LLM Model Selection

(Factors to consider)

Sandeep Arora : Principal Architect

www.trilyen.com

https://www.linkedin.com/in/sandeeparora/

LLM Model Selection

There is no one size fits all. Model selection will depend on your industry & business use case. The priority for each business use case is different. For ex, an **AI Assistant** prioritizes low latency. An Advanced analytics app may prioritize advanced reasoning.

Cohere
Meta
Mistral AI
AI21 Labs
Amazon
Anthrop
Stability AI
Gemini
OpenAI

Selecting a LLM model

Factors to consider

- Use Case
- Cost
- Model Size
- Performance/Latency
- Advanced Reasoning
- Safety & Reliability
- Deployment options

By Sandeep Arora
www.trilyen.com

Business use case

Your business use case and its complexity significantly influences the selection of the appropriate model.

Narrow use cases

Customer support automation: simple FAQ and standard queries
Content moderation: Filter & flag inappropriate content or spam
Text classification: Sentiment analysis or topic labelling
Educational applications: Quiz , training and flashcard apps

Intermediate

Educational applications: Quiz , training and flashcard apps
Marketing: Generate personalized SMS campaigns
Content Generation: Write product description, summarize documents

Complex use cases

Advanced analytics: summarize multi-modal content
Legal and contract analysis: analyze large volume of legal content
Healthcare analytics: assist in diagnostics
Field service : multi-modal field service assistants

complexity

Narrow AI use cases (points to consider)

- Small language model(SLM) may be ok
- Prompt engineering may be sufficient to adapt model for your use case. This saves time & cost since fine-tuning is time consuming & expensive
- Various deployment options (Open-source vs deploy on-prem vs API pay-per-use model)
- Little or no regulatory constraints

Business use case

Your business use case and its complexity significantly influences the selection of the appropriate model.

Narrow use cases

Customer support automation: simple FAQ and standard queries
Content moderation: Filter & flag inappropriate content or spam
Text classification: Sentiment analysis or topic labelling
Educational applications: Quiz , training and flashcard apps

Intermediate

Educational applications: Quiz , training and flashcard apps
Marketing: Generate personalized SMS campaigns
Content Generation: Write product description, summarize documents

Complex use cases

Advanced analytics: summarize multi-modal content
Legal and contract analysis: analyze large volume of legal content
Healthcare analytics: assist in diagnostics
Field service : multi-modal field service assistants

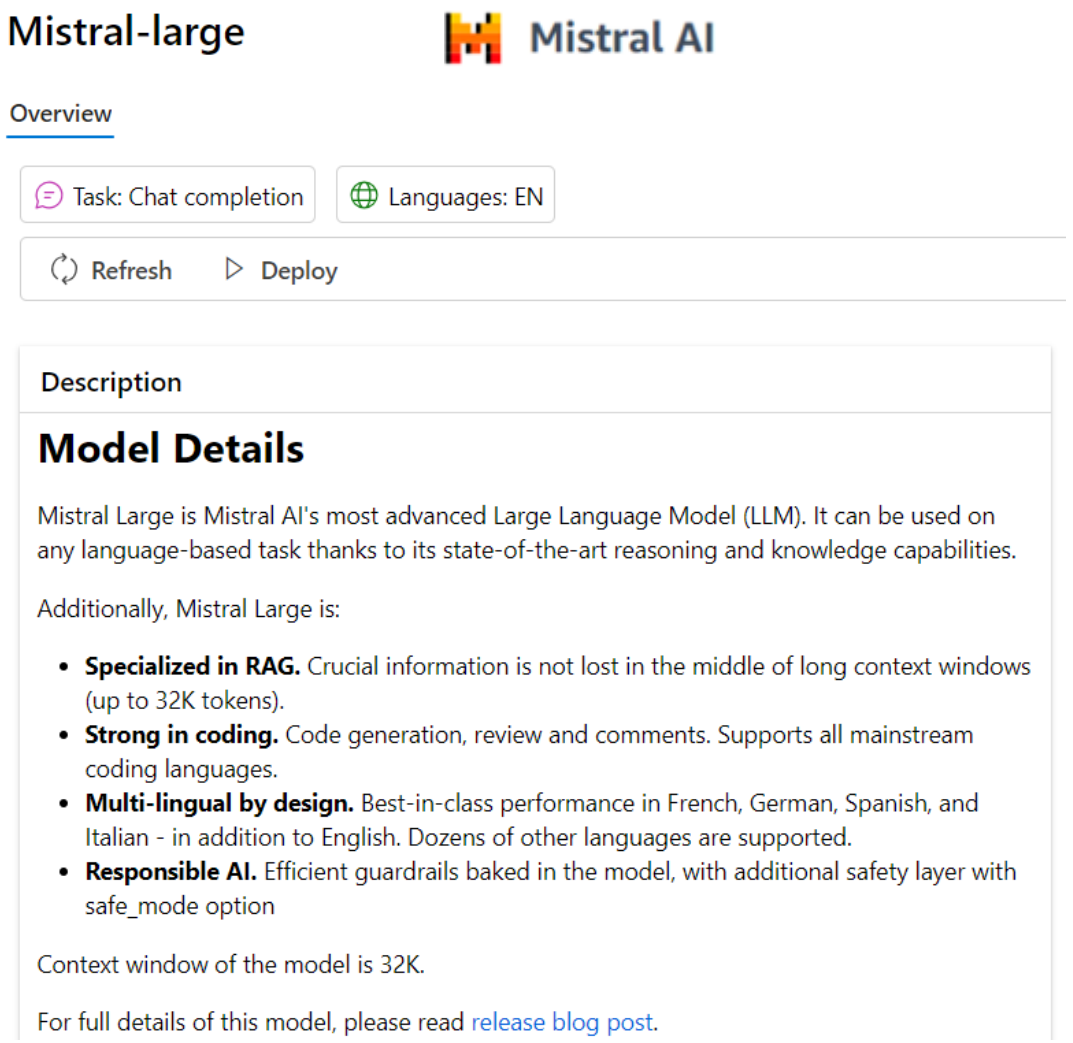
complexity

Complex AI use cases (points to consider)

- Large size model may be needed if advance reasoning is a priority
- Fine tuning may be a **requirement**
- Multi modality may is a requirement
- Strict legal compliance requirement

Match model to use case

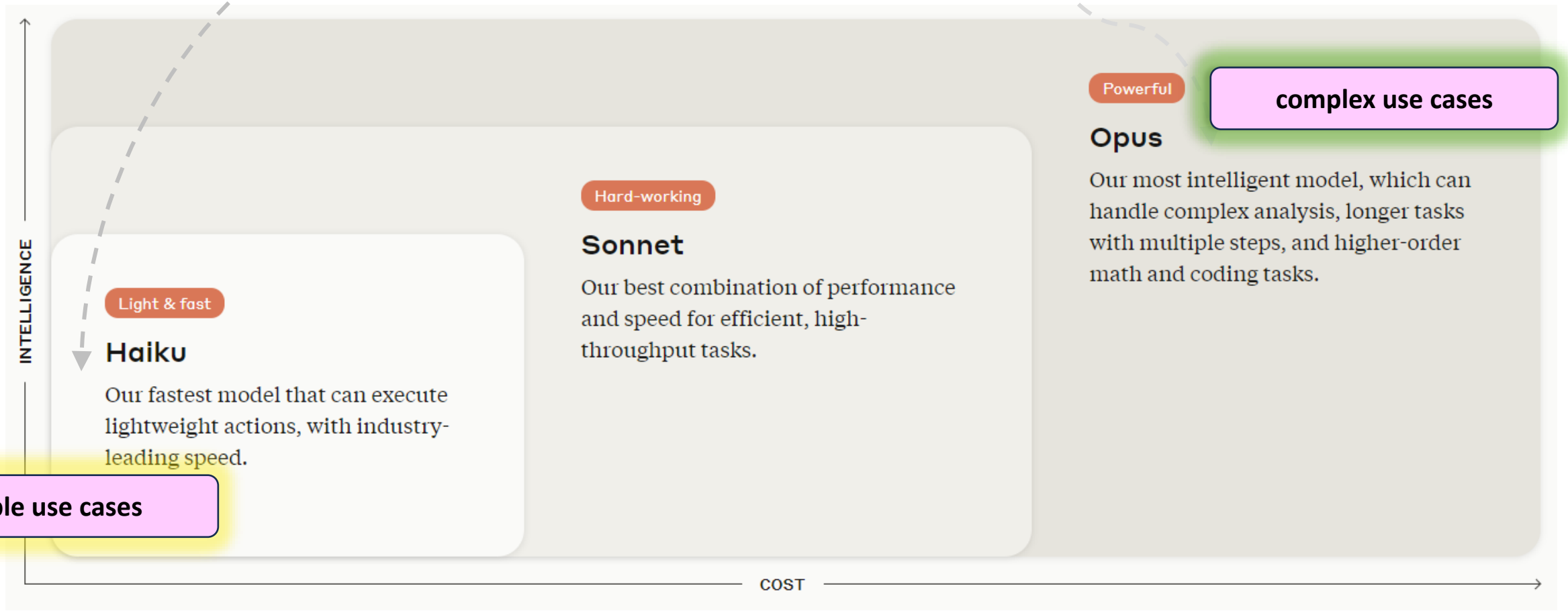
Understand the **capabilities** of different models to pick the right ones.



Mistral large capabilities

Anthropic's family of Claude models.

Different models for different use cases

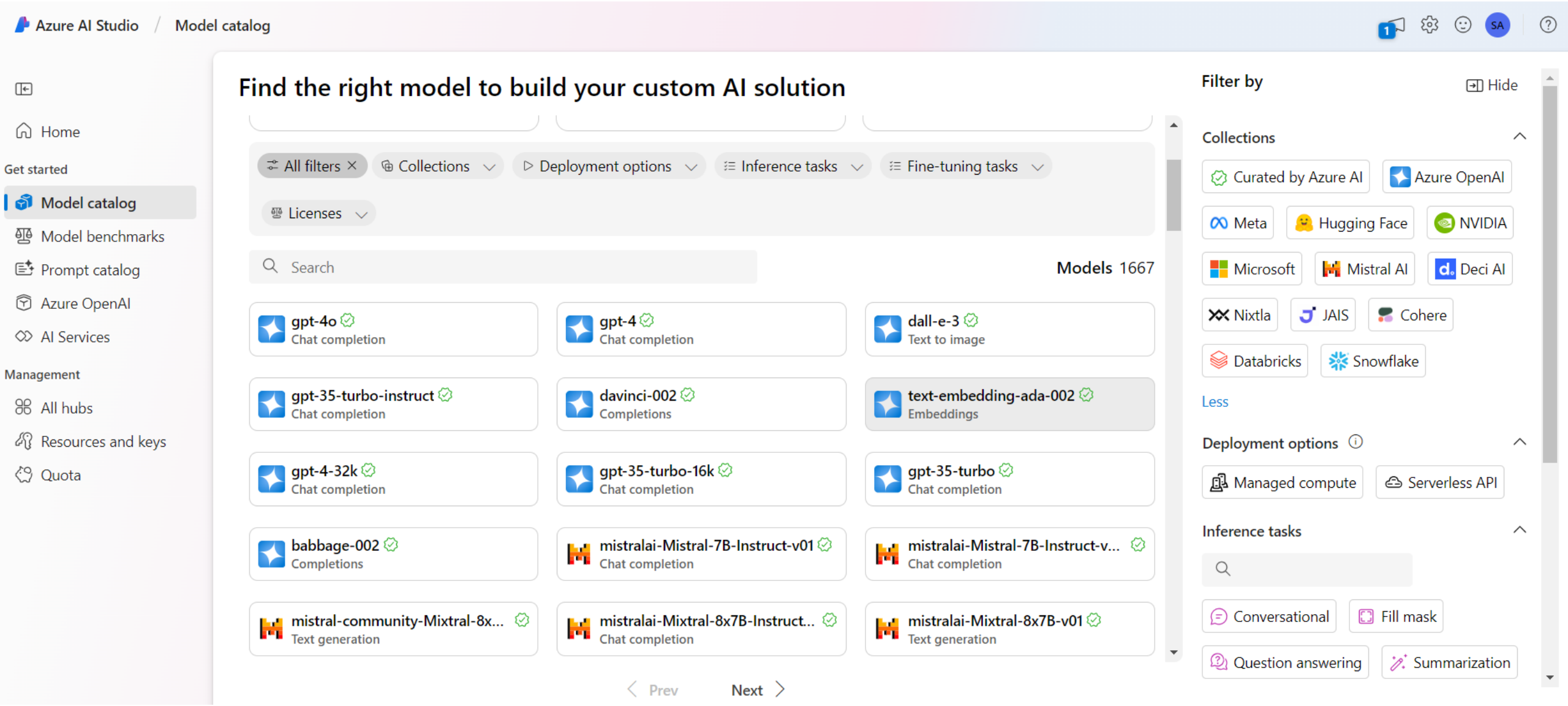


simple use cases

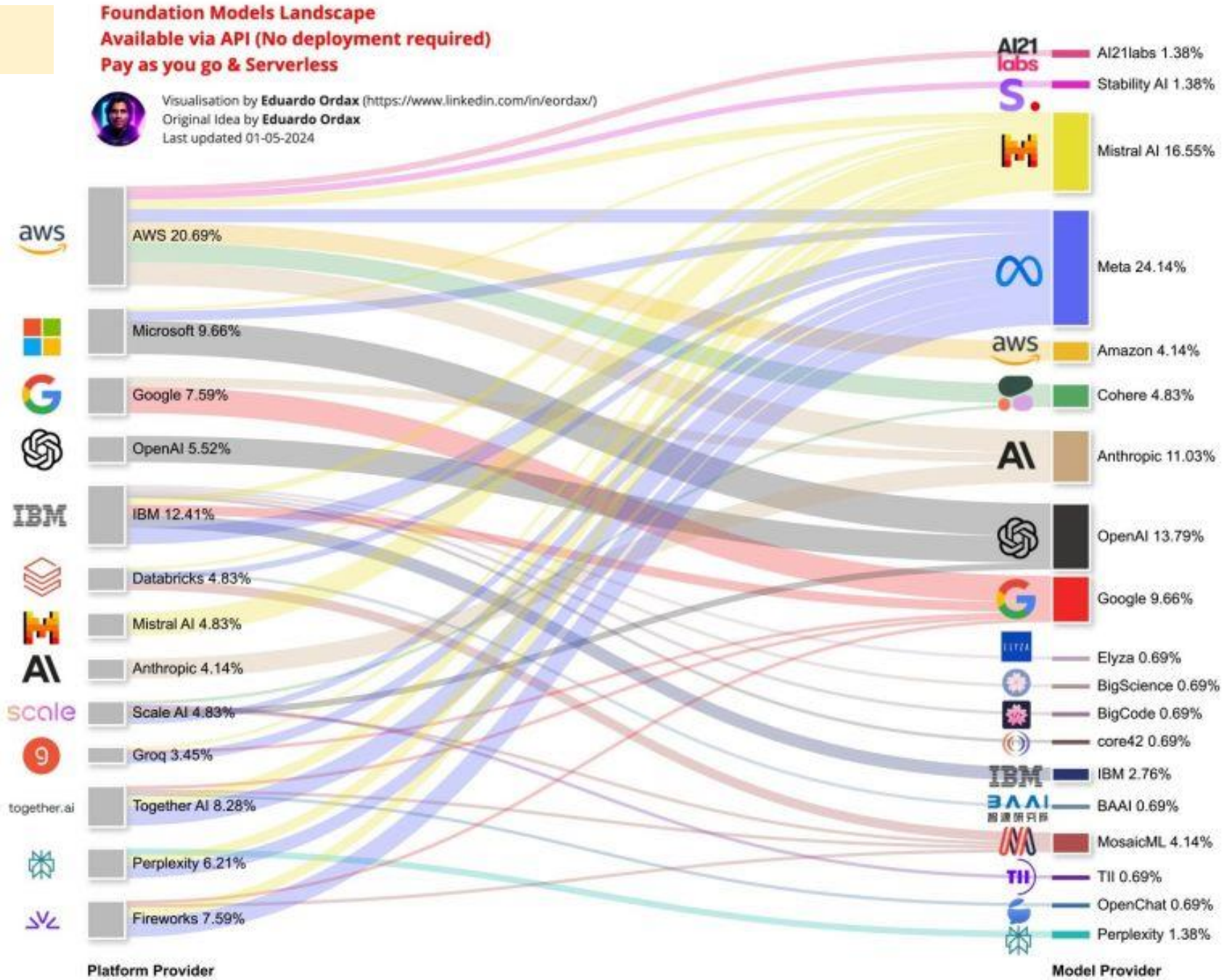
complex use cases

Sourcehttps://www.anthropic.com/claude

Azure AI Studio has 1667 models. Research and have a point-of-view on the models which match your use case. Do your due diligence. This will take time, but will be beneficial in the build stage.

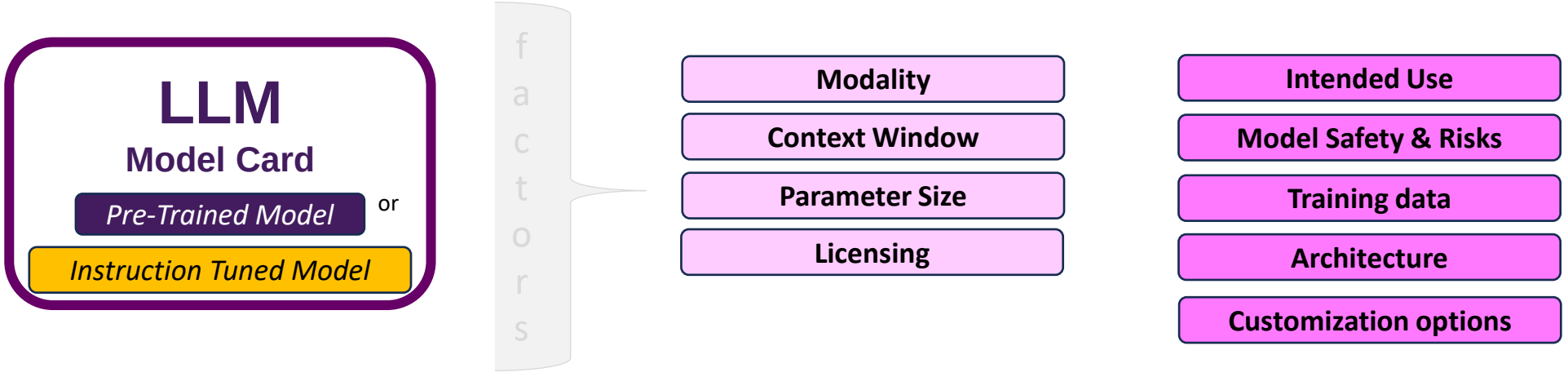


Foundation models landscape



LLM Model evaluation

Pay attention to LLM model properties and evaluate for your use case. Compare **model cards** of LLM models to understand model intended use, properties, architecture, risks, strengths & weakness of each.



- **Modality**: Some are unimodal, focused on text (GPT-3, BERT), while others are multimodal, handling text and images (Claude 3 Sonnet, GPT-4).
- **Customization**: Can the model be customized for your domain. How easy is it to customize?
- **Model Safety & Risks**: How does the model process & store data. Compliance with GDPR, HIPPAA & other regulations. Can decisions be **explained**

- **Context Window**: The amount of context they can process varies, with GPT-3 handling up to 2048 tokens and models like Jurassic-1 Jumbo handling over 1 million tokens. Larger context windows allow better understanding of long-form inputs. (Claude 3 is 200k tokens roughly translates to 375 pages)

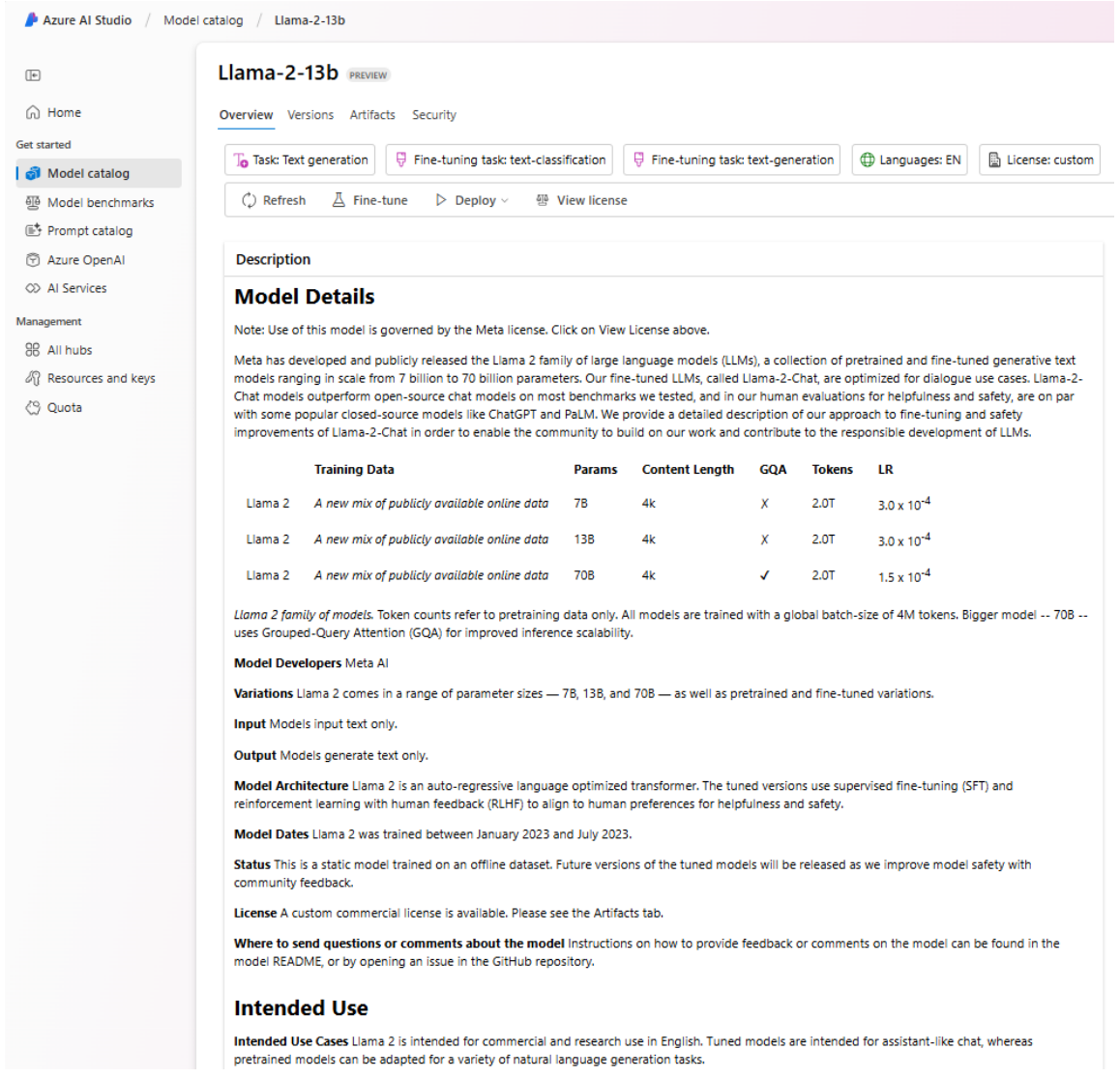
- **Parameter Size**: They range from hundreds of millions to over a trillion parameters (LLaMA available at several sizes 7B, 13B, 33B, and 65B). Generally, larger models perform better on more complex tasks, but more expensive and can result in higher latency

Open Source models are publicly available and can be used, modified, and distributed freely. These models promote transparency and collaboration, often benefiting from community contributions and continuous improvements. Examples: Meta (LLaMA), Mistral AI

Proprietary models are developed and owned by organizations. They are typically accessible through paid licenses or subscriptions, offering exclusive features, support, and optimized performance tailored to specific use cases. Examples: OpenAI (GPT-4), Google (PaLM), Amazon (Titan)

LLM Model Card

Review the model card, with focus on intended use, training data, size, security, code samples and evaluation results on internal evaluations library. You can get this information of all Hyperscalers and LLM marketplaces like HuggingFace



Source: Azure AI studio

LLM Model Card example

Home
Get started
Model catalog
Model benchmarks
Prompt catalog
Azure OpenAI
AI Services
Management
All hubs
Resources and keys
Quota

with a meticulous approach to filtering out undesirable documents and images. To safeguard privacy, we carefully filtered various image and text data sources to remove or scrub any potentially personal data from the training data.

More details can be found in the Phi-3 Technical Report.

Benchmarks

To understand the capabilities, we compare Phi-3 Vision-128K-Instruct with a set of models over a variety of zero-shot benchmarks using our internal benchmark platform.

Benchmark	Phi-3 Vision-128K-In1	LlaVA-1.6 Vicuna-7B	QWEN-VL Chat	Llama3-Llava-Next-8B	Claude-3 Haiku	Gemini 1.0 Pro V	G 4 T
MMMU	40.4	34.2	39.0	36.4	40.7	42.0	5
MMBench	80.5	76.3	75.8	79.4	62.4	80.0	8
ScienceQA	90.8	70.6	67.2	73.7	72.0	79.7	7
MathVista	44.5	31.5	29.4	34.8	33.2	35.0	4
InterGPS	38.1	20.5	22.3	24.6	32.1	28.6	4
AI2D	76.7	63.1	59.8	66.9	60.3	62.8	7
ChartQA	81.4	55.0	50.9	65.8	59.3	58.0	6
TextVQA	70.9	64.6	59.4	55.7	62.7	64.7	6
POPE	85.8	87.2	82.6	87.0	74.4	84.2	8

Responsible AI Considerations

Like other models, the Phi family of models can potentially behave in ways that are unfair, unreliable, or offensive. Some of the limiting behaviors to be aware of include:

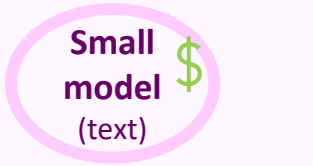
- Quality of Service: The Phi models are trained primarily on English text. Languages other than English will experience worse performance English language varieties with less representation in the training data might experience worse performance than standard American English.
- Representation of Harms & Perpetuation of Stereotypes: These models can over- or under-represent groups of people, erase representation of some groups, or reinforce demeaning or negative stereotypes. Despite safety post-training, these limitations may still be present due to differing levels of representation of different groups or prevalence of examples of negative stereotypes in training data that reflect real-world patterns and societal biases.
- Inappropriate or Offensive Content: These models may produce other types of inappropriate or offensive content, which may make it inappropriate to deploy for sensitive contexts without additional mitigations that are specific to the use case.
- Information Reliability: Language models can generate nonsensical content or fabricate content that might sound reasonable but is inaccurate or outdated.
- Limited Scope for Code: Majority of Phi-3 training data is based in Python and use common packages such as "typing, math, random, collections, datetime, itertools". If the model generates Python scripts that utilize other packages or scripts in other languages, we strongly recommend users manually verify all API uses.

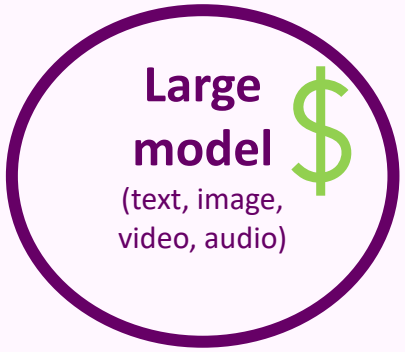
Developers should apply responsible AI best practices and are responsible for ensuring that a specific use case complies with relevant laws and regulations (e.g. privacy, trade, etc.). Important areas for consideration include:

- Allocation: Models may not be suitable for scenarios that could have consequential impact on legal status or the allocation of resources or life opportunities (ex: housing, employment, credit, etc.) without further assessments and additional debiasing techniques.

Model Size effects performance, cost and capabilities of a LLM model

Performance: Larger models generally perform better on complex tasks that require advanced reasoning
Latency: Larger models have longer inference times because processing inputs requires more computation
Resource Requirements: Bigger models may necessitate the use of specialized hardware, like GPUs
Cost: The cost of training, deploying and inferencing is high for large size models.





Llama 2

MODEL SIZE (PARAMETERS)	PRETRAINED	FINE-TUNED FOR CHAT USE CASES
7B	Model architecture:	Data collection for helpfulness and safety:
13B	Pretraining Tokens: 2 Trillion	Supervised fine-tuning: Over 100,000
70B	Context Length: 4096	Human Preferences: Over 1,000,000

Source: https://llama.meta.com/llama2/

Cost

Cost is an important factor when deciding which LLM model to choose. You would need to balance cost & performance for your business use case. Broadly speaking here are a few factors to consider

Specific use case & scale

Smaller, more specialized models may be sufficient for specific tasks and can be more cost-effective than larger, more general-purpose models

Pretrained vs custom model

Using pre-trained models can significantly reduce costs compared to training a model from scratch. Many providers offer models that are pre-trained on diverse data, which can be fine-tuned for specific tasks at a lower cost.

Usage Estimates

Estimate how much you will use the model. Most providers charge by the number of tokens exchanged (input & output). Consider on-premise options and open-source models if usage is very high.

On-premise vs API

Evaluate and consider on-premise options using open-source models for high usage smaller use cases. That will save you on API usage-based costs

Cost: usage estimates (GPT)

GPT-4o New

Our fastest and most affordable flagship model

- Text and image input, text output
- 128k context length
- Input: $5 | Output: $15*

GPT-4 Turbo

Our previous high-intelligence model

- Text and image input, text output
- 128k context length
- Input: $10 | Output: $30*

GPT-3.5 Turbo

Our fast, inexpensive model for simple tasks

- Text input, text output
- 16k context length
- Input: $0.50 | Output: $1.50*

** prices per 1 million tokens*

Advanced Reasoning & Performance

Custom AI applications which need to perform complex cognitive tasks over long context, that go beyond simple tasks(*sentiment analysis, NER, support chatbots, etc*.) need models which excel at various metrics including advanced reasoning. Evaluate and compare how models perform on various metrics.

Depending on your use case, make sure the model excels in language nuances, contextual understanding, can handle multi-step tasks, has low hallucination & false refusal rates and is safe.

Compare models on benchmarks such as MMLU, BBH, HelloSwag, Math, GSM8K, etc.

Advanced Reasoning & Performance

Model Benchmarks

Understand and evaluate model benchmarks against your business use case

CAPABILITY	BENCHMARK	DESCRIPTION	Gemma 7B	Llama-2 7B	Llama-2 13B
General	MMLU 5-shot, top-1	Representation of questions in 57 subjects (incl. STEM, humanities and others)	64.3	45.3	54.8
Reasoning	BBH -	Diverse set of challenging tasks requiring multi-step reasoning	55.1	32.6	39.4
	HellaSwag 0-shot	Commonsense reasoning for everyday tasks	81.2	77.2	80.7
Math	GSM8K maj@1	Basic arithmetic manipulations (incl. Grade School math problems)	46.4	14.6	28.7
	MATH 4-shot	Challenging math problems (incl. algebra, geometry, pre-calculus, and others)	24.3	2.5	3.9
Code	HumanEval pass@1	Python code generation	32.3	12.8	18.3

Source: https://blog.google/technology/developers/gemma-open-models/

Advanced Reasoning & Performance

Understand the various benchmarks and which one matters most to you based on your business use case.

MMLU (Massive Multitask Language Understanding) benchmark — *for* General Capability

MMLU is designed to test the language model's ability to understand and process information across a vast array of topics. It covers 57 distinct subjects ranging from science, technology, engineering, and mathematics (STEM) to the humanities and beyond. This benchmark evaluates the model's breadth of knowledge and its capacity to handle questions that require understanding complex, nuanced, or highly specialized content. Tasks often include multiple-choice questions where the model must select the correct answer from several options based on the given context.

BBH (Broad-Based Reasoning) benchmark — *for* Reasoning Capability

BBH assesses a language model's reasoning abilities across a variety of contexts. This benchmark includes tasks that require the model to engage in multi-step reasoning, problem-solving, and logical deduction. Tasks might involve puzzles, strategy games, or complex scenarios where the model must infer the most logical outcomes or decisions based on the information provided.

HellaSwag benchmark — *for* Reasoning Capability

HellaSwag is designed to evaluate commonsense reasoning within everyday contexts. It challenges models to predict the continuation of a narrative or to choose the most plausible ending among several options. This benchmark tests the model's ability to use everyday knowledge and implicit understandings of the world to make predictions about how scenarios typically unfold.

Advanced Reasoning & Performance (cont.)

GSM8K (Grade School Math 8K) benchmark

for Math capability

GSM8K focuses on basic arithmetic and elementary mathematical concepts, mimicking the level of math problems encountered in a typical grade school curriculum. This benchmark includes tasks such as addition, subtraction, multiplication, and division, as well as simple word problems where the model must apply these operations to solve practical questions.

MATH benchmark:

for Math capability

The MATH benchmark poses more advanced mathematical challenges that cover high school and early college-level topics. These tasks include problems in algebra, geometry, trigonometry, pre-calculus, and basic calculus. The benchmark tests the model's ability to not only perform calculations but also to understand and manipulate mathematical expressions and concepts to find solutions.

HumanEval benchmark:

for Code capability

HumanEval is a coding benchmark that tests the model's ability to generate functional Python code. It presents the model with programming problems, typically requiring the creation of a function that meets specified criteria. The tasks assess the model's understanding of coding logic, syntax, and its ability to implement algorithms that correctly solve the given problems.

Latency

Latency is an important factor to consider when selecting which LLM to use, as it directly affects the user experience & the efficiency of the operations that rely on the model . The choice of an LLM for your application depends significantly on your specific latency requirements and the nature of the tasks it needs to perform

Applications where Latency is critical

- Chatbots and Virtual Assistants
- Language translation or speech recognition
- Real time content moderation
- Multi-modal AI Agents

Applications where accuracy is more important than latency

- Advanced data analysis & research
- Offline agents
- Content generation (say for products) where immediate response is not needed

Latency

Latency refers to the delay between a user's input and the model's response. This delay can be influenced by several factors, such as

- size of the model
- complexity of the input
- efficiency of the model's architecture
- performance of the hardware it's running on
- network speed if the model is hosted remotely

Latency Metrics

Few important metrics to monitor :

•**TTFT (Time To First Token)**: time it takes for the model to generate the first token of a response

•**E2E Latency (End-to-End Latency)**: total time it takes for the model to generate a complete response

•**ITL (Inter-Token Latency)**: known as Time Per Output Token (TPOT), measures the average time the client waits between consecutive tokens in a response

Responsible AI

LLM potentially behave in ways that are unfair, unreliable, or offensive. Make sure the model you select has safety built into it & you evaluate the model for your business use case.

Bias & fairness
LLMs can inherit and even amplify biases present in their training data. This can lead to outputs that are unfair or discriminatory, particularly in sensitive applications involving gender, race, or other personal characteristics.

Hallucination
instances where the model generates text that is factually incorrect, misleading, or entirely fabricated, despite being presented in a confident and plausible manner. This behavior can range from minor inaccuracies to completely erroneous statements.

Explainability
Understanding why an LLM produces a specific output is often difficult, as the models are typically "black boxes" with complex internal workings. This lack of transparency can be problematic, especially in applications requiring clear audit trails or explanations of decisions, such as in healthcare or finance. Make sure you can evaluate model answers if accuracy is a critical requirement for your industry and use case.

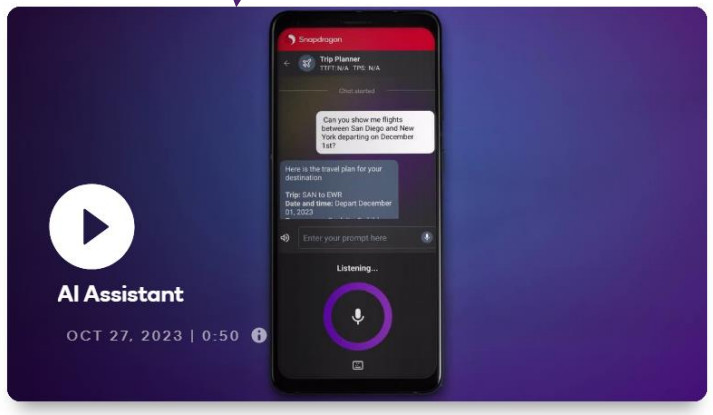
Offensive content
LLM models may generate other types of inappropriate or offensive content, which may make it inappropriate to deploy for sensitive contexts without additional mitigations that are specific to the use case.

Deployment Options: API, On-premise or On-device

Your business use case and choice of deployment will play a key role in model selection process.

On-Device

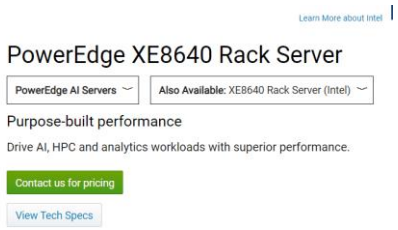
Smaller models for on-device processing like LLMA 2 7B. Ex voice assistants



Source:
www.qualcomm.com/products/mobile/snapdragon/smartphones/mobile-ai

On-premise

Open-source models like LLMA 3 which you can download, modify & setup on a server in your company data center



Source
www.dell.com/en-us/shop/ipovw/poweredge-xe8640?hve=shop+now

Accessed via API

Models like GPT-4 which can be accessed only via API calls