Picking the right foundation model

Generative AI Foundations on AWS

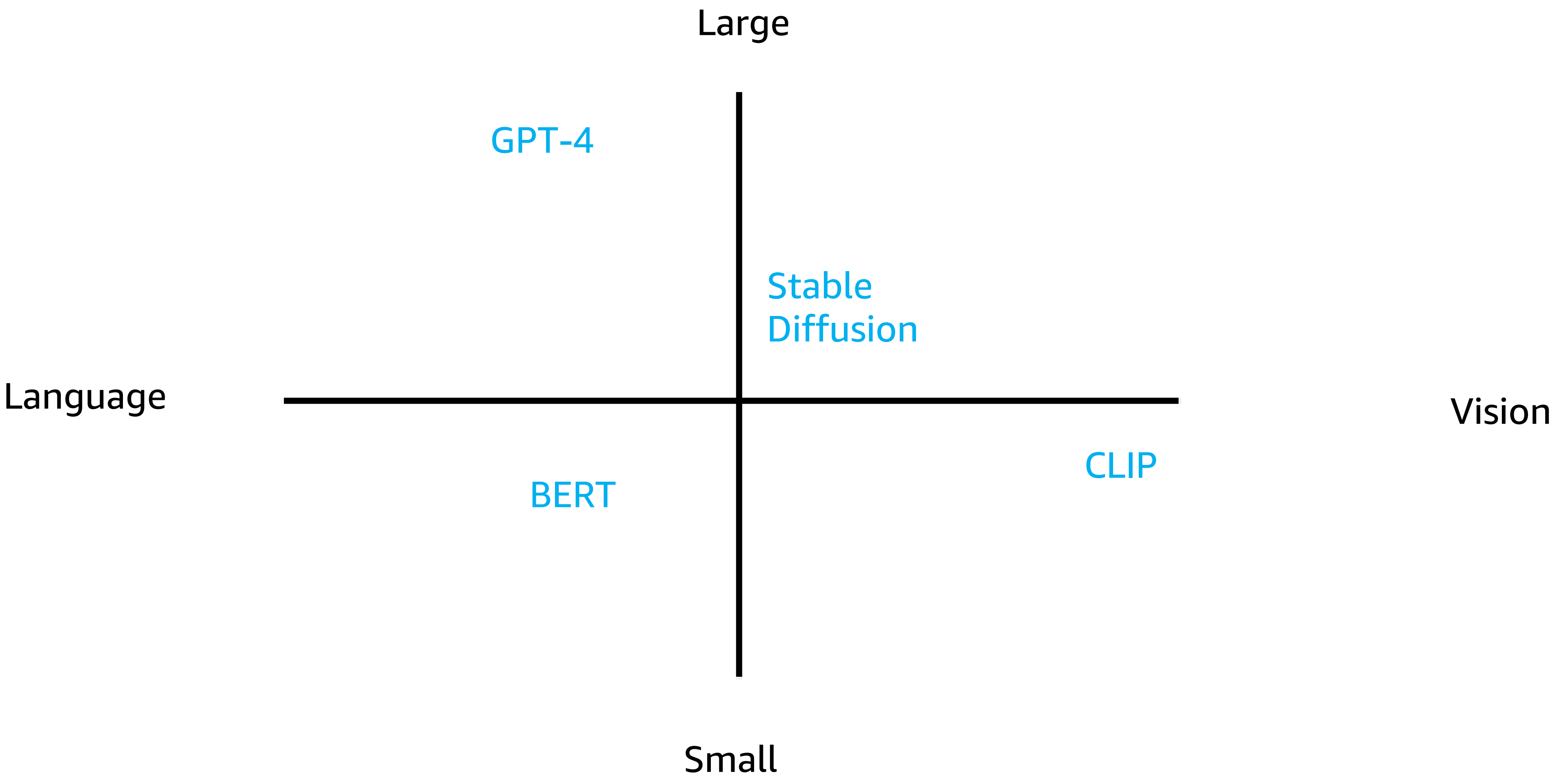
Emily Webber, Principal ML Specialist SA at AWS

Lesson 2 – Level 300

Today's activities

- The basics of foundation model selection
- Considerations: modalities, task, size, accuracy, ease-of-use, licensing, previous examples, external benchmarks
- Does starting with the right model matter
- Hands-on walk through: evaluate FM's on SageMaker

The basics of foundation model selection



Large

GPT-4

Stable Diffusion

Language

Vision

BERT

CLIP

Small

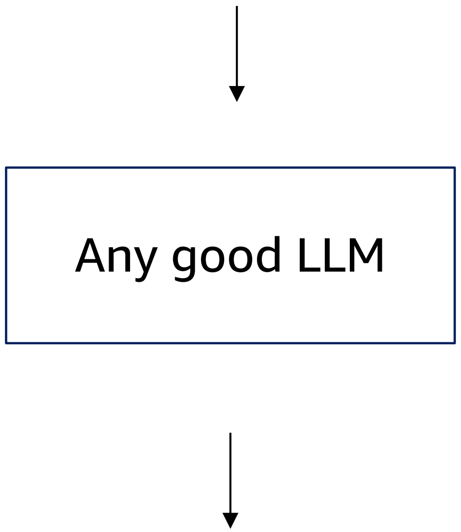
Picking the right foundation model

Modality

The modality of your model drives its output

Text-to-text

Input: If I have two mimosas and one samosa, how many vehicles do I have?

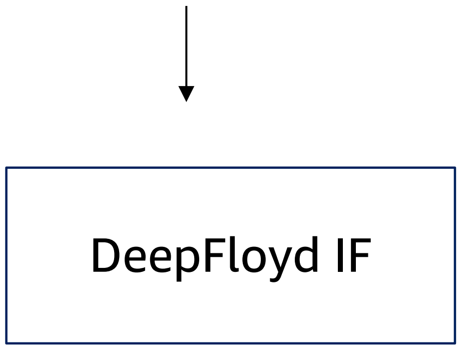


Any good LLM

Output: Having two samosas and one mimosa does not imply ownership of any vehicles. The number of samosas and mimosas you have is unrelated to the number of vehicles you own.

Text-to-image

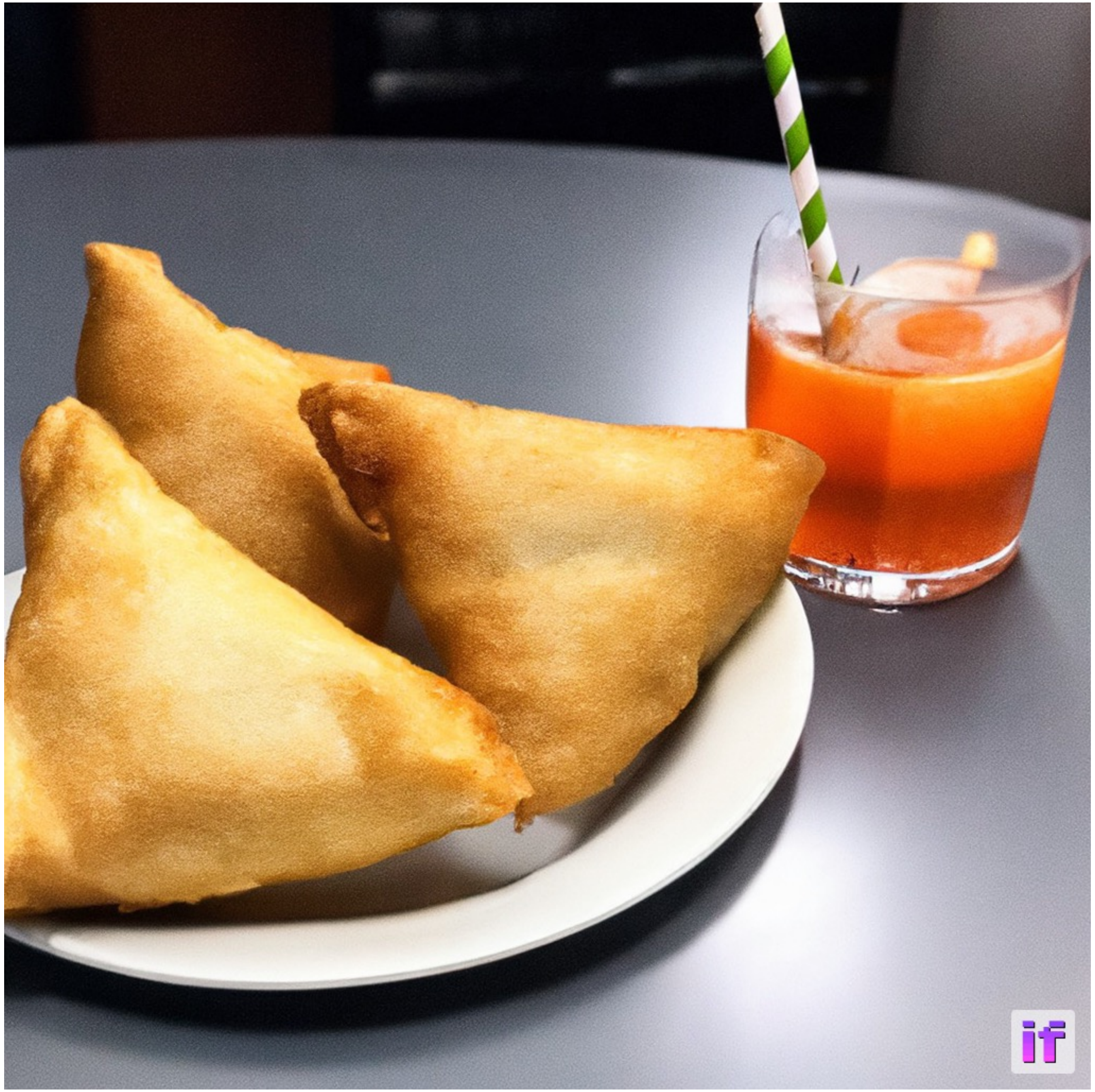
Input: A samosa sitting next to a mimosa on a table. **Negative prompt:** curry, flowers



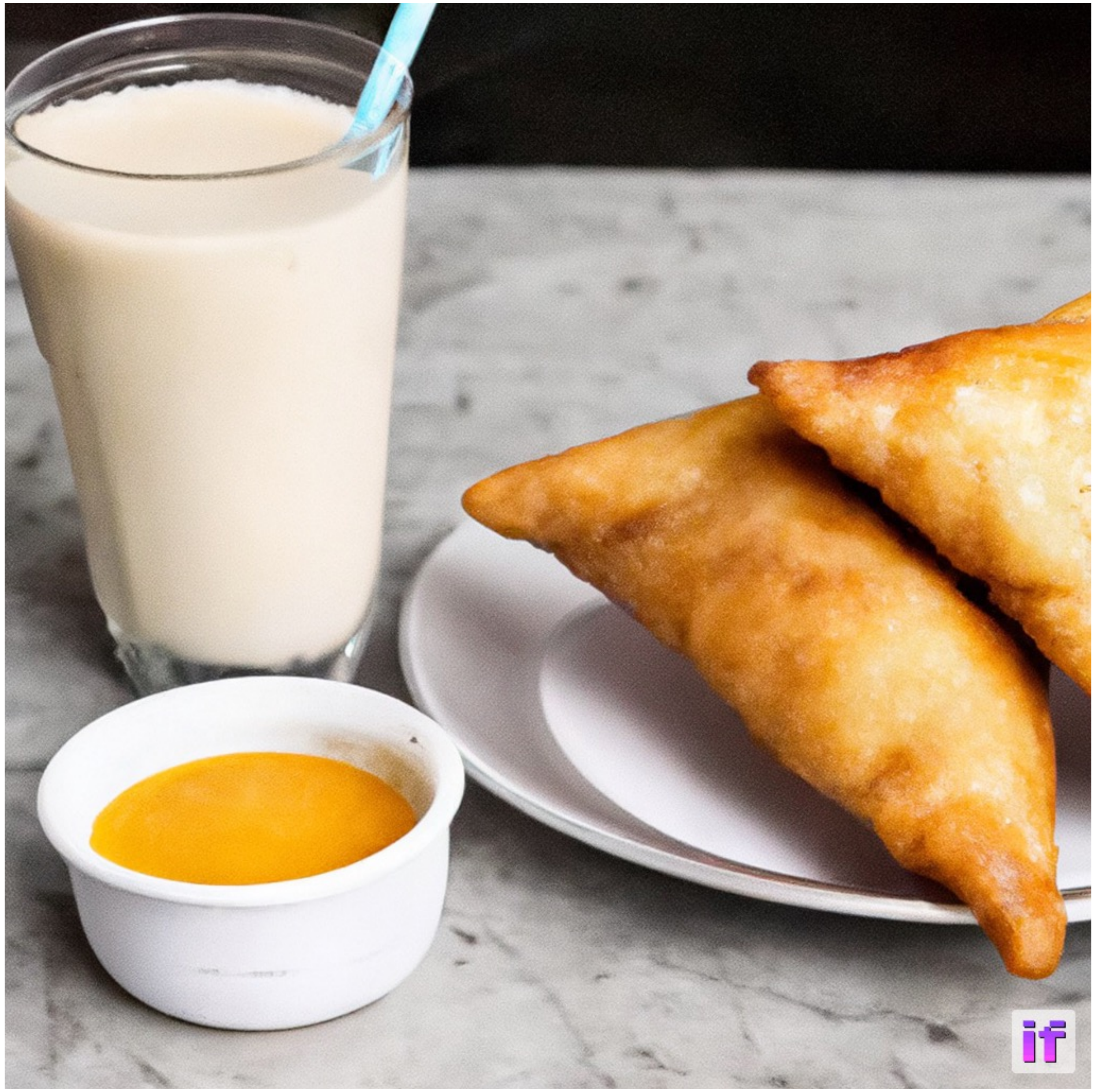
DeepFloyd IF

DeepFloyd IF’s rendition of samosas with a mimosa, with upscaling

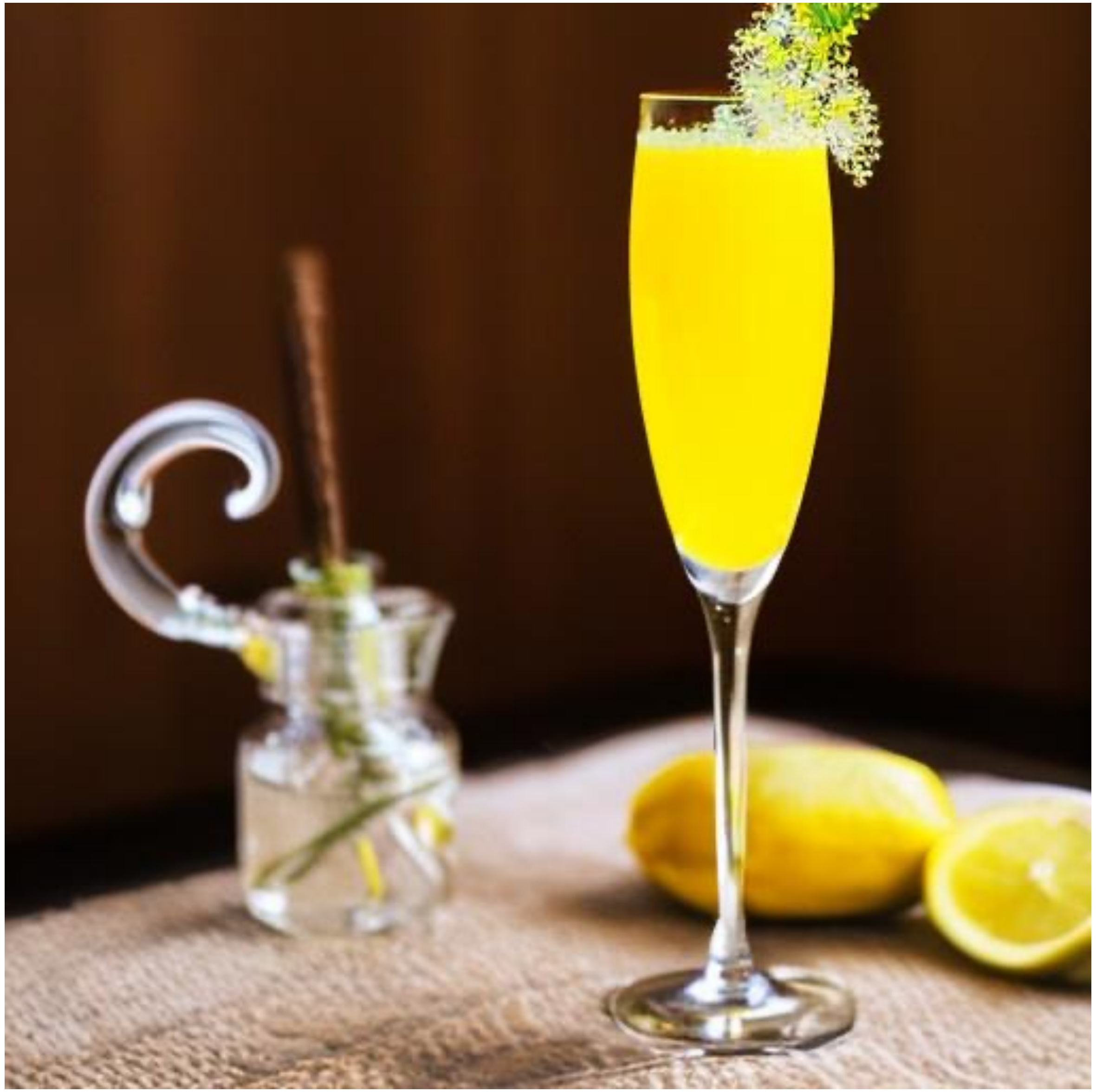
https://huggingface.co/spaces/DeepFloyd/IF



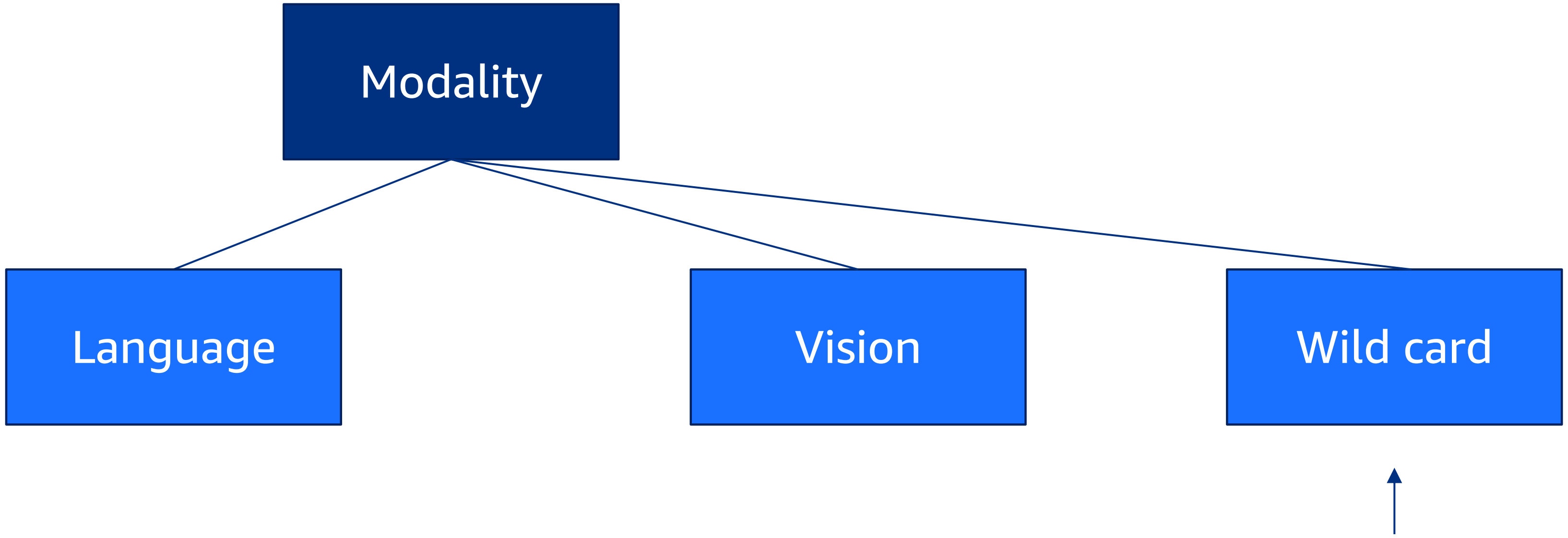
A samosa sitting next to a mango lassi on a table



A mimosa (with Stable Diffusion, not DeepFloyd)



Picking the right foundation model

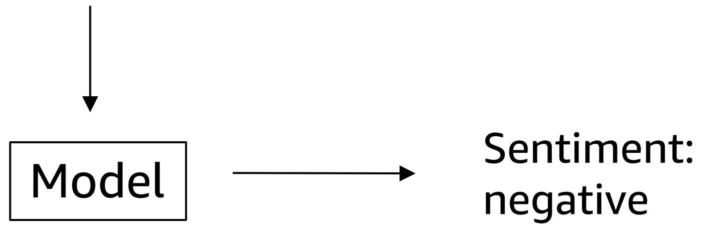


You can train or use a foundation model with any digital modality. Quality and ease-of-execution will vary with widespread adoption.

Proceed with caution.

Recasting your ML task as generative

Text: I am not into this house; it’s way too expensive and too far from the train line!



Model

Sentiment: negative

Traditional classification

Text: I am not into this house; it’s way too expensive and too far from the train line!

Classify this sentence into positive or negative sentiment:

Agent: Negative sentiment

Using generation to classify text

Task – generative or not

Try recasting your ML tasks with a generative model

- Classification
- Forecasting
- Recommendation
- Anomaly detection
- Translation
- Style transfer
- Visual search

Remember, you can still use a foundation model to benefit your ML project *even if you aren't targeting a generative use case.*

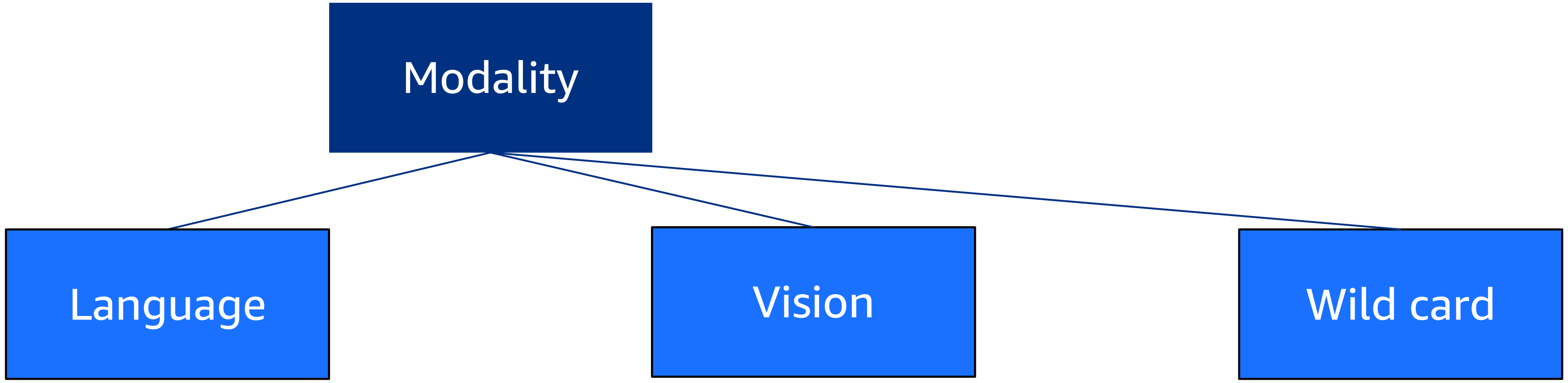
Pros:

- Streamline operations through single model for many use cases, rather than many models for many use cases
- Software, people, processes, datasets
- Accuracy may jump in some cases!
- Can use familiar evaluation metrics, like precision, recall, AUC, etc.

Cons:

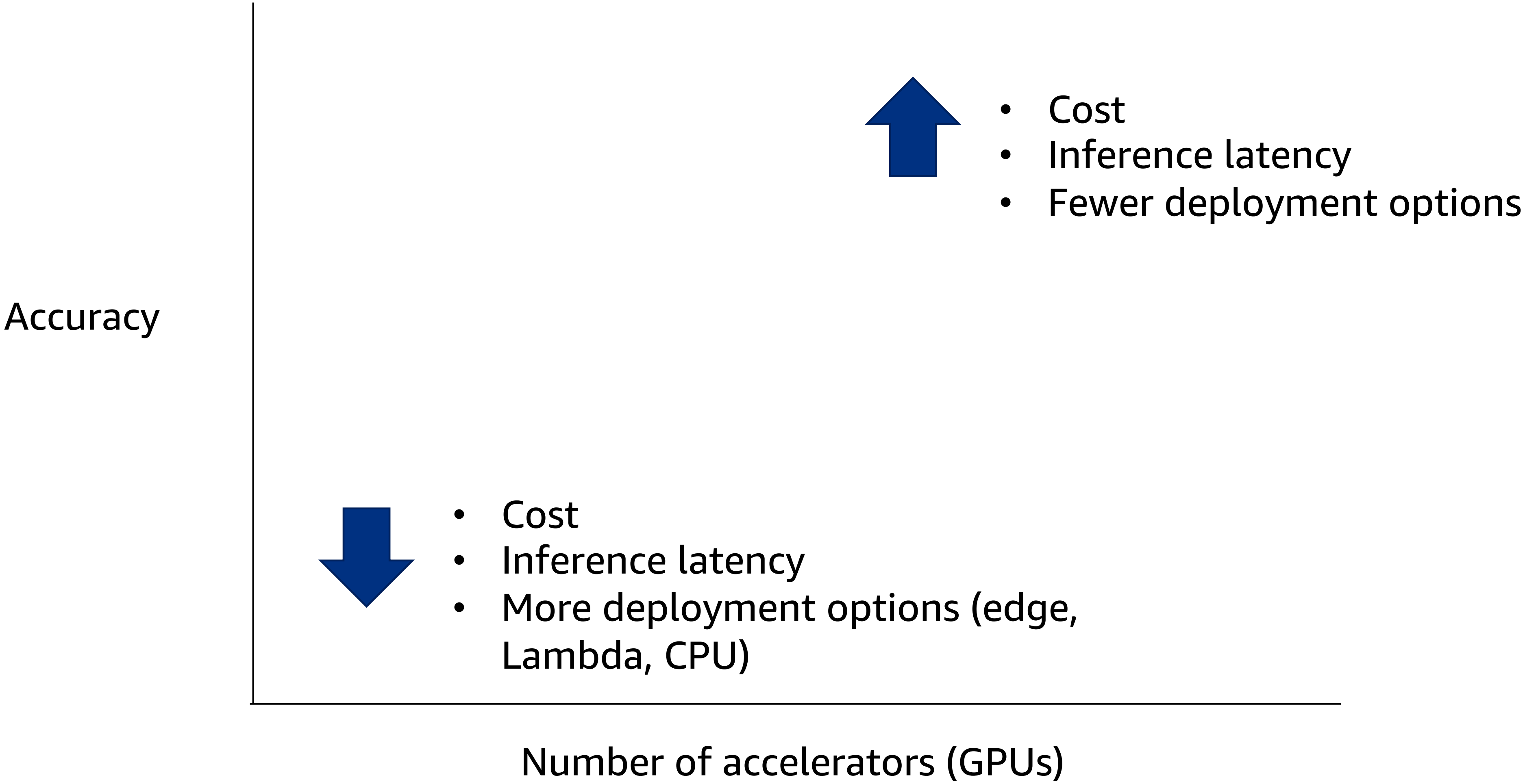
- Model inference runtime may *increase* for extremely large models
- Accuracy may tank in edge cases

Picking the right foundation model



Is generative? If yes, use decoder-based autoregressive models

Impact of size on foundation models



- Cost
- Inference latency
- Fewer deployment options

Accuracy

- Cost
- Inference latency
- More deployment options (edge, Lambda, CPU)

Number of accelerators (GPUs)

Impact of size on foundation models

Case for large foundation models

- Human brain has 86B neurons
- PaLM, Megatron-LM, GPT-3, GPT-J
- Larger datasets should logically hold more information, if deduplication is well handled
- Larger models should perform better with larger datasets
- Larger projects seem to be inherently more inspiring in tech field

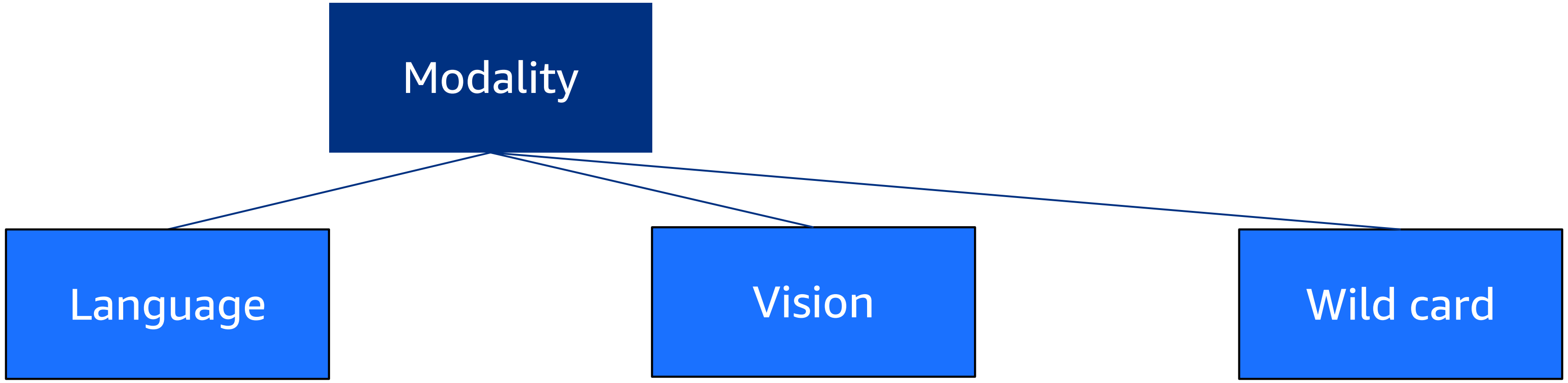
This is your creative space to innovate

🔥

Case against large foundation models

- Stable Diffusion is both single-accelerator *and* highly accurate
- AlexaTM (20B) outperforms GPT-3 (175B) in some cases
- InstructGPT (1.3B) outperformed GPT-3 with 1% of parameter count
- CNN's and encoders may outperform in some cases
- Lower costs are more accessible
- Lower carbon emission is always preferred

Picking the right foundation model



Is generative? If yes, use decoder-based autoregressive models

~ 1B parameters

~ 7B parameters

~ 11B parameters

~ 20B parameters

~ 50B parameters

> 100B parameters

Impact of accuracy on foundation models

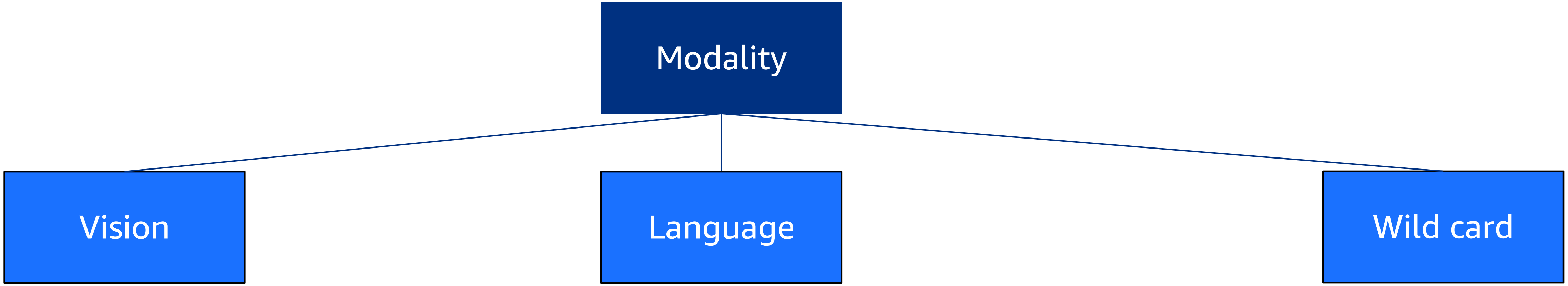
Accuracy can be misleading

Use labelled data with standard metrics

Human feedback always wins

Higher quality means less work for you

Picking the right foundation model



Modality

Vision

Language

Wild card

Is generative? If yes, use decoder-based autoregressive models

~ 1B parameters

~ 7B parameters

~ 11B parameters

~ 20B parameters

~ 50B parameters

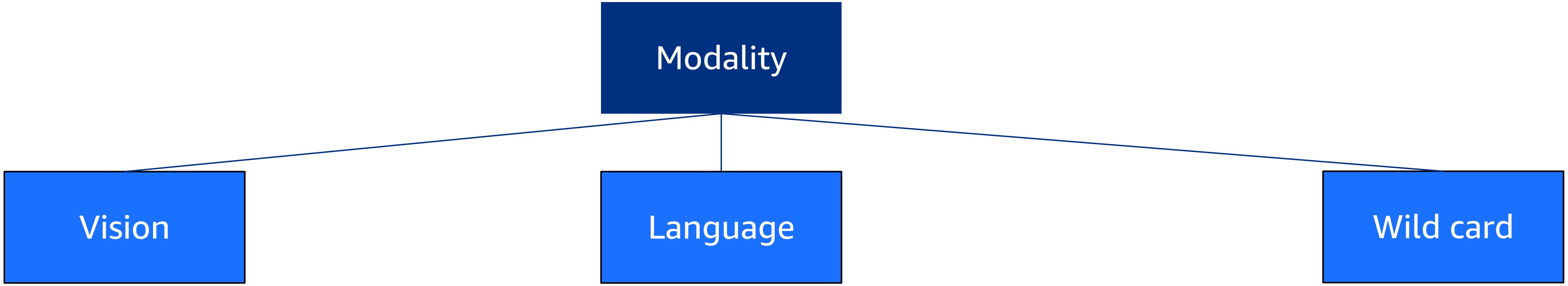
> 100B parameters

Set your accuracy thresholds

Navigating the open source / proprietary continuum

- Vibrant and creative economy of foundation models available
- Each offer unique pros and cons
- Picking open-source frees you from relying on a vendor
- Picking proprietary models gives you access to possibly more performant models more quickly

Picking the right foundation model



Modality

Vision

Language

Wild card

Is generative? If yes, use decoder- autoregressive models

~ 1B parameters

~ 7B parameters

~ 11B parameters

~ 20B parameters

~ 50B parameters

> 100B parameters

Set your accuracy thresholds

? Open source / proprietary

Pro tip – find a working example and start there

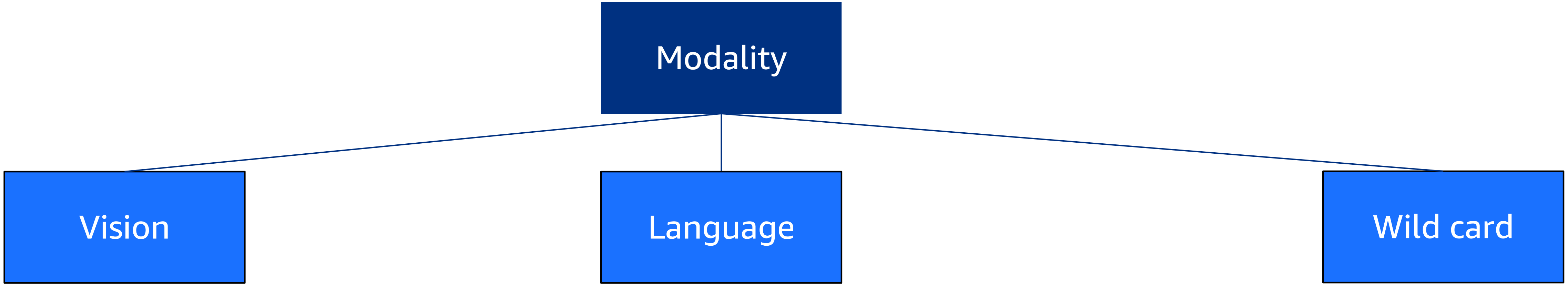
The Internet is a powerful tool!

- Papers on arxiv.org
- Work published at top ML conferences
- Blog posts, recorded sessions, tutorials
- GitHub repositories
- Online curriculum and coursework
- Benchmarks on *Papers with Code*

A working example trumps a concept

- Find a model that fits your use case and industry
- Even if it isn't perfect or overly novel, it may be a great starting point
- Building working content helps you win stakeholder buy-in early and often

Picking the right foundation model



Modality

Vision

Language

Wild card

Is generative? If yes, use decoder autoregressive models

~ 1B parameters

~ 7B parameters

~ 11B parameters

~ 20B parameters

~ 50B parameters

> 100B parameters

Set your accuracy thresholds

Open source / proprietary

Working example to start

Pro tip – find, use, master external benchmarks.

Develop unit tests and edge cases for your domain.

HELM

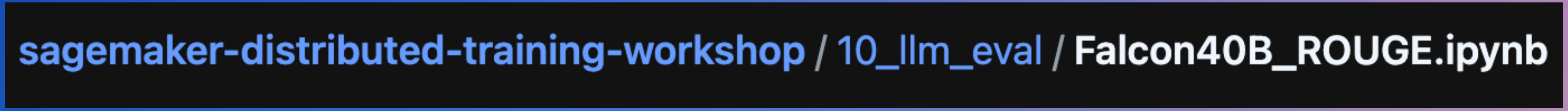
EleutherAI

Hugging Face

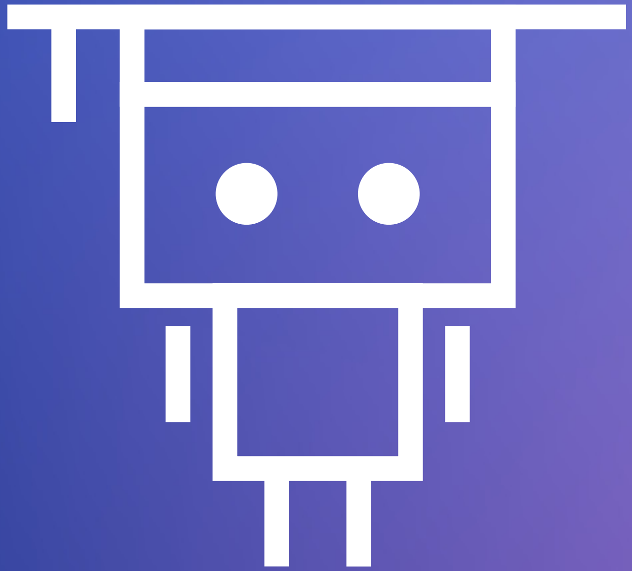
What if a better model comes along tomorrow?

https://bit.ly/sm-nb-2

Hands-on demo



sagemaker-distributed-training-workshop / 10_llm_eval / Falcon40B_ROUGE.ipynb



Thank you!

Type: Corrections, feedback, or other questions?
Contact us at https://support.awsamazon.com/#/contacts/aws-academy.
All trademarks are the property of their owners.