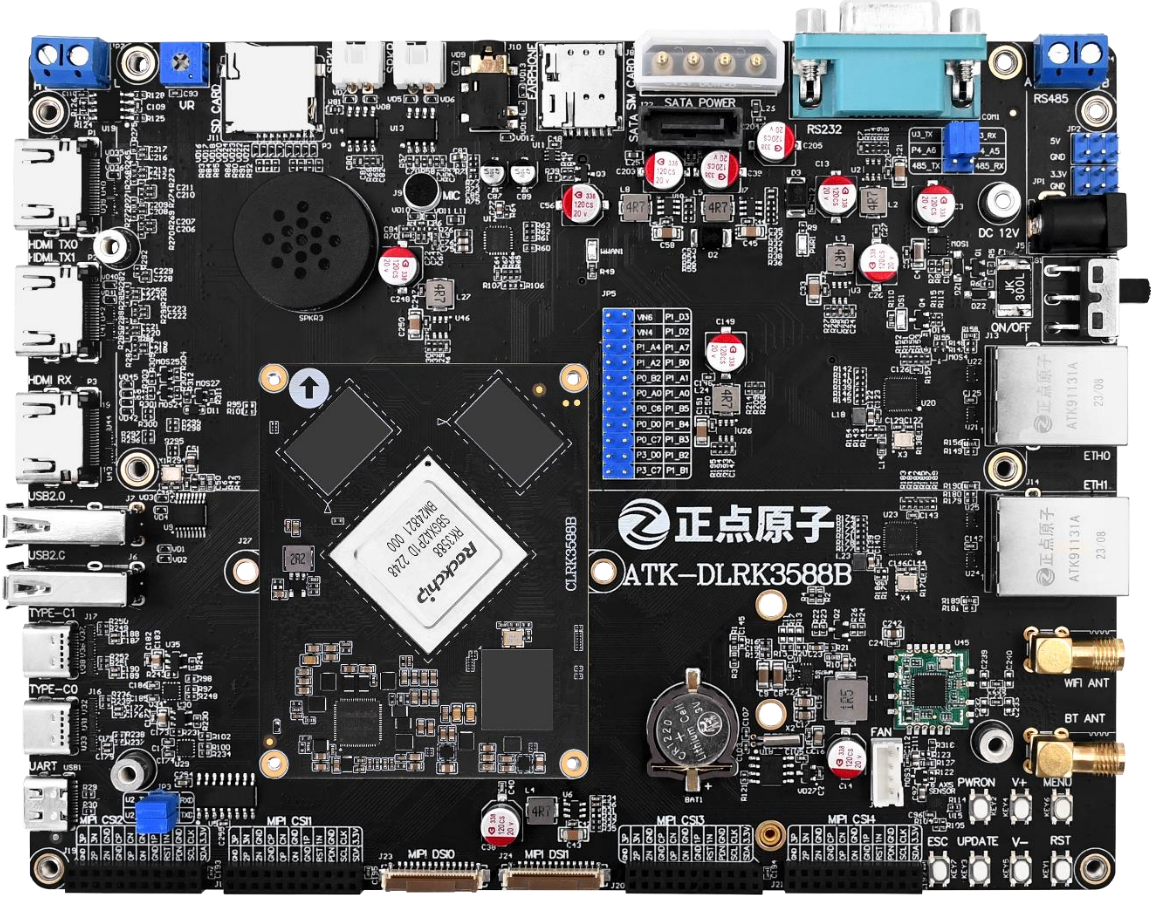
ATK-DLRK3588

Development Board AI Test Manual

V1.0



1. Shopping:

TMALL：https://zhengdianyuanzi.tmall.com

TAOBAO：https://openedv.taobao.com

2. Download

Address：http://www.openedv.com/docs/index.html

3. FAE

Website : www.alientek.com

Forum : http://www.openedv.com/forum.php

Videos : www.yuanzige.com

Fax : +86 - 20 - 36773971

Phone : +86 - 20 - 38271790

Disclaimer

The product specifications and instructions mentioned in this document are for reference only and subject to update without prior notice; Unless otherwise agreed, this document is intended as a product guide only, and none of the representations made herein constitutes a warranty of any kind. The copyright of this document belongs to Guangzhou Xingyi Electronic Technology Co., LTD. Without the written permission of the company, any unit or individual shall not be used for profit-making purposes in any way of dissemination.

In order to get the latest version of product information, please regularly visit the download center or contact the customer service of Taobao ALIENTEK flagship store. Thank you for your tolerance and support.

Revision History：

Version	Version Update Notes	Responsible person	Proofreading	Date
V1.0	release officially	ALIENTEK	ALIENTEK	2023.06.15

Catalogue

Chapter 1. Deployment and testing of the rkllm model

1.1 Introduction to the RKLLM Model

The RKLLM-Toolkit provides functions for model conversion and quantization. As one of the core functions of the RKLLM-Toolkit, it enables users to convert large language models in Hugging Face format into RKLLM models, thereby loading and running the RKLLM model on the Rockchip NPU.

Background and Positioning:

The RKLLM software stack is designed to help users quickly deploy AI models onto Rockchip chips, especially models such as RK3576/RK3588.

Core Components:

- **RKLLM-Toolkit:** This is a development kit that enables users to perform operations such as quantization and conversion of large language models on the computer. Through the Python interface, users can easily convert large language models in Hugging Face format (such as LLaMA, Qwen/Qwen2, Phi2, etc.) into RKLLM models.
- **RKLLM Runtime:** This component is mainly responsible for loading the RKLLM models obtained by the RKLLM-Toolkit on the RK3576/RK3588 board, and implementing the inference of the RKLLM models on the Rockchip NPU by calling the NPU driver. Main Function:
- **Model Conversion:** Supports conversion of various mainstream large language model formats to ensure that the model can be loaded and used on the Rockchip NPU platform.
- **Quantization Function:** Allows users to quantize floating-point models into fixed-point models to improve running efficiency and reduce resource usage. Currently supported quantization types include w4a16 and w8a8.
- **Inference Settings:** When inferring the RKLLM model, users can customize the inference parameter settings, define different text generation methods, and obtain the model's inference results through predefined callback functions.

Development process:

The overall development process of RKLLM mainly consists of two steps: model conversion and board-end deployment and running. First, the large language model in Hugging Face format provided by the user will be converted into the RKLLM model; then, the converted model will be deployed to the RK3576/RK3588 board-end for inference.

Hardware support:

The RKLLM model has been specifically optimized for Rockchip's RK3576/RK3588 and other models, and can fully utilize the performance and characteristics of these chips.

Application scenarios:

The RKLLM model is suitable for various scenarios that require efficient and low-power AI large model inference, such as intelligent voice assistants, natural language processing, text generation, etc.

The RKLLM model is a powerful and user-friendly AI large model deployment solution. It combines the advantages of Rockchip chips with the flexibility of the RKLLM software stack, providing users with efficient and low-power AI inference capabilities. The following figure shows the framework of the RKLLM-toolkit.

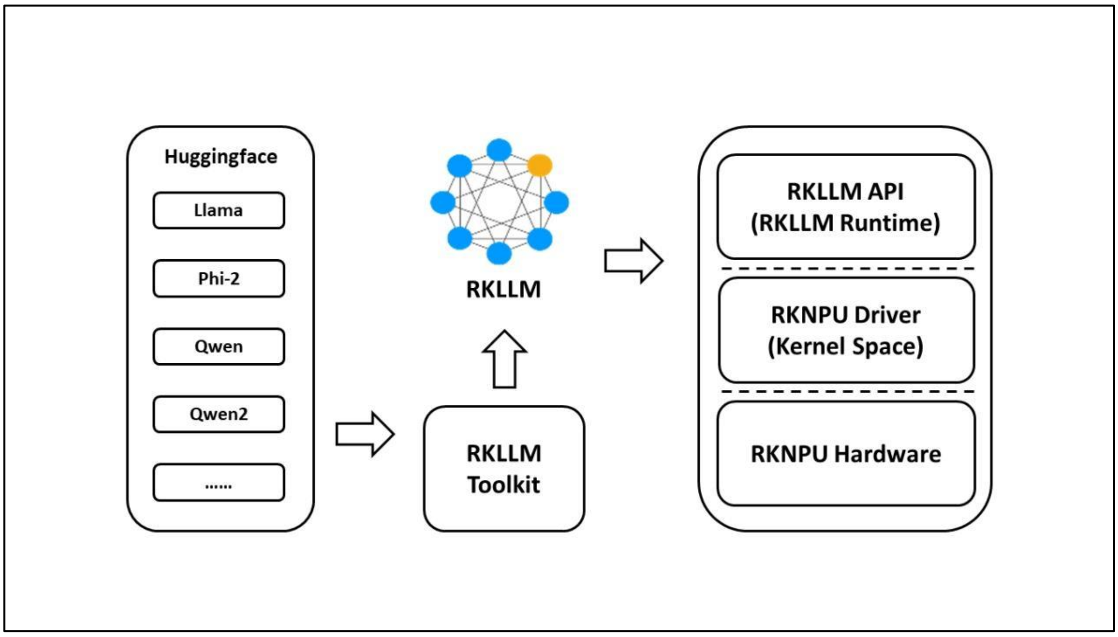


Figure 1.1-1 The framework of the RKLLM-toolkit

In this test, the Qwen model was used for the testing. The Qwen model is an advanced language model launched by the Tongyi Qianwan project. It aims to achieve knowledge understanding and task execution in a wide range of fields through powerful natural language processing capabilities and cross-modal interaction. Here is a brief introduction to the Qwen model:

Model architecture and scale:

The Qwen basic model is the core foundation of the Tongyi Qianwen project, and it is constructed based on the Transformer architecture.

Currently, Qwen has developed five different-sized versions, including Qwen-1.8B, Qwen-7B, Qwen-14B, Qwen-72B, etc. These models are all open-sourced for the public.

These models have undergone sufficient training with 2 to 3 trillion tokens, ensuring a profound knowledge reserve and wide adaptability when handling natural language tasks.

Multilingual capabilities and context handling:

The Qwen model has multilingual processing capabilities, performing particularly well in English and Chinese, and can effectively handle other languages such as Spanish, French, and Japanese.

To enhance multilingual processing efficiency, Qwen employs a proprietary efficient tokenizer and pays special attention to the expansion of context length during the pre-training stage. The open-source Qwen model typically supports context windows of up to 32K word tokens, ensuring the model maintains accuracy and coherence when processing long texts.

Performance and competitiveness:

The evaluation benchmark shows that the open-source Qwen-72B model and its larger private version are on par with industry-leading models such as Llama 2, GPT-3.5, and GPT-4 in terms of performance. This confirms the strong competitiveness of the Qwen base model in terms of original language understanding and generation capabilities.

Model Optimization:

Building upon the base model, Yiting Qianwen has undergone deep optimization of Qwen through post-training techniques such as SFT (Supervised Fine-Tuning) and RLHF (Reinforcement Learning with Human Feedback) to achieve higher-level alignment and interactivity.

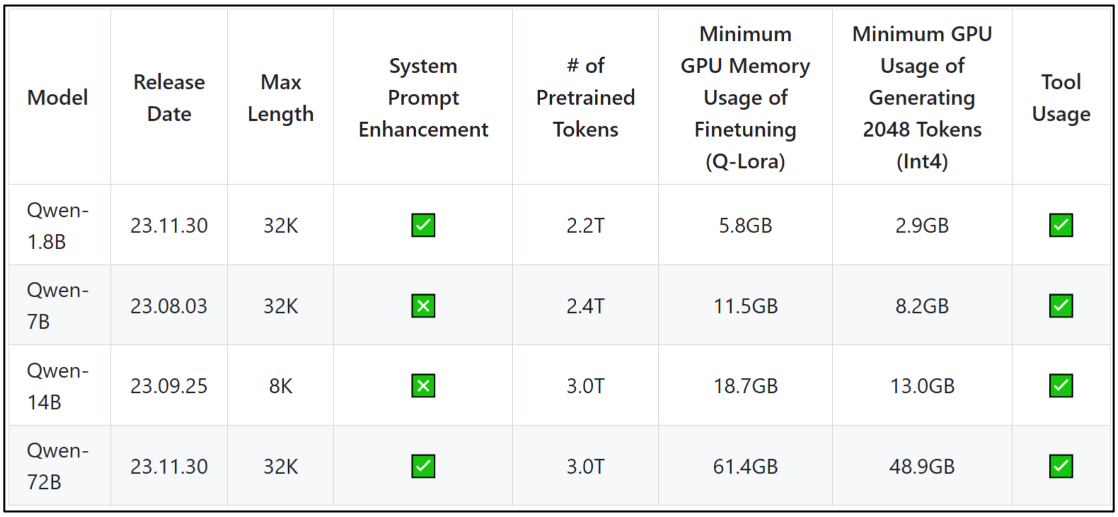
Multimodal Capabilities:

Qwen-VL is a multimodal large model in the Qwen series. It supports input in the form of images, text, and bounding boxes, and outputs in the form of text and bounding boxes. Qwen-VL achieves the best results in the standard English evaluations of the four major types of multimodal tasks, with a comparable model size, and naturally supports multilingual conversations in English, Chinese, and other languages.

Latest Developments:

Alibaba Cloud has recently released the world's most powerful open-source model, Qwen2-72B, which outperforms the most powerful open-source model in the United States, Llama3-70B, and also surpasses many closed-source large models in China. The Qwen2 series of models have significantly enhanced capabilities in code, mathematics, reasoning, instruction compliance, and multi-language understanding, and have won championships in multiple international authoritative evaluations.

In summary, the Qwen model is a powerful and high-performing language model with multi-language processing, long text processing, and multimodal interaction capabilities, and has strong competitiveness in the industry. With the continuous advancement of technology, the Qwen model will continue to be optimized and developed to provide users with higher-quality AI services.



Model	Release Date	Max Length	System Prompt Enhancement	# of Pretrained Tokens	Minimum GPU Memory Usage of Finetuning (Q-Lora)	Minimum GPU Usage of Generating 2048 Tokens (Int4)	Tool Usage
Qwen-1.8B	23.11.30	32K	☑	2.2T	5.8GB	2.9GB	☑
Qwen-7B	23.08.03	32K	☒	2.4T	11.5GB	8.2GB	☑
Qwen-14B	23.09.25	8K	☒	3.0T	18.7GB	13.0GB	☑
Qwen-72B	23.11.30	32K	☑	3.0T	61.4GB	48.9GB	☑

Figure 1.1-2 Qwen performance information

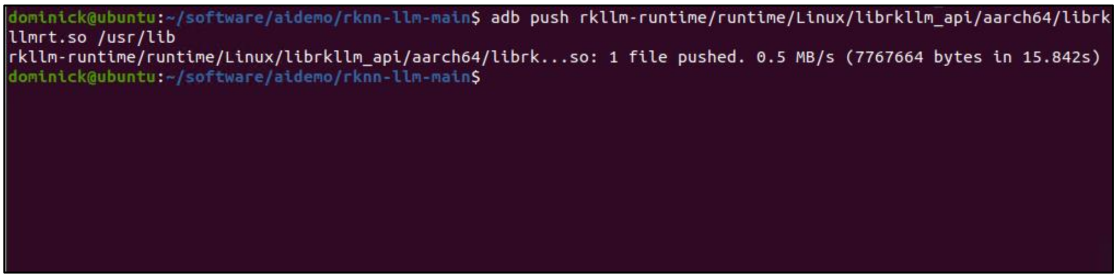
Qwen Open Source Website: https://github.com/QwenLM/Qwen

Hugging Face repository: https://huggingface.co/

1.2 rkllm model deployment test

Before testing the large model, we need to update the inference library librkllmrt.so of the large model to the development board. First, copy the files 01 - Basic Data → 01_codes → 01_AI_Routine → 19_rknn_llm → rknn-llm-main.zip in the directory 01 of the A disk of the development board to the Ubuntu system and decompress them. Enter the corresponding directory after decompression and open the terminal. Execute the following commands to push the inference library librkllmrt.so to the /usr/lib directory of the development board.

```
adb push rkllm-runtime/runtime/Linux/librkllm_api/aarch64/librkllmrt.so /usr/lib
```



```
dominick@ubuntu:~/software/aidemo/rknn-llm-main$ adb push rkllm-runtime/runtime/Linux/librkllm_api/aarch64/librk
llmrt.so /usr/lib
rkllm-runtime/runtime/Linux/librkllm_api/aarch64/librk...so: 1 file pushed. 0.5 MB/s (7767664 bytes in 15.842s)
dominick@ubuntu:~/software/aidemo/rknn-llm-main$
```

Figure 1.2-1 Push the librkllmrt.so library to the development board.

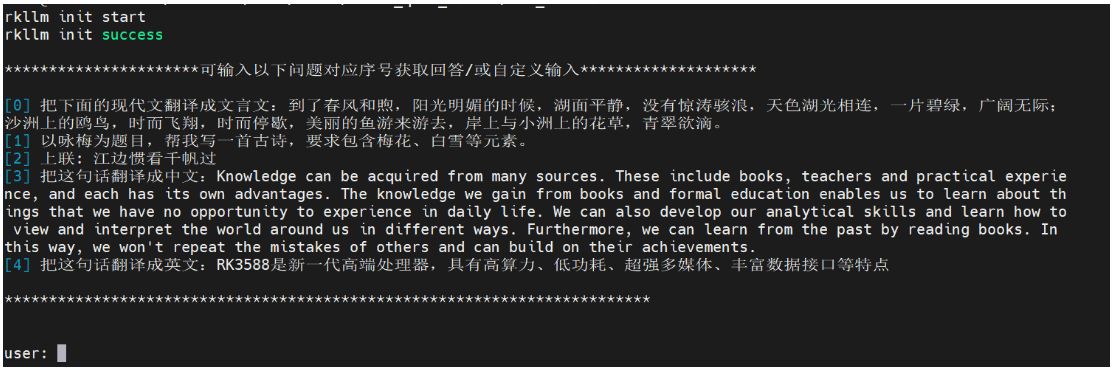
Since the memory required for the conversion model is relatively large, there is a relatively high probability of conversion failure. Therefore, in the development board materials, a converted qwen-1.8b rkllm model has been provided for everyone to test. Copy the rkllm_qwen_1.8.zip in the 01_ codes/AI routine/sourc code/19_rknn_llm of the A disk of the development board to the ubuntu. Through adb (or you can use network transmission which will be faster), push it to the userdata/aidemo directory of the development board. Execute the following commands in the ubuntu terminal.

```
adb push rkllm_qwen_1.8.zip /userdata/aidemo
```

Enter the /userdata/aidemo directory on the development board's serial port terminal and execute the following command.

```
cd /userdata/aidemo
unzip rkllm_qwen_1.8.zip
chmod +x llm_start.sh
./llm_start.sh
```

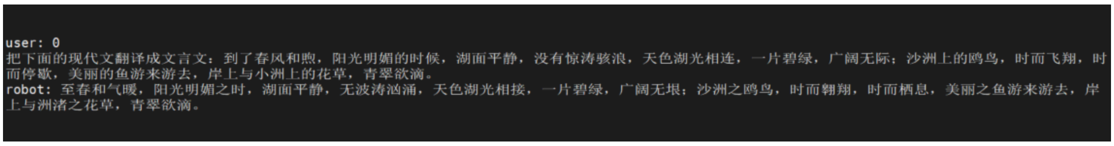
Loading the large model will be a bit slow. Please wait for a while. Once it is successfully loaded, the result will be as shown in the picture below.



```
rkllm init start
rkllm init success

*********************可输入以下问题对应序号获取回答/或自定义输入*******************

[0] 把下面的现代文翻译成文言文：到了春风和煦，阳光明媚的时候，湖面平静，没有惊涛骇浪，天色湖光相连，一片碧绿，广阔无际；
沙洲上的鸥鸟，时而飞翔，时而停歇，美丽的鱼游来游去，岸上与小洲上的花草，青翠欲滴。
[1] 以咏梅为题目，帮我写一首古诗，要求包含梅花、白雪等元素。
[2] 上联：江边惯看千帆过
[3] 把这句话翻译成中文：Knowledge can be acquired from many sources. These include books, teachers and practical experie
nce, and each has its own advantages. The knowledge we gain from books and formal education enables us to learn about th
ings that we have no opportunity to experience in daily life. We can also develop our analytical skills and learn how to
 view and interpret the world around us in different ways. Furthermore, we can learn from the past by reading books. In
this way, we won't repeat the mistakes of others and can build on their achievements.
[4] 把这句话翻译成英文：RK3588是新一代高端处理器，具有高算力、低功耗、超强多媒体、丰富数据接口等特点

************************************************************************


user:
```

Figure 1.2-2 Successfully loaded the rkllm model of qwen.

You can press Enter after entering the number 0 to obtain an answer based on the preset question, or you can input a Chinese or English question yourself and press Enter to get the answer.



```
user: 0
把下面的现代文翻译成文言文：到了春风和煦，阳光明媚的时候，湖面平静，没有惊涛骇浪，天色湖光相连，一片碧绿，广阔无际；沙洲上的鸥鸟，时而飞翔，时
而停歇，美丽的鱼游来游去，岸上与小洲上的花草，青翠欲滴。
robot: 至春和气暖，阳光明媚之时，湖面平静，无波涛汹涌，天色湖光相接，一片碧绿，广阔无垠；沙洲之鸥鸟，时而翱翔，时而栖息，美丽之鱼游来游去，岸
上与洲渚之花草，青翠欲滴。
```

Figure 1.2-3 The qwen's rkllm model has been successfully loaded.

Through testing, it can be observed that some responses are quite comprehensive, while others are of relatively ordinary quality. This is related to the model. In the future, we can consider using a larger-scale model for testing.