Alibaba Cloud

Vibe Coding

Jaru Nartboon

Lead Solution Architect, Alibaba Cloud Thailand

> Immerse yourself fully in the atmosphere & simply forget about the existence of code.
>
> **Andrej 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💜**Iteration over perfection** – Perfection is obsolete if you can always reroll.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.

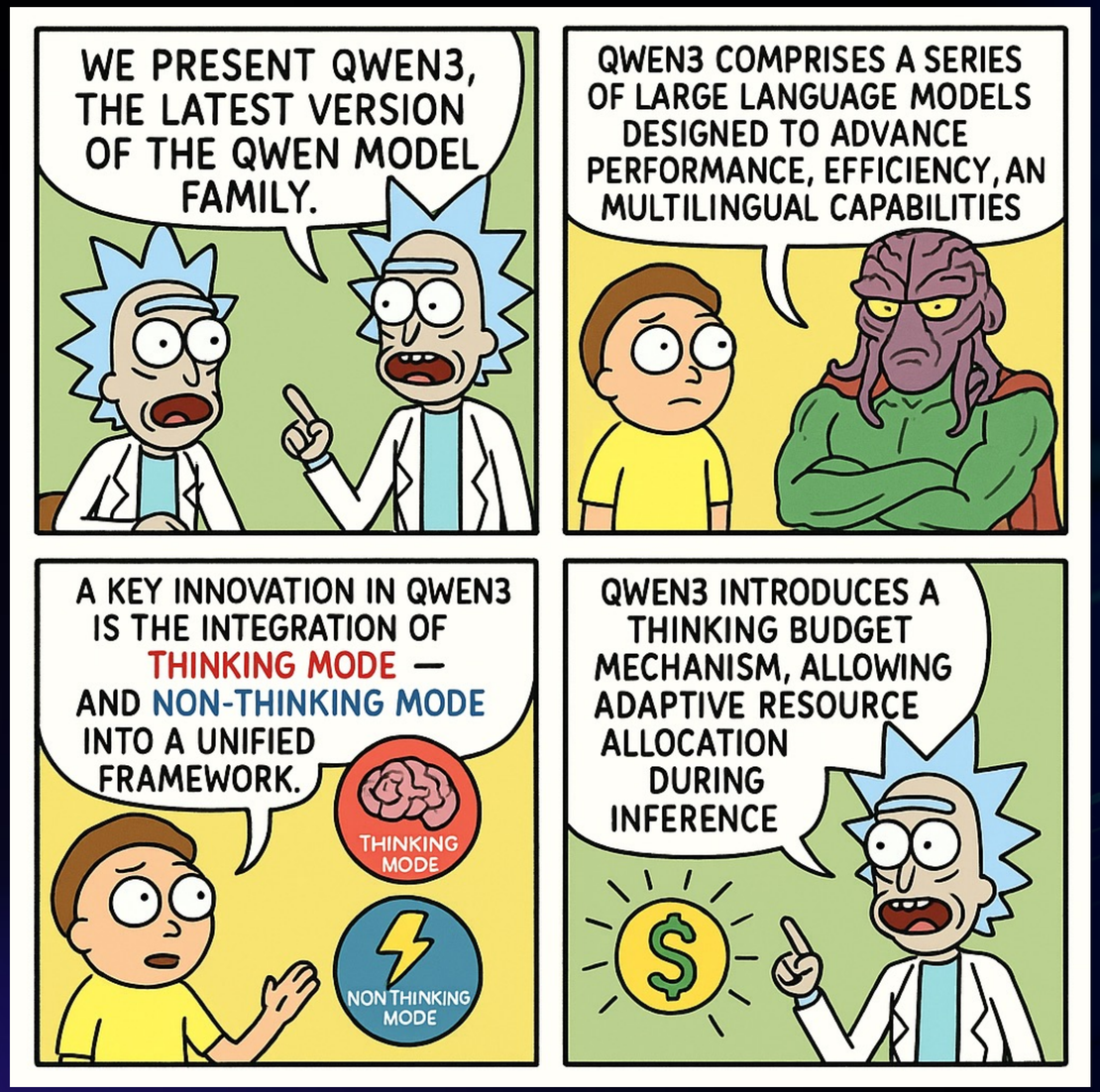
💜**Human taste over technical constraints** – The best tech serves great taste, not the other way around.

https://vibemanifesto.org/

Alibaba Cloud

Model Architecture – Qwen3 : Mixture of Expert (MoE)

Launched 2025/04/29



ARXIV: 2505.09388

Qwen3-235B-A22B

	Dense Architecture	MoE Architecture
Active Param	All parameters (100%)	Subset of experts (25%)
Inference Speed	Slower for large sizes	Much faster for same total size
VRAM Req	Equal to total parameters	Equal to **total** parameters
Specialization	Generalist (all weights learn all)	Specialist (experts learn specific tasks)

Alibaba Cloud

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
MultiIF 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. **SWEBench**: Real-world software engineerinng

Alibaba Cloud

Agentic Code Model - Qwen3-Coder

Launched 2025/06/23

huybery@U-V76490MY-0133:~/coding-agent

23% 16 GB 3.1 kB↓

❯ QWEN3-CODER

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	-

Download
🤗 Hugging Face • 🤖 ModelScope

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

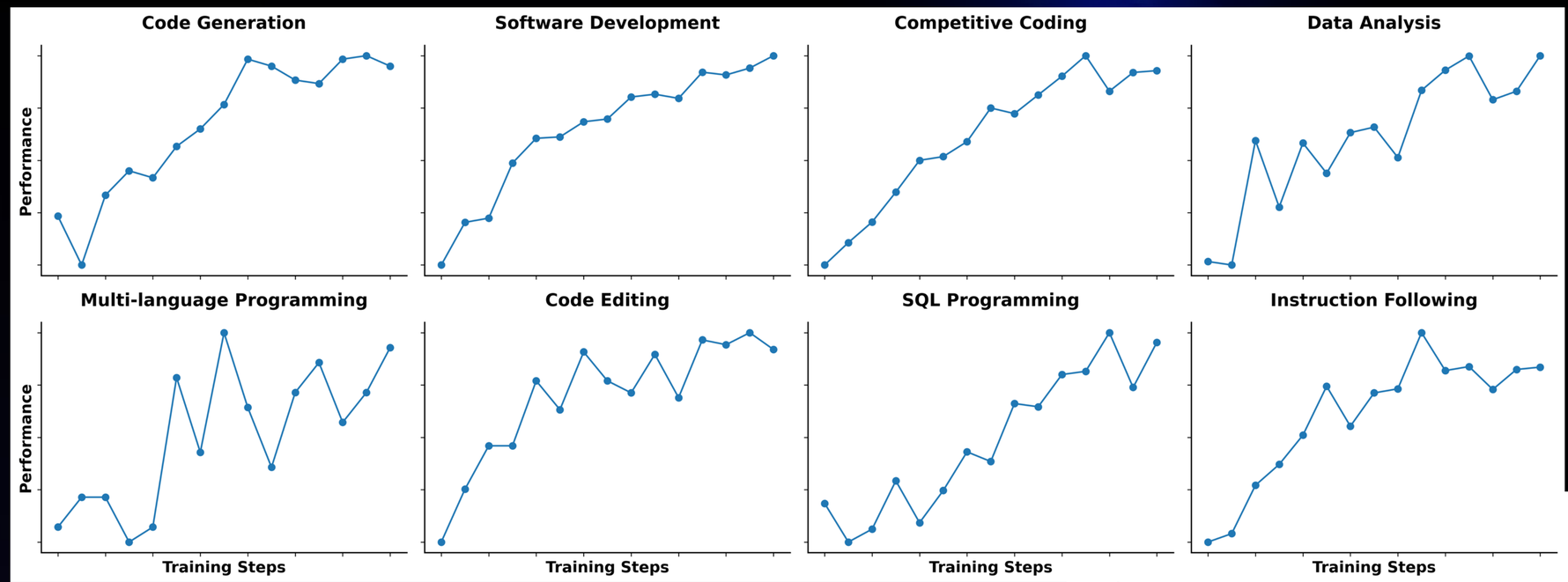
✨ Supporting long context understanding and generation with the context length of 256K tokens;

✨ Supporting 358 coding languages;

✨ Agentic like

Alibaba Cloud

Reinforce Learning Pre/Post Training – Qwen3-Coder

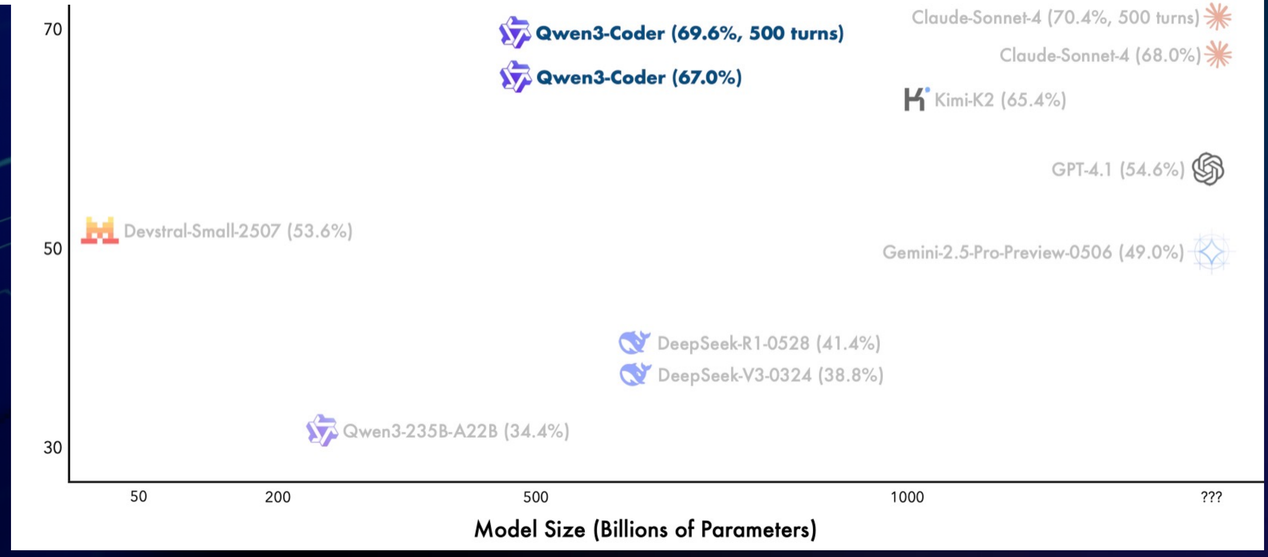


1 **Pre Training :Scaling Code RL:** Algorithmic logic/Optimization

✨ 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.

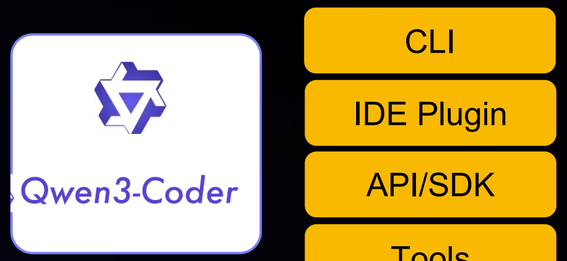
Plan → Edit → Run → Observe → Debug → Repeat

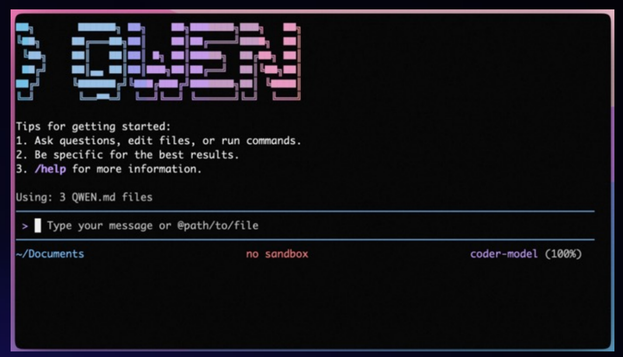


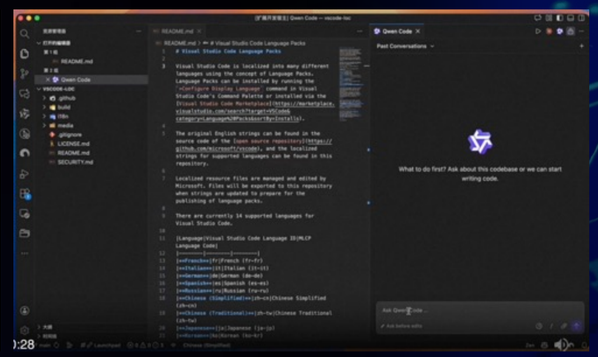
Alibaba Cloud

Integration Patters with – Qwen3-Coder

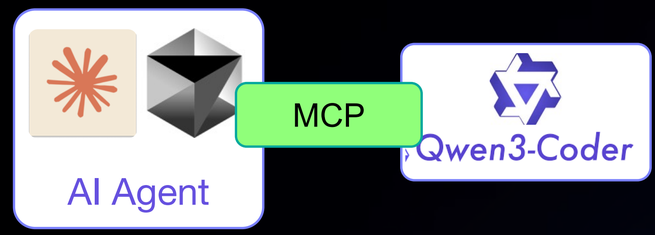
Qwen3-Coder as Brain + Specialist







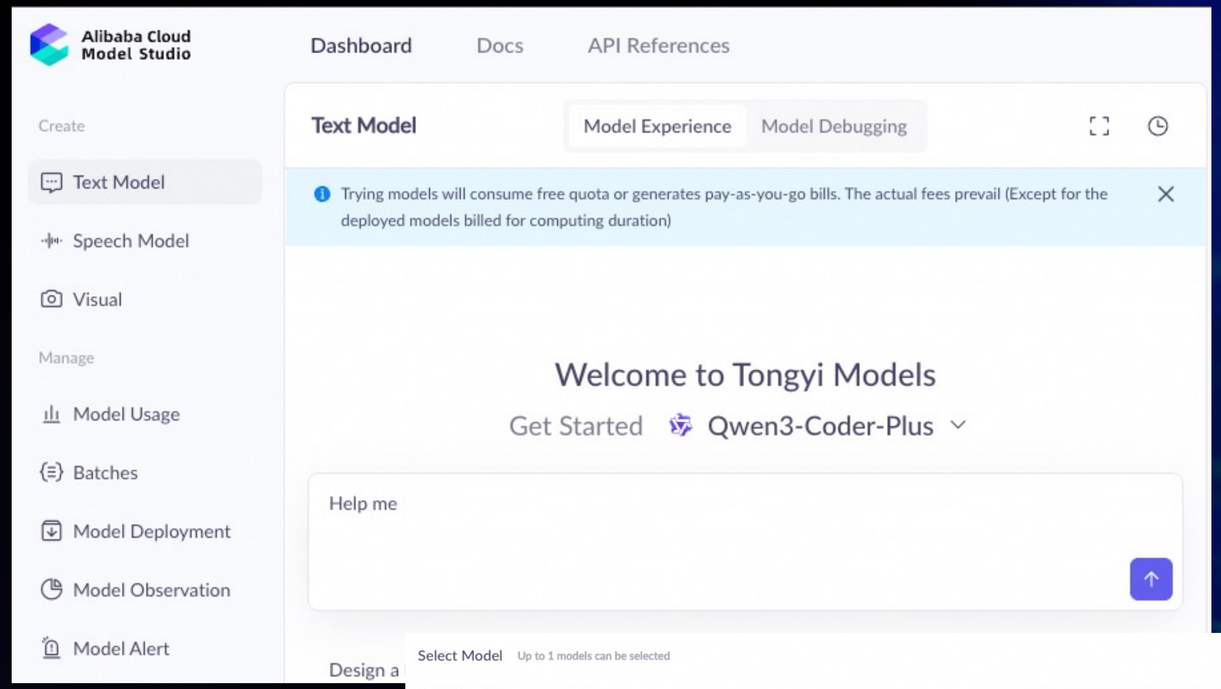
Qwen3-Coder as Specialist

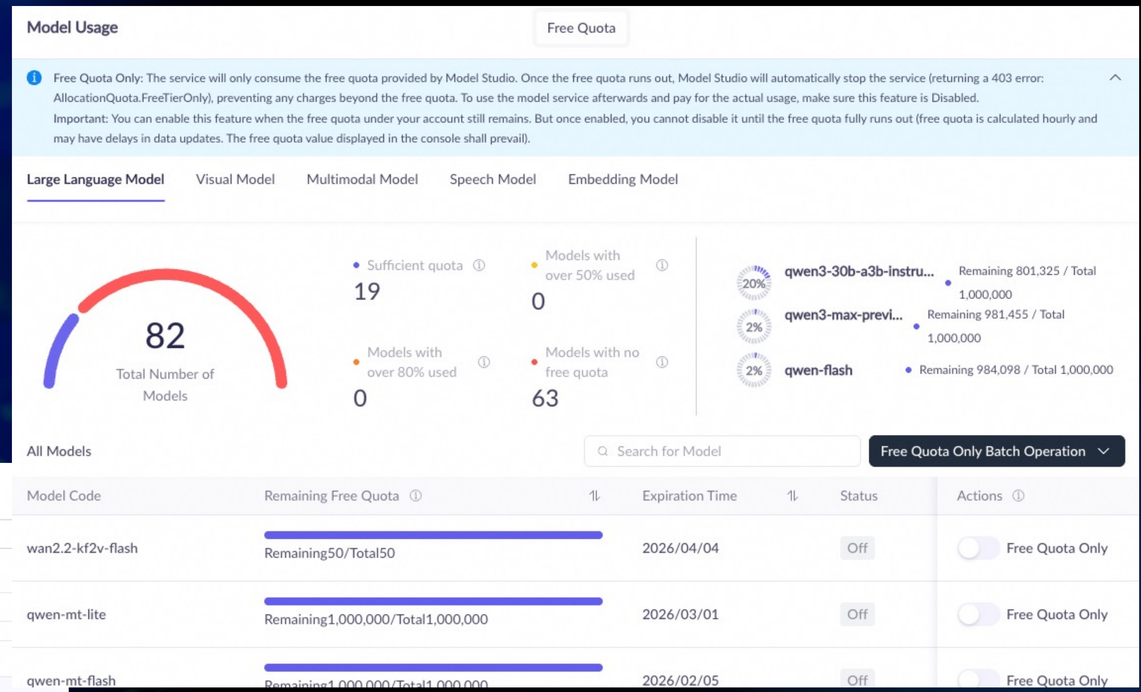


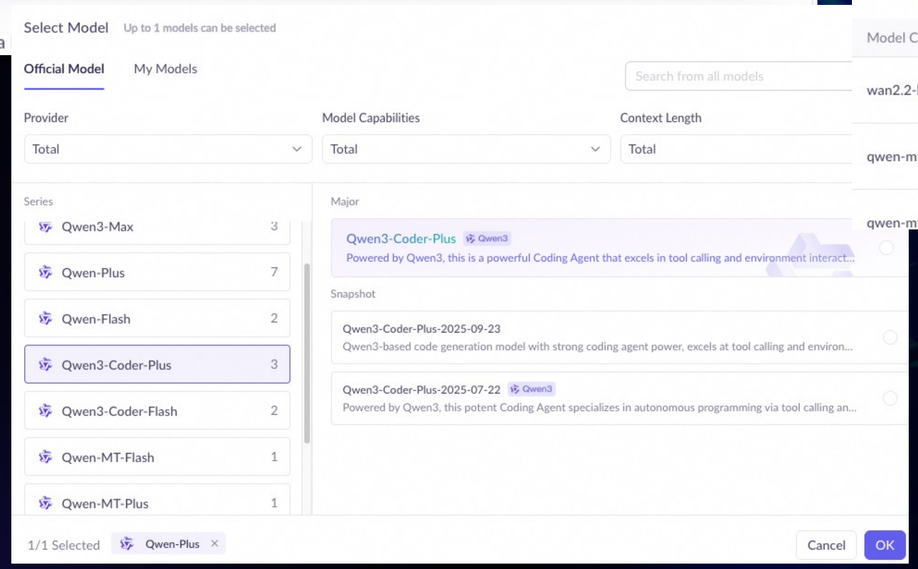
I will plan everything for you, but please help me execute my tasks with your specialized coding capabilities at lower cost

Alibaba Cloud

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

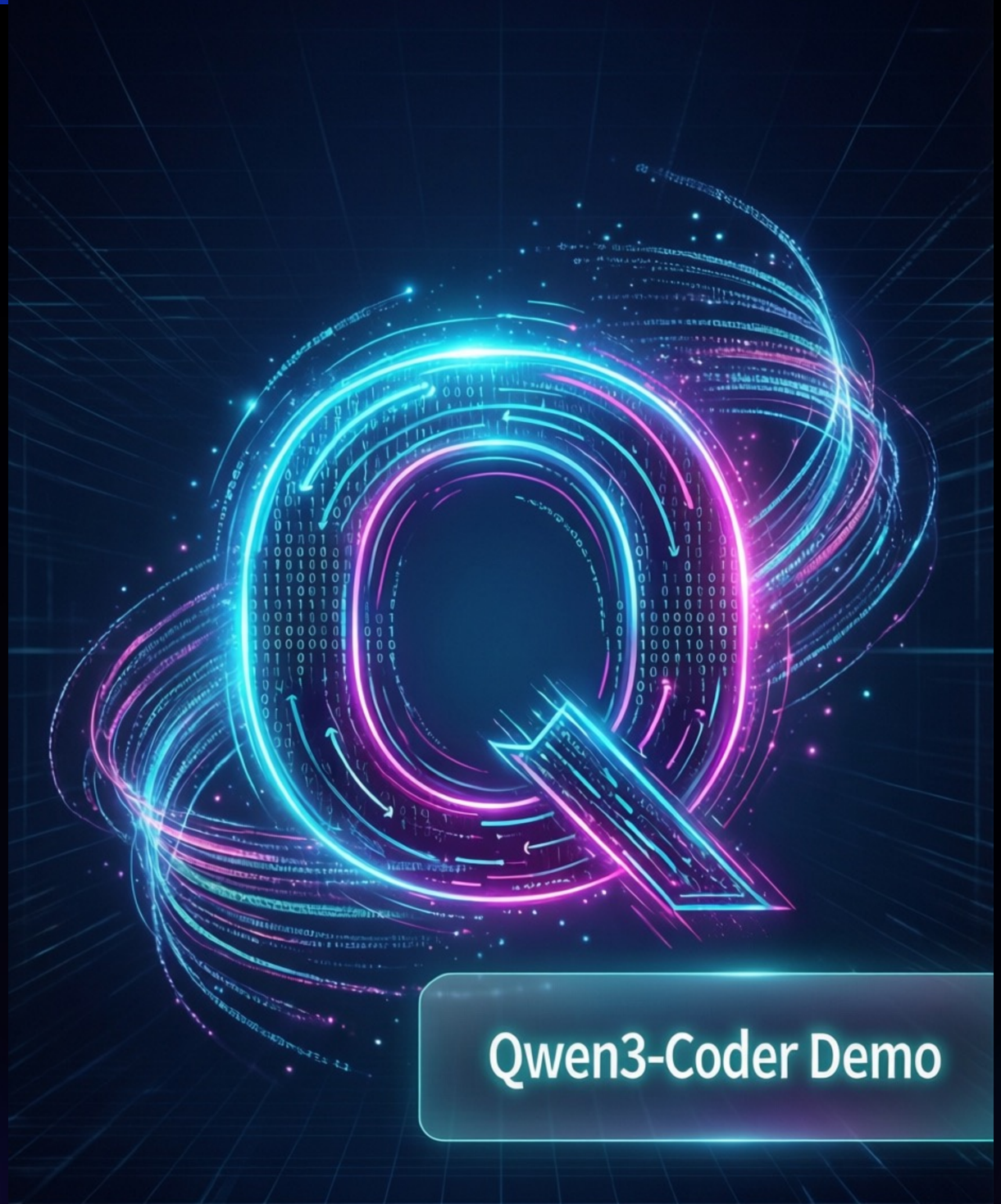






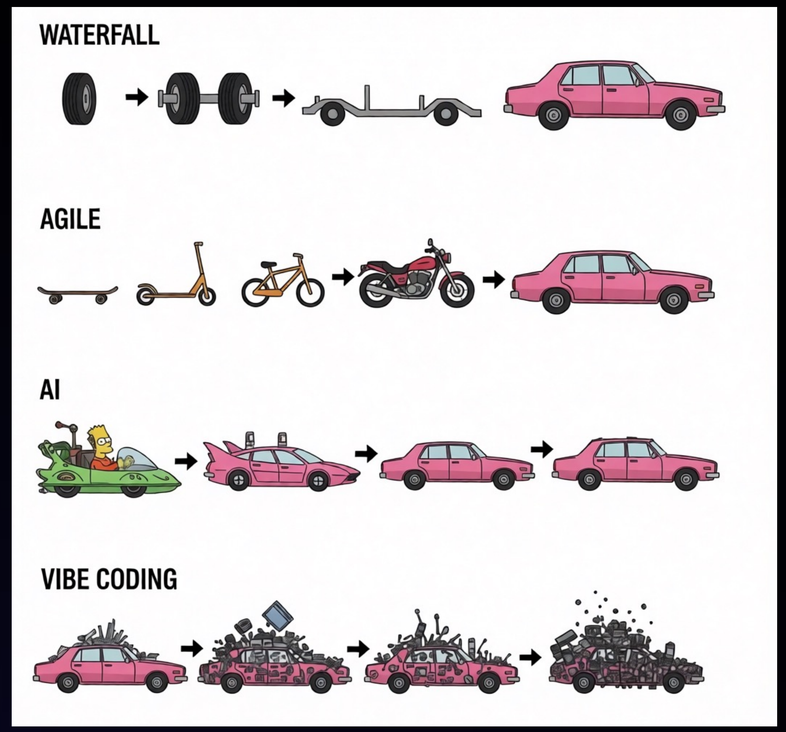
1. **UI & API Access with Security Controls**
2. **Various Models**
3. **Built-in Monitoring & Alerting**

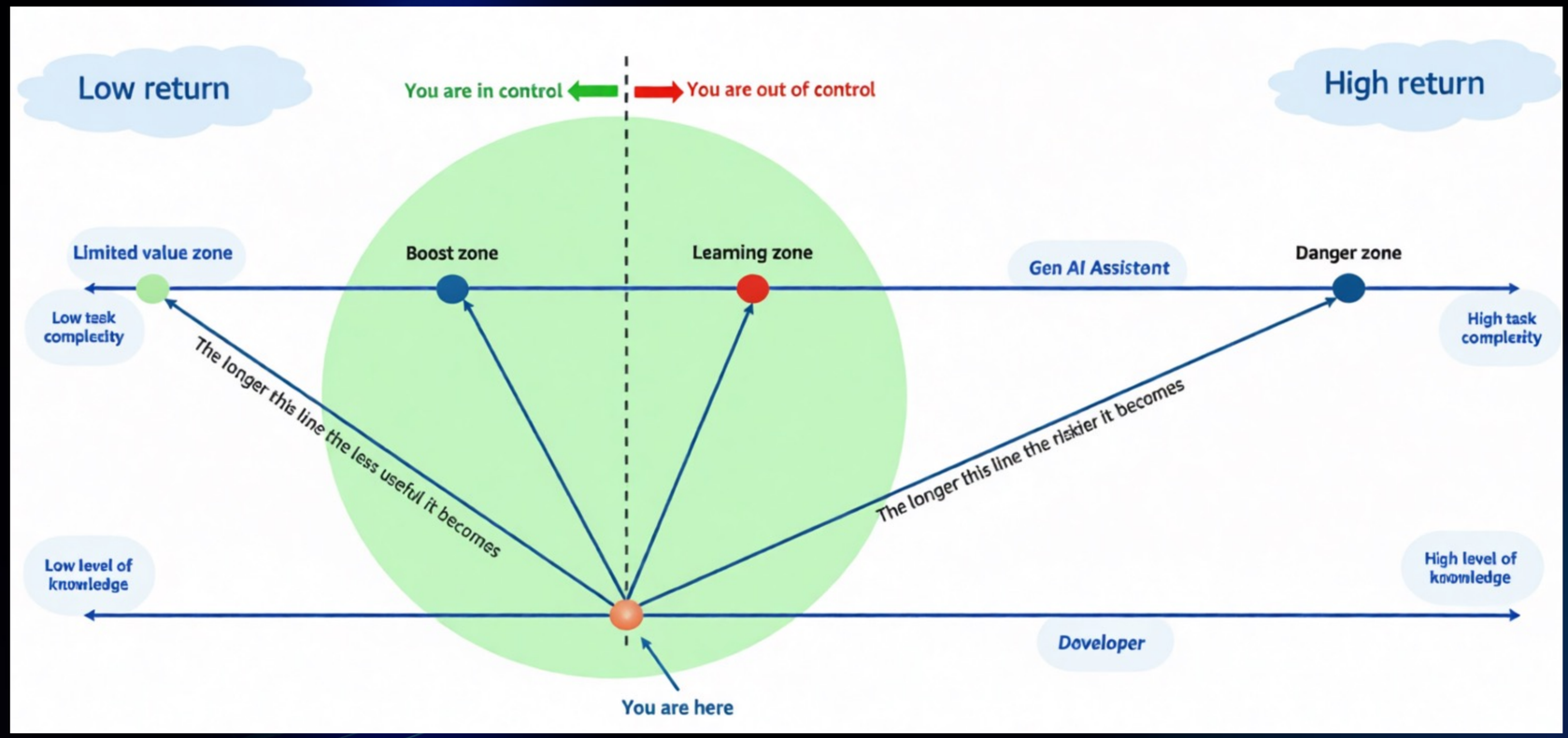
Alibaba Cloud



Alibaba Cloud

Know about Vibe Coding – Framework for GenAI Adoption





GenAI Assistant Adoption Framework

Reference

Alibaba Cloud

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-p
supports context cache. Input text that results in an **implicit cache** hit is billed at 20% of the price.

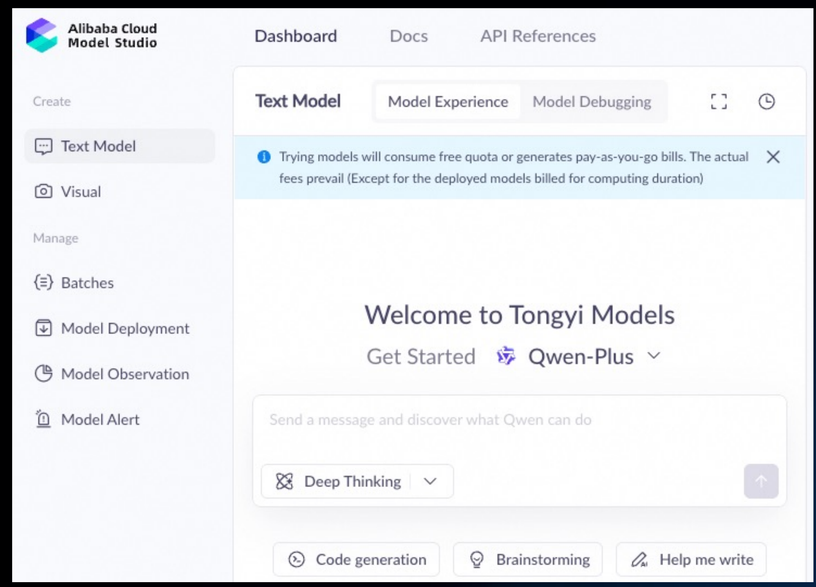
Input tokens per request	Input cost (Million tokens)	Output cost (Million tokens)
0 < tokens ≤ 32K	$1	$5
32K < tokens ≤ 128K	$1.8	$9
128K < tokens ≤ 256K	$3	$15
256K < tokens ≤ 1M	$6	$60

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.
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

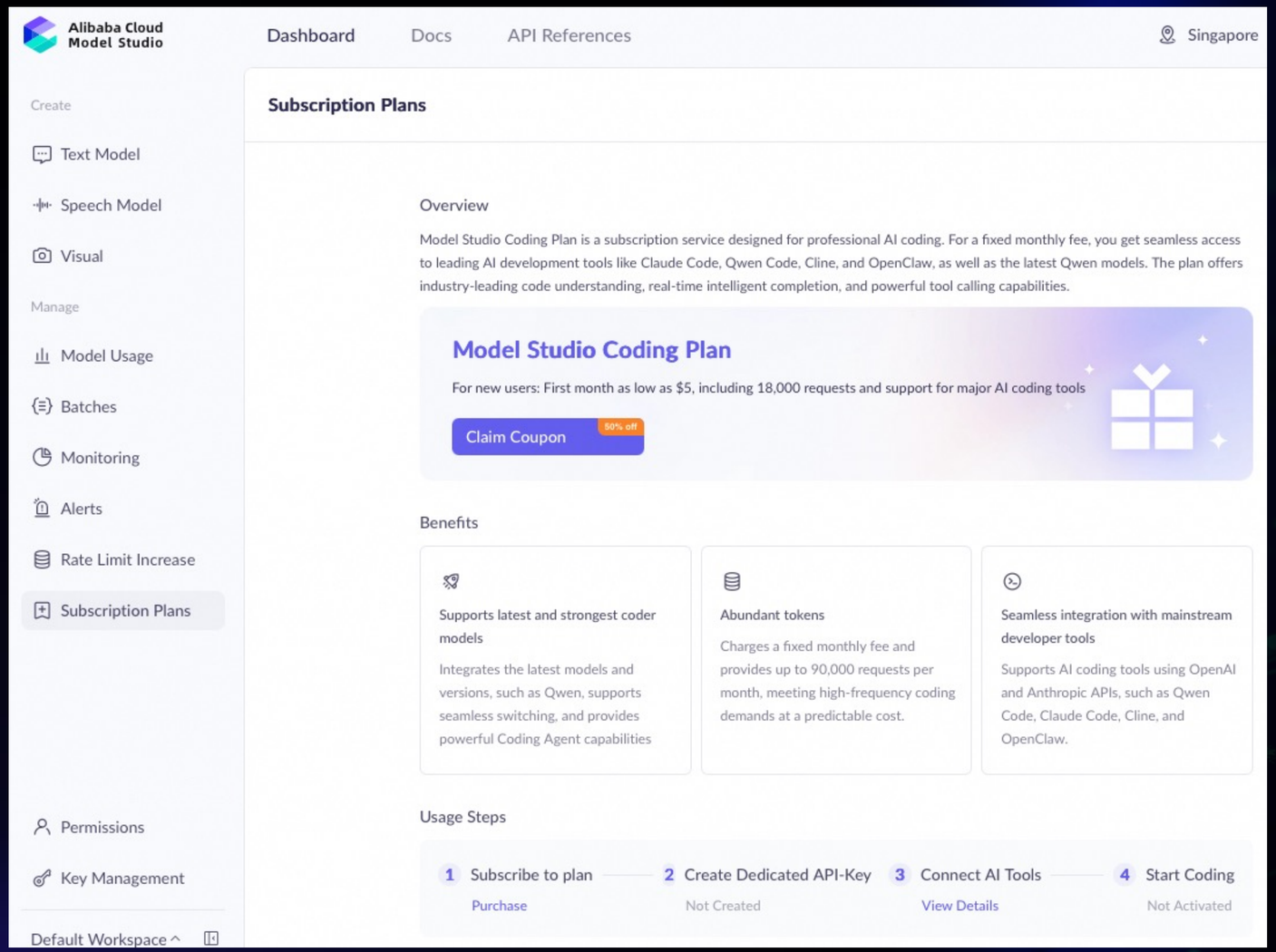


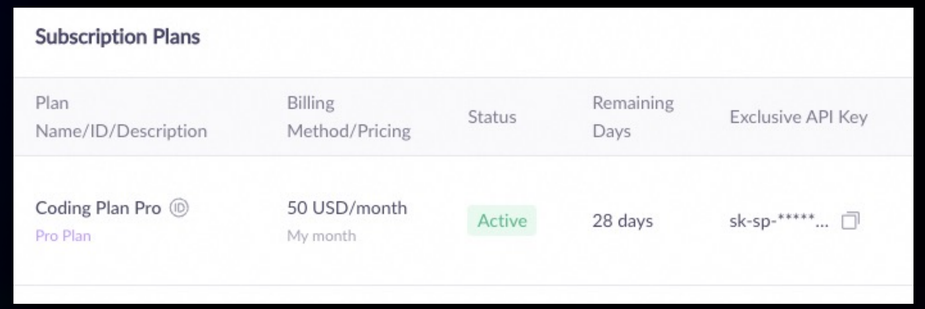
bash

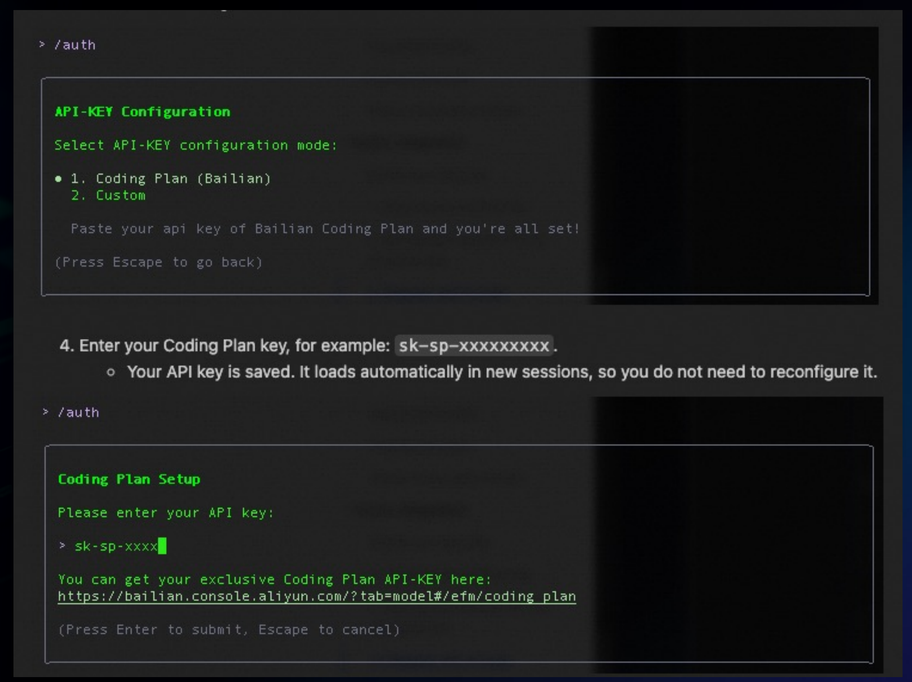
```bash
export OPENAI_API_KEY="your_api_key_here"
export OPENAI_BASE_URL="https://dashscope-intl.aliyuncs.com/compatible-mode/v1"
export OPENAI_MODEL="qwen3-coder-plus"
```

Alibaba Cloud

Pricing– Qwen3-Coder Discount Campaign







LINK

Alibaba Cloud

THANK YOU

ALIBABA CLOUD

AINNOVATION | ANYSTACK | ANYTIME | ANYWHERE