# Llama-2 Report

Llama-2 is a LLM (large language model) that has a pre-built foundation of data to work with and is open for commercial and private use for free. It is capable of understanding and generating natural language and other types of content to perform a wide range of tasks. Llama-2 is the successor to the previous non-commercial model Llama-1 having much more training and information to work with.

The base model of Llama-2 has already been fine-tuned for any specific use-case such as summarising text, generating images, or if a chatbot is needed Llama-2 comes pre-built with optimisation.

While Llama-2 itself is not open source, a few developers have made open sourced models based off the freely available model build for more specific use. These include:

* ***Alpaca:***A version of Llama for instruction-following by Stanford University researchers.
* ***Vicuna***: A chat assistant model from LMSYS Org, which performed better than Alpaca.
* ***Orca***: A fine-tuned version of Llama 2, trained by Microsoft using a teacher-student method, where a more powerful LLM is used to create examples of useful behaviour for the smaller model to follow.
* ***WizardLM***: Fine-tuned using Evol-Instruct, a method to create large amounts of instruction data using LLMs.

Just as the way it had been intended Llama-2 is much more of a foundation for more purposeful models to be built on. This, in turn provides a great help toward android developers. Who don’t have to go toward paid closed-sourced models like GPT-4 or Google’s PaLM 2.

It also has the advantage of being small enough to run smoothly on mobile applications, as smaller variants of Llama-2 (7B and 13B variants) can run even when processing power may be limited. The model also has the advantage of much more privacy compared to commercial models. Businesses can run local model instances without needing to spend large sums of money in infrastructure or cloud computing. Running a local model makes sure that code, training data is used for the model without being loaded onto the commercial server or being used in future training of other closed-sourced models.

Llama-2 is readily made available in many ways:

* Llama-2-13B-chat and Llama-2-70B-chat can be found in WatsonX, among other foundation models.
* Other model weights and starting code can also be found and downloaded from GitHub.
* Llama 2 is also available through open-source providers like Hugging Face and enterprise providers like Microsoft Azure and Amazon SageMaker and Bedrock.

For mobile developers specifically Llama-2 can be used in a few specific ways.

* It can be used as a chatbot that analyses and responds to text.
* Summarise english texts without losing critical information.
* Create poems, stories, blog posts. When given a prompt.
* Track time spent on certain activities. Improving time-consumption.
* Report and analytics, Llama-2 is able to generate good decision based on the dataset it is given.
* Able to manage & organise files, as well as being able to access any important file eliminating the need for most platforms and ensuring files are made readily available.

Overall, Llama-2 is a great language model for smaller startups or individual use. While it may lack computing power compared to other larger models it still maintains its own niche of useful features at the cost of nothing.