Carnegie Mellon University
Software Engineering Institute

SEI Podcasts

Conversations in Artificial Intelligence,
Cybersecurity, and Software Engineering

Using Large Language Models in the National Security Realm

Featuring Shannon Gallagher as Interviewed by Rachel Dzombak

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.

Rachel Dzombak: Hi everyone, and welcome to the SEI podcast series. My name is Dr. Rachel Dzombak, and I am a senior advisor to the director of the SEI's AI [Artificial Intelligence] Division. Today, we are here to talk about the potential uses of large language models [LLMs] in the intelligence community. Earlier this year, the Office of the Director of National Intelligence [ODNI], spurred by a request from the White House, reached out to the SEI's AI Division to explore use cases for large language models and national security, specifically within the intelligence community. Between May and September of 2023, a team of researchers in the SEI's AI Division attempted to determine if and how tools built on top of large language models might be customized for use by the intelligence community. The team also explored the trustworthiness of such tools in this realm. Joining me today to talk about this work is Dr. Shannon Gallagher, who is the AI engineering team lead. Welcome, Dr. Gallagher.

Shannon Gallagher: Thank you, Rachel. It is a pleasure to be here today.

Rachel: I am so excited to jump into this conversation. Let's begin by having you tell our audience a little bit about yourself and what brought you to the SEI.

Shannon: I have been working at the SEI for two years now. I am a machine learning research scientist here. I am also the team lead of the research scientists in the AI engineering center, within the AI Division. What brought me here? Like many people at SEI, I have kind of a roundabout way into public and government research, but it was really quite wonderful. I did my PhD in statistics at CMU and did that in 2019. My thesis was actually about modeling infectious diseases. So, as you can imagine, it took a little way to get to uses for government but made it here to SEI. It has been a very, very exciting place to be, especially working with large language models and generative AI in general.

Rachel: Great. Well, we definitely today will talk about generative AI and large language models. Maybe just for fun, we will talk a little bit about the connections you see between infectious disease and computer science as well. Let's start out by jumping into our exploration of large language models for national security. Since OpenAI released ChatGPT last year, there has been a lot of attention and press around the use of large language models in different contexts. Can we set the stage for our audience by explaining what is a large language model?

Shannon: Yes. Rachel, one of the things that we really wanted to take a look at and set about in Mayflower, which is the project that we have been calling about this large language model's research, is how can I, as a layperson or [if I] have some knowledge of AI, what can my mental model of a large language model be? What we really want to emphasize is that a large language model isn't just one AI system. It is a collection of tools that includes AI systems along with more traditional tools that are combined together in a way that usually ends with some little interface that we can chat and input-output in a typical chat response.

Rachel: Give me an example of what interacting with a large language model looks like.

Shannon: I grew up during the AOL Instant Messenger days. In a lot of ways it reminds me of that. I put in some input. *Hi, how are you doing?* The output takes in the context of what I gave it to the large language model, which will be able to take that and respond to that in a contextual manner. It is *I am doing well. Thanks, how are you?* The neat part about these is that they are really extended since those way back when days of just in the early 2000s and can really take in the context over a long period of time and be able to respond in a seemingly very intelligent manner, definitely a very fluent conversation.

Rachel: I love that you just use the, and put special attention on, a seemingly intelligent manner. What is actually happening behind the scenes there? Why is it seemingly intelligent and not just pure intelligence?

Shannon: These models are very exciting pieces of technology and exploring them has been a lot of fun. I think at the end of the day, I am really grateful for my background in statistics, which really makes me think that everything is…there is a lot of randomness involved. That is very true for these large language models. These large language models are really awesome in the sense that they have seen so much information. They pretty much read the entire Internet. But at the end of the day, how they respond is based on those patterns of information that they have learned. So they are very, very good at predicting what the next word or even sequence of words is going to be. But that is completely based upon all the previous text that it has seen before, along with some guardrails and other reinforcement learning that has been added to really make these systems appeal to humans.

Rachel: Fantastic. I just think breaking it down a little bit is always helpful as we set the stage for a conversation. You also mentioned, or you alluded to Mayflower, the project that our team has recently pursued that was about examining the potential use of large language models in national security, and in particular the intelligence community. Could you talk about the motivation for that work?

Shannon: It was an exciting time, and as you said. GPT-4 had just been released, before that, just a few months prior to that, ChatGPT was released. So large language models were really in the mainstream. There was a lot of buzz about them, about, *What can we do with them*? *How can we use them? How can we make them work?* This included the White House as well. They were interested in how they could use large language models for government use cases. The White House asked ODNI, the Office of the Director of National Intelligence, to figure this problem out. As a trusted collaborator of the U.S. government, the SEI was then tasked with trying to figure out how we can use large language models for intelligence reporting specifically.

Rachel: You are sitting at your desk, I am picturing, and the prompt comes in saying, *Hey Shannon, we need your team to start to understand how can you use large language models in the context of national security*. Where do you even start with a prompt like that? What were the initial steps that you took to get that work going?

Shannon: This was a massive effort, and it really at the time was an all-hands-on-deck effort. Because as I hope to explain, or not explain but discuss during this conversation, this is an interdisciplinary effort. This is not just one person or even a few people getting this to work. It took our infrastructure engineers, our software engineers, many, many research scientists, and also our responsible AI researchers as well. First of all, we had to coordinate everyone, so getting everyone on a schedule and then devising a plan. One of the most important things for our purposes at the time, there weren't very many publicly open foundational models available, which are the large language models that companies, organizations, or institutions have poured in either sometimes millions of dollars, sometimes even much, much more than that, to create these very fluent large language models. We wanted to be able to research and experiment with them, so step one was trying to figure out which foundational model we wanted to use for our purposes.

Rachel: My two follow-up questions on that are, how did you make that selection? Then also, what does experimentation look like? What were the types of experiments that you and the team were running?

Shannon: The first one of what did we end up using and how did we get there? It took a while, and we did a lot of research. We read a lot of articles both on the mainstream, on computer blogs, and also in the academic literature to really see what was available that would suit our purposes. We had a number of constraints. For example, we wanted to be able to tinker with the architecture of the system, which meant that we needed a model that was publicly available. Already that meant a lot of the proprietary models like ChatGPT were not possible. We had to look at the ones that were publicly available. So that was a big constraint. After that, we looked at a variety of both infrastructure—what GPUs we had to run these models and how much space we needed. We looked at cost projections for if we wanted to train these models further about how much we would expect that to cost. Then finally, how well they did compared to other large language models. Fortunately for us, there were a few websites with leaderboards basically of how different publicly available large language models performed on different little tasks. From these, we were able to discern which of these at the time was best to start to use.

Rachel: As you are thinking about from the research angle, engineering concerns, cost concerns, really technology adoption concerns for the government customer, I am sure you are also holding in your mind potential use cases. The design-thinking professional in me is always thinking about being human-centered. As you were starting the research, what were the use cases that you saw for large language models in the context of national security?

Shannon: That is a good question. It is something that really developed over time. Fortunately, we were able to start with that use case of, *How can we use this for intelligence reporting?* ODNI was incredibly helpful in that idea generation. But talking with other government customers, with other researchers, we were able to determine a number of avenues that we thought were useful for large language models. That included code generation, synthetic data generation, interfacing with knowledge management systems, and writing, querying, modifying, and summarizing documents, which is especially the use case that we found with ODNI. Besides that, we also found there was a lot of interest in government communities about war gaming potentials with large language models.

Rachel: As a follow-up to that, I know in the work we do with government customers, there is a myth, right, that is perpetuated that the government is slow to adopt things. But I am already seeing a number of folks not just experimenting with the use of large language models but actively integrating them into their workflows. I was curious what your perspective is, to what extent from your research, did you see that large language models were currently being used in government?

Shannon: Yes. We found that as well. It is definitely very exciting. Maybe compared to previous technologies, large language models are something that I think are quite accessible, and people really have an appetite for them. They want to be able to use these technologies in their current work. Informally, these tools are very, very easily accessible, usually requiring an email or even less just to be able to use. I have personally talked with a number of colleagues who use them for idea generation, for writing simple emails, and helping with writing code. Right away those are ways that people can immediately use these large language models. I think in more formal ways, those have also been integrated fairly rapidly. There are a number of tools that are available for government purposes. There is a very big question of what sort of information can I input into these large language models. That is obviously of very big concern, especially because these large language models can touch national security purposes. There are a few tools, a few that I am aware of are Ask Sage, which integrates a number of AI platforms including Microsoft's Azure and OpenAI. So you can use them for CUI, which is controlled unclassified information. That allows for different documents that these foundational models haven't seen before and to allow for customizations in them. There are also platforms like AWS's GovCloud in which, for example, that we have been utilizing to be able to run these large language models on the cloud. I don't think there is one standard way right now, at the moment, to use these for government purposes. But a lot of development has occurred between May and now, and we expect that to continue, especially with accessibility. I wouldn't say that they are the easiest to use right now for government purposes but are definitely being developed to get there.

Rachel: Let's go into that a little bit deeper. You made a comment saying that there are a lot of questions right now about what types of information can I input into a large language model or tools built on top of large language models. I know that you and I in the past have talked about that our customer needs sometimes conflict with the qualities of most large language models. Could you give me an example of why that is true?

Shannon: Yes, so one thing about large language models is that they generally remember what you tell them. If you put in sensitive information that is going to go, usually go into its database, and perhaps we'll even learn from it. That means that it is quite possible that that large language model can use what you put into it to tell other people. You can easily imagine there could be all sorts of concerns with sensitive or PII [personally identifiable information]. It is very important to only put in information that you are sure is at the level that the system can handle.

Rachel: Absolutely. What are some of the other concerns that folks should be thinking about when it comes to large language models and trust in the context of national security?

Shannon: There are a lot of them. As we talked about a little before, there is a lot of randomness involved with these large language models. They appear very fluent, but as we know in our personal testing, we see a lot of what is commonly known as hallucinations that occur, which is information that the large language model is just kind of making up. We have seen a lot of alternative histories being written by large language models. More concretely, for example, we will ask about a certain researcher who is fairly publicly known, and it will start making up titles of papers of this researcher that don't actually exist. It is very convincing sounding. It sounds like this person could have written these papers, but they do not exist. It is very fluent and can mimic human speech. But to actually verify the contents of its information requires much more work.

Rachel: Yes, one of the things that we of course know is, the field is moving so quickly that practices for test and evaluation really are in a reactive mode because the technology is changing so quickly. That is just one of the engineering challenges that exist with large language models today. You mentioned before that you lead a lot of our AI engineering efforts. Of course, ODNI is one of the primary supporters in our work to grow the field of AI engineering. I was wondering if we could turn now and talk a little bit about some of the engineering challenges in leveraging large language models for national security use cases.

Shannon: Yes, I am glad you brought that up because one of the principal engineers on our effort, Andrew Mellinger, he has a laundry list for us. At the heart of this, every step in the AI stack basically, from all the way down at the ground level of the hardware and then all the way to the top where people are interacting with this, there are challenges to determine. First of all, can we trust the output of this system? Is it working as intended? How if any—because it isn't always AI that is causing potential issues with the system—how can that potentially add to the number of evaluations in the verifications that we have to add to the system?

Rachel: To build on that bit and talk about some of the other engineering concerns, I want to go in two directions. One is a technical question, which is in the summary paper that you published, which of course we will link to in our transcript, you recommended that government agencies should consider augmentation of foundational models instead of fine-tuning them. Could you talk a little bit more about that recommendation? What are the implications it has for different stages of that AI stack or different components of the AI stack?

Shannon: Like you said, one thing that we recommended currently is to use augmentation or orchestration over fine-tuning. The reason behind that is very practical. Right now, we do not have the eval, the available test, and metrics to really determine the effectiveness of a large language model after it has been introduced to new data. The result is that we are right now, without those metrics and evaluations, we know that fine-tuning generally is a much more costly endeavor than being able to use orchestration, for example, retrieval augmented generation, which is called RAG, which is sometimes you can think of it as a search plus prompt engineering. That is a very less costly tool that doesn't require further training of the model. Those are two different avenues to customize a model. We know they both work to some degree, but trying to determine that quantitatively right now is very difficult. So instead of potentially wasting money on something that we cannot verify at the moment, it is more practical to use the less costly procedure of augmentation.

Rachel: I think what you are talking about too is that, yes, there is a need to have the metrics evolve over time, but you also had said, which of course we know because we live in the world, that the field has advanced since when this work started in May. I was curious if you could talk a little bit about some of the main technical improvements that you are seeing, or engineering improvements, and how that would make you rethink any of the elements of the work done in the project.

Shannon: There have been a number of improvements that have occurred over these efforts. About midway in the project, we had a new model that was released every week. That was an effort to look at. But definitely augmentation and orchestration is really starting to play a big role. We see a lot of companies really starting to endorse this method of customizing the models. I think in a lot of ways, it is because currently, it is a lot more easily verifiable because it implements tools that we traditionally understand and can verify and just uses the large language model on top of them to create improved results. I think because we trust those, the parts within it more, it is easier to handle having that result as an output.

Rachel: I think what you are hitting on here is so many of the intricacies and interdependencies between different engineering components of the system. Before you had mentioned this was a whole-team interdisciplinary effort. I was wondering, also in light of the technology changes, could you talk a little bit about how you are seeing the skills needed to engineer large language models evolve over time.

Shannon: Yes, one thing that we learned is that large language models can be very persnickety in the sense where that they can give you a very good answer, but usually, we require a lot of coaxing. I think that more formally is known as prompt engineering of what to tell the large language model, so you get that output that you want out of it. That is something that is both a concern, I think, to our government customers. But also just from a research perspective, it is a very interesting of, *How do I educate the user, and how do I assist the user in creating prompts that eventually elicit the response that is most helpful to the person*. That is something that has also really become apparent.

You can even see job resumes now for prompt entities.

Rachel: Which I always find so fascinating because in in my world, framing and reframing is the heart of the innovation process. Being able to see and ask questions from a variety of different perspectives. I think it is interesting there is a lot of experimentation happening with prompt engineering today, and I completely agree with you. It is going to be an area of growth as people think about, *How are the questions I am asking driving different meanings and driving*... I love that you use the word coaxing because it creates different responses based on how you frame the question. Within that response, you also brought up users. Ultimately, if large language models are to be used in the context of government, they also have to be usable. Could you talk a little bit about some of the barriers today to adopting large language models in our everyday workflows?

Shannon: In some ways, I think there is an adage that goes for large language models pretty well; that they are pretty simple to use and very difficult to master. It doesn't take a lot to be able to say, *Hey, how is it going, large language model*? But to be able to tell you about the current political climate or recent events in the city can require a lot of special wording to be used. In that way, there is some learning curve. So pretty easy for anyone, especially for recreational purposes, *Can you write this funny poem?* to fairly easy to do. But to have it to be able to trust the output of that, I think, is where the majority of the work is going to be with the large language models. When we actually need to know is if it says, *This and this happened on this date and in this location*, we need to be able to confirm that as a result. Right now, that takes a lot of human manual verification.

Rachel: Absolutely. If you were a government leader, say you ascended, right, and you had a team of government folks working for you. Or, if you had the opportunity, which we do all the time, to counsel government leaders saying, *Hey, what should I be telling my team? What should I be telling my managers*? *What do they need to know about large language models?* What are the one or two factors or elements about large language models that you wish the general public, and especially government employees, understood better?

Shannon: I think there is a lot of overlap with cybersecurity here, in that, I think they can be incredibly useful tools but require a healthy dose of skepticism with the responses. Don't believe everything that the large language model says at face value. Try to verify it from other sources and perhaps even other large language models. Another thing is I really do think they are going to be helpful technology, even if it ends up just being something as simple as writing emails or scheduling events. I think that still would be a very incredible time saver. There are definitely more potential applications for large language models. I guess that is two things. Right now, I think, use but cautiously. But also, I don't think they are going away.

Rachel: I mean we are all using them every day whether we realize it or not, in our Google searches, our email responses. Yes, I think people forget that the DoD is the largest employer in the U.S., if not the world. So some small change in that person's day-to-day workflow, even just generating some texts or simple response. At scale, that can have transformative productivity effects. I love thinking about the implications in that way. For the Mayflower paper that we mentioned, that of course we will link in our resources, you had a bunch of recommendations included in that guidance for government agencies starting in this field. I am curious, any others that have popped up recently? What are you reading, learning from, etcetera?

Shannon: One thing that we have noticed in the past few months is that there is a big question to the cost of large language models. I think the initial number can be really daunting, especially when you hear about GPT-4 and such, where those very, very hundreds of millions of dollars sums are used. But what we have found due to a lot of the community of researchers and other individuals who just like to tinker around, for at least for a research perspective, a lot of these techniques and abilities are very, very accessible on quite small computers and only need a few people. That includes document collection as well. If your organization has, just a few documents, you don't need thousands of documents. It is great if you do, but if you have 25 documents that you wanted to be able to explore with, this is still something that is quite accessible. I think that we were originally thinking that we were going to spend most of our budget on computer resources. But with some careful thought, because I think it still is easy to spend that money very quickly, but with some thought about what we want to use it on and trying to come up with the research question first, we actually think this is very accessible for smaller agencies as well.

Rachel: I love that. I think there is such an intimidation factor often, both in terms of the usability but also, yes, in the resourcing. Knowing that experimentation is accessible is huge for organizations, especially ones that are not technology-first organizations. So glad you shared that. As you mentioned, of course, the field is evolving every single day. What are the big questions that you are pursuing in your research right now about large language models?

Shannon: Part of the AI division, one thing that we are very excited about is that we have a new trust lab. I think ultimately what we want for large language models is for their implementation in a trusted manner. I think towards that, we are working with our infrastructure engineers and the other scientists and trust researchers is how can we make better metrics for these large language models for our government customers specifically. So right now, a lot of metrics. I have referenced a few of them, but right now they are, they are like tests basically. That can be awesome if your large language model can perform very well on the SATs. But maybe that is not directly applicable for our government customers. So how to evaluate these. Both mathematically, in a rigorous manner, how can we visualize these results better? Ultimately, how can we better explain the outputs from these large language models to our customers? I think a large part of that is going to be through augmentation where we force the large language model to quote show its work and either cite its sources or reason step by step instead of just outputting an answer without any context.

Rachel: Awesome. Shannon, thank you for talking with us today. We will include links in the transcript to resources mentioned during this podcast. Finally, a reminder to our audience that our podcasts are available every place you download podcasts as well as the SEI's YouTube channel. If you like what you heard today, please give us a thumbs up. Thanks again for joining us.

Thanks for joining us. This episode is available where you download podcasts, including SoundCloud, Spotify, 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 email us at info@sei.cmu.edu. Thank you.