Unicoder AI Models – Complete Reference

This document provides a comprehensive list of AI models available in Unicoder, your AI-powered IDE. It includes model details, performance metrics, and suggested usage.

OpenAI Models

Model	Description	Uptime	Throughput	Latency	Suggested Usage / Label
GPT-5	Advanced reasoning & multimodal support; ideal for complex coding & analysis.	95.23%	74.8 ms	2–6 s	Power / Complex Tasks
GPT-4o	Multimodal (text, vision, audio); suitable for AI applications needing multiple input types.	94.87%	80 ms	3–10 s	Multimodal / Pro Tasks
GPT-5 Nano	Small, cost-effective, low-latency autocomplete model.	95.23%	74.8 ms	2–6 s	Fast / Autocomplete
GPT-4.1	Advanced large language model for coding & reasoning.	91.64%	133 ms	2–7 s	Reasoning / Coding

---

Google AI Models

Model	Description	Uptime	Throughput	Latency	Suggested Usage / Label
Gemini 2.5 Pro	Enhanced performance with large context window for complex tasks.	95.00%	95 ms	4–9 s	Power / Complex Tasks
Gemini 2.0 Flash	Streamlined for quick responses; ideal for low-latency tasks.	97.44%	150 ms	0.17–2.5 s	Fast / Autocomplete

---

Moonshot AI Models

Model	Description	Uptime	Throughput	Latency	Suggested Usage / Label
Kimi-K2	Preview model with 63k context window; suitable for large-scale coding.	N/A	N/A	N/A	Experimental / Long Context

Zhipu AI Models

Model	Description	Uptime	Throughput	Latency	Suggested Usage / Label
GLM-4.5 AirL	Lightweight mixture-of-experts model; optimized for speed.	97.59%	135 ms	0.6–1.5 s	Fast / Coding

DeepSeek AI Models

Model	Description	Uptime	Throughput	Latency	Suggested Usage / Label
DeepSeek R1	671B parameters (37B active), open-source, strong reasoning.	96.23%	120 ms	1–4 s	Reasoning / Open-Source
DeepSeek V3	Advanced LLM with efficient architecture and high performance.	97.23%	33 ms	3–6 s	Performance / Complex Tasks
DeepSeek V3.1 Turbo	High-performance, low-latency model for autocomplete & coding.	97.88%	155 ms	0.4–1.1 s	Turbo / Fast
DeepSeek-v3.1-turbo	High-performance turbo model from DeepSeek; optimized for speed.	97.88%	155 ms	0.4–1.1 s	Turbo / Fast

Microsoft AI Models

Model	Description	Uptime	Throughput	Latency	Suggested Usage / Label
MAI-DS-R1-FP8	Reasoning-focused model; fast and reliable.	98.77%	82 ms	1–2.5 s	Reasoning / Fast

Tencent AI Models

Model	Description	Uptime	Throughput	Latency	Suggested Usage / Label
Hunyuan A13B Instruct	Instruction-tuned 13B model; optimized for guided coding.	98.90%	165 ms	0.4–1.1 s	Instruction-Tuned / Turbo

LLaMA AI Models

Model	Description	Uptime	Throughput	Latency	Suggested Usage / Label
LLaMA-4 Maverick	Mixture-of-experts architecture; excels in reasoning & coding.	98.93%	50 ms	0.5–1.5 s	Reasoning & Coding / MoE

Qwen AI Models

Model	Description	Uptime	Throughput	Latency	Suggested Usage / Label
Qwen 2.5-72B	72.7B params; strong long-context understanding, coding & math.	96.00%	80 ms	4–9 s	Large Context / Complex Tasks
Qwen2.5 Coder 7B Instruct	Instruction-tuned coding model; fast autocomplete & suggestions.	99.43%	205 ms	0.3–0.9 s	Instruction-Tuned / Turbo
Qwen-3 Coder Plus	Advanced code generation & development tasks.	97.10%	90 ms	2.5–6 s	Advanced Coding / Reasoning

Alibaba AI Models

Model	Description	Uptime	Throughput	Latency	Suggested Usage / Label
Tongyi DeepResearch 30B A3B	30B parameter model for deep research applications.	95.99%	90 ms	0.8–2 s	Research / Reasoning

Anthropic AI Models

Model	Description	Uptime	Throughput	Latency	Suggested Usage / Label
Claude-3 7 Sonnet	High-performance model for reasoning & coding.	99.95%	150 ms	0.5–2 s	Fast / Reasoning / Coding
Claude Opus 4	High-performance, stable model for interactive coding.	99.95%	150 ms	0.5–2 s	Fast / Reasoning / Coding
Claude Sonnet 4	Balanced performance & cost; high-volume production workloads.	90.00%	60 ms	2–6 s	Production / Reasoning
Claude Sonnet 4.5	Upgraded Sonnet; reliable and high performance.	98.70%	85 ms	2.5–6 s	Production / Reasoning

Notes / Recommendations

- **Turbo / Fast models**: Great for autocomplete, live coding, and suggestions.
- **Reasoning / Complex models**: Best for debugging, code generation, and heavy logic tasks.
- **Instruction-Tuned models**: Ideal for following user instructions precisely.
- **Large Context models**: Suitable for long scripts, notebooks, or multi-file projects.
- **Preview / Experimental models** (e.g., Kimi-K2) should be labeled clearly for advanced users.
- Consider **latency and uptime** when suggesting models to users for interactive tasks.