Coding

Dal prompt alla produzione:
prova, confronta, valida, integra con GitHub Models

#4<DEV>

Massimo Bonanni

AGIC

CONSOFT
INFORMATICA

present
exprivia group

KEY2

Did you know that you can play, test, and use tons of generative models for free?

No!?!?!

Then follow me to discover GitHub Models!!!

GitHub Models is a suite of developer tools that take you **from AI idea to ship**, including a **model catalog**, **prompt management**, and **quantitative evaluations**.

Public Preview

The Four Phases of GitHub Models

Prototype

- Access the Playground directly from GitHub, no API keys needed.
- Experiment with multiple LLMs (OpenAI, Anthropic, Mistral, etc.).
- Adjust parameters (temperature, max tokens, etc.) in real time.
- Quickly iterate and refine your prompt design.

Compare

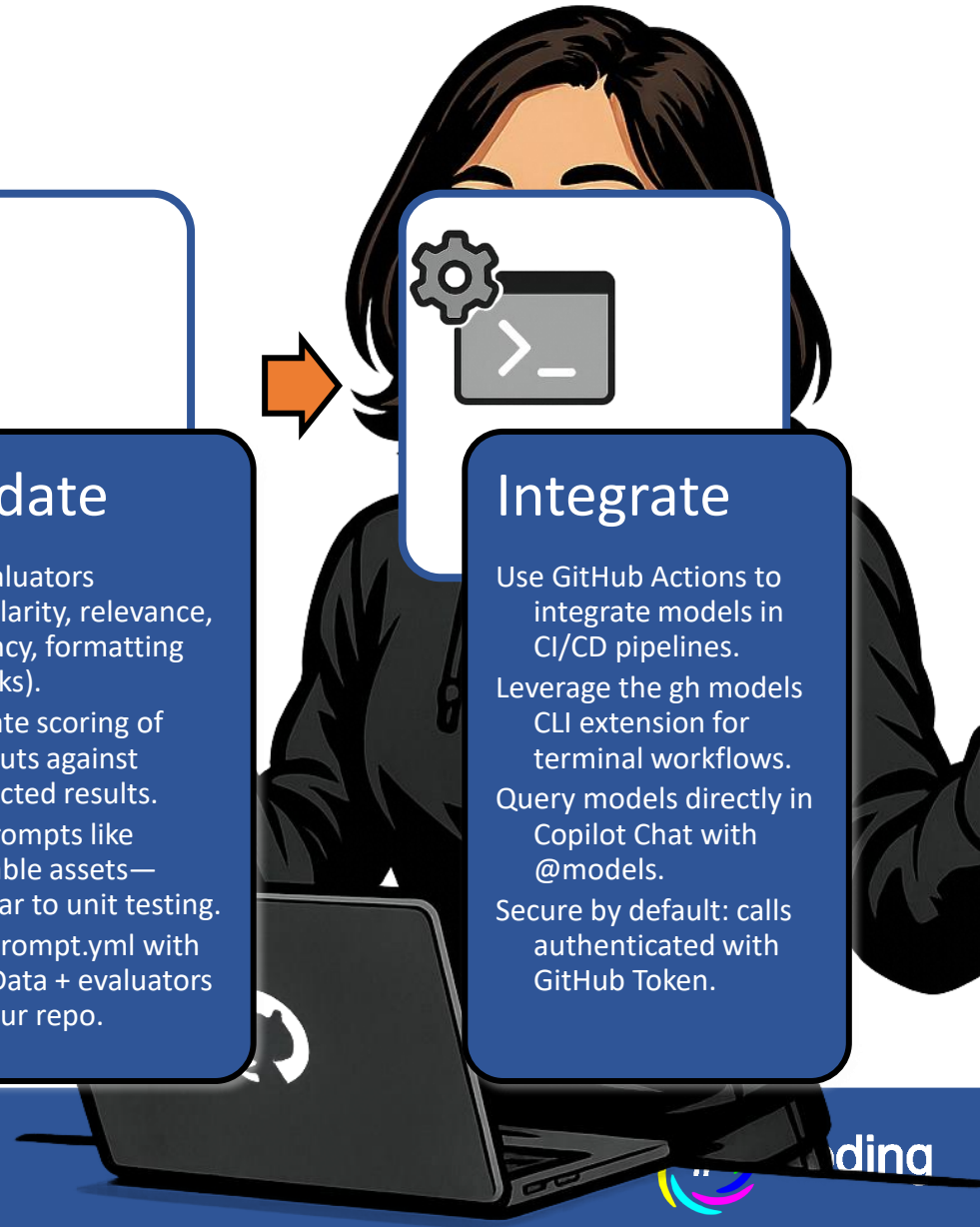
- Run the same prompt across multiple models side-by-side.
- Use test datasets for reproducible input/output comparisons.
- Measure differences in latency, token usage, and style.
- Make evidence-based choices on the best model/prompt combo.

Validate

- Add evaluators (similarity, relevance, fluency, formatting checks).
- Automate scoring of outputs against expected results.
- Treat prompts like testable assets—similar to unit testing.
- Store .prompt.yml with testData + evaluators in your repo.

Integrate

- Use GitHub Actions to integrate models in CI/CD pipelines.
- Leverage the gh models CLI extension for terminal workflows.
- Query models directly in Copilot Chat with @models.
- Secure by default: calls authenticated with GitHub Token.



DEMOS

GitHub Models Pricing – Free use

All GitHub accounts have rate-limited access to GitHub Models at no cost.

The playground and free API usage are rate limited by requests per minute, requests per day, tokens per request, and concurrent requests.

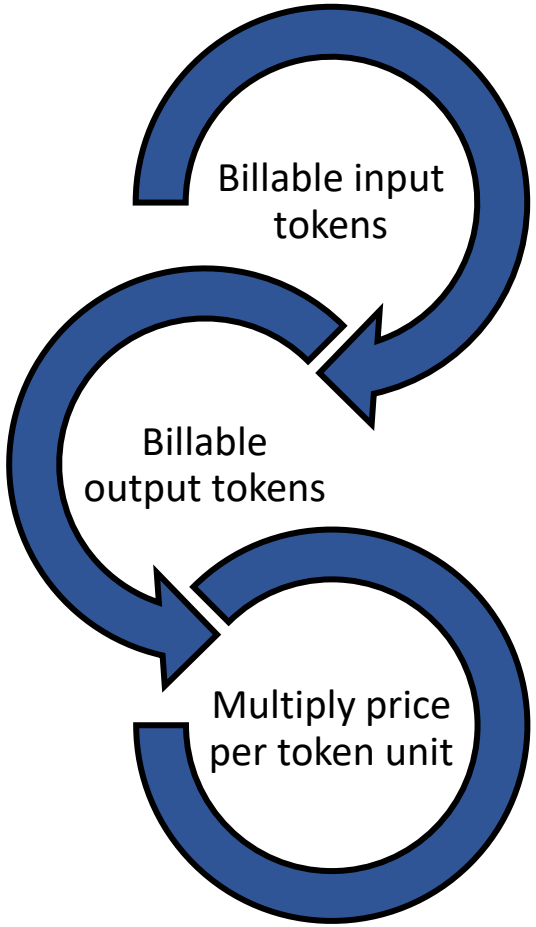
If you get rate limited, you will need to wait for the rate limit that you hit to reset before you can make more requests.

Rate limit tier	Rate limits	Copilot Free	Copilot Pro	Copilot Business	Copilot Enterprise
Low	Requests per minute	15	15	15	20
	Requests per day	150	150	300	450
	Tokens per request	8000 in, 4000 out	8000 in, 4000 out	8000 in, 4000 out	8000 in, 8000 out
	Concurrent requests	5	5	5	8

GitHub Models Pricing - DeepSeek-R1

Model name	Input multiplier	Cached input multiplier	Output multiplier
DeepSeek-R1	0.135	N/A	0.54

Input tokens = 1.000.000
Output tokens = 1.000.000



$$\boldsymbol{Input\ Tokens} \times \boldsymbol{Input\ Multiplier} = 1000000 \times 0.135 = 135000$$

$$\boldsymbol{Output\ Tokens} \times \boldsymbol{Output\ Multiplier} = 1000000 \times 0.54 = 540000$$

$$\boldsymbol{Total\ Billable\ Tokens} \times \boldsymbol{Token\ Unit\ Price} = 675000 \times 0.00001 = 6.75\$$$

GitHub Models Pricing

🛑 Usage stops if no payment method and free quota exhausted.

🏢 **Enterprises/Orgs**: Billing disabled by default; must be enabled by enterprise admin.

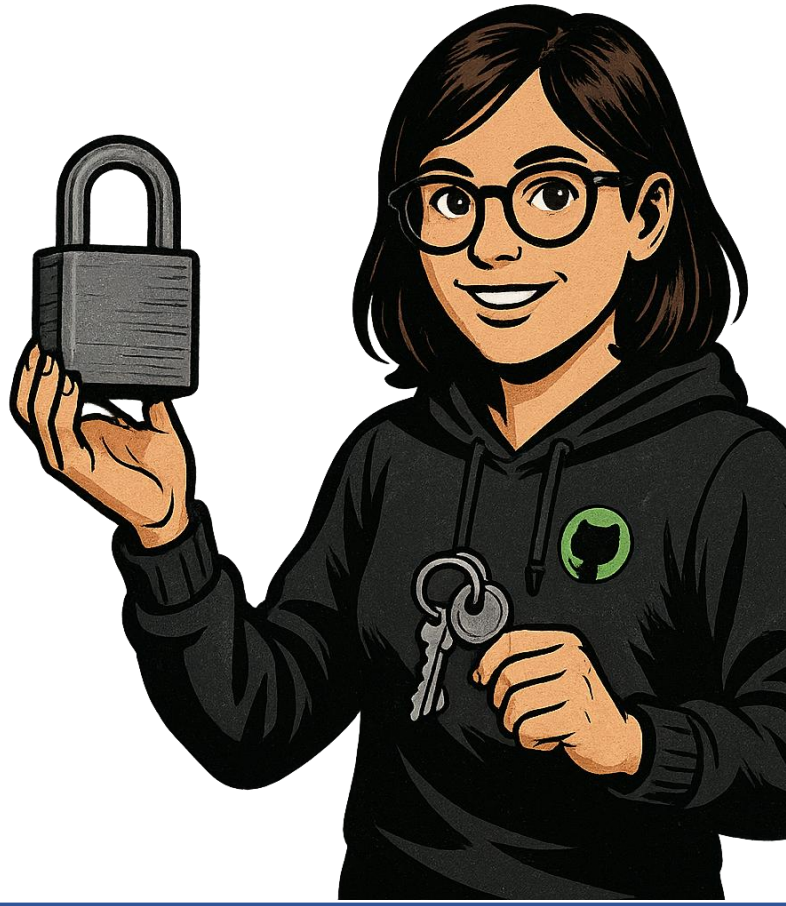
👤 **Personal Accounts**: Users choose to enable paid usage (unless managed by an enterprise).

💰 **Budget Controls**: Enterprises, orgs, and personal accounts can set spending limits.

🔑 **BYO API Keys**: Using custom models with your own provider key → billed externally (e.g., Azure), not via GitHub.

Security

- **Built-in Governance:** Enterprises/orgs can control which models are available and enforce usage policies.
- **No External API Keys:** Access is authenticated via **GitHub Token** – secure by default, no secrets management needed.
- **Prompts as Code:** .prompt.yml files live in your repo → version control, reviews, and full audit trail.
- **Data Privacy:** Requests stay within GitHub's infrastructure; your code and prompts aren't shared with providers outside the GitHub framework.
- **Visibility & Auditability:** Centralized monitoring of usage, costs, and compliance across enterprise accounts.



Takeways

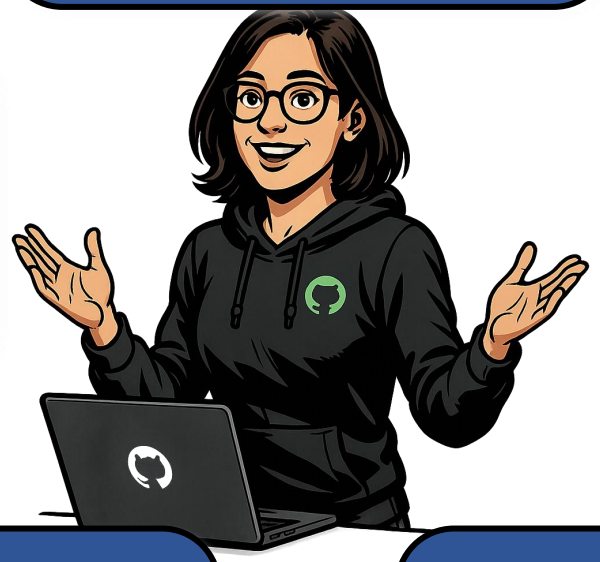
Prototype Fast
Use the Playground to experiment with prompts quickly, no setup needed.

Compare & Choose
Run side-by-side comparisons to find the right model + prompt combo.

Validate Quality
Treat prompts like code: test them with evaluators, version them in your repo.

Integrate Seamlessly
Bring AI into your workflow with GitHub Actions, CLI, or Copilot Chat.

Secure & Governed
Enterprise-grade governance and privacy controls are built in by default.



Massimo Bonanni

Sr Technical Trainer @ Microsoft

massimo.bonanni@microsoft.com

aka.ms/maxlinkedin