

# Vibe Coding



## Jaru Nartboon

Lead Solution Architect, Alibaba Cloud Thailand

Immerse yourself fully in the atmosphere & simply forget about the existence of code.

**Andrey Karpathy**  
**Co-founder of OpenAI**

- ♥ **Flow over friction** – Ride the wave, don't fight it.
- ♥ **Iteration over perfection** – Perfection is obsolete if you can always reroll.
- ♥ **Augmentation over automation** – AI is a collaborator, not a replacement.
- ♥ **Product thinking over code crafting** – What matters is what you build, not how you write it.
- ♥ **Rerolling over debugging** – If fixing takes too long, regenerate.
- ♥ **Human taste over technical constraints** – The best tech serves great taste, not the other way around.

<https://vibemanifesto.org/>

# Model Architecture – Qwen3 : Mixture of Expert (MoE)

Launched 2025/04/29



Qwen3-235B-A22B		Dense	MoE
Active Param	All parameters (100%)		
Inference Speed	Slower for large sizes		Much faster for same total size
VRAM Req	Equal to total parameters		Equal to <b>total</b> parameters
Specialization	Generalist (all weights learn all)		Specialist (experts learn specific tasks)

# Model Architecture – Qwen3 : Mixture of Expert (MoE)

	Qwen3-235B-A22B MoE	Qwen3-32B Dense	OpenAI-o1 2024-12-17	Deepseek-R1	Grok 3 Beta Think	Gemini2.5-Pro	OpenAI-o3-mini Medium
ArenaHard	95.6	93.8	92.1	93.2	-	96.4	89.0
AIME'24	85.7	81.4	74.3	79.8	83.9	92.0	79.6
AIME'25	81.5	72.9	79.2	70.0	77.3	86.7	74.8
LiveCodeBench v5, 2024.10-2025.02	70.7	65.7	63.9	64.3	70.6	70.4	66.3
CodeForces Elo Rating	2056	1977	1891	2029	-	2001	2036
Aider Pass@2	61.8	50.2	61.7	56.9	53.3	72.9	53.8
LiveBench 2024-11-25	77.1	74.9	75.7	71.6	-	82.4	70.0
BFCL v3	70.8	70.3	67.8	56.9	-	62.9	64.6
MultilF 8 Languages	71.9	73.0	48.8	67.7	-	77.8	48.4

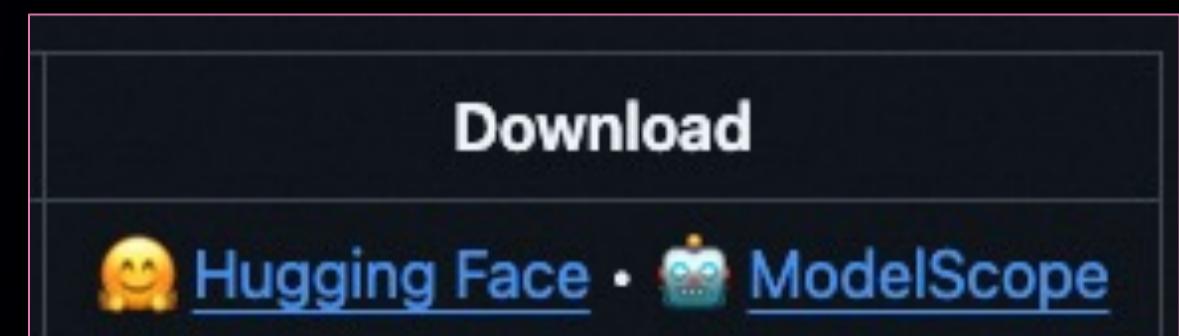
1. AIME 24/25: We sample 64 times for each query and report the average of the accuracy. AIME'25 consists of Part I and Part II, with a total of 30 questions.  
 2. Aider: We didn't activate the think mode of Qwen3 to balance efficiency and effectiveness.  
 3. BFCL: The Qwen3 models are evaluated using the FC format, while the baseline models are assessed using the highest scores obtained from either the FC or prompt formats.

- 1 **CodeForce:** Algorithmic logic/Optimization
- 2 **LiveBench:** Coding Proficiency
- 3 **LiveCodeBench:** General coding & Anti-leakage
- 4 **SWE Bench:** Real-world software engineering

# Agentic Code Model - Qwen3-Coder

Launched 2025/06/23

huybery@U-V76490MY-0133:~/coding-agent					
Benchmarks	Open Models Qwen3-Coder 480B-A35B-Instruct	Kimi-K2 Instruct	DeepSeek-V3 0324	Proprietary Models Claude Sonnet-4	OpenAI GPT-4.1
Agentic Coding					
Terminal-Bench	37.5	30.0	2.5	35.5	25.3
SWE-bench Verified w/ OpenHands, 500 turns	69.6	-	-	70.4	-
w/ OpenHands, 100 turns	67.0	65.4	38.8	68.0	48.6
w/ Private Scaffolding	-	65.8	-	72.7	63.8
SWE-bench Live	26.3	22.3	13.0	27.7	-
SWE-bench Multilingual	54.7	47.3	13.0	53.3	31.5
Multi-SWE-bench mini	25.8	19.8	7.5	24.8	-
Multi-SWE-bench flash	27.0	20.7	-	25.0	-
Aider-Polyglot	61.8	60.0	56.9	56.4	52.4
Spider2	31.1	25.2	12.8	31.1	16.5
Agentic Browser Use					
WebArena	49.9	47.4	40.0	51.1	44.3
Mind2Web	55.8	42.7	36.0	47.4	49.6
Agentic Tool Use					
BFCL-v3	68.7	65.2	64.7	73.3	62.9
TAU-Bench Retail	77.5	70.7	59.1	80.5	-
TAU-Bench Airline	60.0	53.5	40.0	60.0	-



model name	type	length
Qwen3-Coder-480B-A35B-Instruct	instruct	256k
Qwen3-Coder-480B-A35B-Instruct-FP8	instruct	256k
Qwen3-Coder-30B-A3B-Instruct	instruct	256k
Qwen3-Coder-30B-A3B-Instruct-FP8	instruct	256k

Alibaba Cloud  
Model Studio



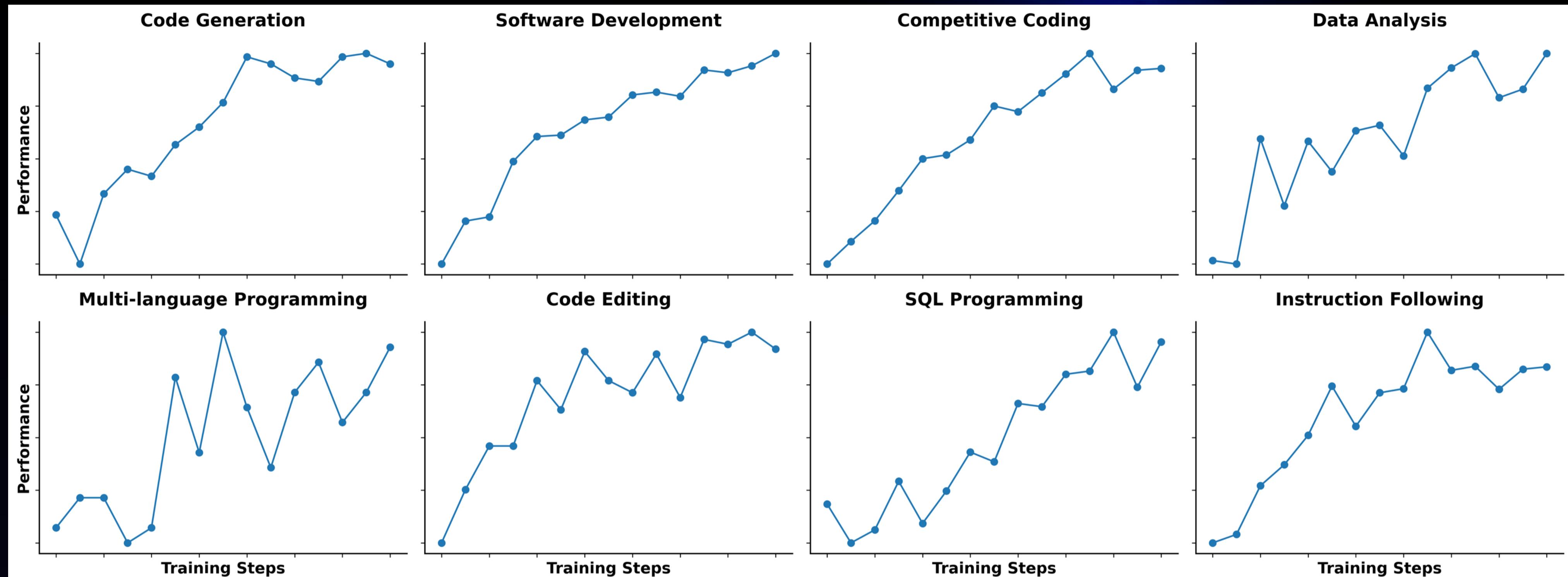
Qwen3-coder-plus  
Qwen3-coder-flash

✨ Supporting long context understanding and generation with the context length of 256K tokens;

✨ Supporting 358 coding languages;

✨ Agentic like

# Reinforce Learning Pre/Post Training – Qwen3-Coder

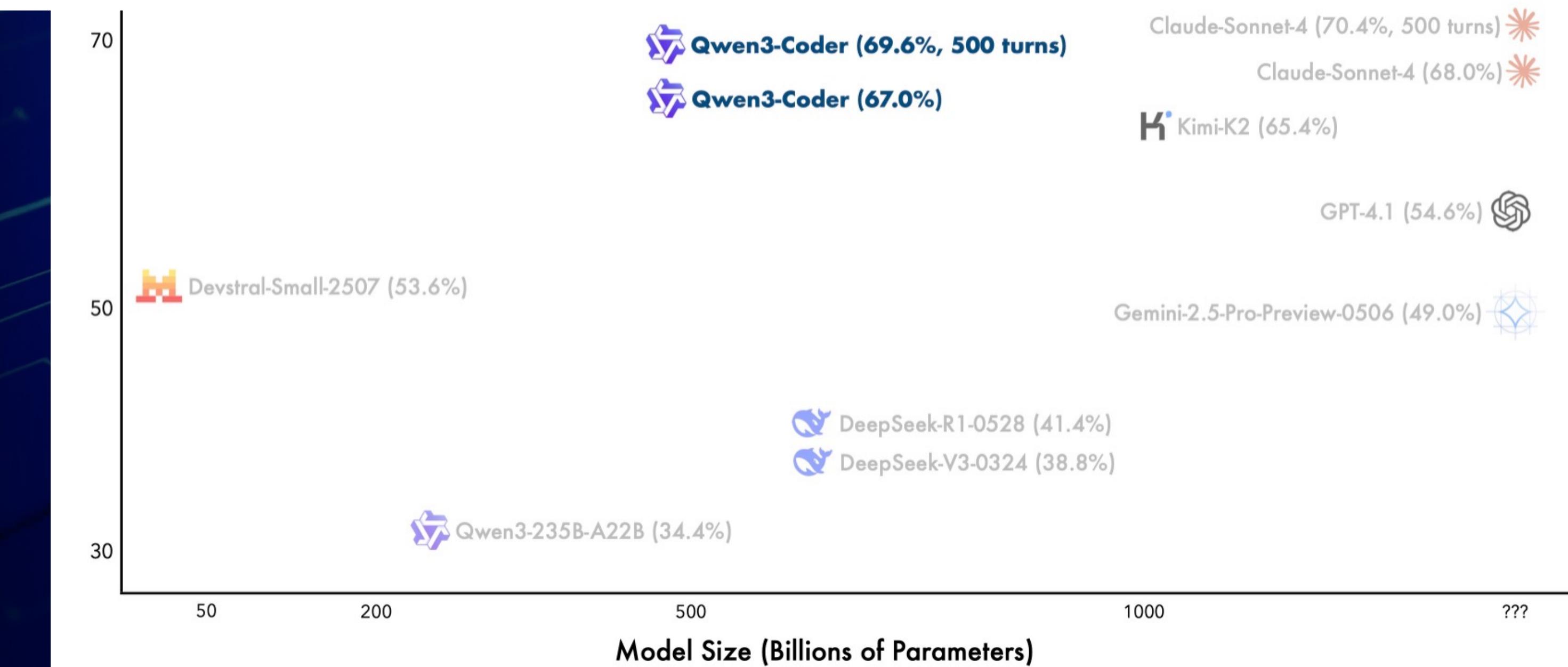


1 **Pre Training :Scaling**  
**Code RL:**  
 Algorithmic logic/Optimization

★ 7.5T tokens trained  
 with 70% code data;

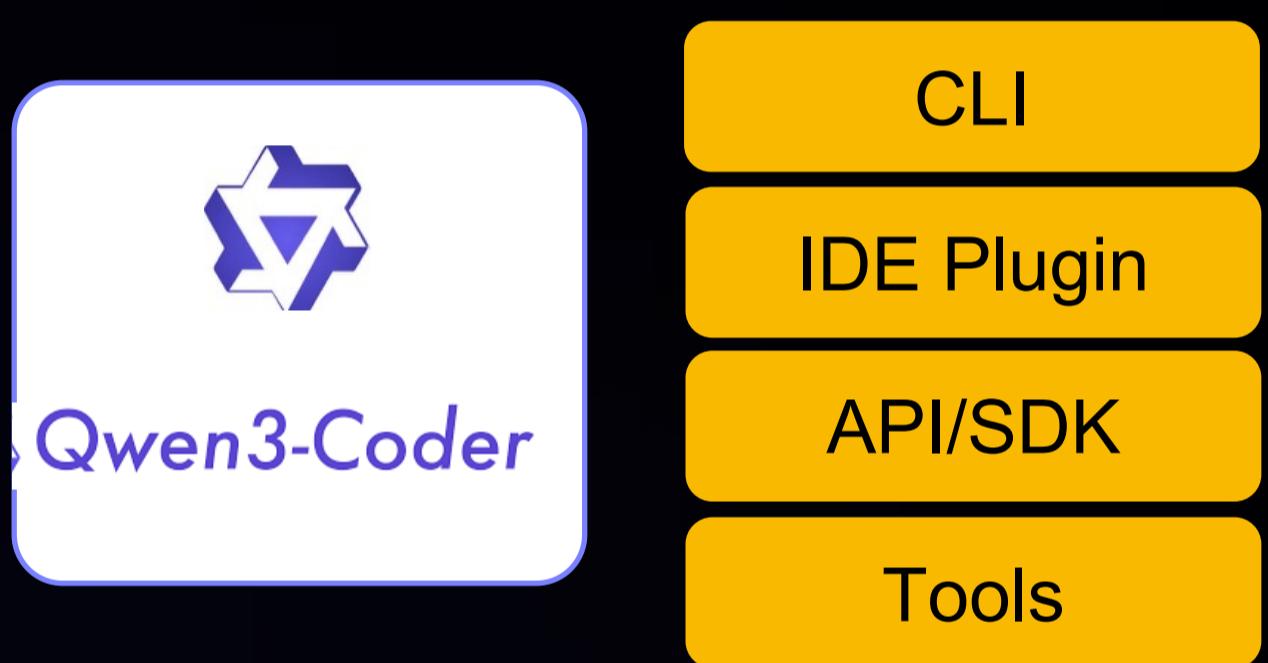
2 **Post Training - Scaling Long-Horizon**  
**RL(Agent RL) :**  
 Solve real-world task in multi-turn interaction with tools.

Plan → Edit → Run → Observe → Debug → Repeat

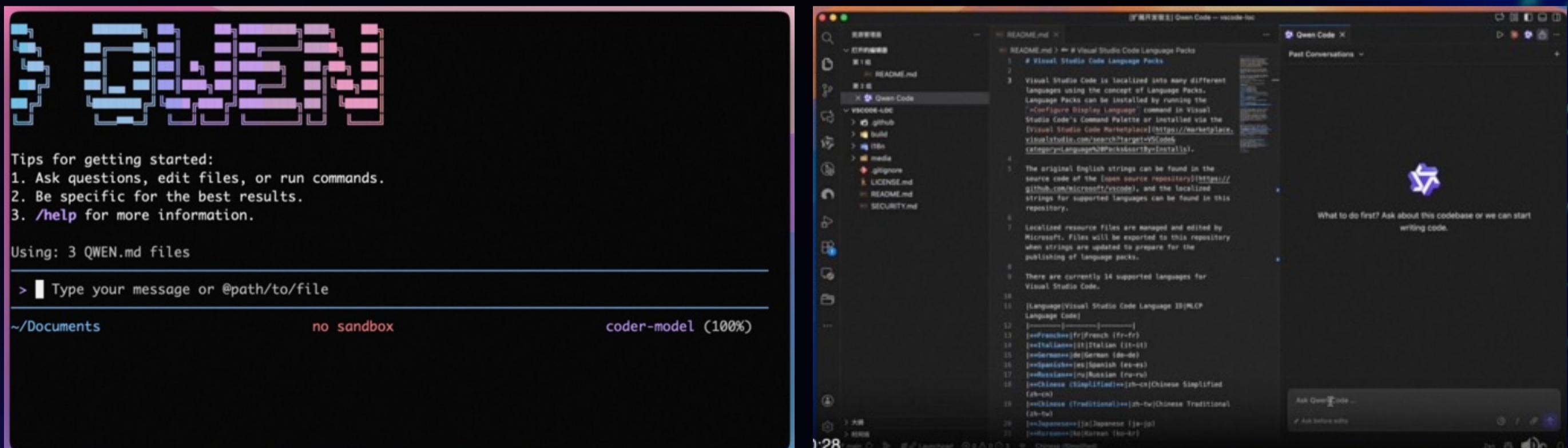
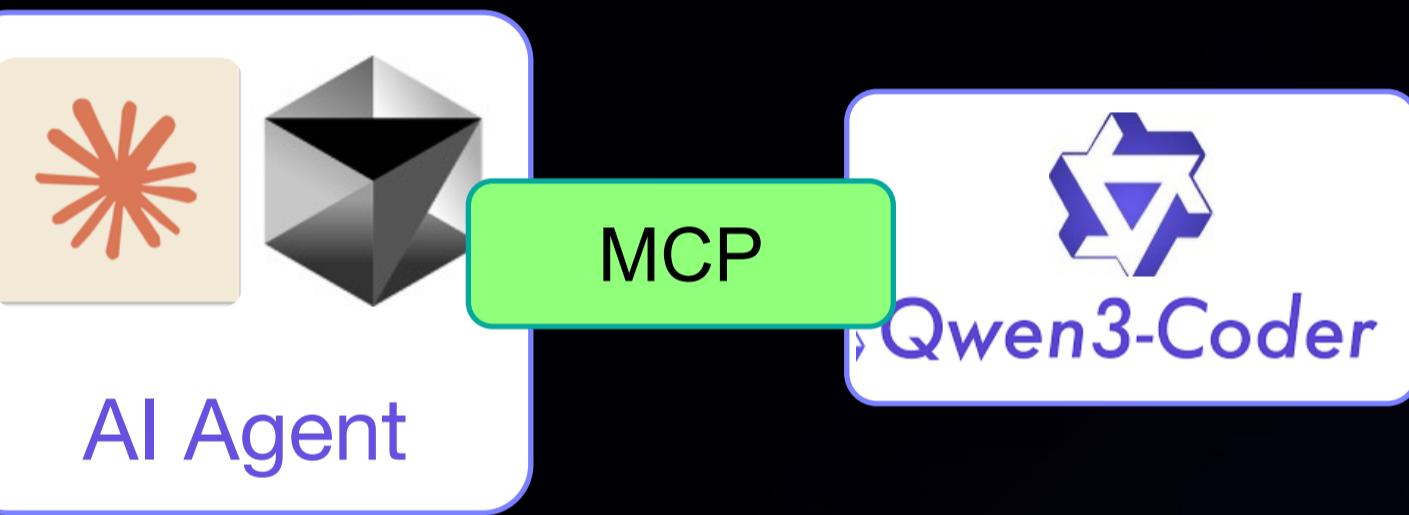


# Integration Patterns with – Qwen3-Coder

Qwen3-Coder as Brain + Specialist



Qwen3-Coder as Specialist



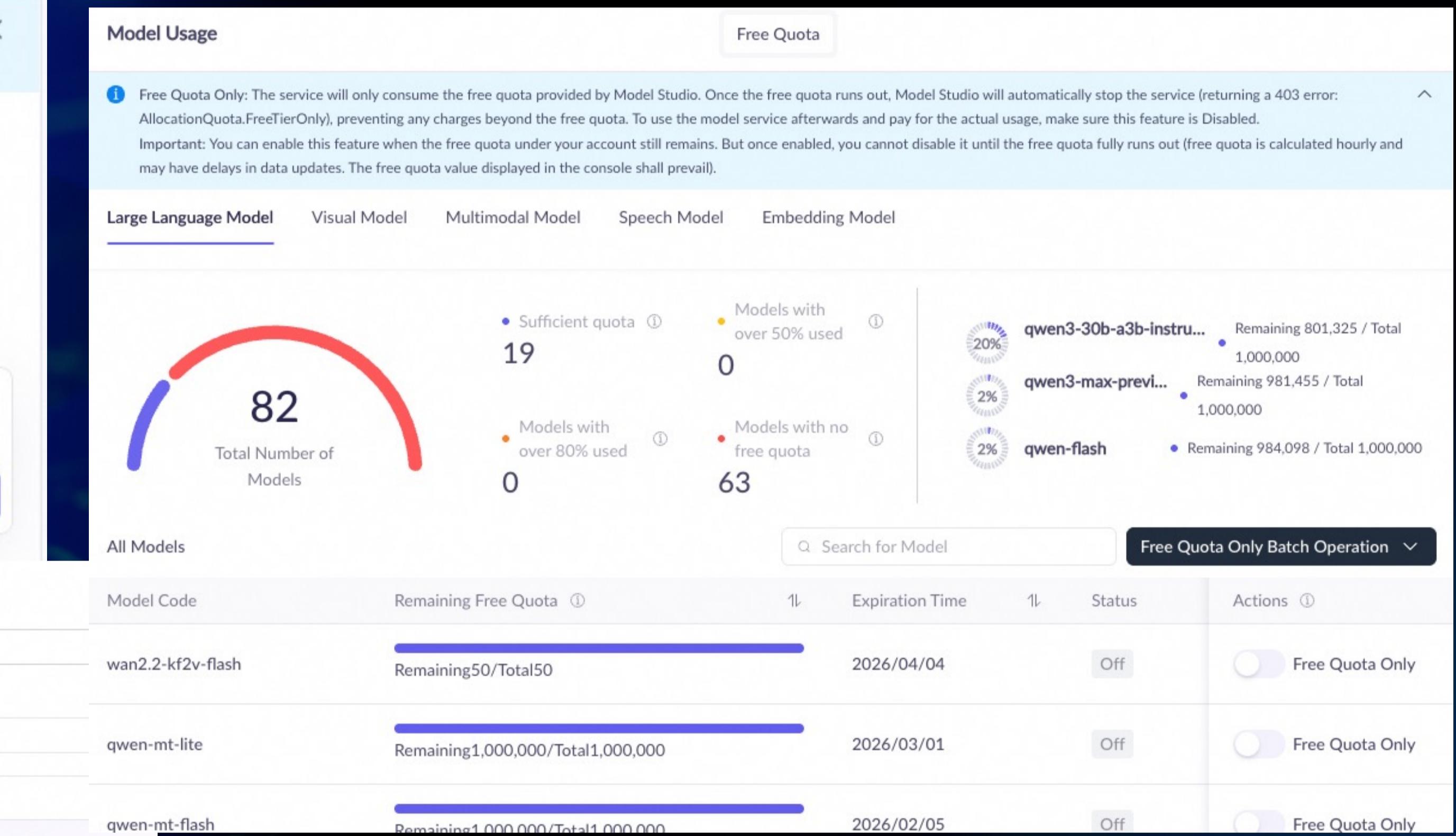
I will plan everything for you, but  
please help me execute my tasks  
with your specialized coding  
capabilities at lower cost

# Model Studio – Model-as-Service for Qwen3-Coder & More

The screenshot shows the Alibaba Cloud Model Studio interface. On the left, there's a sidebar with icons for Create, Text Model (selected), Speech Model, Visual, Manage, Model Usage, Batches, Model Deployment, Model Observation, and Model Alert. The main area has tabs for Text Model, Model Experience, and Model Debugging. A message at the top says: "Trying models will consume free quota or generates pay-as-you-go bills. The actual fees prevail (Except for the deployed models billed for computing duration)". Below this is a "Welcome to Tongyi Models" section with a "Get Started" button and a dropdown for "Qwen3-Coder-Plus". A "Help me" button is also present. At the bottom, there's a modal for "Select Model" showing a list of "Official Model" and "My Models". The "Official Model" section includes fields for Provider (Total), Model Capabilities (Total), and Context Length (Total). The "My Models" section lists "Series": Qwen3-Max (3), Qwen-Plus (7), Qwen-Flash (2), Qwen3-Coder-Plus (3), Qwen3-Coder-Flash (2), Qwen-MT-Flash (1), and Qwen-MT-Plus (1). A note says "Powered by Qwen3, this is a powerful Coding Agent that excels in tool calling and environment interact...".

1 UI & API Access with Security Controls

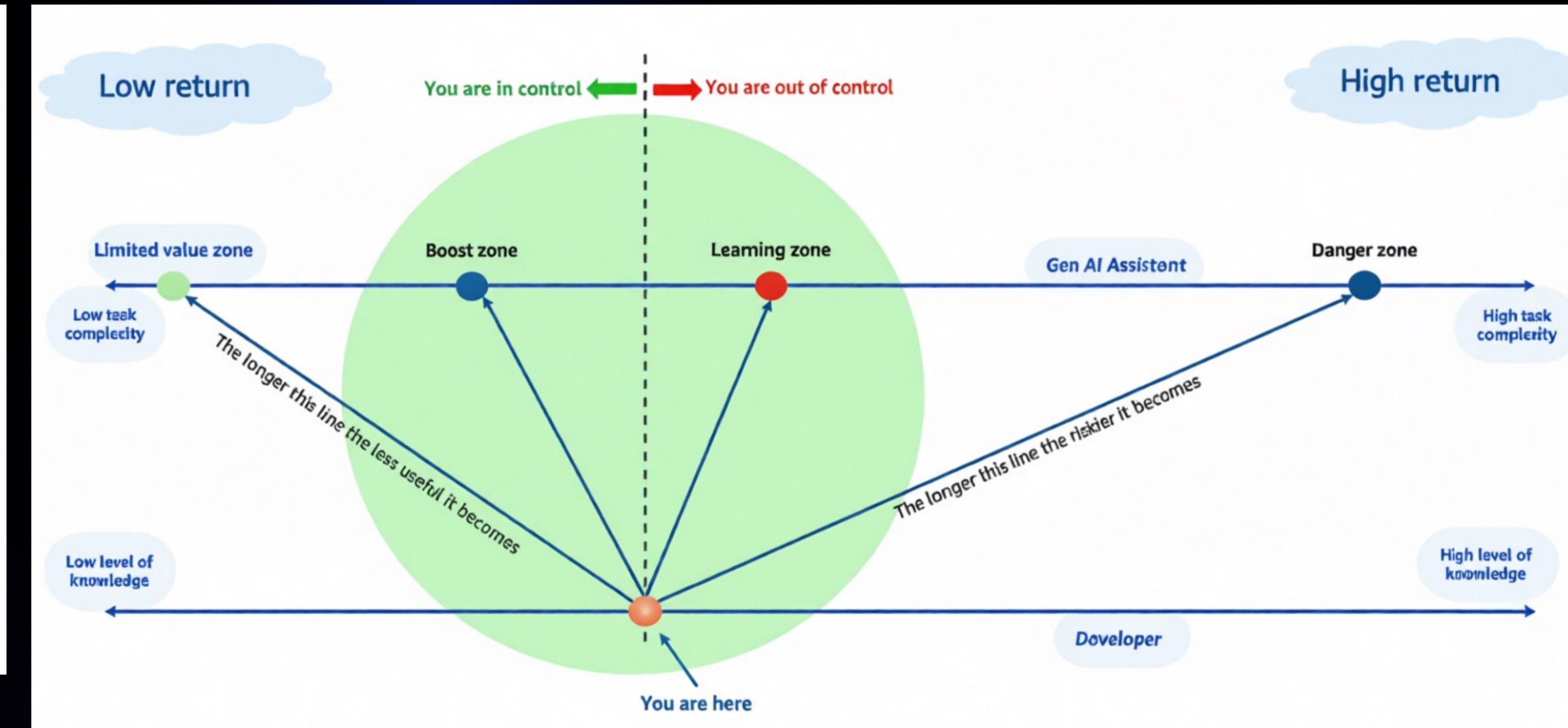
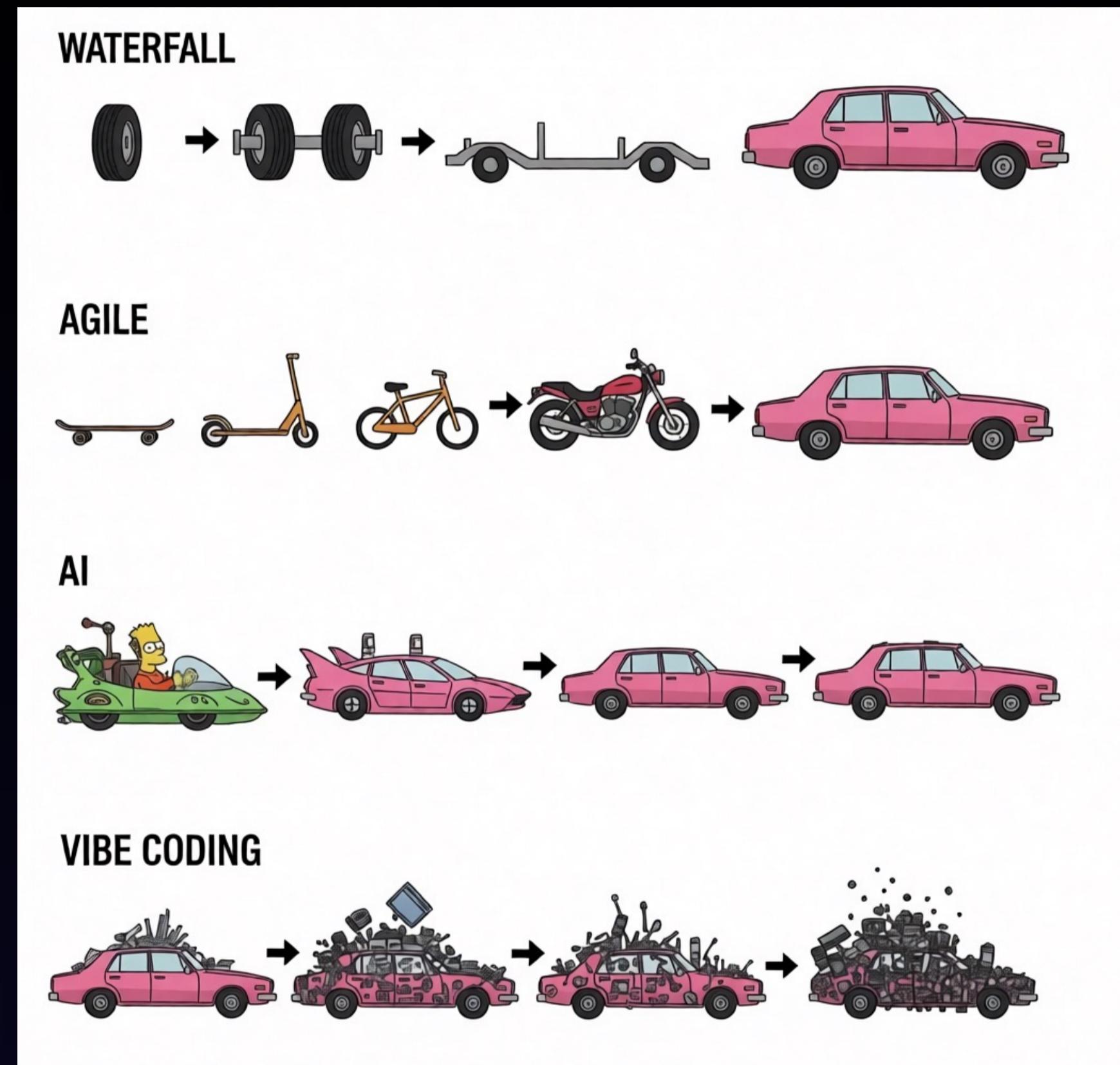
2 Various Models



3 Built-in Monitoring & Alerting



# Know about Vibe Coding – Framework for GenAI Adoption



GenAI Assistant Adoption Framework

# Pricing– Qwen3-Coder

[qwen3-coder-plus series](#)    [qwen3-coder-flash series](#)

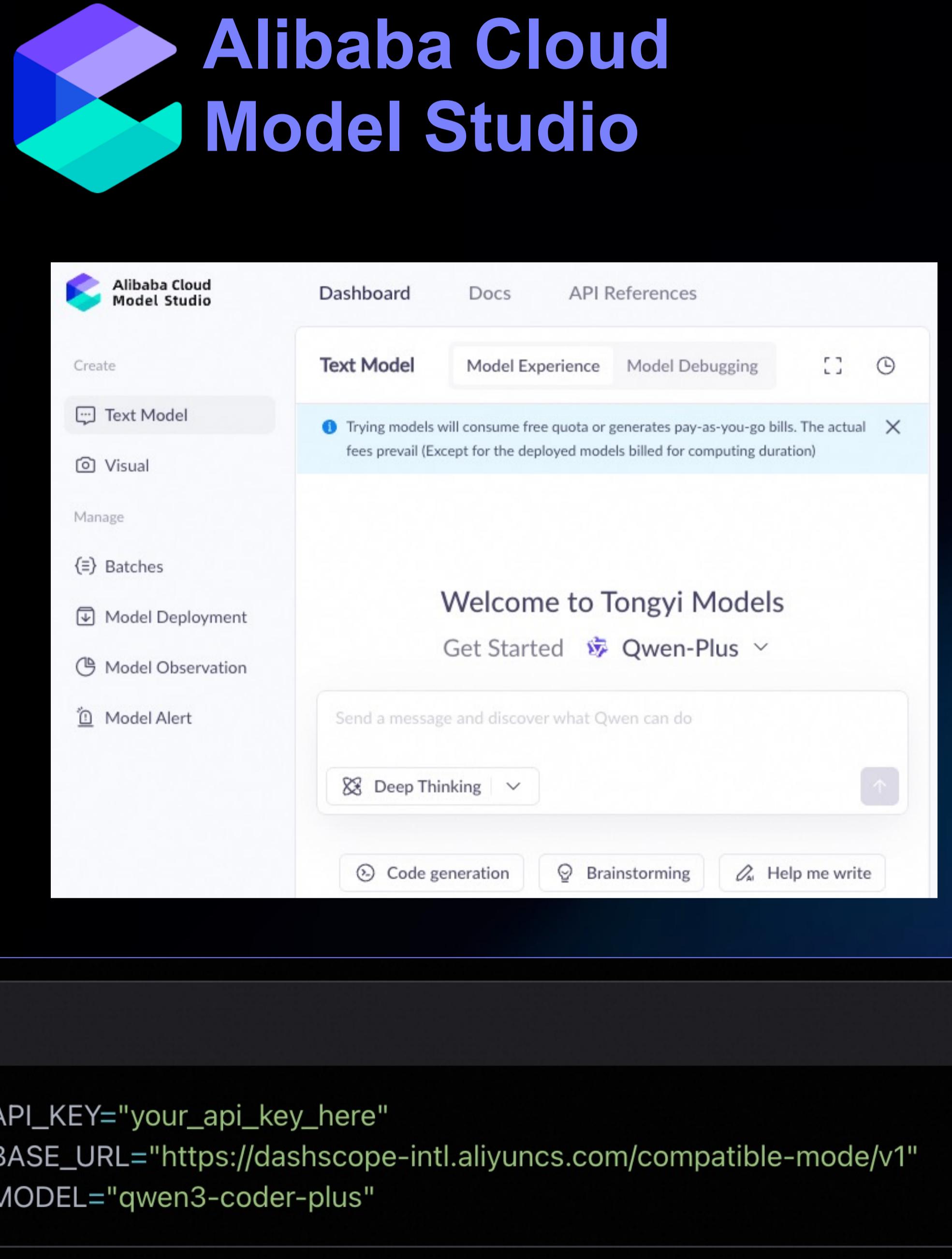
The prices for qwen3-coder-plus, qwen3-coder-plus-2025-09-23, and qwen3-coder-plus-2025-07-22 are as follows. qwen3-coder-plus supports [context cache](#). Input text that results in an **implicit cache** hit is billed at 20% of the price.

Input tokens per request	Input cost (Million tokens)	Output cost (Million tokens)
0 < tokens ≤ 32K	\$1	\$5
32K < tokens ≤ 128K	\$1.8	\$9
128K < tokens ≤ 256K	\$3	\$15
256K < tokens ≤ 1M	\$6	\$60

[qwen3-coder-plus series](#)    [qwen3-coder-flash series](#)

The prices for qwen3-coder-flash and qwen3-coder-flash-2025-07-28 are as follows. qwen3-coder-flash supports [context cache](#). Input text that results in a cache hit is billed at 20% of the price.

Input token count	Input cost (per 1M tokens)	Output cost (per 1M tokens)
0 < Tokens ≤ 32K	\$0.3	\$1.5
32K < Tokens ≤ 128K	\$0.5	\$2.5
128K < Tokens ≤ 256K	\$0.8	\$4
256K < Tokens ≤ 1M	\$1.6	\$9.6



**Alibaba Cloud Model Studio**

Dashboard    Docs    API References

**Text Model**    Model Experience    Model Debugging

Try models will consume free quota or generates pay-as-you-go bills. The actual fees prevail (Except for the deployed models billed for computing duration)

Welcome to Tongyi Models

Get Started    Qwen-Plus

Send a message and discover what Qwen can do

Deep Thinking    ↑

Code generation    Brainstorming    Help me write

```
bash
1 export OPENAI_API_KEY="your_api_key_here"
2 export OPENAI_BASE_URL="https://dashscope-intl.aliyuncs.com/compatible-mode/v1"
3 export OPENAI_MODEL="qwen3-coder-plus"
```

# Pricing- Qwen3-Coder Discount Campaign

**Subscription Plans**

**Overview**

Model Studio Coding Plan is a subscription service designed for professional AI coding. For a fixed monthly fee, you get seamless access to leading AI development tools like Claude Code, Qwen Code, Cline, and OpenClaw, as well as the latest Qwen models. The plan offers industry-leading code understanding, real-time intelligent completion, and powerful tool calling capabilities.

**Model Studio Coding Plan**

For new users: First month as low as \$5, including 18,000 requests and support for major AI coding tools

**Benefits**

- Supports latest and strongest coder models
- Abundant tokens
- Seamless integration with mainstream developer tools

**Usage Steps**

- 1 Subscribe to plan
- 2 Create Dedicated API-Key
- 3 Connect AI Tools
- 4 Start Coding

Purchase Not Created View Details Not Activated

Subscription Plans				
Plan Name/ID/Description	Billing Method/Pricing	Status	Remaining Days	Exclusive API Key
Coding Plan Pro <small>Pro Plan</small>	50 USD/month My month	Active	28 days	sk-sp-*****... <input type="button" value="Copy"/>

> /auth

**API-KEY Configuration**

Select API-KEY configuration mode:

- 1. Coding Plan (Bailian)
- 2. Custom

Paste your api key of Bailian Coding Plan and you're all set!  
(Press Escape to go back)

4. Enter your Coding Plan key, for example: sk-sp-xxxxxxxxx.

- Your API key is saved. It loads automatically in new sessions, so you do not need to reconfigure it.

> /auth

**Coding Plan Setup**

Please enter your API key:

> sk-sp-xxxx

You can get your exclusive Coding Plan API-KEY here:  
[https://bailian.console.aliyun.com/?tab=model#/efm/coding\\_plan](https://bailian.console.aliyun.com/?tab=model#/efm/coding_plan)

(Press Enter to submit, Escape to cancel)

[LINK](#)

THANK YOU  
**ALIBABA CLOUD**

**AINNOVATION | ANYSTACK | ANYTIME | ANYWHERE**

