New technical deep dive course: Generative AI Foundations on AWS

by Emily Webber | on 26 JUL 2023 | in Amazon Machine Learning, Amazon SageMaker, Amazon SageMaker JumpStart, Artificial Intelligence, Generative AI* | Permalink | Comments | Share

Generative AI Foundations on AWS is a new technical deep dive course that gives you the conceptual fundamentals, practical advice, and hands-on guidance to pre-train, fine-tune, and deploy state-of-the-art foundation models on AWS and beyond. Developed by AWS generative AI worldwide foundations lead Emily Webber, this free hands-on course and the supporting GitHub source code launched via AWS Youtube. If you are looking for a curated playlist of the top resources, concepts, and guidance to get up to speed on foundation models, and especially those that unlock generative capabilities in your data science and machine learning projects, then look no further.

During this 8-hour deep dive, you will be introduced to the key techniques, services, and trends that will help you understand foundation models from the ground up. This means breaking down theory, mathematics, and abstract concepts combined with hands-on exercises to gain functional intuition for practical application. Throughout the course, we focus on a wide spectrum of progressively complex generative AI techniques, giving you a strong base to understand, design, and apply your own models for the best performance. We'll start with recapping foundation models, understanding where they come from, how they work, how they relate to generative AI, and what you can to do customize them. You'll then learn about picking the right foundation model to suit your use case.

Once you've developed a strong contextual understanding of foundation models and how to use them, you'll be introduced to the core subject of this course: pre-training new foundation models. You'll learn why you'd want to do this as well as how and where it's competitive. You'll even learn how to use the scaling laws to pick the right model, dataset, and compute sizes. We'll cover preparing training datasets at scale on AWS, including picking the right instances and storage techniques. We'll cover fine-tuning your foundation models, evaluating recent techniques, and understanding how to run these with your scripts and models. We'll dive into reinforcement learning with human feedback, exploring how to use it skillfully and at scale to truly maximize your foundation model performance.

Finally, you'll learn how to apply theory to production by deploying your new foundation model on Amazon SageMaker, including across multiple GPUs and using top design patterns like retrieval augmented generation and chained dialogue. As an added bonus, we'll walk you through a Stable Diffusion deep dive, prompt engineering best practices, standing up LangChain, and more.

More of a reader than a video consumer? You can check out my 15-chapter book "Pretrain Vision and Large Language Models in Python: End-to-end techniques for building and deploying foundation models on AWS," which released May 31, 2023, with Packt publishing and is available now on Amazon. Want to jump right into the code? I'm with you—every video starts with a 45-minute overview of the

key concepts and visuals. Then I'll give you a 15-minute walkthrough of the hands-on portion. All of the example notebooks and supporting code will ship in a public repository, which you can use to step through on your own. Feel free to reach out to me on Medium, LinkedIn, GitHub, or through your AWS teams. Learn more about generative AI on AWS.

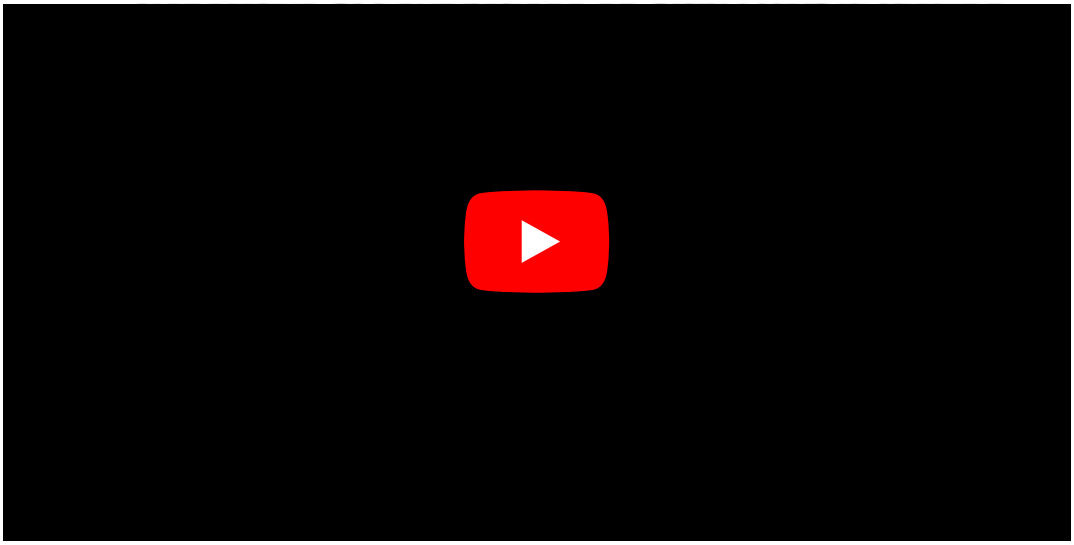
Happy trails!

Course outline

1. Introduction to Foundation Models

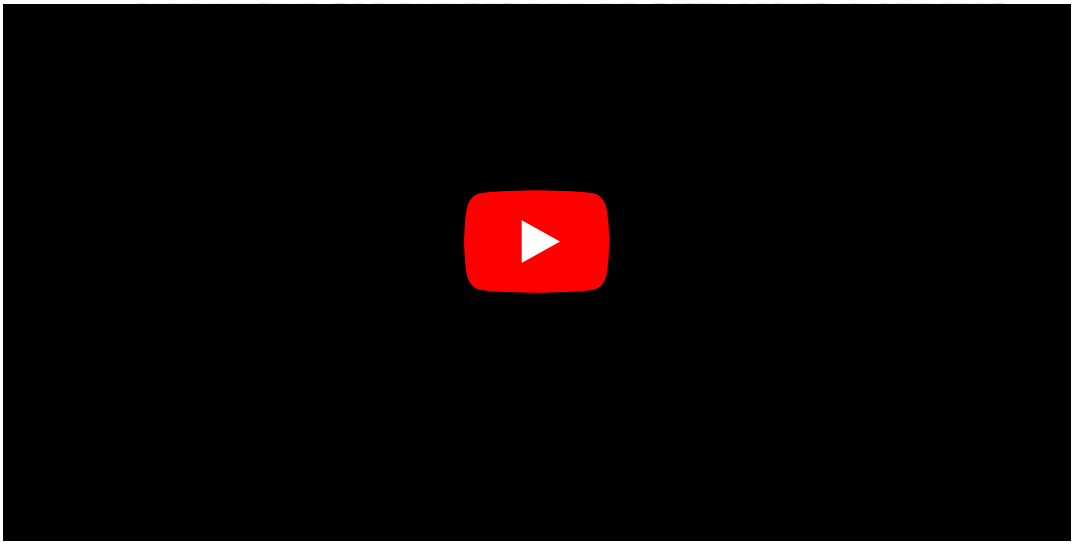
• What are large language models and how do they work? • Where do they come from? • What are other types of generative AI? • How do you customize a foundation model? • How do you evaluate a Generative model? • Hands-on walk through: Foundation Models on SageMaker	Lesson 1 slides Lesson 1 hands-on demo resources

2. Picking the right foundation model



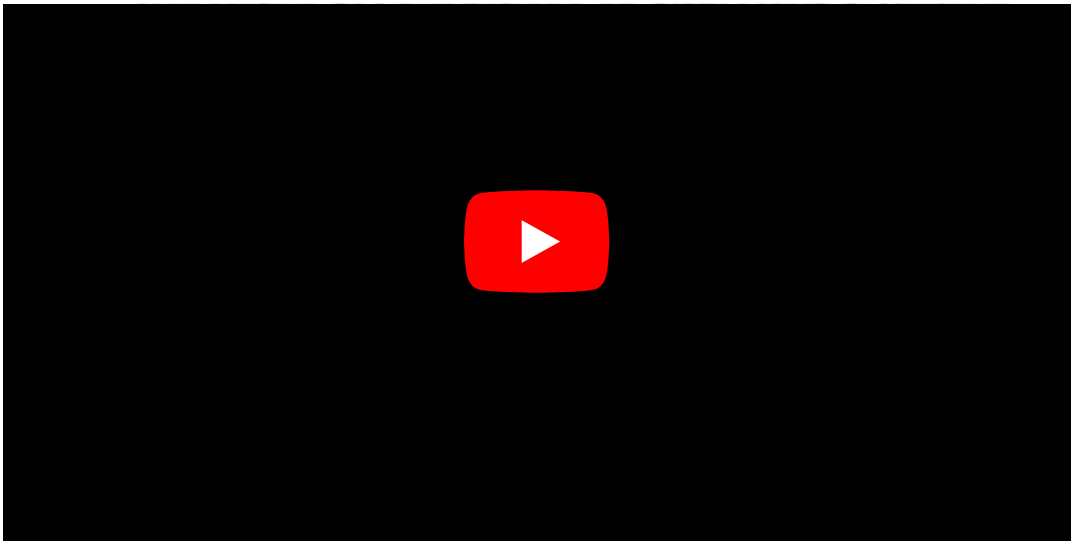
• Why starting with the right foundation model matters
• Considering size
• Considering accuracy
  ◦ Considering ease-of-use
• Considering licensing
• Considering previous examples of this model working well in your industry
  ◦ Considering external benchmarks
	Lesson 2 slides
	
	Lesson 2 hands-on demo resources

3. Using pretrained foundation models: prompt engineering and fine-tuning



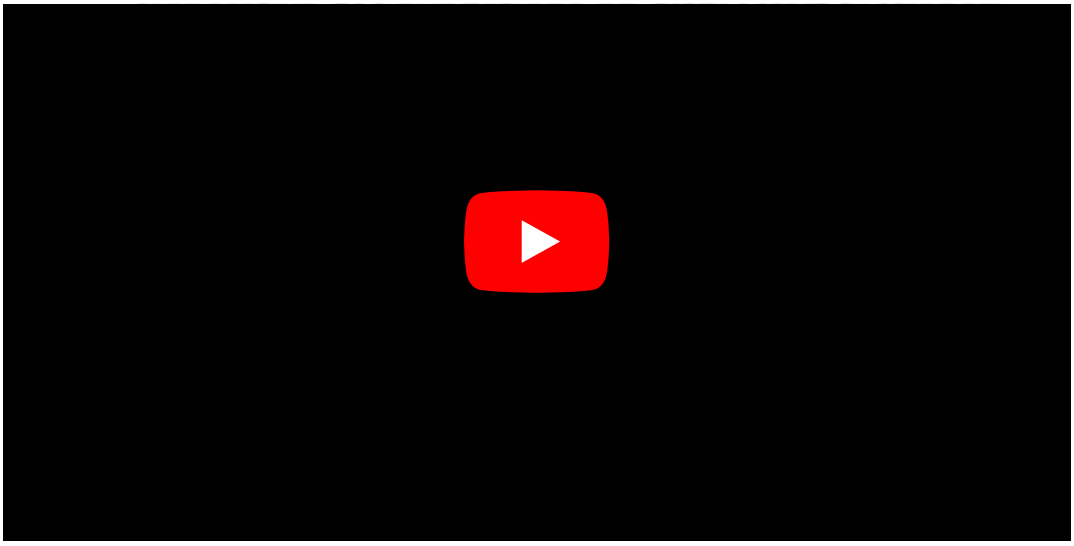
• The benefits of starting with a pre-trained foundation model
• Prompt engineering:
  ◦ Zero-shot
  ◦ Single-shot
  ◦ Few-shot
  ◦ Summarization
  ▪ Classification
  ◦ Translation
• Fine-tuning
  ◦ Classic fine-tuning
  ◦ Parameter efficient fine-tuning
  ◦ Hugging Face's new library
  ◦ Hands-on walk through: prompt engineering and fine-tuning on SageMaker
	Lesson 3 slides
	
	Lesson 3 hands-on demo resources

4. Pretraining a new foundation model



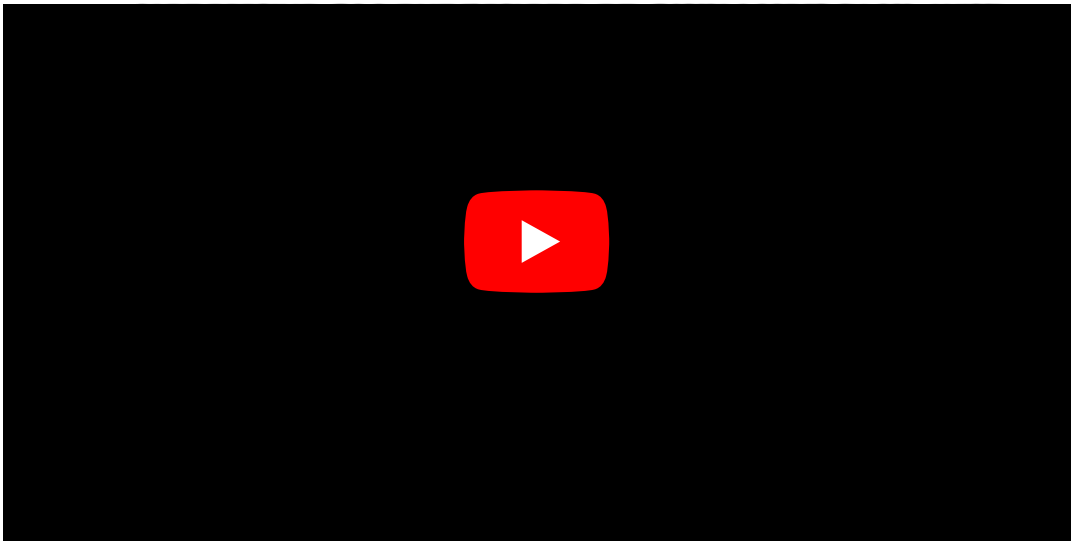
• Why would you want or need to create a new foundation model? ◦ Comparing pretraining to fine-tuning • Preparing your dataset for pretraining • Distributed training on SageMaker: libraries, scripts, jobs, resources • Why and how to adapt a new script to SageMaker distributed training	Lesson 4 slides Lesson 4 hands-on demo resources

5. Preparing data and training at scale



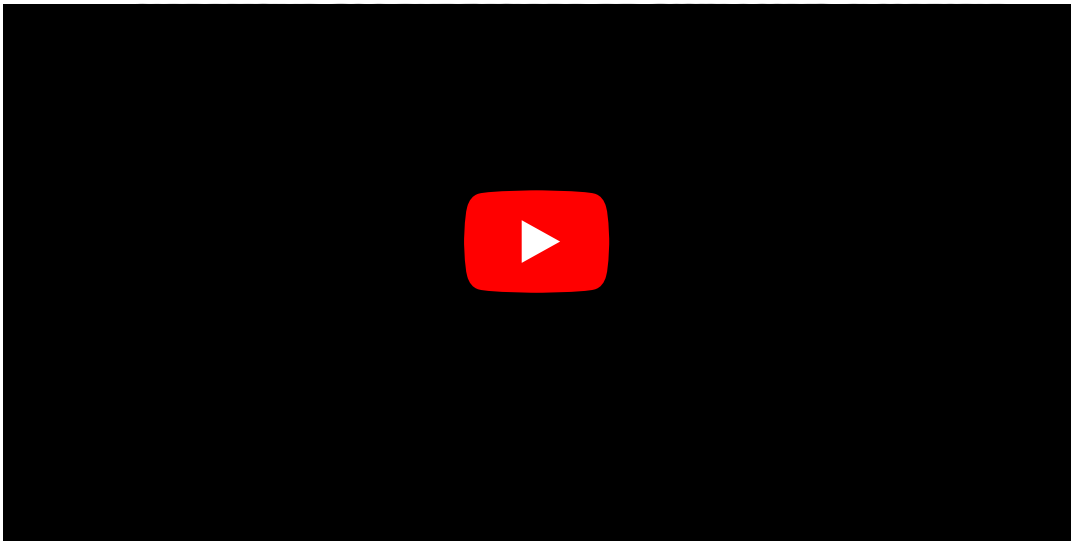
• Options for prepping data at scale on AWS • Explain SageMaker job parallelism on CPU instances • Explain modes of sending data to SageMaker Training • Introduction to FSx for Lustre • Using FSx for Lustre at scale for SageMaker Training • Hands-on walk through: configuring Lustre for SageMaker Training	Lesson 5 slides Lesson 5 hands-on demo resources

6. Reinforcement learning with human feedback



• What is this technique and why do we care about it • How it gets around problems with subjectivity and objectivity through ranking human preferences at scale • How does it work? • How to do this with SageMaker Ground Truth • Updated reward modeling • Hands-on walk through: RLFH on SageMaker	Lesson 6 slides Lesson 6 hands-on demo resources

7. Deploying a foundation model



• Why do we want to deploy models? • Different options for deploying FM's on AWS • How to optimize your model for deployment • Large model deployment container deep dive • Top configuration tips for deploying FM's on SageMaker • Prompt engineering tips for invoking foundation models • Using retrieval augmented generation to mitigate hallucinations • Hands-on walk through: Deploying an FM on SageMaker	Lesson 7 slides Lesson 7 hands-on demo resources

Summary

Generative AI Foundations on AWS is one of seven new free and low-cost AWS courses available to help you use generative AI for people of all roles and experience levels. Whether you're a business leader interested in how generative AI can transform your business or a developer seeking to use generative AI to boost your productivity, we have training to help build your knowledge and practical skills with Amazon's generative AI services. Find the right training for your skill level and use case in this blog post: 7 free and low-cost AWS courses that can help you use generative AI.

About the author

Emily Webber joined AWS just after SageMaker launched, and has been trying to tell the world about it ever since! Outside of building new ML experiences for customers, Emily enjoys meditating and studying Tibetan Buddhism.

Comments

What do you think?

56 Responses

 Awesome. Love it!

 Helpful

 tl;dr

7 Comments 1 Login ▼

Join the discussion...

LOG IN WITH OR SIGN UP WITH DISQUS

Name

♡ • Share Best Newest Oldest

Peter Zhang
2 months ago edited

Amazing, thank you for sharing !!! Does the YouTube video have the same content as the book ?

0 0 Reply • Share ›

Bárbara Silveira Fraga
5 months ago

Excellent!!

0 0 Reply • Share ›

Márcio Zacarias
5 months ago

Thanks, this seems to be a great follow up on the Coursera course I finished last week!

0 0 Reply • Share ›

Emily Webber ■ → Márcio Zacarias

5 months ago

Hey thanks!

1 0 Reply • Share ›

isaac imitini ■ → Márcio Zacarias

4 months ago

Hi, is it the Machine learning with Python course?

0 0 Reply • Share ›

Márcio Zacarias ■ → isaac imitini

4 months ago

I believe it's different, it's more focused in Generative AI things, but it does have a good amount of machine learning and python content

0 0 Reply • Share ›

isaac imitini ■ → Márcio Zacarias

4 months ago

Thank you, I just completed a machine learning course as well so I believe this would be nice to add to my knowledge as well.

1 0 Reply • Share ›