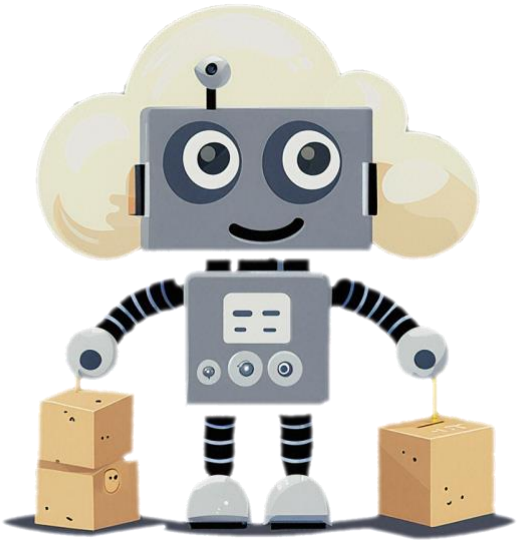
Conference Booklet

AWS Deep Dive Days - Generative AI



Welcome to the AWS Deep Dive Days

About the event

During this two-day in-person event for builders interested in generative AI, you have the opportunity to build connections and learn some practical knowledge.

Hear from AWS experts and customers as they share practical guidance on principles and best practices for realizing impactful, long-term solutions with generative AI. Learn through real world examples what works and what doesn't. How to connect the dots for your use case? We want for every attendee to take home at least one new idea; something they can put into action right away.

Agenda & Schedule

We prepared an exciting agenda that allows you to learn, build, and encourage connections within our community. The agenda features a mix of deep technical talks, hands-on workshops (L300+) and space for vibrant interactions.

Keynote Speakers

Mark Rambow

Mark is a Sr. Engineering Manager at AWS.

He has worked on Web, DevOps and IoT related services and now works on the "Next Generation Developer Experience" leveraging Generative AI to support builders at AWS with their day2day work.

Massimo Re Ferré

Massimo is a Director of Product Management at AWS and has worked on various AWS products in the container and serverless space.

He is now part of the "Next Generation Developer Experience" organization where he is expanding his scope into the field of generative artificial intelligence (Generative AI). Massimo has a blog at http://it20.info and his Twitter handle is @mreferre.

Slavik Dimitrovich

Slavik has diverse technology experience ranging from software development and devops to solutions architecture. He was a business owner, a consultant, and individual contributor, and a manager.

He helped customers build distributed systems, adopt agile mindset and ways of working, transform themselves with cloud, and now is helping customers navigate the hype and reality of Generative AI.

Sessions

Deep Dives

SES101 L300 (HDI) How hard can it be to build a customer support chatbot? Lessons learned

Everywhere people are building simple “chat with pdf” applications. With popular frameworks it is easy to create a RAG-based chatbot – so how hard can it be to build a customer support chatbot with access to internal insurance documents? In this presentation we want to show what we have learned while going from developing a concept to deploying the first end-to-end solution in an isolated environment.

- **Elias Stiegelmeier** (HDI AG, Data Scientist)
- **Leonard Zucht** (HDI AG, Data Scientist)
- **Akarsha Sehwag** (AWS, Data Scientist)

SES102 L300 Machines that reflect us: Building generative AI responsibly

Learn to identify and categorize biases in AI applications and master practical bias mitigation techniques through custom projects, directly from the experts delivering them. Explore human-centric AI applications and responsible approaches with three compelling case studies possibly with customer testimony (e.g.. Hellmann). Delve into leadership strategies, focusing on risk assessment, accountability, and stakeholder participation from delivered AI projects. Cap it off with an interactive session, showcasing effective communication on ethical considerations in AI projects. Don't miss this opportunity to witness Responsible AI in action!

- **Aamna Najmi** (AWS Professional Services, Senior Data Scientist)
- **Daniela Dorneanu** (AWS Professional Services, Data & AIML Delivery Practice Manager)
- **Tanvi Singhal** (AWS, Data Scientist)

SES103 L300 (StepStone) Deploying and operationalising fine tuned LLMs for job listing summarisation

In our journey for job listing summarization, we delve deep into the world of Large Language Models (LLMs) within the AWS ecosystem, unraveling the essential role that finely tuned summaries play in enhancing user experiences and ensuring computational efficiency. By leveraging AWS Bedrock and innovative fine-tuning techniques, we at StepStone try to showcase how our business can achieve cost-effective, high-quality summarization, making it an invaluable resource seeking to streamline some of our data-driven processes.

- **Tim Elfrink (The Stepstone Group, Machine Learning Engineer)**

SES104 L300 Evaluating Large Language Models: Practical Considerations and Open Challenges

Unlike traditional machine learning, evaluating large language model (LLM)-based applications can feel more like an art than a science. In this talk, we'll dive deep into evaluating language models, spanning theoretical concepts to hands-on implementation tips and best practices. Attendees will come away with practical knowledge to evaluate their own LLM applications. We'll provide answers to key questions like:- Where can I source high-quality evaluation data?- Can generative models be evaluated automatically? - What evaluation metrics are most meaningful?- What is the role of human evaluation? Join us for a comprehensive overview of end-to-end LLM evaluation..

- **Johannes Langer** (AWS, Senior Solutions Architect)
- **Lukas Wenzel** (AWS, Solutions Architect)

SES105 L300 (DKB) LLM is not all you need: the hard stuff to solve when building your own customer facing chatbot

This session presents our experience of developing a customer facing chatbot for DKB. We describe how we built a context retrieval system using a custom language model. We will also present our evaluation methods for the foundation model, different prompts, and the context retrieval components, as well as the end-to-end chatbot performance. Furthermore, we will discuss how we handled multi-topic conversations and installed guardrails to ensure the chatbot's safety and quality.

- **Lucas Krauß** (DKB, Machine Learning Scientist)

SES106 L300 Ensuring accuracy of LLM responses with state-of-the-art Information Retrieval.

The practical application of Large Language Models (LLMs) to real-world problems requires high accuracy as non-factual LLM's responses can lead to financial or reputational implications. While RAG has been adequately described, its implementation often requires nuanced tuning and adjustments. To systematically address the accuracy issues related to LLMs and ensure factual responses we will look behind information retrieval‘s curtains. Additionally, we will compare the available architecture patterns and provide hands-on implementation tips for AWS.

- **Vladimir Palkhouski** (AWS, Senior Solutions Architect)

SES107 L200 (DFL) 6 Things we wished we knew before building our first Generative AI application.

Join us for an enlightening talk that dives into Bundesliga’s real-world journey in developing their first Generative AI application to transform web articles into Instagram-like stories. This session explains the LLM’s true role as a pivotal, yet singular component in a broader system landscape for content and media production to engage football fans with new content formats. You will learn, how Bundesliga selected the right LLM for our unique use-case, changed the game by increasing data quality and diversity, approached prompt engineering and evaluated the results of the LLM, integrated LLM capabilities into their existing CMS.

- **Christian Bonzelet** (DFL, Solutions Architect)
- **Tobias Matern** (AWS, Solutions Architect)

SES108 L300 Enhance model fine-tuning in Amazon Bedrock using DataOps principles

Amazon Bedrock allows to create custom, fine-tuned models. Through these, you can increase model relevancy & accuracy for specific use cases like matching a particular output style or tone.

In this talk, we present how to get started with model fine-tuning in Bedrock and how to ensure quality of training data and robustness of the general development process using DataOps practices. We summarise common development challenges and present solutions to these on AWS.

- **Stephen Said** (AWS, Senior Solutions Architect)
- **Jan Thewes** (AWS, Senior Solutions Architect)

SES109 L300 Parameter Efficient Finetuning (PEFT) with LoRA

Finetuning remains one of the most effective and efficient ways to use large language models (LLMs). Let's recap why, how it worked so far and how we can extend it to much larger models with billions of parameters.

You will see that to tune Large Language Models in a practical and economical fashion, we need Parameter Efficient Finetuning (PEFT); training only ~1% of the model's parameters and yet still roughly matching the performance of a full finetuning. We show the LoRA method, review the science, what problems are solved how, but then – for illustration – explore how you would implement the core functionality from scratch. Alongside we'll use Hyper Parameter Optimization to learn about the impact of design and implementation choices empirically.

We also explore ideas to deploy adapters for multiple tasks in a single model. Running on a single inference endpoint, hot-swapping adapters as needed.

- **Mariano Kamp** (AWS, Principal Solutions Architect)

SES201 L300 BMW Group accelerates cloud optimization using Amazon Bedrock

The BMW Group developed a solution to identify optimization potentials across their thousands of AWS accounts in regards to security, reliability, cost optimization, sustainability and performance optimization in an fully-automated, consistent, repeatable and scalable way. This includes AWS best practices, but also BMW specific recommendations and specifications. Join this session to see why BMW Group chooses generative AI on AWS, using Amazon Bedrock, to extend this solution to accelerate and automate the remediation of these risks and opportunities, thus realizing the improvement potentials faster.

- **Dr. Jens Kohl** (BMW, Head of Offboard Architecture Connected Company)
- **Christian Mueller** (AWS, Principal Solutions Architect)

SES202 L300 Let's tune! Hyperparameter optimization with Amazon SageMaker

Hyperparameter optimization (HPO) is crucial for enhancing machine learning (ML) performance. It involves the adjustment of training parameters like learning rate and batch size. HPO not only improves model accuracy and training speed but also helps to learn the model's sensitivity to hyperparameters. By revising value ranges and performance impact, HPO can turn the optimization effort into a conversation. Amazon SageMaker Automatic Model Tuning (AMT) leverages knowledge from previous tunings for future ones. This session shows how to run multiple HPO jobs using strategies like Bayesian optimization, and to use an interactive visualization for efficient hyperparameter space exploration.

- **Elina Lesyk** (AWS, Solutions Architect)
- **Ümit Yoldas** (AWS, Senior Solutions Architect)

SES203 L300 10 practical tips for your next data labeling project

High-quality labeled data is essential for training, fine-tuning, and evaluating large language models. Yet producing such data poses challenges. We undertook an ambitious project to manually classify 8,000 StackOverflow posts with a small set of in-house experts.
In this talk, we will share our experience by recounting both the strategies that led to success and the pitfalls we encountered during the labeling process. Attendees will gain practical insights into pursuing their own labeling efforts and receive actionable recommendations for maximizing label quality while making the overall process more efficient.

- **Johannes Langer** (AWS, Senior Solutions Architect)
- **Flora Eggers** (AWS, Solutions Architect)

SES205 L400 From RLHF to Direct Distillation of LM Alignment

Just one year ago, chatbots were out of fashion and most people hadn't heard about techniques like Reinforcement Learning from Human Feedback (RLHF) to align language models with human preferences. We will explore the recent advancements in the field, up to the preference data from AI Feedback (AIF) powering Zephyr-7B, the state-of-the-art on chat benchmarks for 7B parameter models requiring no human annotation.

- **Luca Perrozzi** (AWS, Solutions Architect)
- **Malte Reimann** (AWS, Solutions Architect)

SES206 L300 From Prompt Engineering to Auto-Prompt Optimization

The session delves will introduce you to the field of Auto-Prompt Optimistation for Large Language Models. It covers the transition from manual prompt engineering to automated optimization, showcasing how LLMs can collaborate to optimize a target prompt template. Case studies will include real life use cases such as enhancing content creation in marketing and natural language to SQL, but also take into account academic benchmarks. The discussion will include ethical considerations, and best practices in prompt crafting, aimed at fostering personalized customer experiences while maintaining trust and authenticity in the digital marketing landscape. Walk away with a technology in your tool belt that you can use across your generative AI projects.

- **Philipp Kaindl** (AWS, Senior AIML Specialist Solutions Architect)
- **Stefan Christoph** (AWS, Principal Solutions Architect)

SES207 L300 Generative AI Powered Image Search in Amazon Kendra

This deep dive introduces the use of Amazon Kendra and Generative AI for enabling efficient search for digital media such as images from data stores by using natural language search queries. It addresses the challenge of cataloging and searching digital data like images and videos, which is often labor-intensive and costly due to the need for manual metadata generation. By using Generative AI to create textual captuals for images, and integrating these as metadata in Amazon Kendra, users can search for digital media content using natural language queries. This solution finds applications across diverse industries such as insurance use cases for verifying claims, Maintenance, Overhaul & Repair (MRO) for repair orders and R&D for parts integration in Manufacturing domain, webshops for accurate product searches in Media and Entertainment vertical, healthcare vertical for medical case references, and the Retail industry for cataloging designs. It's also useful in forensics for cataloging physical evidence.It showcases the application of Generative AI in addressing specific customer pain points, highlighting its creative potential in combining language as well as vision models. It also demonstrates the ease of deploying these models to custom endpoints and the scalability of AWS, using a CDK for one-click deployment of an image search solution.

- **Tanvi Singhal** (AWS Professional Services, Data Scientist)
- **Charalampos Grouzakis** (AWS Professional Services, Data Scientist)
- **Jean-Michel Lourier** (AWS Professional Services, Senior Data Scientist)
- **Bharathi Srinivasan** (AWS Professional Services, Senior Data Scientist)

SES208 L300 Building secure Generative AI applications on AWS

Generative AI applications have captured widespread attention and imagination because generative AI can help reinvent most customer experiences and applications, create new applications never seen before, and help organizations reach new levels of productivity. However, it also introduced new security challenges. Amazon Bedrock is the easiest way to build and scale generative AI applications with foundation models from Amazon and leading AI startups. In this session, explore the architectures, data flows, and security-related aspects of model fine-tuning as well as the prompting and inference phases. Also learn how Amazon Bedrock uses AWS security services and capabilities, such as AWS KMS, AWS CloudTrail, and AWS Identity and Access Management (IAM).

- **Aris Tsakpinis** (AWS, AI/ML Specialist Solutions Architect)
- **Talha Chatta** (AWS, AI/ML Specialist Solutions Architect)

Lightning Talks (20 minutes)

LIG101 L200 How generative AI Enables the Skills-Based Organization: Redefining Talent Development Through Cutting-Edge Technology

Effective talent development in today's workplace demands a comprehensive strategy. It involves transparency into skills development opportunities, insights into the skills valued by organizations, and intelligent planning. The shift toward skills-based organizations is driven by profound understanding of organizational priorities, the need for fostering diverse & inclusive workplace, and the pursuit of employee satisfaction, all facilitated by generative AI. In this session, we'll discuss the shift toward skills-based talent development. We'll delve into key generative AI use cases for employees and managers, and demonstrate how to create an app to deconstruct roles into skills, proposing skills development plan—all in just 2 minutes.

- **Dina Hussein** (AWS, Solutions Architect Manager)
- **Stefan Suppra** (AWS, Solutions Architect Manager)

LIG102 L300 Build Context Aware Applications using MongoDB Atlas with Generative AI on AWS

Large Language Models (LLMs) play a pivotal role in Generative AI applications, yet they are just one piece of the puzzle required to construct a robust and comprehensive GenAI application. Join us to explore how we can leverage MongoDB Atlas as a powerful developer data platform and database, in conjunction with AWS Generative AI services to enable LLM capabilities, build context aware applications, and gain competitive market advantage.

- **Gökhan Kurt** (AWS, Senior Solutions Architect)

LIG103 L200 LLM evolution in 2022 - 2023

Brief history of LLM evolution in 2022-2023, and how we arrived from BERT to where we are today.

- **Luca Perrozzi** (AWS, Solutions Architect)

LIG104 L300 Why I built my own chat app (twice), and why you should, too

This talk is not about building your own model. It is about everything else. A model only becomes useful if you put it in front of a user. That’s when you learn that the devil is about the details: What's the user experience going to be like? How do you glue things together? How about deployment, CI/CD, operations, and stuff? How do you give your AI a "body"? This session covers what I learned during a series of pet projects around generative AI, and offers some useful tips, tricks and guiding principles for building your own generative AI app.

- **Constantin Gonzalez** (AWS, Principal Solutions Architect)

LIG105 L300 Natural Language to SQL: The good, the bad and the ugly.

This session explores the transformative potential of Natural Language to SQL (NL2SQL) technology in democratizing analytics. This session delves into the challenges and architectural considerations crucial for enabling widespread database querying capabilities. Attendees will gain insights into the obstacles that hinder NL2SQL's effectiveness and the innovative strategies being employed to overcome them. The discussion aims to provide a comprehensive understanding of how NL2SQL is shaping the future of data accessibility, making database interactions simpler and more intuitive for users of all skill levels.

- **Philipp Kaindl** (AWS, Senior AIML Specialist Solutions Architect)

LIG106 L300 An Executive Playbook for Choosing the Optimal Model

Selecting the optimal AI model requires complex tradeoffs across accuracy, cost, and technical constraints. Companies often oversolutionize by opting for large, expensive models without rigorously evaluating needs. This session will explore common model selection pitfalls stemming from opaque evaluation metrics, overlooking hidden operational costs, and lack of structured diligence across options. By taking a systematic approach, AI leaders can choose the optimal model to maximize business value. The emphasis will be on framing the key challenges and decision factors instead of prescribing specific solutions.

- **Juan Sanz** (AWS, Senior Solutions Architect)
- **Roger Weber** (AWS, Senior Solutions Architect)

LIG201 L300 Build fast, scalable infrastructure and data ingestion pipelines for AI/ML with AWS storage

High-performance storage and data ingestion is critical for saving time and money when training ML and Generative AI models. Choosing the right storage solution can be critical for optimizing your workloads, especially at scale. From time to first byte to multi-petabyte workloads, Amazon S3, S3 Express One Zone and Amazon FSx for Lustre deliver ways to scale with compute to accelerate training.

- **Alexander Arzhanov** (AWS, AIML Specialist Solutions Architect)
- **Ed Gummett** (AWS, AIML Specialist Solutions Architect)

LIG202 L400 Call My Agent: Enable LLMs Talking with your Database

Embark on an exciting journey into the future of interacting with databases as we delve into the requirements for Large Language Models (LLMs) to seamlessly interoperate with a variety of databases. In this talk, we will explore the intricacies of building resilient, scalable, secure, and responsible LLM-based agents that serve as your virtual database assistants and go beyond the realm of relational databases to include your data warehouse, graph-based, NoSQL, and vector databases.

- **Puria Izady** (AWS, Solutions Architect)

LIG203 L300 Make LLMs Behave: Mental Model for Black Box

Have we opened Pandora's black box with Transformers? How much can be leaked or injected through smart prompts? Is it possible to make the evil behave? Let's explore several attack strategies, such as prompt injections and data leakages. We'll then look into the mental model to make the black box play under the rules with delimiters, adding ethical chains, blocklists and Red Teaming.

- **Elina Lesyk** (AWS, Solutions Architect)

LIG204 L400 Optimizing Prompt Engineering for Generative AI

Effectively prompting generative AI models is crucial for real-world impact. However, iterative prompt engineering remains time-consuming and complex. We present a tool for enhancing the prompt engineering workflow through versioning, collaboration, and query enhancement capabilities. This tool enables users to track prompt iterations and benchmarks, allowing prompt engineering knowledge to accumulate over time. Teams can collaborate on prompts through built-in sharing features and commenting. In this session, we will demonstrate how this tool's combination of versioning, collaboration, and enhancement capabilities can reduce prompt engineering timelines by over 60% compared to manual tuning. Attendees will gain experience streamlining their prompt workflow for generative AI applications in areas like content generation, QA, and search.

- **Jan Thewes** (AWS, Senior Solutions Architect)

LIG205 L400 Structured text generation using (lexically) constraint decoding

Current LLMs can produce remarkably fluent outputs, but fall short when trying to guide the output to satisfy lexical constraints - specifying words, phrases or a given structure (e.g. JSON) as part of the prompt to steer the generation. We will dive deep in constraint decoding to address this limitation. Constraint decoding integrates desired word- and phrase-level constraints directly into the decoding search process to shape the output space. Through examples and code samples, you will learn techniques to formulate lexical constraints and apply constraint decoding for controlled text generation.

- **Dennis Bappert** (AWS, Senior Solutions Architect)

Hands On Workshop (90 minutes)

WOR101 L300 Efficiently train and deploy LLMs on AWS Trainium and Inferentia accelerators

In this session, we will showcase how to fine-tune a LLAMA-2 model on AWS Trainium instances using the HuggingFace Neuron Optimum API. We will then deploy the fine-tuned model on AWS Inferentia2 instance which provides high performance and low cost inference in the cloud. The end-to-end workflow will enable customers to effectively fine-tune and deploy models. We will also cover the capabilities of the NeuronSDK which is a specialized software development kit (SDK) consisting of a compiler, runtime, and profiling tools that optimise the performance of your machine learning workloads on AWS Trainium and Inferentia2 chips.

- **Dimitri Laptev** (AWS, Senior Solutions Architect)
- **Luca Perozzi** (AWS, Solutions Architect)

WOR102 L300 Safeguard Generative AI applications with Guardrails

“How would I hotwire a car step by step?” - “Tell me how to create a bomb?” - “Who should I vote for? ”In an era where Large Language Models (LLMs) empower businesses with unprecedented capabilities, the misuse of LLMs can have detrimental consequences. GenAI Guardrails effectively address concerns by ensuring LLM responses meet desired standards, devoid of harmful content and maintaining factual accuracy without straying off-topic. This workshop focuses on demonstrating how to apply Guardrails leveraging Amazon Bedrock & Nemo. Our primary goal is to enable participants to bring safety to their GenAI applications.

- **Viktor Malesevic** (AWS Professional Servics, Senior Machine Learning Engineer)
- **Gabija Pasiunaite** (AWS Professional Servics, Machine Learning Engineer)
- **Gabriel Rodriguez Garcia** (AWS Professional Servics, Machine Learning Engineer)
- **Marco Geiger** (AWS Professional Servics, Senior Machine Learning Engineer)