aws marketplace

GENERATIVE AI – LARGE LANGUAGE MODELS (LLM)

Not all models are created equal: Key factors to consider when selecting an LLM

Today’s speakers

Priya Arora
Head of Generative AI
Center of Excellence, AWS

David Potes
Senior Manager,
Data and AI, AWS

John Dickerson
Chief Scientist,
Arthur AI

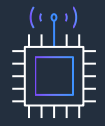
Zach Fry
VP Engineering,
Arthur AI

Innovation can **transform industries**

Everything you need to accelerate your generative AI journey

Choice and flexibility of models

Differentiate with your data



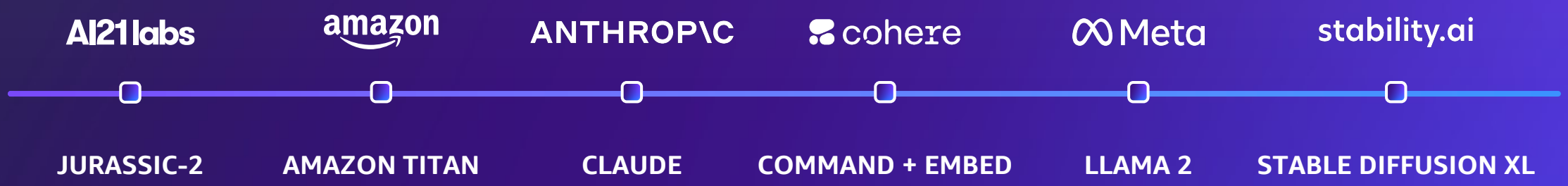
Responsible AI integration

Low cost and performant infrastructure

Generative AI-powered applications

Amazon Bedrock

BROAD CHOICE OF MODELS



Model selection criteria dimensions

Accuracy
Measure of accuracy in response, completeness, and coverage of facts

Speed
Measure of time to first byte and complete results

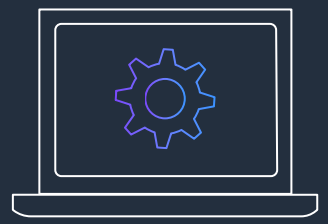
Economics
Measure of the cost to host and invoke LLM

Transparency
Measure of hallucination in responses and accuracy of citations and sources

Responsibility
Measure and management of security, privacy, and governance

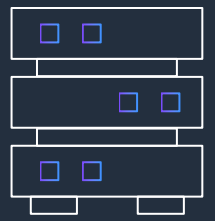
Responsible AI Panel

Customers have questions...



Which model should I use?

How can I move quickly?

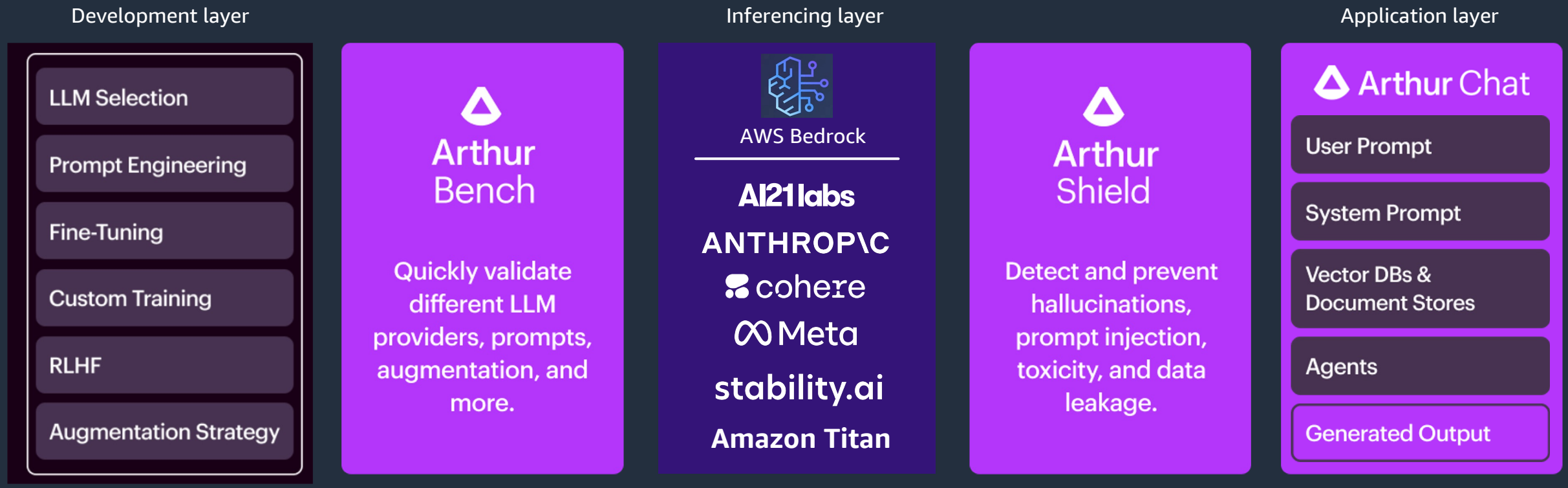


How can I keep my data secure & private?

LLM evaluation vs system evaluation

LLMs are a component of a larger, generative AI-enabled system

Evaluating the model alone is **necessary** but not sufficient for measuring system performance



Needs and realities for our enterprise customers

What we're hearing

- Latency, latency, latency
- Cost per token matters
- ROI

What we're seeing

- LLMs evaluating LLMs
- Leaderboard-ization and reality
- System-level security needs

Not all metrics are created equal: What are they?

Traditional AI/ML/Data science metrics

Example metrics
- Accuracy, precision, recall, F1-score…

Advantages
- Precise; battle-tested implementations exist; efficient (speed, cost)

Disadvantages
- ”Not enough” for newer LLM-enabled tasks (code generation, summarization)

Newer, similarity-based metrics

Example metrics
- BLEU, ROUGE, BERTScore…

Advantages
- Useful when there is no specific value to predict, but you do have sample “good answers” in an eval dataset
- Quick; numeric/”well-defined” metrics

Disadvantages
- Lacks human elements: misses semantics, context for text
- Requires reference data

LLMs evaluating LLMs

Example metrics
- Using chain-of-thought (CoT) reasoning with an LLM in the evaluation loop

Advantages
- Shown to correlate with human evaluation; general purpose

Disadvantages
- Costly, slower, difficult to debug or explain

Standardize the evaluation workflow across tasks

Focus on the system
- For task types, enumerate metrics that track to that task
- RAG systems test
- Remember traditional ML best practices

Evolving landscape – evolving tooling that
- Stays up to date
- Remains backward compatible
- Informs user of new best practices

Leaderboards and other solutions
- Stanford HELM
- Hugging Face Open LLM Leaderboard
- Arthur Bench

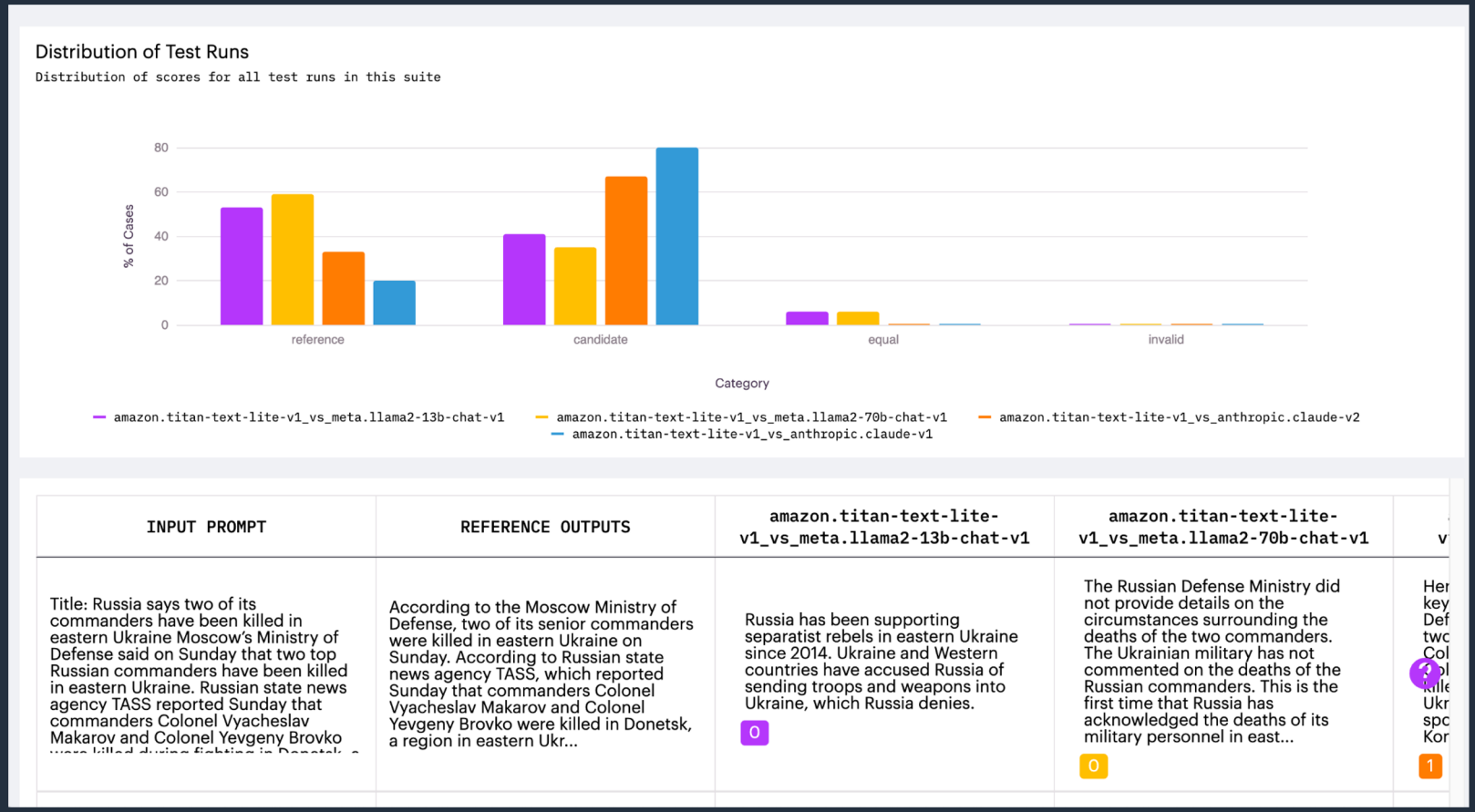
Traditional best practices
- Holdout set
- Cross-validation as appropriate
- Remember class imbalance
- Periodically look at your data
- Collect and correlate quantitative human feedback with metrics

Model evaluation using Arthur Bench

Arthur Bench
- Get working in 10 minutes
- Open source

Start with task definition
- Use or build scoring method
- Support for labeled / unlabeled data

Easy integration with LLM systems
- Bedrock API Client
- Scoring with LLMs evaluating LLMs



Which Bedrock model performs best for short news summarization?

Task definition:

- **Input:** 49 news articles, no labeled data
- **Task:** Summarize in less than 250 words
  Score: Use LLM to pick best summary
- **Compare:** Compute ELO score showing head-to-head comparison

Bench-enabled decision:

- **Pick model** with best performance, latency, cost for your task

Bedrock model name	ELO rating
anthropic.claude-v2:	1345
meta.llama2-70b-chat-v1	1079
amazon.titan-text-lite-v1	898
anthropic.claude-v1	864
meta.llama2-13b-chat-v1	811

Evaluation demo: Arthur Bench

Meta’s Llama2 13b, 70b; AWS’s Titan Text Lite; Anthropic’s Claude v1 and v2

A Case study with Seek AI: Evaluating models for text-to-SQL

Challenge

- **Identifier hallucination** (e.g., database/table/column names) occurs frequently in SQL generation if preventative measures are not taken
- Detecting identifier hallucinations can be difficult since production-ready parsers are difficult to build
- Not well-controllable with traditional methods, such as prompt engineering, fine-tuning, or prompt tuning

Solution

Combination of the following has yielded a **near-0% identifier hallucination rate**:

- Deterministic systems after SQL generation that detect and resolve hallucinations
- Non-traditional generation methods prohibit hallucinations from step 0

Opportunities

- Performance of text-to-SQL models are bounded by amount of available context and database schema ambiguity
- Use of statistical parsers that are adaptable to multiple SQL dialects and don't break because their grammars are too permissive/not permissive enough

seek

Seek what matters:
Seek AI is a natural language interface for structured data.

Products:

- Seek AI
- SEEKER-1: state-of-the-art model for querying structured data
- MiniSeek: Spider leaderboard model

Input: “What is the average purchase price of an item bought in Ohio?”

Expected output: SELECT AVG(purch_price) as avg
FROM purchases WHERE purch_loc = ‘Ohio’;

Hallucination: SELECT AVG(purch_price) as avg
FROM purchases WHERE location = ‘Ohio’;

purchases	
purch_id	integer
purch_sku	integer
purch_price	numeric
purch_loc	varchar(40)

> “When you’re building a product, often the task is a lot more specific than what these leaderboards do. […] Not only should you have very task-specific benchmarks, but you should have a whole suite of them.”
>
> ***- Raz Besaleli, Co-founder & Director of AI Research, Seek AI***

Q&A

Priya Aurora
Head of Generative AI
Center of Excellence, AWS

David Potes
Senior Manager,
Data and AI, AWS

John Dickerson
Chief Scientist,
Arthur AI

Zach Fry
VP Engineering,
Arthur AI

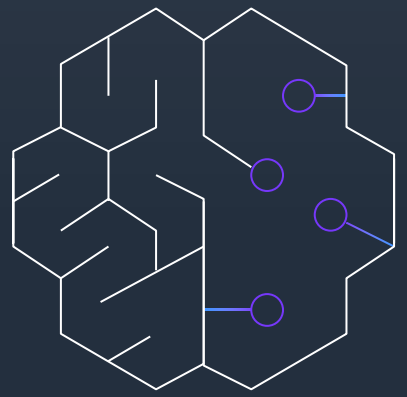
How can you engage Arthur AI

Get better results from your machine-learning models. Our AI performance technology measures and improves models, helping enterprise teams optimize across accuracy, explainability, and fairness.

Model agnostic

Open source

Tiered pricing and private offers



SaaS

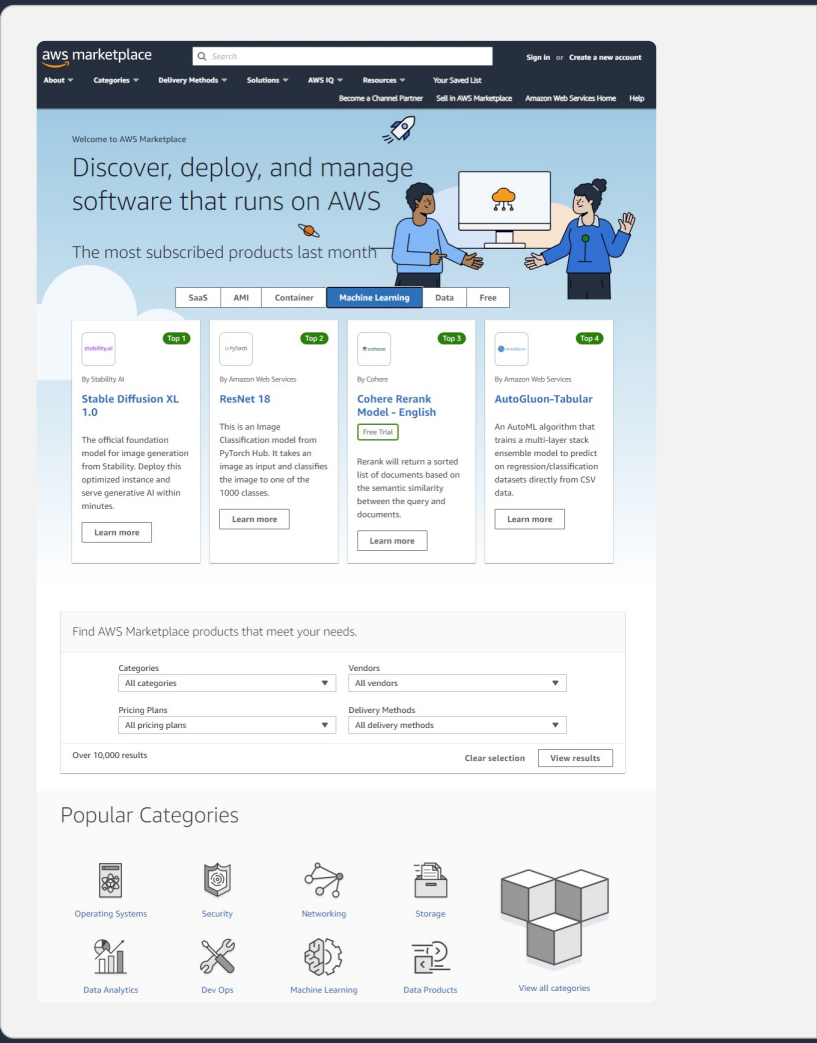
AWS Marketplace listing

What is AWS Marketplace?

aws marketplace

- Over **15K+** listings
- **4K+** ISVs (free, BYOL, or commercial)
- Deployed in **31** regions
- **330K+** monthly active customers
- Over **2.5M** current subscriptions
- Offers **82+** product categories

- Deploy software on demand
- Flexible consumption and contract models
- Easy and secure deployment, almost instantly
- Simplified billing



How can you get started?

Find

A breadth of generative AI solutions

Applications
Foundation models
Tooling
Professional services

Buy

Through flexible pricing options

Free trial
Pay-as-you-go
Budget alignment
Private offers
Billing consolidation
Enterprise Discount Program
Private Marketplace

Deploy

With multiple deployment options

AI/ML models
Amazon Machine Image
Containers
CloudFormation template
Amazon EKS/Amazon ECS
SaaS
AWS Data Exchange

Resources

Model selection and evaluation

- Arthur AI in AWS Marketplace

Foundation model solutions

- AI21 Labs in AWS Marketplace
- Anthropic in AWS Marketplace
- Cohere in AWS Marketplace
- Stability.ai in AWS Marketplace
- NVIDIA in AWS Marketplace

Responsible AI panel

Upcoming spotlight series

AI21 Labs

Anthropic

Cohere

Stability.ai

NVIDIA

Thank you!