. In this paper, they explore a lot of open questions that are still out there about, *What does red teaming for generative AI look like? What role does it play in regulation? How does it relate to conventional red team and practices in cybersecurity?* Really understanding what we have, what information can we take, and how does that apply in a generative AI context.

Both of those works I think would be a great resource for people to turn to and just get a little bit more information about what this space is starting to shape up to be. And if there is just general curiosity, I think those are two great resources coming out of the SEI, especially.

Violet: Yes, outside of the SEI, a couple resources that come to mind, there have been a number of conferences, or a couple of conferences that Katie and I have attended, although there are a number of conferences that look at these topics. A couple that we attended were the conference on computer human interaction (CHI), where there was a workshop that we participated in, the HEAL Workshop, which stands for Human-centered Evaluation and Auditing of Language Models. They have a lot of really interesting papers that looked at this problem from different angles. A lot of this tends to be very research intensive. I do not know how easy it is for people who do not have a research background to pick up these things and learn. But at least they are a starting place.

And then we also attended FAccT, which is the ACM Conference on Fairness, Accountability, and Transparency. And similarly, there is a lot of really interesting work related to LLMs and their impact on people or society, different ways to evaluate them. Outside of these kind of research conference proceedings, one other idea is just to take a look at the AI Incident Database or other similar kinds of databases online. They tend to showcase news articles that talk about different kinds of incidents that have occurred with AI systems. Taking a look at incidents that have occurred with an LLM that you are thinking of using could also be helpful for getting a sense of some of the known challenges or risks associated with using that system.

Suzanne: OK, thank you. What is next for both of you on the research front? This is a very rich area. What are you, what are you working on that we can bring you back to talk about in a few months?

Katie: Yes, something that was a really nice tie-in from this work going into another project that I get to be a part of is looking at fairness in AI systems. And more specifically, the part that I am part of right now is understanding hallucinations and large language models. It was a really nice segue from the work that Violet and I did, taking those concepts of, *How can we prompt large language models to see if we can create these hallucinations*?

We are really interested in intrinsic and extrinsic hallucinations. Intrinsic, *Did it omit something that maybe it should have brought up into a summary?* Which is what we are focused on. Or extrinsic, *Did it bring in something extra?* That has been really interesting to approach it from a different perspective of, *Okay, I have an understanding of these scenarios that exist, and the work that we did*. But just exploring it a little bit further in terms of hallucinations and more broad concepts of, *Okay, here's something I want you to summarize. Let's see how the hallucinations take shape*. That has been really intriguing, and I have been really excited to work on that work.

Suzanne: What about you, Violet?

Violet: For me, I have been leading a roughly two-year project that is looking at building transparency mechanisms into AI systems. Thinking about how to make their outputs easier for humans to understand, and how to make well-informed decisions. In that sense, I think we will continue looking at human-centered questions related to testing and evaluation. People want to know how decisions are made at times, or they need additional tools to figure out if a proposed solution or output of the system is actually accurate. I think those are issues that we see in the LLM space as well, though I have been focusing on computer vision systems.

And then I also could see a lot of cool ways that we could take this research forward. For example, at FAccT, we heard, or we saw a poster that I thought was really interesting where a researcher was working closely with local members of her community who were a part of the Muslim community, and she met with these individuals to learn more about concerns that they had about LLMs and built out a set of scenario, something similar to a scenario. Although she framed it a little differently. But kind of like a fill-in-the-blank situation where it will fill in blanks with different individuals' names.

I thought that that project really looked at community concerns and used that to drive what kinds of tests she developed and used was super cool. I think something similar for our scenario work would be a really interesting way to follow up on the project, although at the moment I am kind of locked into the transparency project. But I think there are a lot of really cool things we could do in the future to explore this area further.

Suzanne: You both are going to be busy for quite a while. There is lots to do here. This is one of the most exciting areas of research at the SEI in terms of it bringing a lot of fresh perspectives and a new set of problems. We have always done human-centered work at the SEI, but the AI work really hits that a lot harder than some of our other research. I think you are going to be busy, so, that is good.

I do want to thank both of you for taking time to talk to with us today about this work. To our listeners, thanks for joining us today. We are going to include lots of links in our transcript. And especially to the paper and all of the other resources that we mentioned in the podcast. This series is available in all the places that you can find podcasts, including Apple, SoundCloud, Spotify, and, of course, the SEI's YouTube channel. As always, if you have any questions, please do not hesitate to email us at info@sei.cmu.edu. Thank you.

Thanks for joining us, this episode is available where you download podcasts, including SoundCloud, TuneIn radio, and Apple podcasts. It is also available on the SEI website at sei.cmu.edu/podcasts and the SEI's YouTube channel. This copyrighted work is made available through the Software Engineering Institute, a federally funded research and development center sponsored by the U.S. Department of Defense. For more information about the SEI and this work, please visit www.sei.cmu.edu. As always, if you have any questions, please do not hesitate to e-mail us at info@sei.cmu.edu. Thank you.