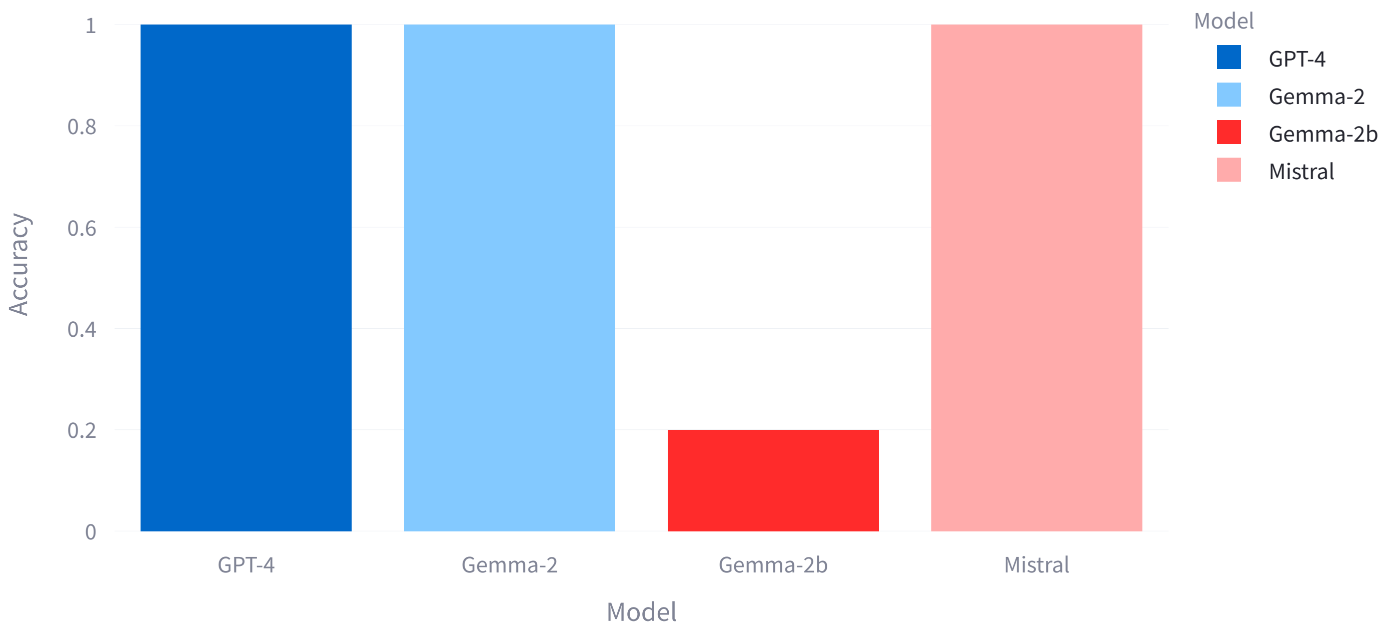
LLM Performance Evaluation

Performance Metrics

	model	accuracy	latency
0	GPT-4	1	1.0553
1	Gemma-2	1	4.76
2	Gemma-2b	0.2	1.2069
3	Mistral	1	1.7759

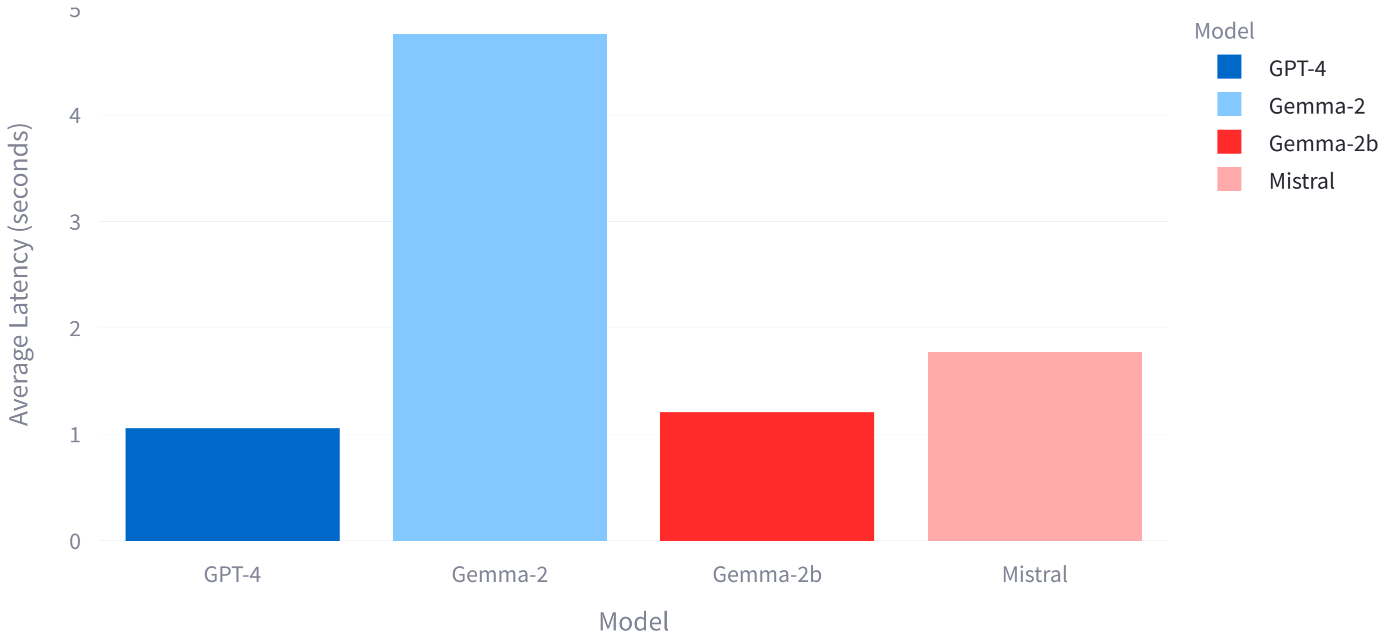
Accuracy Comparison

Model Accuracy Comparison



Latency Comparison

Model Latency Comparison



Detailed Results

	model	message	expected	predicted
0	GPT-4	I was charged twice for my last subscription payment.	Billing	Billing
1	GPT-4	The app keeps crashing when I try to open it.	Technical Issue	Technical
2	GPT-4	What are the differences between your Basic and Premium plans?	Product Inquiry	Product I
3	GPT-4	I've been waiting for support for 3 days and no one has responded!	Complaint	Complain
4	GPT-4	My account password isn't working after the recent update.	Technical Issue	Technical
5	Mistral	I was charged twice for my last subscription payment.	Billing	Billing
6	Mistral	The app keeps crashing when I try to open it.	Technical Issue	Technical
7	Mistral	What are the differences between your Basic and Premium plans?	Product Inquiry	Product I
8	Mistral	I've been waiting for support for 3 days and no one has responded!	Complaint	Complain
9	Mistral	My account password isn't working after the recent update.	Technical Issue	Technical

Mistral: 7B parameter
(it prioritizes speed over accuracy)
Gemma:2b: 2B parameter
(parameter efficiency might not be optimal)
Gemma2: 9B parameter
Gpt-4: hundreds of trillions parameter

parameter count alone doesn't determine performance
- model architecture, training approach, and optimization techniques play crucial roles in the final performance characteristics

Accuracy Comparison:

GPT-4 has the highest (about 0.8)
Gemma-2 and Gemma-2b has same performance (about 0.6)
Mistral is the lowest (about 0.4)

Latency Comparison:

GPT-4 has the highest (about 4-5 seconds)
Gemma-2 and Gemma-2b has mid-level latency (about 2-3 seconds)
Mistral has the lowest (about 1-2 seconds)

mistrali huggingfaceden,
gemmayı googledan bir dene

GPT-4 produces the most accurate results but at the cost of higher processing time. It's ideal for applications where accuracy is crucial and there are no strict time constraints.

Gemma models (2 and 2b) show moderate performance in terms of both accuracy and latency. This makes them suitable for applications requiring a balanced performance/speed ratio.

Mistral is the fastest responding model but has lower accuracy. It can be preferred for real-time applications where high accuracy is not critical.

For high accuracy requirements: GPT-4
For balanced performance: Gemma models
For low latency requirements: Mistral