Gemini Nano

Gemini Nano lets you deliver rich generative AI experiences without needing a network connection or sending data to the cloud. On-device AI is a great solution for use-cases where low cost, and privacy safeguards are your primary concerns.

For on-device use-cases, you can take advantage of Google's Gemini Nano foundation model. Gemini Nano runs in Android's AICore system service (https://android-developers.googleblog.com/2023/12/a-new-foundation-for-ai-on-android.html), which leverages device hardware to enable low inference latency and keeps the model up-to-date.

ML Kit GenAI APIs

ML Kit's GenAI APIs harness the power of Gemini Nano to help your apps perform tasks. These APIs provide out-of-the-box quality for popular use cases through a high-level interface. The ML Kit GenAI APIs are built on top of AICore, an Android system service that enables on-device execution of GenAI foundation models to facilitate features such as enhanced app functionality and improved user privacy by processing data locally. Learn more (https://developers.google.com/ml-kit/genai).

Note: The ML Kit GenAI API Additional Terms of Service (https://developers.google.com/ml-kit/genai-terms) apply to the use of the GenAI APIs. Developers are solely responsible for the safety of their API client and their app's user experience.

Key features

The ML Kit GenAI APIs support the following features:

- **Prompt** (https://developers.google.com/ml-kit/genai/prompt/android): Generate text content based on a custom text-only or multimodal prompt.
- **Summarization** (https://developers.google.com/ml-kit/genai/summarization/android): Summarize articles or conversations as a bulleted list.

- **Proofreading** (https://developers.google.com/ml-kit/genai/proofreading/android): Proofread short chat messages.
- **Rewriting** (https://developers.google.com/ml-kit/genai/rewriting/android): Rewrite short chat messages in different tones or styles.
- **Image Description** (https://developers.google.com/ml-kit/genai/image-description/android): Generate a short description of a given image.
- **Speech Recognition** (https://developers.google.com/ml-kit/genai/speech-recognition/android): Transcribe spoken audio to text.

Architecture through AICore

As a system-level module, you access AICore through a series of APIs in order to run inference on-device. In addition, AICore has several built-in safety features, ensuring a thorough evaluation against our safety filters. The following diagram outlines how an app accesses AICore to run Gemini Nano on-device.

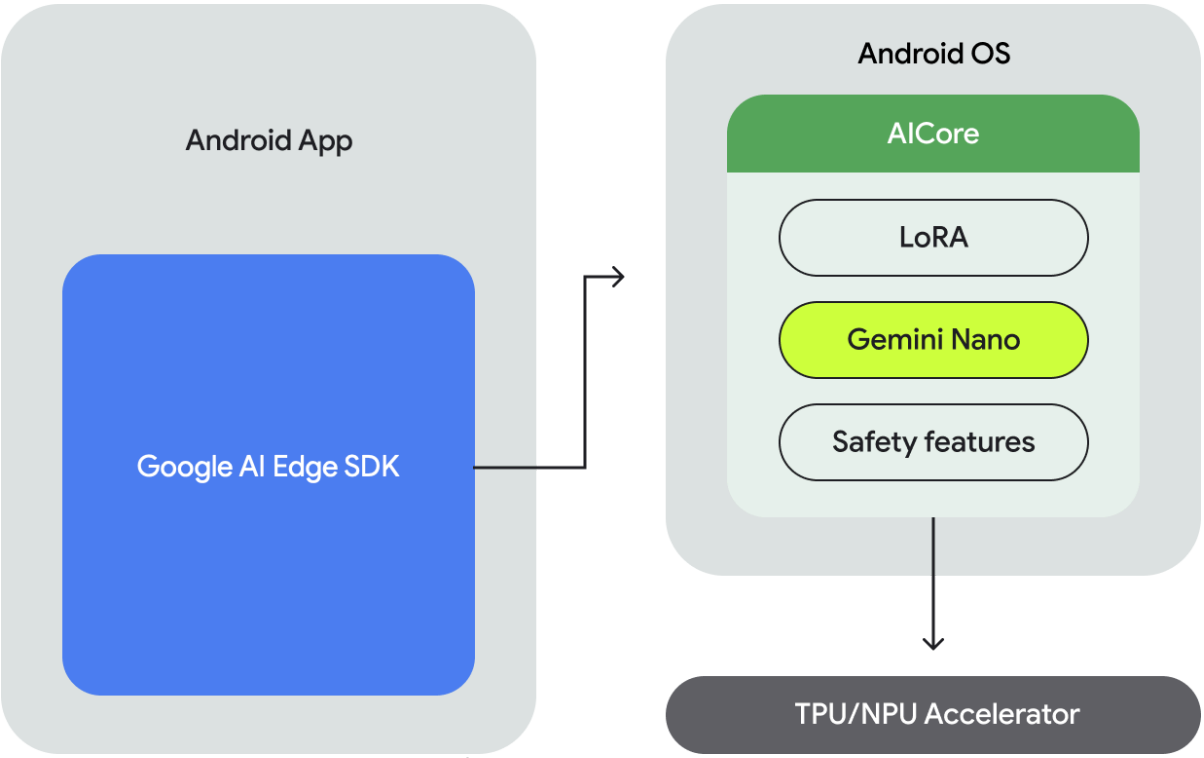


Figure 1. *AICore serves as the interface between your app and the Gemini Nano model, managing model updates and safety while leveraging on-device hardware.*

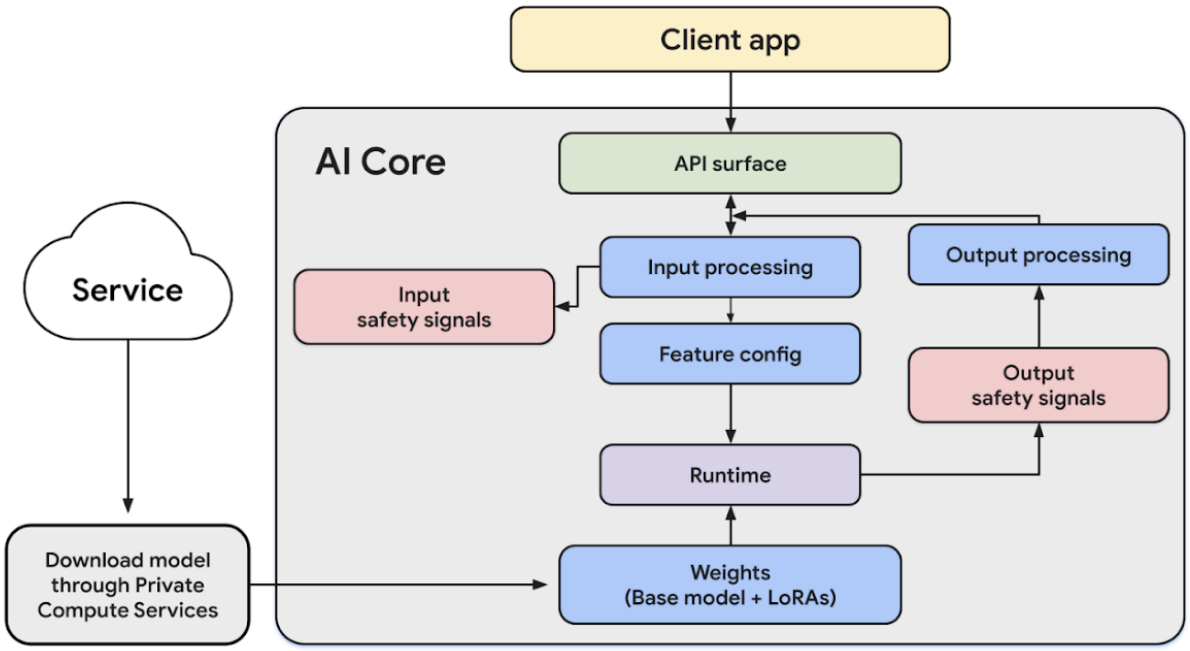
Keep user data private and secure

On-device generative AI executes prompts locally, eliminating server calls. While this removes network latency, inference speed depends on device hardware. This approach enhances privacy by keeping sensitive data on the device, enables offline functionality, and reduces inference costs.

AICore adheres to the Private Compute Core (https://arxiv.org/abs/2209.10317) principles, with the following key characteristics:

- **Restricted Package Binding**: AICore is isolated from most other packages, with limited exceptions for specific system packages. Any modifications to this allowed list can only occur during a full Android OTA update.
- **Indirect Internet Access**: AICore does not have direct internet access. All internet requests, including model downloads, are routed through the open-source Private Compute Services (https://github.com/google/private-compute-services) companion APK. APIs within Private Compute Services must explicitly demonstrate their privacy-centric nature.

Additionally, AICore is built to isolate each request and doesn't store any record of the input data or the resulting outputs after processing them to protect user privacy. Read the blog post An Introduction to Privacy and Safety for Gemini Nano (https://android-developers.googleblog.com/2024/10/introduction-to-privacy-and-safety-gemini-nano.html) to learn more.



Google

Figure 2. *The AICore architecture manages input and output safety, request processing, and model weights to provide a secure environment for on-device AI.*

Benefits of accessing AI foundation models with AICore

AICore enables the Android OS to provide and manage AI foundation models. This significantly reduces the cost of using these large models in your app, principally due to the following:

- **Ease of deployment**: AICore manages the distribution of Gemini Nano and handles future updates. You don't need to worry about downloading or updating large models over the network, nor impact on your app's disk and runtime memory budget.
- **Accelerated inference**: AICore leverages on-device hardware to accelerate inference. Your app gets the best performance on each device, and you don't need to worry about the underlying hardware interfaces.