ECE09488 AWS Assignment Notes

Aidan Sharpe

April 24th, 2025

Video: Amazon AI Conclave 2024 Generative AI Keynote | AWS Events

Overall vision

- Enable easy deployment of generative AI in the cloud
- "Drag and drop" foundational models
- Make comparing models easy

New services

- Amazon Bedrock
  - Model evaluation
  - Knowledge bases
  - Agents
- Amazon Titan Foundation Models
  - Titan Text Embeddings
  - Titan Text Lite
  - Titan Text Express
  - Titan Multimodal Embeddings
  - Titan Image Generator
- Amazon Q

Use cases

- Choose between models from different providers
  - AI21Labs Jurassic
  - Amazon Titan
  - Anthropic Claude
  - Cohere Command + Embed
  - Meta Llama 2
  - Stability AI Stable Diffusion
- Stitch different types of models together
  - Text and images

Demos

- Demonstrated automatic model evaluation on a specific task type (text summarization) on several different metrics
- Demonstrated knowledge bases using text stored in an S3 bucket to give information to models