

LLM API Pricing Awareness

Instructor

Dipanjan Sarkar

Head of Community & Principal AI Scientist at Analytics Vidhya

Google Developer Expert - ML & Cloud Champion Innovator

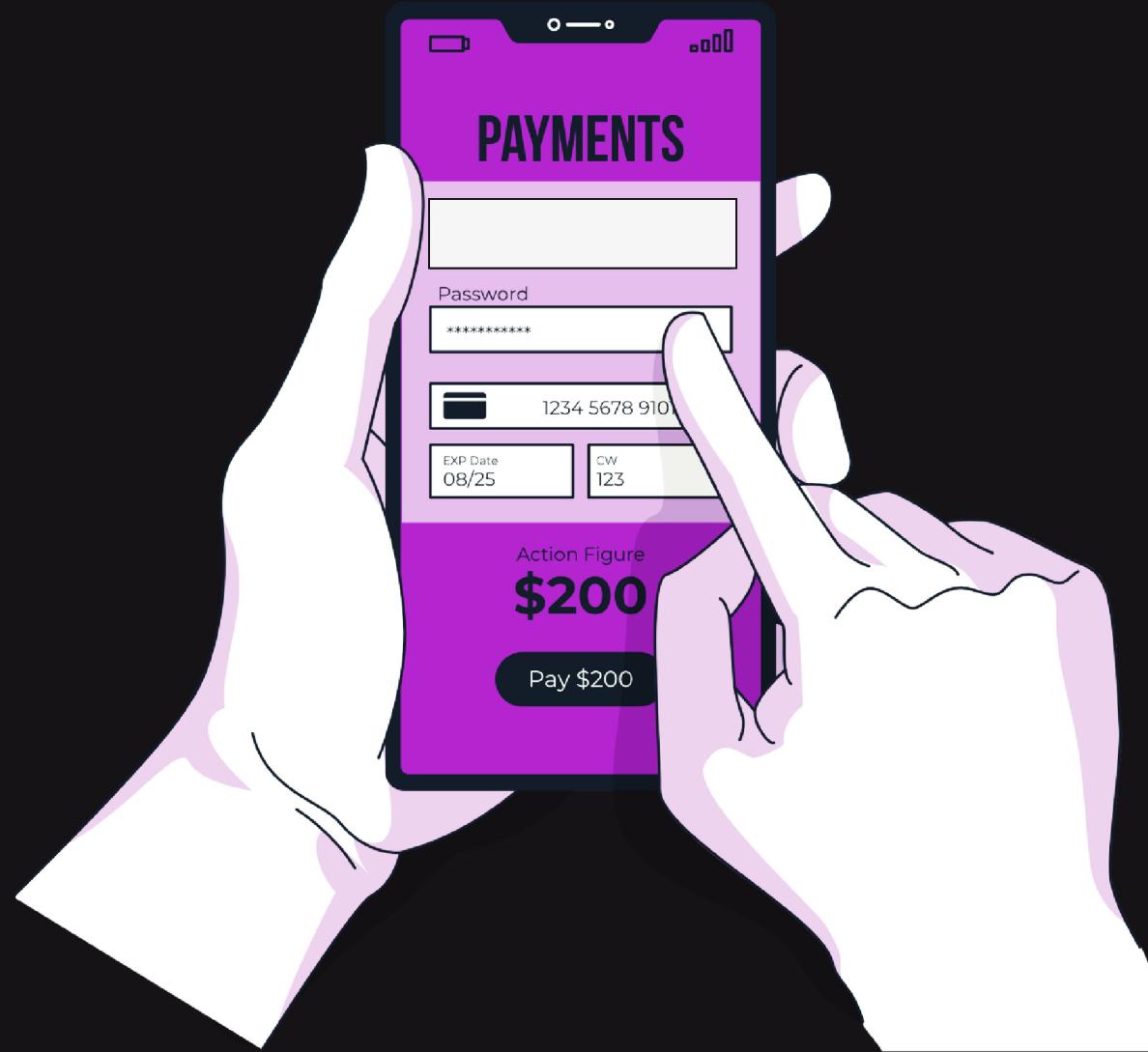
Published Author



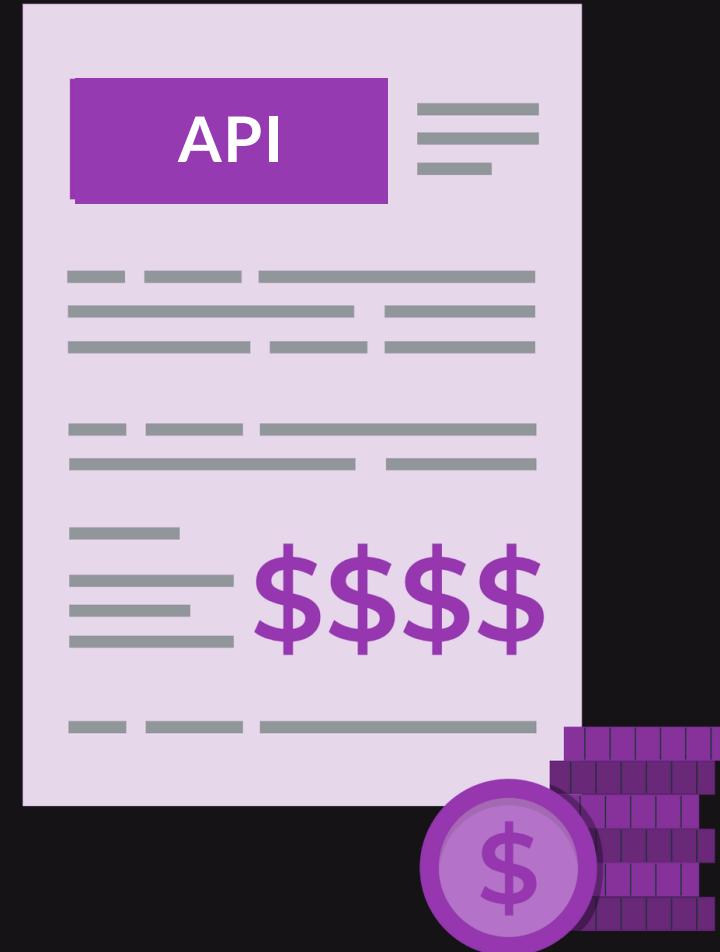
66

LLMs are usually not free and have a cost

- Commercial LLMs typically charge based on the number of tokens or requests made.
- Open-source LLMs typically incur costs for maintaining GPU infrastructure or based on tokens or requests if using a managed platform.



*It's Good to be aware of API
costs to prevent unnecessary
surprises when using them...*



Commercial LLM APIs



OpenAI GPT API



Google Gemini API

OpenAI API Pricing (As of Late 2024)

Model	Input Pricing (\$ per 1K tokens)	Output Pricing (\$ per 1K tokens)
gpt-4o	0.00250	0.0100
gpt-4o-mini	0.00015	0.0006
gpt-3.5-turbo	0.00050	0.0015
o1-preview	0.01500	0.0600
o1-mini	0.00300	0.0120
chatgpt-4o-latest	0.00500	0.0150

Source: <https://openai.com/api/pricing/> for the latest prices

Google Gemini API Pricing (As of Late 2024)

Model	Rate Limits	Input Pricing (\$ per 1K tokens)	Input Pricing (\$ per 1K tokens)
Gemini 1.5 Flash Free	15 RPM, 1M TPM, 1,500 RPD	Free of Charge	Free of Charge
Gemini 1.5 Flash Paid (≤ 128k tokens)	2,000 RPM, 4M TPM	0.000075	0.00030
Gemini 1.5 Flash Paid <td>2,000 RPM, 4M TPM</td> <td>0.00015</td> <td>0.00060</td>	2,000 RPM, 4M TPM	0.00015	0.00060
Gemini 1.5 Pro Free	2 RPM, 32k TPM, 50 RPD	Free of Charge	Free of Charge
Gemini 1.5 Pro Paid (≤ 128k tokens)	1,000 RPM, 4M TPM	0.00125	0.00500
Gemini 1.5 Pro Paid <td>1,000 RPM, 4M TPM</td> <td>0.00250</td> <td>0.01000</td>	1,000 RPM, 4M TPM	0.00250	0.01000

Source: <https://ai.google.dev/pricing> for the latest prices

Open-Source LLM API Platforms



Groq



HuggingFace

HuggingFace APIs

- The **Serverless Inference API** provides fast, free access to thousands of ML models for various tasks, ideal for prototyping and experimentation:
 - **Text Generation:** High-quality responses with large language models.
 - **Image Generation:** Create custom images, including LoRAs for unique styles.
 - **Document Embeddings:** Build advanced search and retrieval systems.
 - **Classical AI Tasks:** Models for classification, speech recognition, and more.



HuggingFace APIs

⚡ Fast and Free to Get Started:

- The Inference API is free with higher rate limits for PRO users.
- For production needs, explore [Inference Endpoints](#) for dedicated resources, autoscaling, advanced security features, and more.



Groq Cloud Free Tier Limits (Late 2024)

ID	Requests per Minute	Requests per Day	Tokens per Minute	Tokens per Day
gemma-7b-it	30	14400	15,000	500,000
gemma2-9b-it	30	14400	15,000	500,000
llama-3.1-70b-versatile	30	14400	18,000	500,000
llama-3.1-8b-instant	30	14400	20,000	500,000
llama-3.2-11b-text-preview	30	7000	7,000	500,000
llama-3.2-11b-vision-preview	30	7000	7,000	500,000
llama-3.2-1b-preview	30	7000	7,000	500,000
llama-3.2-2b-preview	30	7000	7,000	500,000
llama-3.2-90b-text-preview	30	7000	7,000	500,000
llama-3.2-90b-vision-preview	30	7000	7,000	500,000
llama-guard-3-8b	15	3500	7,000	250,000
llama3-70b-8192	30	14400	15,000	500,000
llama3-8b-8192	30	14400	6,000	500,000
llama3-groq-70b-8192-tool-preview	30	14400	30,000	500,000
llama3-groq-8b-8192-tool-preview	30	14400	30,000	500,000
llava-v1.5-7b-4096-preview	30	14400	15,000	500,000
mixtral-8x7b-32768	30	14400	5,000	500,000

Check out [pricing details here](#) for free API and [here for paid API](#)

Use Claude to Improve Existing Prompts Faster

The screenshot shows the Anthropic Workbench interface with the following details:

Left Panel (Prompt):

- Title:** Healthcare AI Report Structure
- Text:** highlighting key information.
2. Content Guidelines:
 - Focus on the most important and impactful points from the report.
 - Use clear, concise language suitable for executive-level readers.
 - Avoid unnecessary details, technical jargon, or repetition.
3. Formatting:
 - Use appropriate headings for each section.
 - Present bullet points using "-" symbols.
Before writing your final summary, wrap your analysis in `<analysis>` tags. In your analysis process:
 1. For each section (Applications, Benefits and Challenges, Future Scope):
 - a. Extract and quote 2-3 key sentences or phrases from the report.
 - b. Formulate 2-3 options for the one-line overview.
 - c. List 4-5 potential bullet points.
 2. Select the best one-line overview and up to three bullet points for each section, ensuring they complement each other without repeating information.
 3. Count the total lines of your draft summary to ensure it stays within the 10-line limit. Adjust if necessary.
After your analysis, present your final summary using `<summary>` tags. Ensure that your summary maintains a consistent structure across all three sections, regardless of the content of the original report.

Here's an example of the desired output structure (using generic content):

`<summary>`
Applications:
One-line overview of applications.
 - Key application point 1
 - Key application point 2
 - Key application point 3

Thank You
