SDS PODCAST EPISODE 784: ALIGNING LARGE LANGUAGE MODELS, WITH SINAN OZDEMIR

Jon Krohn:	00:02	This is episode number 784 with Sinan Ozdemir, Founder and CTO at LoopGenius.
	00:19	Welcome back to the Super Data Science Podcast. Today's Five-Minute Friday episode was filmed live with a guest, the brilliant Sinan Ozdemir at the Open Data Science Conference East, or ODSC East for short, which was held in Boston in late April. Sinan is Founder and CTO of LoopGenius, a generative AI startup. He authored several excellent books, including most recently, the bestselling, "Quick Start Guide to Large Language Models." He's a serial AI entrepreneur, including founding a Y Combinator-backed generative AI startup way back in 2015 that was later acquired. In today's episode, Sinan provides an overview of what it means for an LLM to be aligned, and this overview should be of interest to technical and non-technical folks alike. Ready? Let's jump right into our conversation.
	01:09	Sinan, welcome to your third time, I think, on the podcast. Great to have you here. We're shooting live at ODSC East in Boston. It's been a great conference. You gave a talk on LLMs, which is topical. You're doing a lot of content creation now around LLMs. I guess a lot of people are talking about them in general, including, I'd like to highlight your book, "A Quick Start Guide to LLMs," which people can pick up and get quickly started on training and deploying their own LLMs.
Sinan Ozdemir:	01:40	That's the hope.
Jon Krohn:	01:41	Your talk at ODSC East was about aligning outputs from LLMs?
Sinan Ozdemir:	01:47	Aligning LLMs in general, and the concept of alignment. It's actually one of the major case studies in the book itself is, how do you take these raw, pre-trained LLMs, like Llama 3 or even GPT-2, which is one of the earlier

case studies that I did, and how do you teach them how to hold a conversation? And then how do you teach them to hold a conversation while learning new information? And what does alignment look like outside of instruction? And all of that goes into this big word of alignment that we all use every day.

Jon Krohn: 02:21 Very cool. So, what kinds of techniques are critical for people to know? I know techniques like reinforcement learning from human feedback, reinforcement learning from AI feedback, which Anthropic has popularized. There's also just plain old supervised learning, fine-tuning, which is easier to get up and running. Are those all viable alignment techniques or approaches to alignment?

Sinan Ozdemir: 02:44 They absolutely are. The concept of alignment is a bit nebulous in the sense that most people, when they define the term, they'll say that LLMs that are aligned are performing in line with human expectation. What does that mean? It's answering your question, but is it answering your question well, conversationally? Is it answering your question without bringing up topics that you don't really want to necessarily hear from your chatbot? And you can instill all of that behavior through things like supervised fine-tuning, by providing high quality example conversations; telling it, showing it, these are good conversations. And of course there's reinforcement learning from feedback, in general. The concept of reinforcement learning demands the idea of some kind of reward mechanism or some feedback mechanism, and it doesn't have to be human, to your point. It can be not even an AI providing feedback, it could just be any kind of classifier.

03:41 So one of our other examples was producing summaries of news articles that are just more neutral sounding, and the idea of what is neutral is alignment. I'm expecting

neutrality out of this, let's say, news article summarizer, but how do you quantify neutrality? Well, that's what sentiment classifiers basically do. You take the logit values of the neutral class and say, "Here's your reward on how neutral you are. Optimize for this." So it doesn't have to be, "Oh, the humans prefer this." It doesn't have to be even the AI prefers this. It's just any kind of feedback mechanism that can go into a reinforcement learning is viable.

Jon Krohn: 04:21 Very cool. What would happen if I had an LLM and I trained it without alignment? What would that even mean? How can I train without alignment? And then what would be bad about deploying a system like that?

Sinan Ozdemir: 04:35 True. That's a good question, because again, the term alignment is such a big term when we talk about, what are your expectations of the model? So when people talk about training with or without alignment, it's an interesting question, because when OpenAI talks about alignment, they're generally not just talking about instructional alignment, meaning if you ask it a question, it will answer your question, they're also tacking on the idea of, "Well, it's also going to be helpful and not harmful."

05:04 But really those are two separate ideas. Making an LLM instructional, i.e. answering your question, is one form of alignment and a separate form of alignment would be, "But not on these topics," and, "This is the behavior that I want you to embody. You're still answering questions, but also think about your guardrails and if you're being helpful and not harmful." So, the idea of training an LLM without alignment is more just fine-tuning to a specific task, which in some ways you could consider a form of alignment, because it is performing in some expected manner. But usually we're talking about guardrails and helpfulness and things like that.

Jon Krohn:	05:48	Gotcha. Makes a lot of sense. And so, if somebody is listening to this podcast and they want to be able to ensure that their algorithm is aligned, if they're using an open source LLM, I know that some of the weight providers, I don't want to say open source providers-
Sinan Ozdemir:	06:04	I like that.
Jon Krohn:	06:04	... but companies like Meta, when they put Llama 3 out there, they spend, it seems like, millions of dollars on ensuring alignment. And so that gives you an open source option to feel comfortable taking off the shelf an open source model that is going to be well aligned with ethical and safe answers. In contrast, I believe that if you download a Mistral LLM, they haven't put that kind of alignment effort into it. So, I guess that's one approach that somebody could follow. They could download an open source LLM from a provider that they're confident has put a lot of effort into alignment.
Sinan Ozdemir:	06:49	How do you quantify that?
Jon Krohn:	06:50	Well, how do you quantify that? Yeah, it's hard to know. And, let's say you do a lot of fine-tuning of even a Llama, or you create your own LLM from scratch, how could a listener to this podcast most easily get about getting the right data or the right frameworks for doing alignment themselves?
Sinan Ozdemir:	07:10	Yeah, that's a big question, because to answer the previous one of, how do you even know if your LLM is even aligned or not, the answer is testing, evaluation. And the subset of that would be benchmarks, open standardized validation sets that we all more or less agree that if it performs well, it is aligned.
Jon Krohn:	07:34	And there are benchmarks just for alignment out there.

Sinan Ozdemir:	07:36	There are benchmarks for different specific types of alignment, like, "Are you being factual? Are you being hurtful? Are you being helpful?" These all tend to have their own specific types of benchmark, like TruthfulQA is a benchmark for being factual, but you also have... OpenAI created a data set, actually in one of their first papers on alignment, they talked about a dataset they handcrafted about testing for hurtfulness. So for anyone listening to this who wants to understand, what are the tools and the frameworks, coming back to what we said earlier is; well, what are you aligning to? If you're aligning to factuality, then some kind of supervised fine-tuning where you're actually giving it the answers to questions is going to be really helpful. If you're aligning to something like style, let's take Grok, X's Grok, their fun mode versus their regular mode. There's no difference in factuality per se, at least not intentional, but there's a style difference. Well, how do you test for fun mode? What is fun? Who is defining that? Who's quantifying that?
	08:38	So a lot of it comes down to a really internal look at, "Well, what are we actually aligning this thing to?" And, "Is there a way to quantify that? And if so," and there better be, "What's the data set for doing that and can we test that over time?" That's probably the biggest piece of advice I would give anyone trying to get into just alignment as a whole.
Jon Krohn:	08:57	Nice. All right. Thanks, Sinan. This has been a quick, eye-opening episode on the importance, as well as practical tips, for delivering on alignment with LLMs. Thank you very much and we look forward to your fourth appearance in the future.
Sinan Ozdemir:	09:12	Thanks for having me, as usual.
Jon Krohn:	09:14	All right, that's it for today's practical episode on LLM model alignment. If you enjoyed it, consider supporting

this show by sharing, reviewing, or subscribing, but most importantly, just keep on listening. And until next time, keep on rocking it out there. I'm looking forward to enjoying another round of the Super Data Science Podcast with you very soon.