Glossary

The book uses vocabulary from three communities that don't always agree on terms: Intel NPU hardware, OpenVINO/Hugging Face software, and the agent-design literature. Definitions here are tuned to how the book uses each term, not to general usage. Entries are alphabetical.

ANE (Apple Neural Engine). Apple's fixed-function tensor accelerator integrated into M-series and A-series silicon. Accessible only through Core ML. The M4 family ships 16 cores at 38 TOPS. Discussed in Chapter 1.1 as a comparison point to Intel NPU.

Autoregressive decode. The LLM generation phase where one output token is produced per forward pass, conditioned on all prior tokens via the KV cache. Memory-bandwidth-bound on NPU. Distinct from prefill.

Batch size. The number of independent inference streams processed in one forward pass. On Intel NPU, batch size 1 is the only configuration that makes sense for LLM decode — batching adds latency without throughput because the bandwidth ceiling is already binding.

CACHE_DIR. OpenVINO Core property pointing to a directory where compiled-blob bytecode is persisted across processes. Saves 30–60 seconds of cold-start compile on every subsequent run. Always set in production.

Channel-wise quantization (group-size −1). Per-column weight quantization where each output channel has its own scale. Required for larger LLMs on NPU (>~5 B parameters); finer-grained group-128 is preferred for smaller models.

Chunked prefill. OpenVINO 2025.3+ feature where long prompts are processed on NPU in fixed-size chunks (default `NPUW_LLM_PREFILL_CHUNK_SIZE=1024`) under a `PREFILL_HINT=DYNAMIC` static-shape pipeline. The illusion of dynamic shape, paid for by chunk granularity.

Click to Do. Microsoft's Phi Silica frontend in Windows 11 — a fixed set of prompt templates exposed as right-click actions on selected text. No learned router, no multi-step loop. The canonical single-shot NPU LLM deployment.

Cold start / warm start. Cold start is the first invocation of a compiled NPU model in a process: 30–60 seconds for a 3B–8B LLM at INT4 (longer for larger models). Warm start is every subsequent load from `CACHE_DIR`: 1–3 seconds. The gap is why `CACHE_DIR` matters.

Compute tile. The single piece of silicon on Lunar Lake and Panther Lake holding the CPU, Xe iGPU, and NPU on-die. Distinct from AMD's XDNA, where the NPU is a separate IP block.

Copilot+ PC. Microsoft's certification requiring ≥40 TOPS of NPU performance, 16 GB RAM, and 256 GB SSD minimum. Phi Silica and other on-device Copilot+ features require a Copilot+ machine. Floor is Intel NPU 4 (Lunar Lake) or equivalents.

Core Ultra Series 1 / 2 / 3. Intel's client CPU branding for Meteor Lake (Series 1, Dec 2023), Lunar Lake (Series 2, Sep 2024), and Panther Lake (Series 3, CES 2026).

Cross-attention. Decoder self-attention that also reads encoder output as keys and values. Present in encoder-decoder seq2seq models like M2M-100; absent in decoder-only models like Llama or Phi-3. Doubles the per-layer attention KV footprint on M2M-100.

Decoder-only. Transformer architecture consisting only of an autoregressive decoder, with no separate encoder. Llama, Phi, Qwen, GPT-style models. Compare with encoder-decoder seq2seq.

DRAM bandwidth. The single most important hardware constraint for LLM decode on Intel NPU. Lunar Lake's LPDDR5X-8533 provides 136.5 GB/s shared across CPU, iGPU, and NPU on a 128-bit on-package bus. No per-device quota is published.

Encoder-decoder seq2seq. Transformer architecture with separate encoder (processes input once) and decoder (autoregressive output). M2M-100, T5, BART, NLLB-200. The encoder fits NPU constraints well; the decoder does not.

FP8 (E4M3, E5M2). 8-bit floating-point formats. E4M3 has 4 exponent bits and 3 mantissa bits (wider range, less precision); E5M2 is the opposite. Native FP8 support is NPU 5 (Panther Lake) and later only.

GQA (Grouped-Query Attention). Attention variant where multiple query heads share a smaller set of key/value heads, reducing KV cache memory. Phi-3-mini uses GQA with 8 KV heads against 32 query heads. M2M-100 does not use GQA.

Greedy decoding. Always selecting the highest-probability next token; no sampling, no beam search. The only decoding mode supported on Intel NPU's `LLMPipeline`. Beam search and multinomial sampling require CPU or GPU.

Hexagon NPU. Qualcomm's NPU architecture, descended from the Hexagon QDSP6 phone DSP with bolted-on Tensor Accelerator and Vector eXtensions. Snapdragon X Elite reaches 45 TOPS.

iGPU. Integrated GPU. On Intel: Xe1 (Meteor Lake), Xe2 (Lunar Lake), Xe3 (Panther Lake). Not constrained by the NPU's bandwidth ceiling; typically 2× faster than NPU for LLM decode.

InferRequest. OpenVINO Runtime API primitive representing one in-flight inference. Supports async execution via callbacks; multiple `InferRequest` objects can target the same compiled model.

INT4-sym, group-size 128. The canonical NPU LLM weight quantization recipe: symmetric, 4-bit, with 128-element groups sharing a scale. `--sym --ratio 1.0 --group-size 128` in Optimum-Intel CLI.

ITL (Inter-Token Latency). The decode-phase per-token latency, measured starting from the second output token. Memory-bandwidth-bound on NPU. Llama 2 7B: ~54 ms/token; DeepSeek-Distill-8B INT4: ~163 ms/token.

KV cache. The keys and values from prior tokens' attention computations, retained across decode steps to avoid recomputation. Per-token footprint = batch × num_heads × head_dim × 2 (K+V) × dtype_bytes × num_layers.

LLMPipeline. The OpenVINO GenAI Python class for decoder-only LLMs. Internally selects `StaticLLMPipeline` for NPU. Exposes `start_chat()` / `finish_chat()` for stateful KV management. No equivalent exists for `OVModelForSeq2SeqLM`.

LPDDR5X-8533. The on-package DRAM technology on Lunar Lake. 8,533 MT/s × 128-bit bus / 8 = 136.5 GB/s platform bandwidth. Shared across CPU, iGPU, NPU. No private NPU DRAM.

Lunar Lake. Intel codename for Core Ultra Series 2 (September 2024). First Intel NPU at 48 TOPS (NPU 4). Single-tile integration of CPU + Xe2 + NPU on the compute die.

M2M-100. Facebook AI's many-to-many 100-language translation model (2020). Available in 418M, 1.2B, and 12B parameter sizes. MIT-licensed. The book's primary worked example because it stresses NPU constraints (full MHA, dynamic decode, cross-attention).

MAC (multiply-accumulate). The fundamental NPU operation. Each Intel NPU NCE has a 2,048 INT8 MAC/cycle array; total MACs scale with NCE count.

MAX_PROMPT_LEN. `LLMPipeline` NPU property bounding the maximum prefill input length. Default 1024 tokens; query at runtime for the validated ceiling on a given OpenVINO version and hardware.

Memory-side L4 cache. 8 MB cache on Lunar Lake's compute tile, shared across CPU/iGPU/NPU clients. Sits between the on-die units and the LPDDR5X memory controller.

Meteor Lake. Intel codename for Core Ultra Series 1 (December 2023). First Intel NPU (NPU 3720) at ~11.5 TOPS claimed / 9.5 TOPS measured. INT8 and FP16 only — no NF4 or FP8.

MHA (Multi-Head Attention). "Full" attention with `num_heads == num_kv_heads`. Distinct from GQA and MQA. M2M-100 uses MHA; modern decoder-only LLMs typically do not.

MLPerf Client. Industry-standard client-side ML benchmark suite. Version 0.6 includes Llama 2 7B; Intel's published Core Ultra Series 1 numbers (TTFT 1.09 s, 18.55 tok/s) come from MLPerf Client v0.6.

Movidius. The Irish startup Intel acquired in 2016. Source of the VPU/NPU architecture lineage; the SHAVE DSP descends directly from Movidius silicon.

NCE (Neural Compute Engine). The compute unit within an Intel NPU: one MAC array plus associated SHAVE DSPs. NPU 3720 has 2 NCEs; NPU 4 has 6; NPU 5 has 3 (each ~2× wider).

NF4. 4-bit "normal float" weight quantization format. Channel-wise only on Intel NPU. Supported on NPU 4 (Lunar Lake) and later; not on NPU 3720.

NLLB-200. Meta's 200-language successor to M2M-100. Same `M2M100ForConditionalGeneration` architecture class in Hugging Face Transformers. CC-BY-NC 4.0 license — not usable in commercial products. The book uses M2M-100 instead.

NNCF. Neural Network Compression Framework. Intel's open-source quantization toolkit, invoked by Optimum-Intel's `optimum-cli export openvino` workflow.

NPU. Neural Processing Unit. Domain-specific accelerator for dense matmul and fixed-function activations. Distinct from GPU (general-purpose shader array) and CPU (general-purpose scalar/vector).

OpenVINO. Intel's Apache-2.0-licensed cross-device inference toolkit. Targets CPU, iGPU, NPU, dGPU, and Gaudi from one intermediate representation. Currently at 2026.1 (May 2026).

OpenVINO IR. The OpenVINO intermediate representation: an `.xml` graph file plus a `.bin` weight file. Generated by `optimum-cli export openvino` or `mo` (legacy Model Optimizer).

Optimum-Intel. Hugging Face × Intel integration package providing `OVModel*` classes and the `optimum-cli export openvino` command. The canonical export path from PyTorch to OpenVINO IR.

OVMS (OpenVINO Model Server). Network-attached model server wrapping OpenVINO Runtime. The "process requests sequentially" note in NPU Stateful documentation is an OVMS scheduler policy, not an NPU hardware limit.

OVModelForSeq2SeqLM. Optimum-Intel class for encoder-decoder models. Used for M2M-100. Does not expose per-component `device_map` — splitting encoder and decoder across devices requires subclassing or driving the IR files via `core.compile_model()`.

Panther Lake. Intel codename for Core Ultra Series 3 (announced CES 2026). NPU 5 at ~50 TOPS, native FP8 support, programmable LUT for activations, Intel 18A process.

Phi Silica. Microsoft's Copilot+ on-device LLM. Architecture: CPU tokenizer + embedding + LM-head, NPU transformer blocks, CPU decode with N=64 KV sliding window. Published numbers (TTFT 230 ms, 20 tok/s) are on Snapdragon X — Intel-hardware Phi Silica numbers are unpublished.

Plan-then-execute. Reasoning architecture where one planning LLM call produces a fixed sequence of sub-tasks, and deterministic code executes them. NPU-friendly because the LLM cost is one prefill plus one decode, not a loop.

Prefill. The LLM inference phase that processes the input prompt before generating output. Compute-bound on NPU; the TTFT phase. Distinct from autoregressive decode.

Prefix caching. Cached KV for shared prompt prefixes (e.g., a system prompt reused across many requests). OpenVINO 2025.4+ feature. Enabled on NPU via `NPUW_LLM_ENABLE_PREFIX_CACHING:YES`.

PTQ (Post-Training Quantization). Quantizing a trained model without retraining. The default path on Intel NPU; invoked by NNCF through Optimum-Intel. Compare with QAT.

QAT (Quantization-Aware Training). Quantizing during training, with quantization simulated in the forward pass. Higher quality than PTQ for aggressive bit-widths, but requires the original training pipeline. Supported by NNCF but rarely used on NPU.

ReAct (Reason + Act). Iterative reasoning architecture where the model alternates between "Thought" tokens and tool calls (Yao et al. 2022). The dominant cloud-agent pattern. Infeasible on NPU at 5+ steps; recommended to run on iGPU instead.

Roofline model. Performance model relating arithmetic intensity (FLOPs per byte) to a ceiling that's the minimum of peak compute and peak bandwidth. The basis for the 6.10 tok/s → 24.4 GB/s analysis in Chapter 1.3.

SDPA (Scaled Dot-Product Attention). Fused attention operator in modern OpenVINO IR. Replaces the explicit MatMul + softmax + MatMul sequence. Required for the encoder-side MHA fusions added in OpenVINO 2025.2.

SHAVE / SHAVE-V. Streaming Hybrid Architecture Vector Engine. The VLIW DSP within an Intel NPU NCE, descended from Movidius. SHAVE-V (Lunar Lake and later) is ~4× wider than the original SHAVE.

Single-shot. Reasoning architecture: one prompt, one response, no loop. The NPU-native pattern. Translation, summarization, tone-rewrite all qualify.

Sliding-window KV. KV cache management that retains only the most recent N tokens. Phi Silica uses N=64. Trades recompute for bandwidth — a favorable trade on bandwidth-constrained NPU.

SoC. System on Chip. Intel Core Ultra is a heterogeneous SoC combining CPU, iGPU, NPU, media engines, and I/O in one package (single-die on Lunar/Panther Lake).

Static shape. Compile-time-determined tensor dimensions. The Intel NPU compiler requires fully static shapes for non-LLM workloads. Chunked prefill (2025.3+) softens this for decoder-only LLMs via `LLMPipeline`; there's no equivalent for seq2seq.

Stateful model. OpenVINO model with persistent internal variables (KV cache slots) that survive across `infer()` calls. The substrate for `LLMPipeline.start_chat()` / `finish_chat()`.

TextEmbeddingPipeline. OpenVINO 2026.1 GenAI pipeline for sentence-embedding models on NPU. Enables on-device RAG retrieval without leaving the agent process.

TOPS (Trillions of Operations Per Second). NPU's marketing-friendly throughput metric. Misleading in isolation — Intel NPU TOPS climbs from 11.5 to 50 across generations, but the bandwidth ceiling that bounds real LLM decode performance does not climb proportionally.

TTFT (Time-To-First-Token). The prefill-phase latency: time from prompt submission to the first output token. Compute-bound on NPU. Llama 2 7B at 128-token prompt: 1.09 seconds on Core Ultra Series 1 NPU.

Vector store. External long-term-memory database queried by retrieval against an embedding model. Lives on CPU/disk; the embedding model itself can run on NPU. Standard pattern for long-context agents.

VLIW (Very Long Instruction Word). The instruction format of the SHAVE DSP. Encodes multiple parallel operations per instruction word; relies on the compiler for scheduling.

Whisper. OpenAI's audio transcription model. One of OpenVINO GenAI's validated NPU pipelines (alongside `LLMPipeline` and `VLMPipeline`).

Xe2 / Xe3. Intel's iGPU microarchitecture generations on Lunar Lake (Xe2) and Panther Lake (Xe3). The iGPU sits next to the NPU on the same compute tile but has its own scheduling, separate from the NPU compiler path.

XDNA. AMD's NPU architecture, descended from Xilinx Versal AI Engine tiles arranged in a 2D spatial array. XDNA 2 in Ryzen AI 300 reaches 50 TOPS INT8 plus 50 TOPS Block FP16.

Previous: *Chapter 5: Real-World Case Studies* **Next:** *References*

Revision #1
Created 2026-05-12 19:22:05 UTC by Admin
Updated 2026-05-12 19:22:05 UTC by Admin