

Android On-Device AI Stack Cheat Sheet

Start with the task, not the SDK.

Option	Best for	What your team owns	Complexity	Device considerations
ML Kit task APIs	Common, well-defined vision and language tasks	Product integration, state, errors, and UX	Low	Often broad, but varies by API and delivery mode
ML Kit GenAI APIs + Gemini Nano	On-device summarization, rewriting, prompting, and short generated output	Prompting, evaluation, availability handling, and UX	Moderate	Supported devices only; runtime status must be checked
MediaPipe Tasks	Customizable vision, audio, gesture, and live media pipelines	Pipeline configuration, model choice, and performance	Moderate	Depends on the task, model, input mode, and hardware
LiteRT / LiteRT-LM	Custom ML or generative models that need deeper runtime control	Model selection, optimization, delivery, compatibility, and lifecycle	High	Depends heavily on the model and target devices

A Quick Decision Path

- Common, well-defined task?** (OCR, barcode, translation, face detection) → Start with an **ML Kit task-specific API**.
- Short on-device generated text or structured output?** (summarization, rewriting, guided insight) → Evaluate **ML Kit GenAI APIs powered by Gemini Nano**.
- Customizable vision, audio, gesture, or live media?** (pose, segmentation, gestures) → Evaluate **MediaPipe Tasks**.
- Bring your own model or deep runtime control?** Traditional ML → **LiteRT**. On-device LLM → **LiteRT-LM**.
- One path cannot cover all requirements?** Use a **hybrid architecture**, with each implementation behind the same product interface.