Analytics Vidhya

Choosing the Right LLM

Instructor

Dipanjan Sarkar

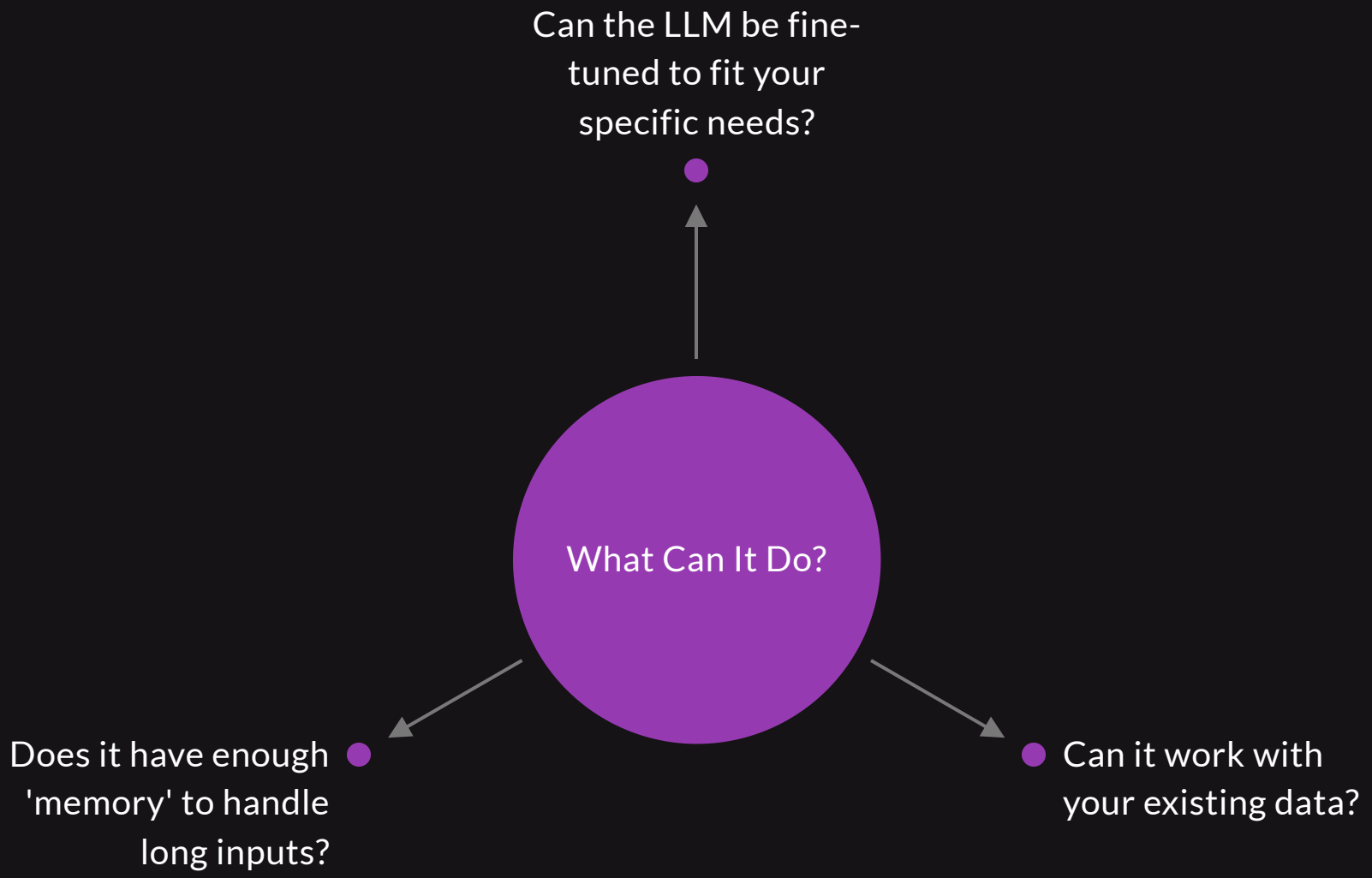
Head of Community & Principal AI Scientist at Analytics Vidhya

Google Developer Expert - ML & Cloud Champion Innovator

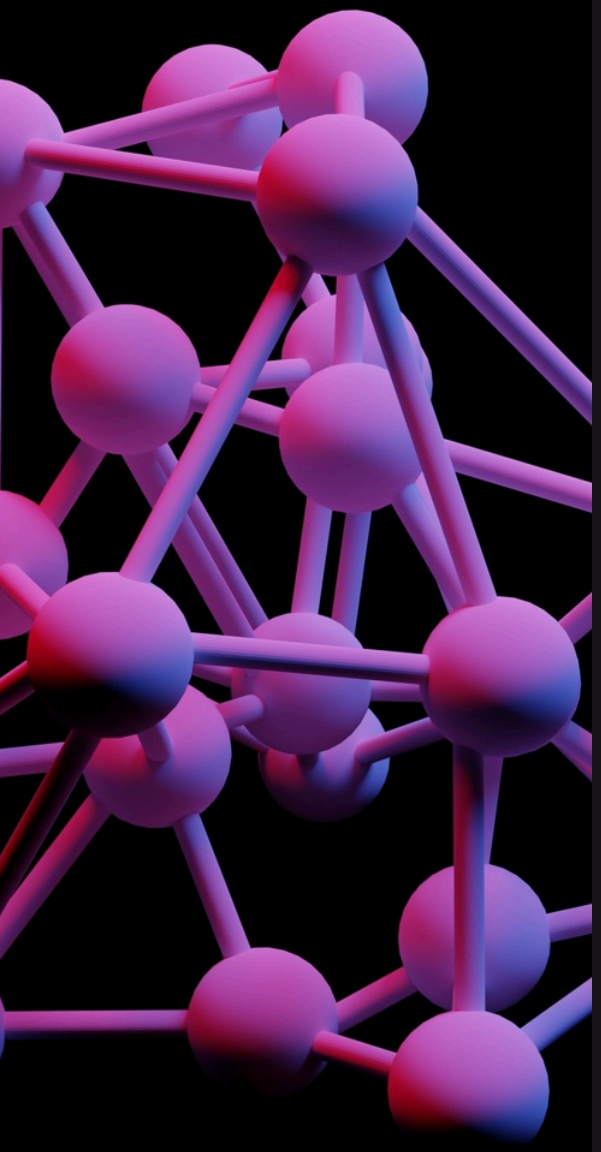
Published Author

Choosing the Right LLM - Factors

1. What Can It Do? (Capability)



Analytics Vidhya

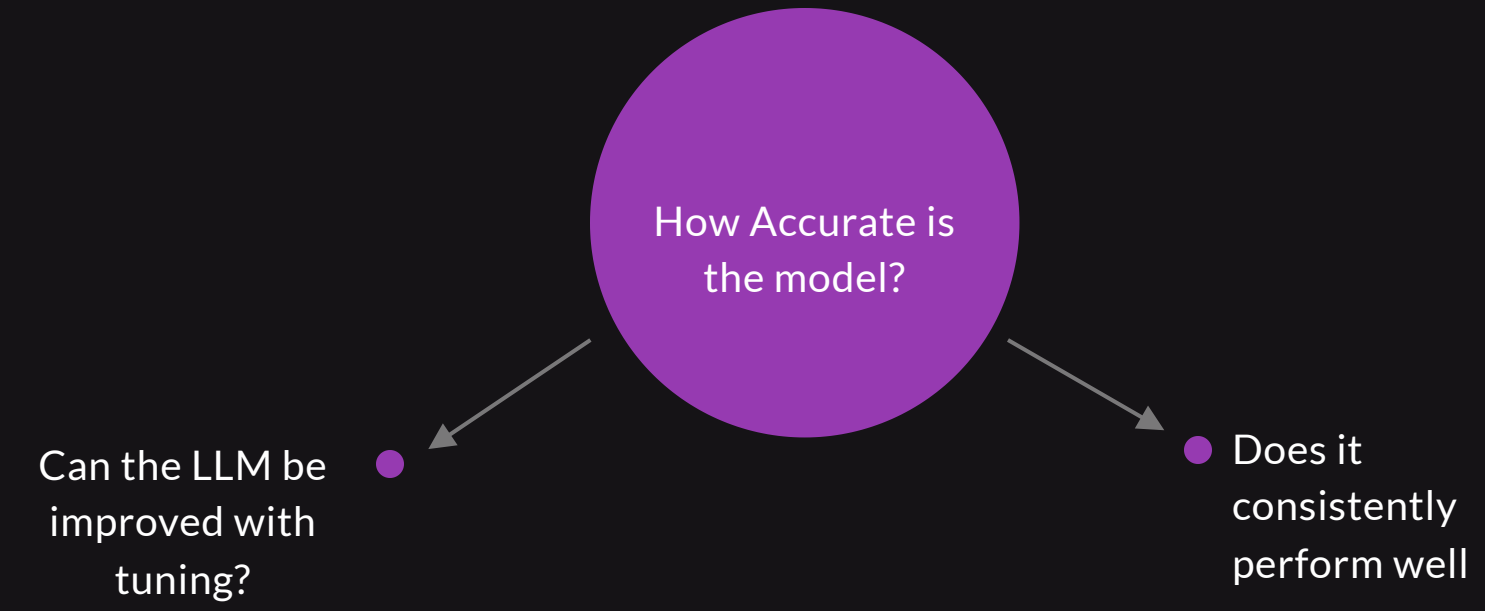


Capability Comparison - LLMs

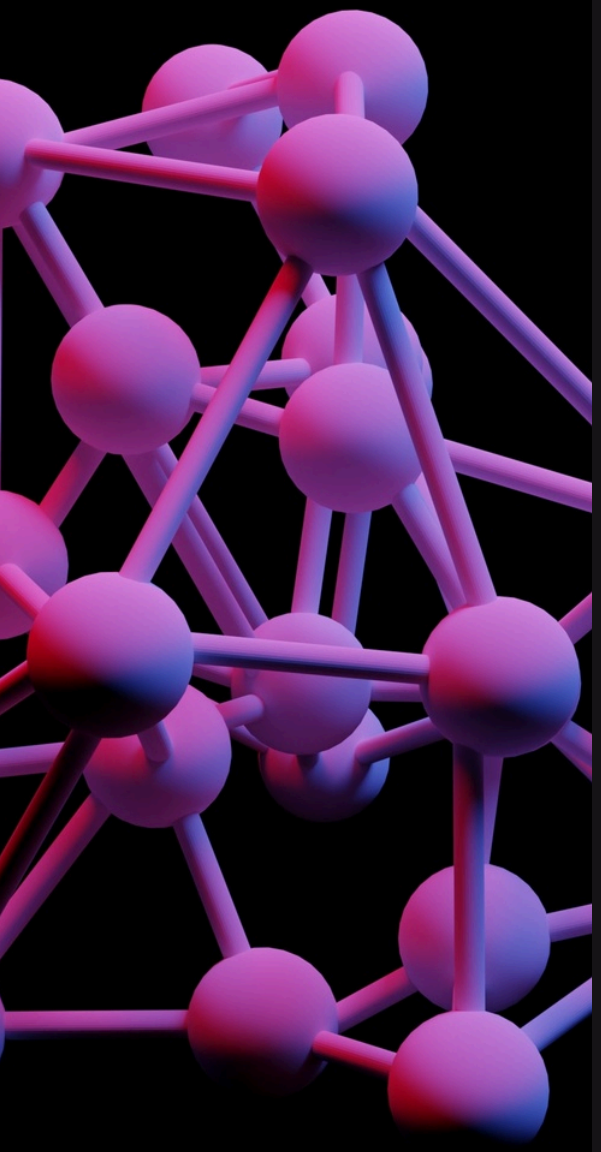
LLM	Can be Fine-Tuned	Works with Custom Data	Memory (Context Length)
LLM 1	✓	✓	2048 tokens
LLM 2	✗	✓	4096 tokens
LLM 3	✓	✗	1024 tokens

Choosing the Right LLM - Factors

2. How Accurate Is It?



Analytics Vidhya



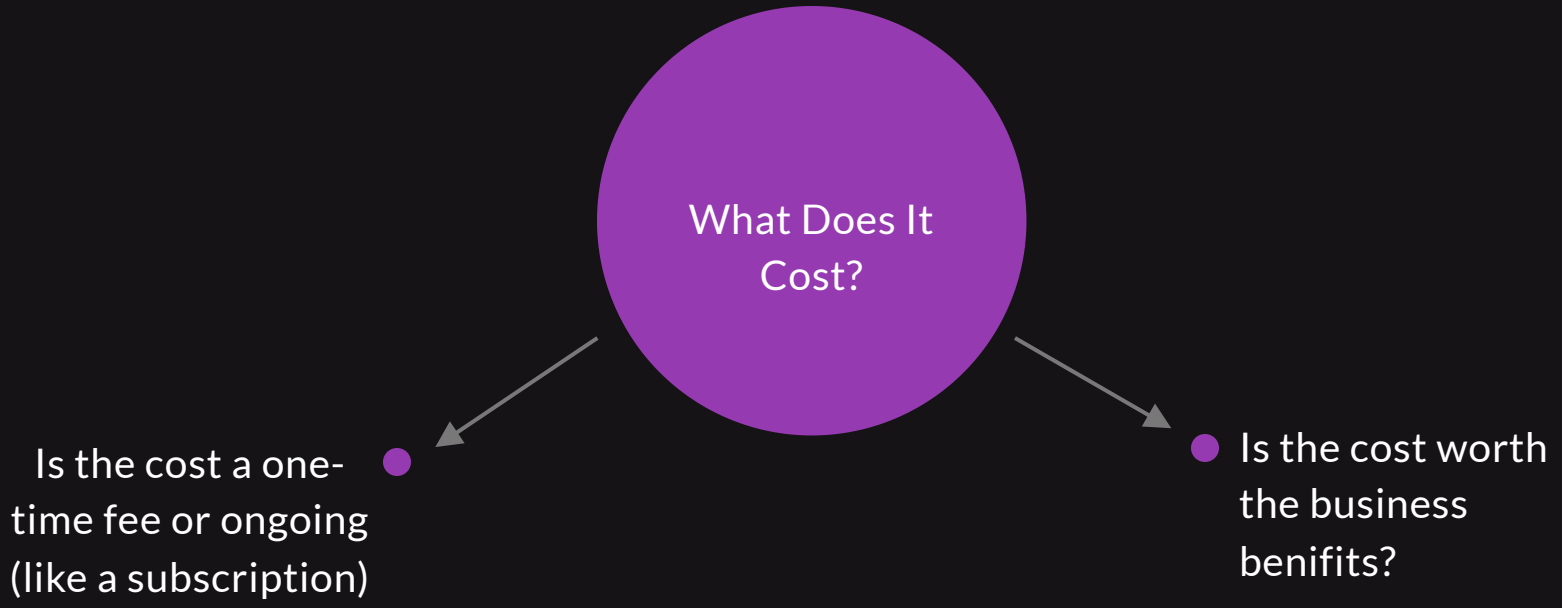
Accuracy Comparison - LLMs

LLM	General Accuracy	Accuracy with Custom Data
LLM 1	90%	85%
LLM 2	85%	80%
LLM 3	88%	86%

Analytics Vidhya

Choosing the Right LLM - Factors

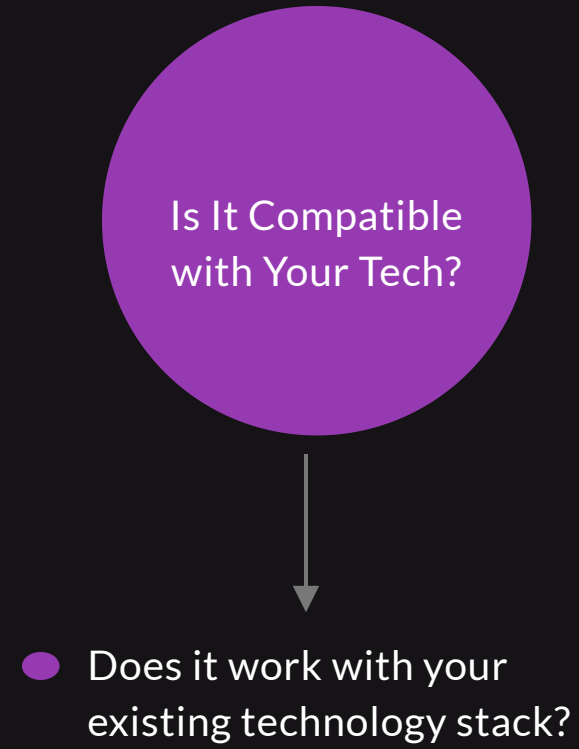
3. What Does It Cost?



Analytics Vidhya

Choosing the Right LLM - Factors

4. Is It Compatible with Your Tech?



Make sure the LLM fits with your current tech setup. Most LLMs use Python, but your business might use something different.

Analytics Vidhya

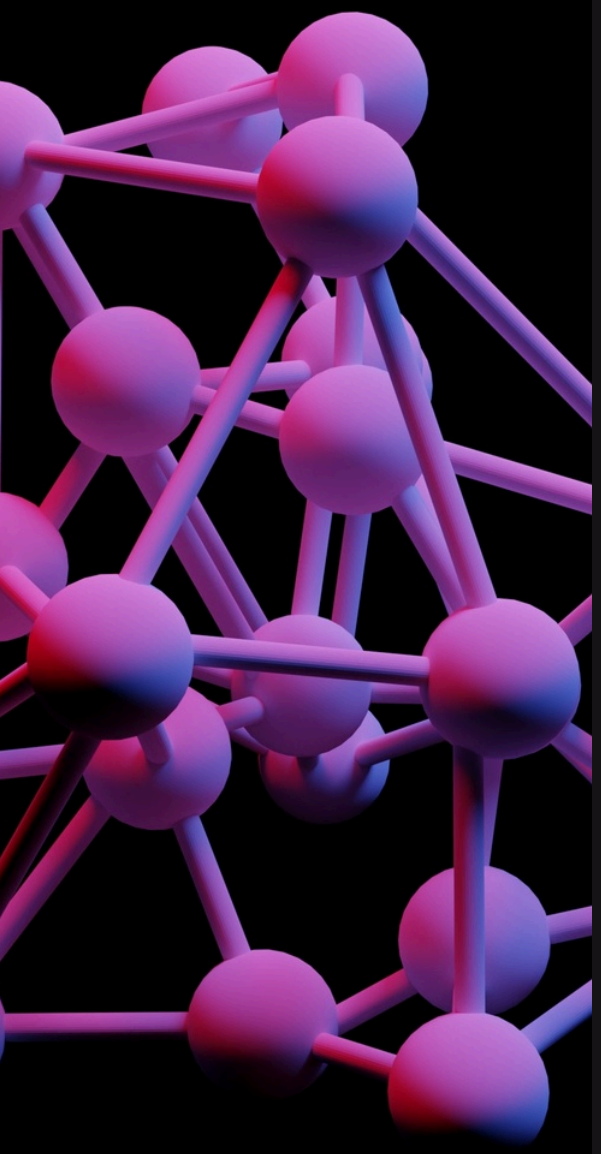
Choosing the Right LLM - Factors

5. Is It Easy to Maintain?

- Does the LLM have good support and clear documentation?

Maintenance is often overlooked. Some LLMs need more updates or come with limited documentation, which could make things harder in the longer run.

Analytics Vidhya

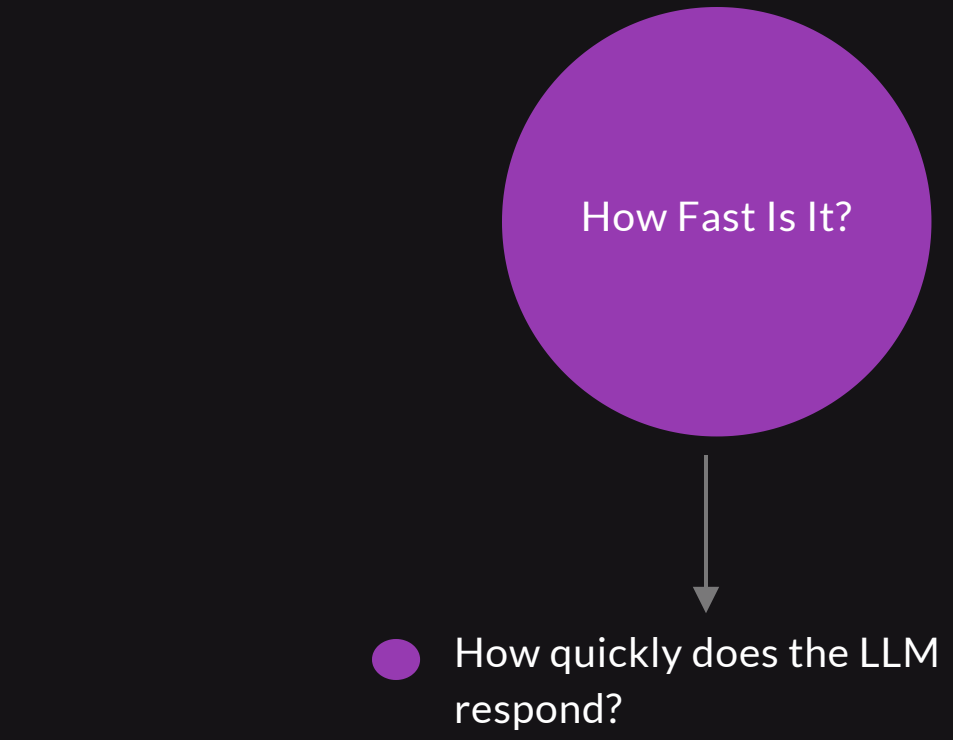


Maintenance Comparison - LLMs

LLM	Maintenance Level	Documentation Quality
LLM 1	Low (Easy)	Excellent
LLM 2	Medium (Moderate)	Limited
LLM 3	High (Difficult)	Inadequate

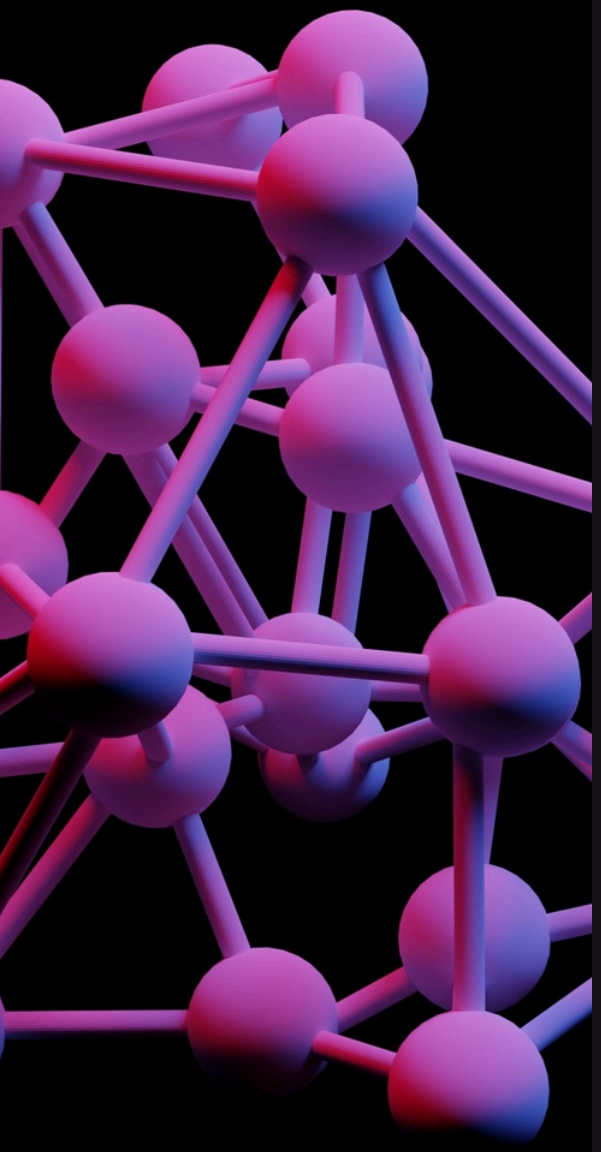
Choosing the Right LLM - Factors

6. How Fast Is It? (Latency)



Latency is the time taken by the LLM to respond. Speed is important for some applications (like customer service), while for others, it might not be a big deal.

Analytics Vidhya

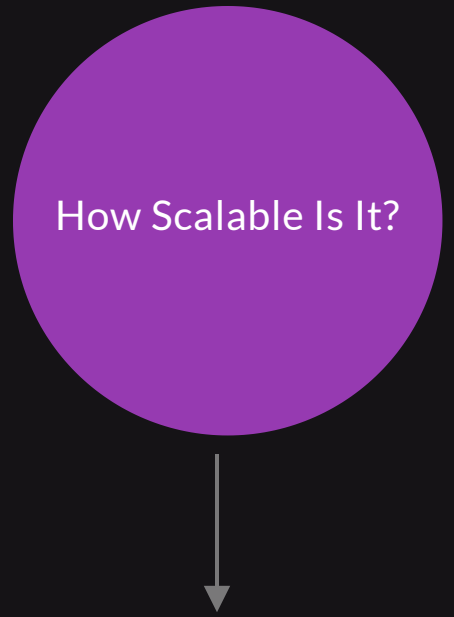


Latency Comparison - LLMs

LLM	Response Time	Can It Be Optimized?
LLM 1	100ms	Yes (80ms)
LLM 2	300ms	Yes (250ms)
LLM 3	200ms	Yes (150ms)

Choosing the Right LLM - Factors

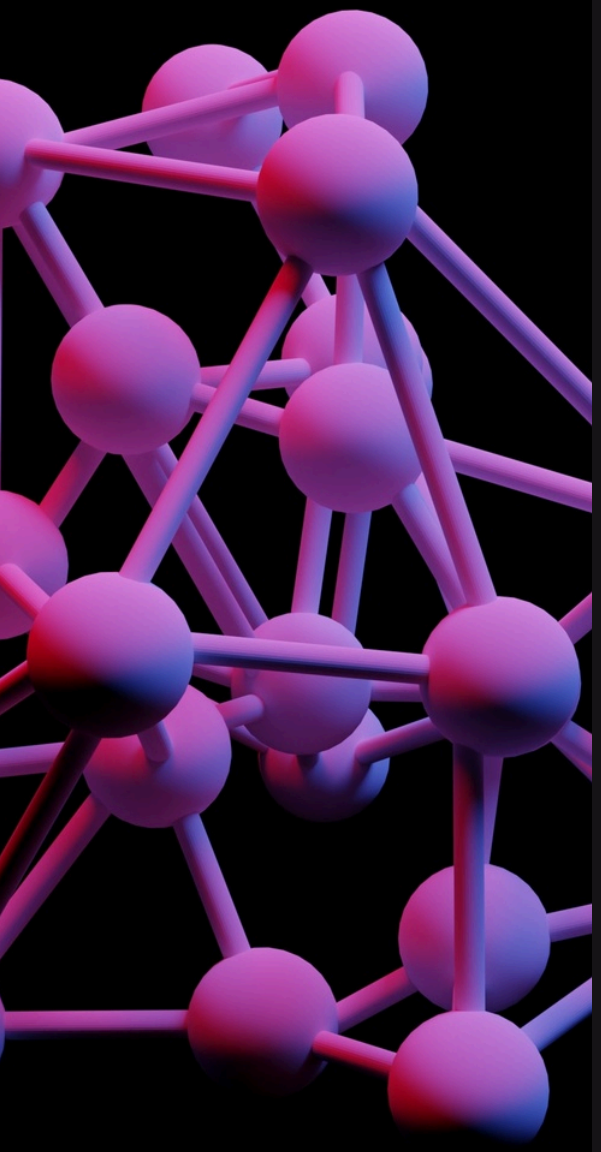
7. Can It Scale?



- Can it scale up to handle more users or data?

LLM needs to handle multiple people or lots of data simultaneously.

Analytics Vidhya

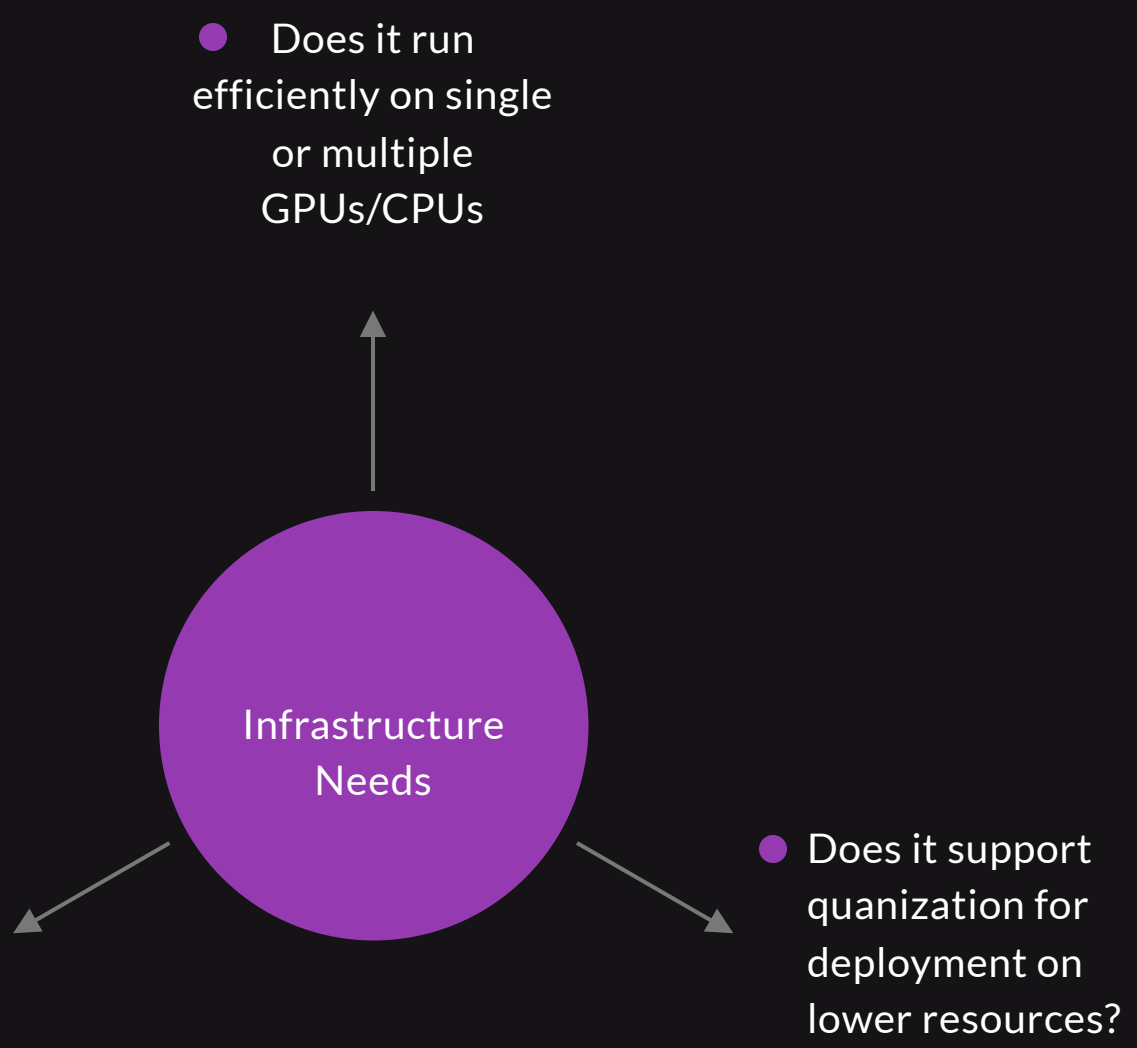


Scalability Comparison - LLMs

LLM	Max Users	Scalability Level
LLM 1	1,000	High
LLM 2	500	Medium
LLM 3	1,000	High

Choosing the Right LLM - Factors

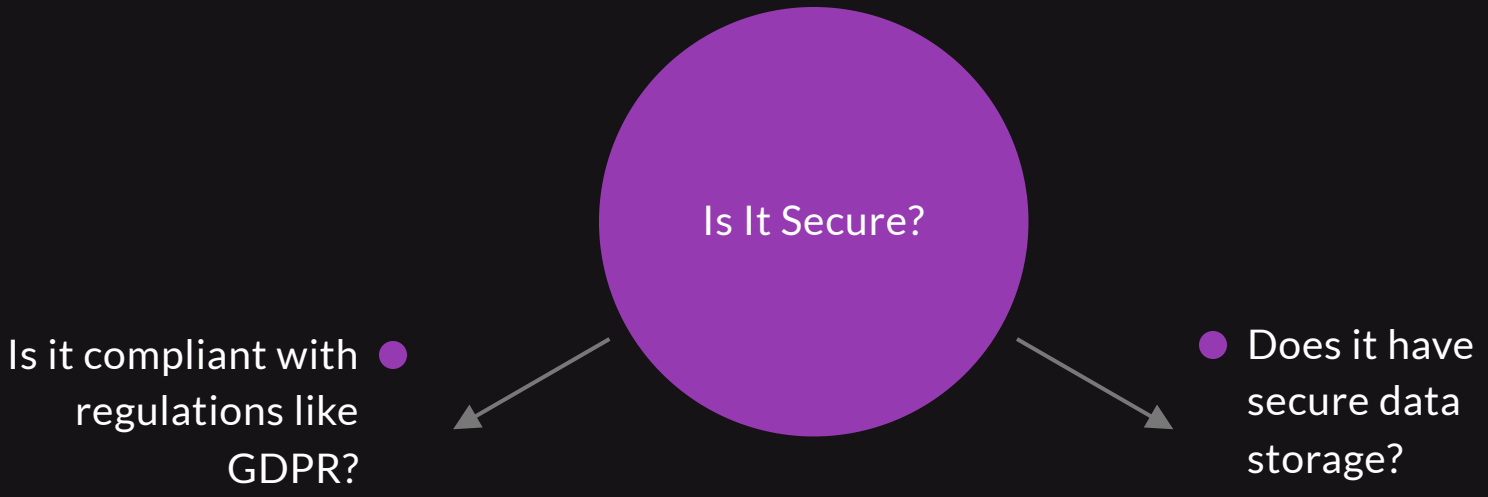
8. Infrastructure Needs



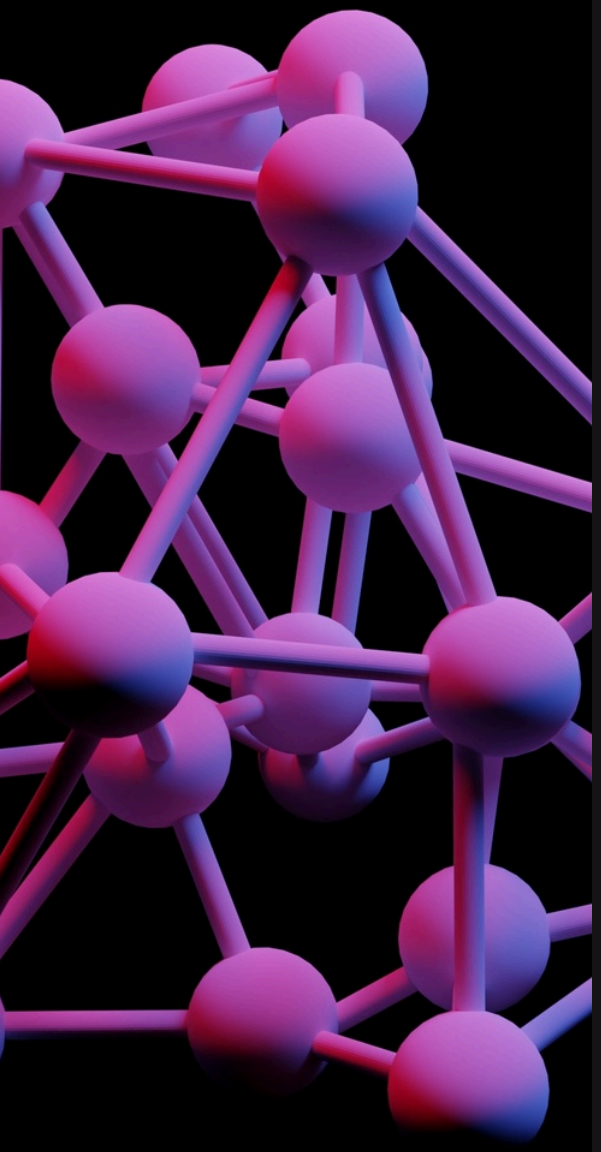
Analytics Vidhya

Choosing the Right LLM - Factors

9. Is It Secure?



Analytics Vidhya



Security Comparison - LLMs

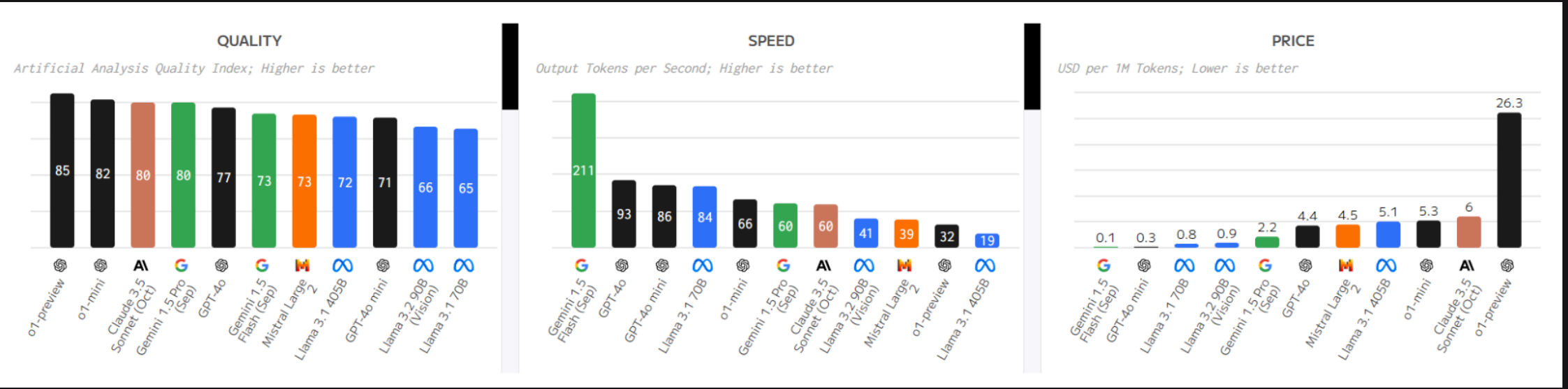
LLM	Security Features	Compliance to Regulations
LLM 1	High	Yes
LLM 2	Medium	No
LLM 3	Low	Yes

Choosing the Right LLM - Factors

10. What Kind of Support Is Available?

Analytics Vidhya

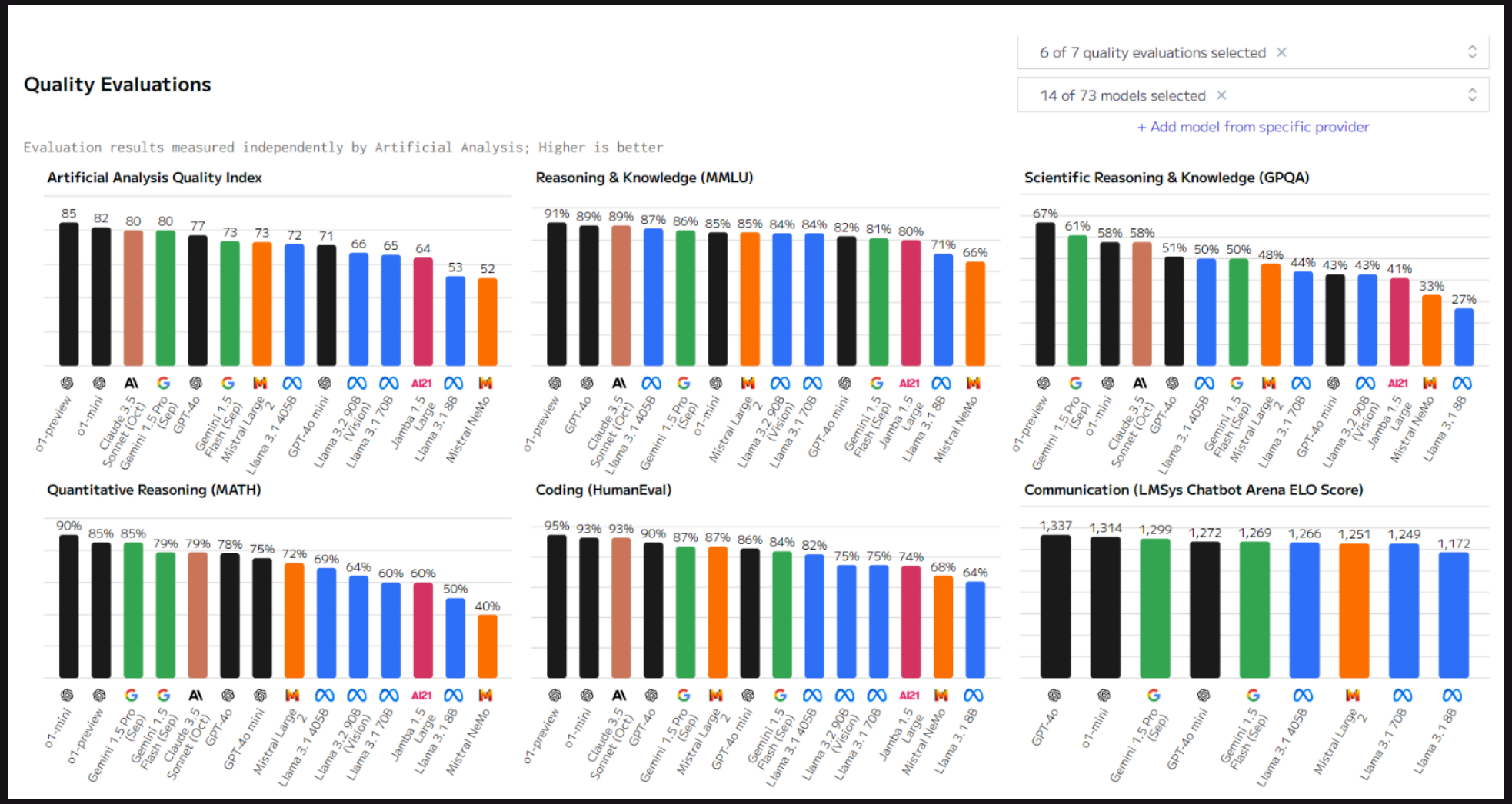
Choosing the Right LLM - Evaluation Benchmark



Source: https://artificialanalysis.ai

Analytics Vidhya

Choosing the Right LLM - Evaluation Benchmark



Source: https://artificialanalysis.ai

Analytics Vidhya

Thank You

Analytics Vidhya