

Local OpenVINO Coding Model — 2026-09-14

OpenVINO provides local inference on the Intel iGPU for coding and other bounded workloads. The current documented route is `openvino/qwen2.5-coder-7b`; GPT-5.6 Luna remains the default OpenAI model. A larger 14B model was not practical within the available shared memory.

Configuration

- Backend: OpenVINO GenAI on Intel integrated graphics.
- Model: Qwen2.5-Coder 7B Instruct, INT4 OpenVINO format.
- Client contract: OpenAI-compatible local endpoint.
- OpenClaw model route: `openvino/qwen2.5-coder-7b`.
- Generation: serialized requests, approximately 8K prompt budget and 512-token output cap.
- OVMS: optional alternative serving layer; use only when separately configured and verified.
- Credentials: local endpoint details and any access token belong in protected runtime configuration, never in this page.

Relationship to other systems

OpenClaw selects this route for local coding tasks while hosted OpenAI remains the default. Skills may request the local route when workload, privacy or latency makes it appropriate; BookStack documents the relationship and safe configuration, not live secrets.

Revision #4

Created 2026-09-14 05:23:44 UTC by Admin

Updated 2026-09-14 07:12:02 UTC by Admin