OpenRouter

OpenRouter Kya Hai?

OpenRouter ek platform hai jo developers ko different AI models (LLMs - Large Language Models) ko ek hi API ke zariye access karne ka easy way deta hai.

Yani agar aapko:

Kabhi OpenAI ka GPT model use karna hai,

Kabhi Anthropic, Mistral, ya LLaMA jaise open-source models try karne hain...

...to aapko har baar alag code likhne ki zarurat nahi. Bas API key aur model name change karo — code same rahega!

Use Karne Ka Faida

Single API → Bohat saare models

Cost effective → Ye khud best/cheap/available model choose karke route kar deta hai

Fast switching → Code na badlay, model change ho jaye

Tool Calling Support → AI ke through external tools (jaise weather, calculator, etc.) use kiye ja sakte hain

User Interface aur Developer Tools

OpenRouter 2 cheezein deta hai:

Chatroom Interface
Jahan aap alag alag models ko try kar sakte ho (jaise playground).

Developer API
Jisse aap apni applications me AI ka power daal sakte ho (Next.js, Python, TypeScript etc.).

OpenAI Compatible

Agar aap pehle se OpenAI ka chat completion API use kar rahe hain, to sirf:

API key

aur base URL change karke
OpenRouter ke saath same code chala sakte ho. Super easy!

Kya OpenRouter Models Host Karta Hai?

Nahi, OpenRouter khud models host nahi karta — ye proxy ka kaam karta hai. Yani aap ka request OpenRouter ko jaata hai, aur wo usay best provider (OpenAI, Anthropic etc.) tak bhejta hai — bina aapko infrastructure ka jhanjhat diye.

User Interface (UI) Features

OpenRouter sirf backend API nahi hai — is ka front-end UI (user interface) bhi kaafi powerful hai:

Key Features:

Chatroom:
Aap https://openrouter.ai/chat par jakar direct multiple AI models se baat kar sakte ho. Jaise:

GPT-4

Claude (Anthropic)

Mistral

LLaMA, etc.

Interactive Testing:
Yahan aap AI responses ko test kar sakte ho without writing any code.

Token Management:
Aap dekh sakte ho:

Kitne input/output tokens use hue

Kitni cost lagi

Kis model ne kitna response diya

Account Management:
Aap apni usage history, credits, aur billing info track kar sakte ho.

API (Developers ke liye)

OpenRouter sirf chat karne ke liye nahi — ye developers ke liye bhi full support deta hai.

✅ API Key Features:
Standardized endpoint: Ek hi endpoint se multiple models use kar lo.

OpenAI-Compatible:
Agar aap already OpenAI ka code likh chuke ho (like using openai.ChatCompletion.create()), to aap sirf:

base URL

aur API key change karke
OpenRouter ke through same code chala sakte ho.

Integration Support:
Aap TypeScript, Python, Node.js jaise environments me asaani se integrate kar sakte ho.

OpenRouter Ka Kaam Karne Ka Tareeqa

Hosting kaise hoti hai?

OpenRouter khud models host nahi karta

Ye ek proxy ki tarah kaam karta hai — yani aapka request leta hai aur best model provider tak usay route karta hai.

Ye dekhta hai:

Konsa model available hai

Konsa sasta hai

Konsa fast hai

Is tarah aapko best performance milti hai, bina har model ke API se individually deal kiye.

OpenRouter ka OpenAI Chat Completion API ke Saath Compatibility

Kya Matlab?
OpenRouter ka API system bilkul OpenAI Chat API jaisa hai. Yani agar aapne pehle https://api.openai.com/v1/chat/completions use kiya hai, to aap OpenRouter pe migrate karne ke liye sirf:

✅ API key change karo
✅ Base URL change karo → https://openrouter.ai/api/v1

Baqi code jaisa ka taisa chalega!

Example Parameters:
model

messages

temperature

max_tokens

✅ OpenAI SDKs (Python, Node.js etc.) ke saath bhi ye API smoothly kaam karta hai.

Function Calling Support (Tool Calling)

Kya Hai Function Calling?

Function calling ka matlab hai ke AI suggest kar sake ke kisi external tool ko kab aur kaise use karna hai. Jaise:

"Mujhe Lahore ka weather batao" → AI suggest kare ke getWeather(city: "Lahore") function ko call karo.

OpenRouter me Kaise Kaam Karta Hai?

Aap JSON schema ke through tool define karte ho.

AI model suggest karta hai ke konsi function call kare.

Developer (aap) decide karte ho ke us suggestion ke mutabiq kya karna hai.

Konsay Models Support Karte Hain?

✅ GPT-4, GPT-4o (OpenAI)

✅ Claude (Anthropic)

❌ Sabhi models nahi karte — OpenRouter automatically unhi models tak request bhejta hai jo function calling support karte hain.

Proxy vs Hosting: OpenRouter Kya Hai?

Kya OpenRouter khud models host karta hai?

Nahi! OpenRouter ek proxy hai.

Iska matlab:

Aapka request OpenRouter pe jaata hai

OpenRouter best provider (OpenAI, TogetherAI, AWS, Fireworks, etc.) ko choose karta hai

Wahan se response aapko milta hai

Kya Fayda Hai?

✅ Aapko infrastructure maintain nahi karna

✅ Har model ke liye API key manage nahi karni

✅ Sasta, fast aur reliable model milta hai — because of routing

Additional Features

Playground: Test karne ke liye GUI chatroom

Pay-per-use pricing: Sirf jitne tokens use karte ho, utna pay karo

Free models bhi available hote hain

⚡ Streaming support: Live response jaise OpenAI ke streaming mode

Prompt logging (optional): Debugging ke liye prompts track kar sakte ho

Comparison aur User Experience

Feature	OpenRouter ka Benefit
Model Variety	200+ models from OpenAI, Meta, Mistral etc.
Code Compatibility	Same as OpenAI APIs
Cost Optimization	Cheapest and fastest provider selected
Free Access	Kuch models bilkul free
Centralized Management	1 API key → multiple models
Credits Required	Free tier jaldi khatam ho sakta hai

Free Models ka kya matlab hai?

OpenRouter par kuch models bilkul free hain – matlab unka use karne par aap se koi token charges nahi liya jata.

Lekin:

Aapko account banana parta hai

Limitations hoti hain, jaise:

20 requests per minute

200 requests per day

Yeh free models usually IDs mein :free likha hota hai (e.g., openchat-3.5:free)

March 2025 ke mutabiq kya haal hai?

50 free models available hain

Un mein se 6 models ka context window 1 million+ tokens ka hai (yani ek baar mein bohat zyada data samajh sakte hain)

Yeh baat confirm hui hai recent X (Twitter) posts se:

Official OpenRouter ne kaha: “50 free models including 6 with 1M+ context windows!”

Doosre users ne bhi yehi bataya

Free Models kis kaam ke liye ache hain?
Agar aap personal projects, testing, ya light usage ke liye AI use kar rahe ho, to yeh perfect hain

Lekin agar aapka use zyada heavy hai, to phir rate limits aapko restrict kar sakti hain

Kuch important points:
Kuch models jo pehle free the (e.g., openchat-3.5) ab ho sakta hai paid ho gaye hon – iska status kabhi kabhi change hota rehta hai

Gemini models bhi ek waqt tak free the, with same 20/min and 200/day limits

Gemini Compatibility

Gemini models (especially Google ke) OpenRouter pe available to hain, lekin woh fully OpenAI-style chat/completions endpoint jaisa behave nahin karte — unka format slightly different hota hai, aur kuch limitations bhi hoti hain.

"Agar Gemini OpenAI-style code accept nahi karta, to phir OpenRouter ka 'unified code' claim kis had tak sach hai?"

✅ **Short Answer (Truthfully):**
OpenRouter mostly unified hai — lekin 100% nahi.

Most models (OpenChat, Mixtral, Claude, Mistral, LLaMA, etc.) follow OpenAI-style chat/completions API.

Kuch models, jaise Google Gemini, Anthropic Claude (older versions), aur kuch specialized APIs — unka behavior thoda different hota hai.

Is wajah se, "unified" code ka matlab yeh nahi ke har model bilkul same way se kaam karega — balke:

Aap same endpoint + same API key + same headers use kar sakte hain — magar prompt formatting model-specific hoti hai.

✅ Example: What's Unified
Feature Unified (✅)
Same API URL (/v1/chat/completions) ✅
Same HTTP method (POST) ✅
Same Auth Header (Bearer API key) ✅
Same base code for calling ✅

What's Not Unified (Model Specific)
Feature Varies Model to Model
messages[] vs prompt ❌

Support for streaming ✗
System prompt handling ✗
Token limit & behavior ✗
Role awareness (user/assistant) ✗