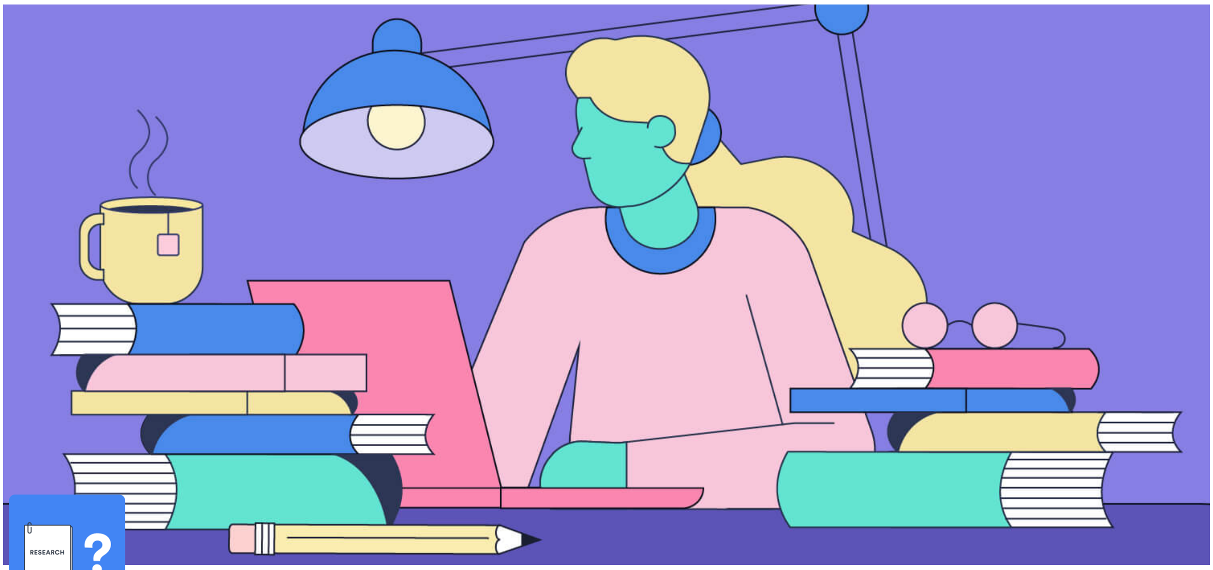


Course conclusion and ongoing research

> Source: Coursera

▼ Responsible AI with Dr Sephus

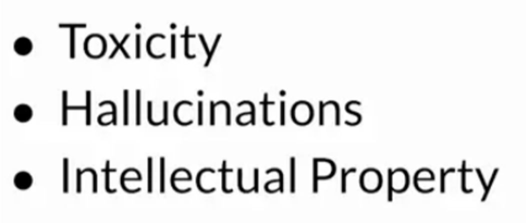
LLM-powered applications are still in their infancy, and researchers announce new techniques or strategies to improve performance and reliability almost daily. This course can only include what is known or understood at launch, but we are confident this field will continue to evolve quickly. Let's highlight a few areas of active research. With the growth of AI comes the recognition that we must all use it responsibly.

Below are the answers from Dr Nashlie Sephus about responsible AI in generative AI with large language models.

Dr. Sephus is a principal technology evangelist for Amazon AI at AWS. In this role, she focuses on fairness and accuracy, identifying and mitigating potential biases in Artificial Intelligence. She formally led the Amazon Visual Search team as an applied scientist in Atlanta. Fun fact: the team launched the Visual Search for placement parts on the Amazon shopping app.

▼ Special challenges of responsible generative AI

What are some new risks and challenges of responsible AI, specifically in the context of generative AI with large language models?



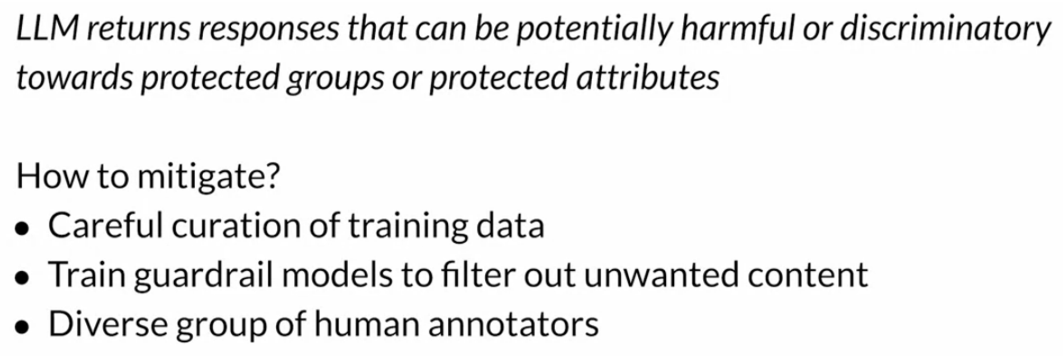
- Toxicity
- Hallucinations
- Intellectual Property

Focus on three challenges: toxicity, hallucinations, and intellectual property.

▼ Toxicity

Let's start with the first one, toxicity. What is the challenge in detail? And what advice do practitioners have for mitigating it?

So toxicity at its core, meaning toxic, implies certain language or content that can be harmful or discriminatory towards certain groups, especially towards marginalized groups or protected groups.



LLM returns responses that can be potentially harmful or discriminatory towards protected groups or protected attributes

How to mitigate?

- Careful curation of training data
- Train guardrail models to filter out unwanted content
- Diverse group of human annotators

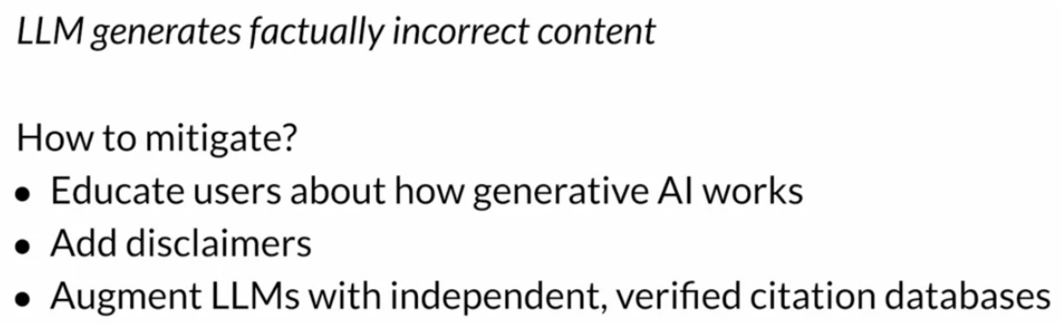
Let's start with the training data. That is the basis of every AI, so we can start by curating the training data. The guardrail models can also be trained to detect and filter out unwanted content in the training data. We also consider how much human annotation is involved in training data and training annotations.

We want to ensure that we provide enough guidance to those annotators and educate a very diverse group of annotators so that they can understand how to pull out certain data or how to mark certain data. This is a great point about the importance of having diversity among human annotators.

▼ Hallucinations

In Hallucinations, we think about things that are not true or that seem like they could be true, but they aren't. They have no basis. This is exactly what it means with generative AI due to how we train large language models or neural networks in general.

We often don't know what the model is learning, so it will sometimes try to fill gaps where it has missing data. This often leads to false statements or hallucinations.



LLM generates factually incorrect content

How to mitigate?

- Educate users about how generative AI works
- Add disclaimers
- Augment LLMs with independent, verified citation databases

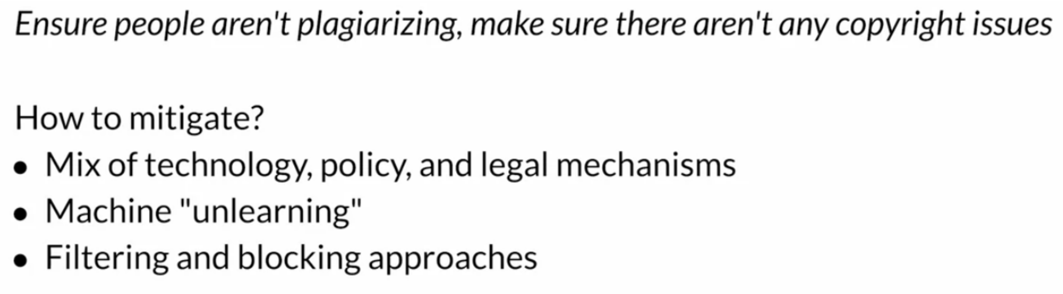
We can educate the users that this is the reality of this technology and add a disclaimer so they know that this is something you should be able to look out for.

We can also augment large language models with independent and verified sources to double-check against the data we’re getting back. Ensure that we develop methods for attributing generated output to particular training data so that we can always trace back to where we got the information.

Last, we always want to define the intended and unintended use cases. Again, because these things can cause hallucinations, we want to ensure the user is aware and transparent about how these things operate.

▼ Intellectual Property

This will have to be addressed because people are using data regarding AI that has been returned from these models. This can be plagiarizing someone's previous work, or you can have copyright issues for pieces of work and existing content.



Ensure people aren't plagiarizing, make sure there aren't any copyright issues

How to mitigate?

- Mix of technology, policy, and legal mechanisms
- Machine "unlearning"
- Filtering and blocking approaches

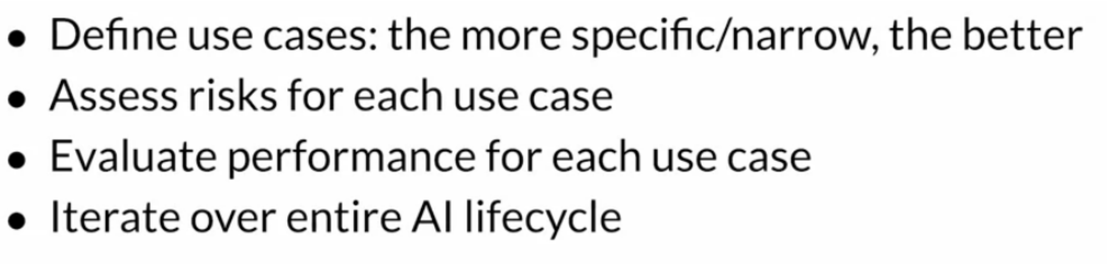
Various technologies, policymakers, and other legal mechanisms will likely address this over time. We also want to incorporate a system of governance to ensure that every stakeholder is doing what they need to do to prevent this from happening in the near term.

There's a new concept of machine unlearning in which protected content or its effects on generative AI outputs are reduced or removed. So, this is just one very primitive approach in research today.

We can also do filtering or blocking approaches that compare generated content to protected content and training data and suppress or replace it if it's too similar before presenting it to the user.

Responsibly build and use generative AI models

Defining use cases is very important. The more specific, the more narrow, the better.



- Define use cases: the more specific/narrow, the better
- Assess risks for each use case
- Evaluate performance for each use case
- Iterate over entire AI lifecycle

One example where we use gentive AI to test and evaluate the robustness of a system is when it comes to face ID systems. We use journeys of AI to create different versions of a face. For example, if I'm trying to test a system that uses my face to unlock my phone, I want to make sure I test it with different versions of my face, with long hair, with short hair, with glasses on, with makeup on, with no makeup on. We can also use gentle AI to do this at scale. This is an example of how we use that to test its robustness.

Also, we want to make sure we access the risk because each use case has its own set of risks. Some may be better or worse.

Also, evaluating performance is a function of the data and the system. The same system may perform poorly or terribly when tested with different data types.

Also, we want to make sure we iterate over the AI lifecycle. It's never one and done. Creating AI is a continuous Iterative cycle in which we implement responsibility at the concept and deployment stages and monitor that feedback over time.

Lastly, we want to issue governance policies throughout the lifecycle and accountability measures for every stakeholder involved.

Some active topics that the research community is working on

Watermarking and fingerprinting are ways to include almost like a stamp or signature in a piece of content or data so that we can always trace back.

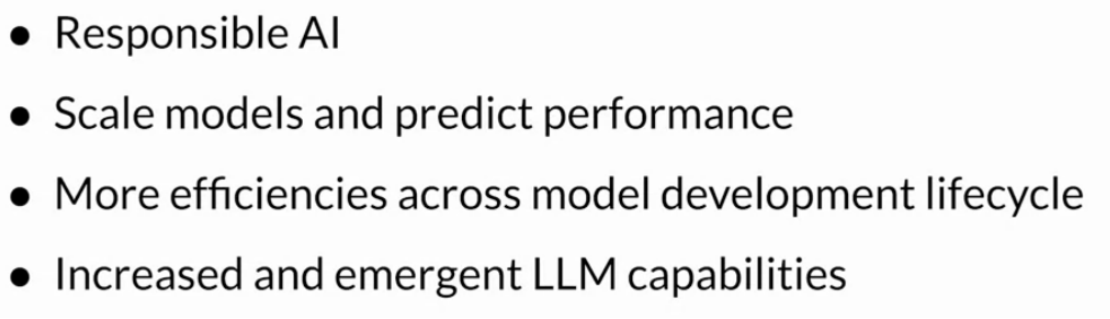
Creating models that help determine if the content was created with gentle AI is also a budding field of research.

It's a very exciting time. The future of AI is accessible inclusive, and innovation is comming.

Course conclusion

On-going research

Along with responsible AI, researchers are looking into techniques to align models with human values and preferences, increase model interpretability, and implement efficient model governance. As model capabilities increase, we'll also need more scalable techniques for human oversight, such as constitutional AI.



- Responsible AI
- Scale models and predict performance
- More efficiencies across model development lifecycle
- Increased and emergent LLM capabilities

Researchers continue to explore scaling laws for all steps of the project lifecycle, including techniques that better predict model performance so that the resources are guaranteed to be used efficiently, for example, through simulations.

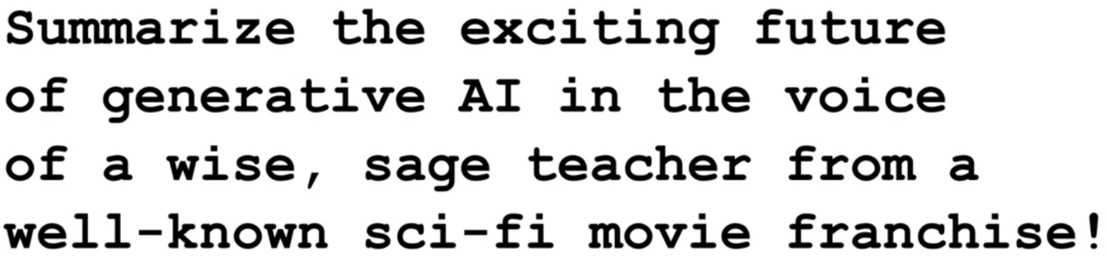
The scale doesn't always mean bigger; research teams work on model optimizations for small devices and edge deployments. For example, llama.cpp is a C++ implementation of the LLaMA model using four-bit integer quantization to run on a laptop.

Similarly, we'll see advancements and efficiencies across the model development lifecycle, especially more efficient pre-training, fine-tuning, and reinforcement learning techniques. We'll see increased and emergent LLM capabilities. For example, researchers are looking into developing models that support longer prompts and contexts, such as summarizing entire books.

In fact, during the development of this course, we saw the first announcement of a model supporting a 100,000-token context window. This corresponds roughly to 75,000 words and hundreds of pages. Models will also increasingly support multi-modality across language, images, video, audio, etc. This will unlock new applications and use cases and change how we interact with models.

We've seen the first amazing results of this with the latest generation of text-to-image models, where natural language becomes the user interface to create visual content. Researchers are also trying to learn more about LLM reasoning and are exploring LLMs that combine structured knowledge and symbolic methods. This research field of neurosymbolic AI explores the model's abilities to learn from experience and the ability to reason from what has been learned.

▼ What the future holds - Answer from a LLM model



Summarize the exciting future of generative AI in the voice of a wise, sage teacher from a well-known sci-fi movie franchise!

```
Exciting, the future of generative AI is, hmm?

A realm where machines converse like humans, they shall.
With finesse, their words flow, blurring lines between
beings. Masterpieces they create, paintings and stories
that astound. Boundaries they shatter, unlocking mysteries,
advancing knowledge.

The force of AI, a powerful ally it shall be, pushing
humanity forward.

In awe, we stand, as generative AI shapes a future
bright and full of wonder, hmm.
```

😉

▼ Reading

Community platform: here.

NOTE: If you don't have an account yet, please follow the instructions here and then return to this page.

Copyright Notice

These slides are distributed under the Creative Commons License.

DeepLearning.AI makes these slides available for educational purposes. You may not use or distribute these slides for commercial purposes. You may make copies of these slides and use or distribute them for educational purposes as long as you cite DeepLearning.AI as the source of the slides.

For the rest of the details of the license,
see https://creativecommons.org/licenses/by-sa/2.0/legalcode