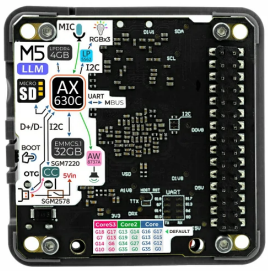
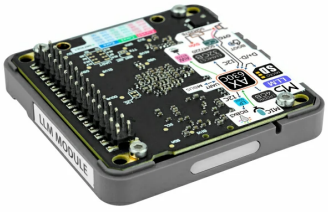
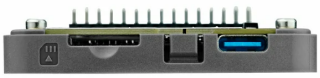
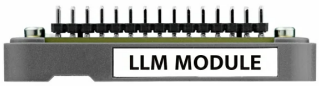
Module LLM

SKU:M140









Description

Module LLM is an integrated offline Large Language Model (LLM) inference module designed for terminal devices that require efficient and intelligent interaction. Whether for smart homes, voice assistants, or industrial control, Module LLM provides a smooth and natural AI experience without relying on the cloud, ensuring privacy and stability. Integrated with the **StackFlow** framework and **Arduino/UiFlow** libraries, smart features can be easily implemented with just a few lines of code.

Powered by the advanced **AX630C** SoC processor, it integrates a 3.2 TOPs high-efficiency NPU with native support for Transformer models, handling complex AI tasks with ease. Equipped with **4GB LPDDR4** memory (1GB available for user applications, 3GB dedicated to hardware acceleration) and **32GB eMMC** storage, it supports parallel loading and sequential inference of multiple models, ensuring smooth multitasking. The main chip's runtime power consumption of approximately 1.5W, making it highly efficient and suitable for long-term operation.

It features a built-in microphone, speaker, TF storage card, **USB OTG**, and RGB status light, meeting diverse application needs with support for voice interaction and data transfer. The module offers flexible expansion: the onboard SD card slot supports cold/hot firmware upgrades, and the **UART** communication interface simplifies connection and debugging, ensuring continuous optimization and expansion of module functionality. The USB port supports master-slave auto-switching, serving as both a debugging port and allowing connection to additional USB devices like cameras. Users can purchase the LLM debugging kit to add a 100 Mbps Ethernet port and kernel serial port, using it as an SBC.

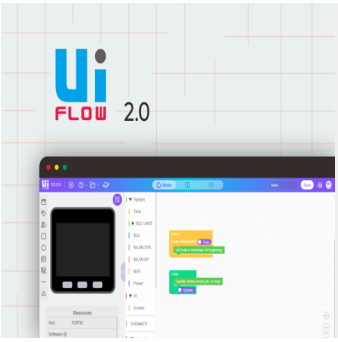
The module is compatible with multiple models and comes pre-installed with the **Qwen2.5-0.5B** language model. It features **KWS** (wake word), **ASR** (speech recognition), **LLM** (large language model), and **TTS** (text-to-speech) functionalities, with support for standalone calls or **pipeline** automatic transfer for convenient development. Future support includes Qwen2.5-1.5B, Llama3.2-1B, and InternVL2-1B models, allowing hot model updates to keep up with community trends and accommodate various complex AI tasks. Vision recognition capabilities include support for CLIP, YoloWorld, and future updates for DepthAnything, SegmentAnything, and other advanced models to enhance intelligent recognition and analysis.

Plug and play with **M5 hosts**, Module LLM offers an easy-to-use AI interaction experience. Users can quickly integrate it into existing smart devices without complex settings, enabling

smart functionality and improving device intelligence. This product is suitable for offline voice assistants, text-to-speech conversion, smart home control, interactive robots, and more.

Arduino IDE

This tutorial introduces how to program and control the Module LLM device using the Arduino IDE



UIFlow2.0

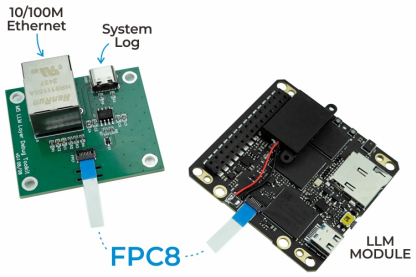
This tutorial introduces how to control the Module LLM device using the UIFlow2.0 graphical programming platform

Features

- Offline inference, 3.2T@INT8 precision computing power
- Integrated KWS (wake word), ASR (speech recognition), LLM (large language model), TTS (text-to-speech generation)
- Multi-model parallel processing
- Onboard 32GB eMMC storage and 4GB LPDDR4 memory
- Onboard microphone and speaker
- Serial communication
- SD card firmware upgrade
- Supports ADB debugging
- RGB indicator light
- Built-in Ubuntu system
- Supports OTG functionality
- Compatible with Arduino/UIFlow

Includes

- 1x Module LLM



> Debug board included with the product (limited to initial release only)

Applications

- Offline voice assistants
- Text-to-speech conversion
- Smart home control
- Interactive robots

Specifications

Specifications	Parameter
Processor SoC	AX630C@Dual Cortex A53 1.2 GHz
	MAX 12.8 TOPS @INT4 and 3.2 TOPS @INT8

	MAX:12.8 TOPS @INT4 and 3.2 TOPS @INT8
Memory	4GB LPDDR4 (1GB system memory + 3GB dedicated for hardware acceleration)
Storage	32GB eMMC5.1
Communication	Serial communication default baud rate 115200@8N1 (adjustable)
Microphone	MSM421A
Audio Driver	AW8737
Speaker	8Ω@1W, Size:2014 cavity speaker
Built-in Units	KWS (wake word), ASR (speech recognition), LLM (large language model), TTS (text-to-speech)
RGB Light	3x RGB LED@2020 driven by LP5562 (status indication)
Power	Idle: 5V@0.5W, Full load: 5V@1.5W
Button	For entering download mode for firmware upgrade
Upgrade Port	SD card / Type-C port
Working Temp	0-40°C
Product Size	54*54*13mm
Packaging Size	133*95*16mm
Product	17.4g

Weight	
Packaging Weight	32.0g

Related Links

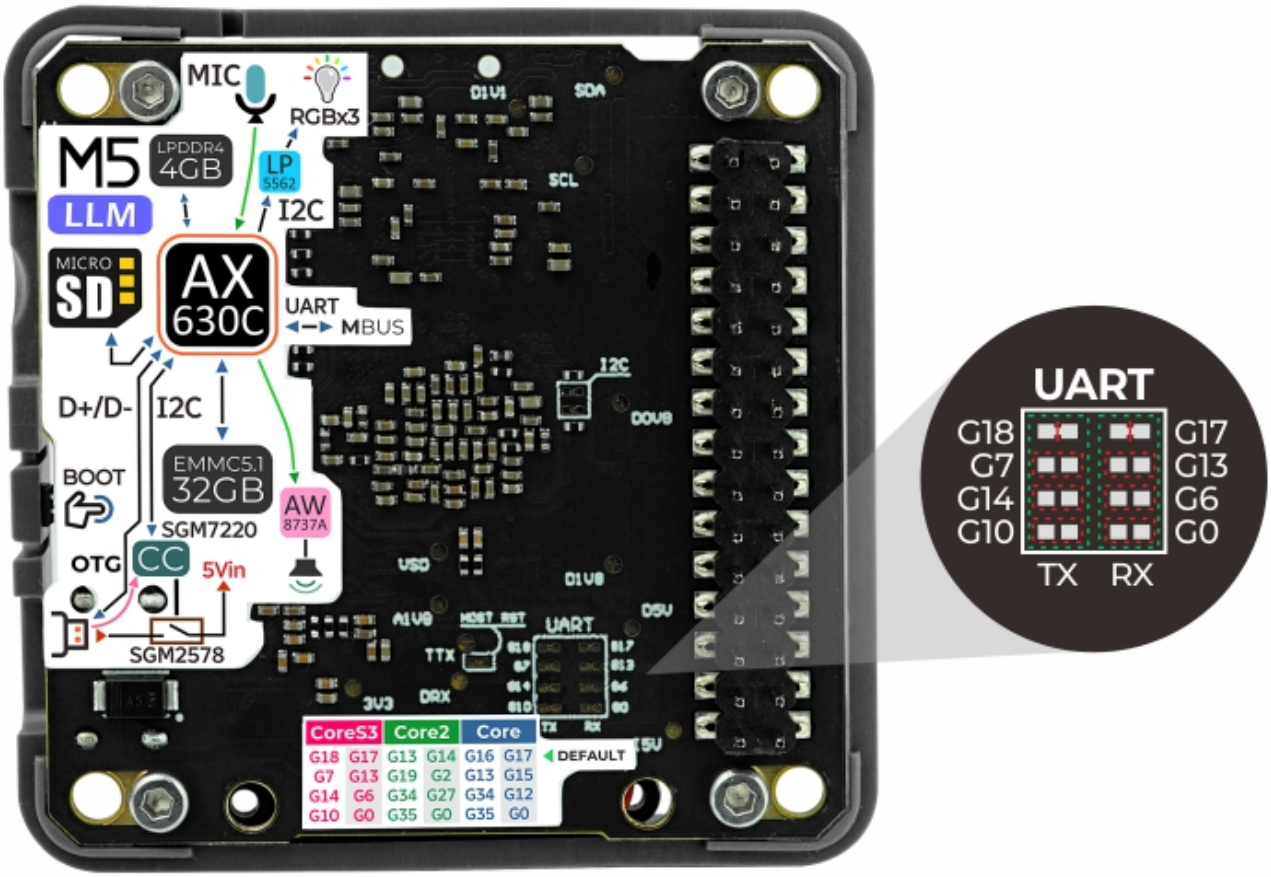
- AX630C

PinMap

Module LLM	RXD	TXD
Core (Basic)	G16	G17
Core2	G13	G14
CoreS3	G18	G17

LLM Module Pin Switching

LLM Module has reserved soldering pads for pin switching. In cases of pin multiplexing conflicts, the PCB trace can be cut and reconnected to other sets of pins.



Tx Rx

CoreS3		Core2		Core		
G18	G17	G13	G14	G16	G17	DEFAULT
G7	G13	G19	G2	G13	G15	
G14	G6	G34	G27	G34	G12	
G10	G0	G35	G0	G35	G0	

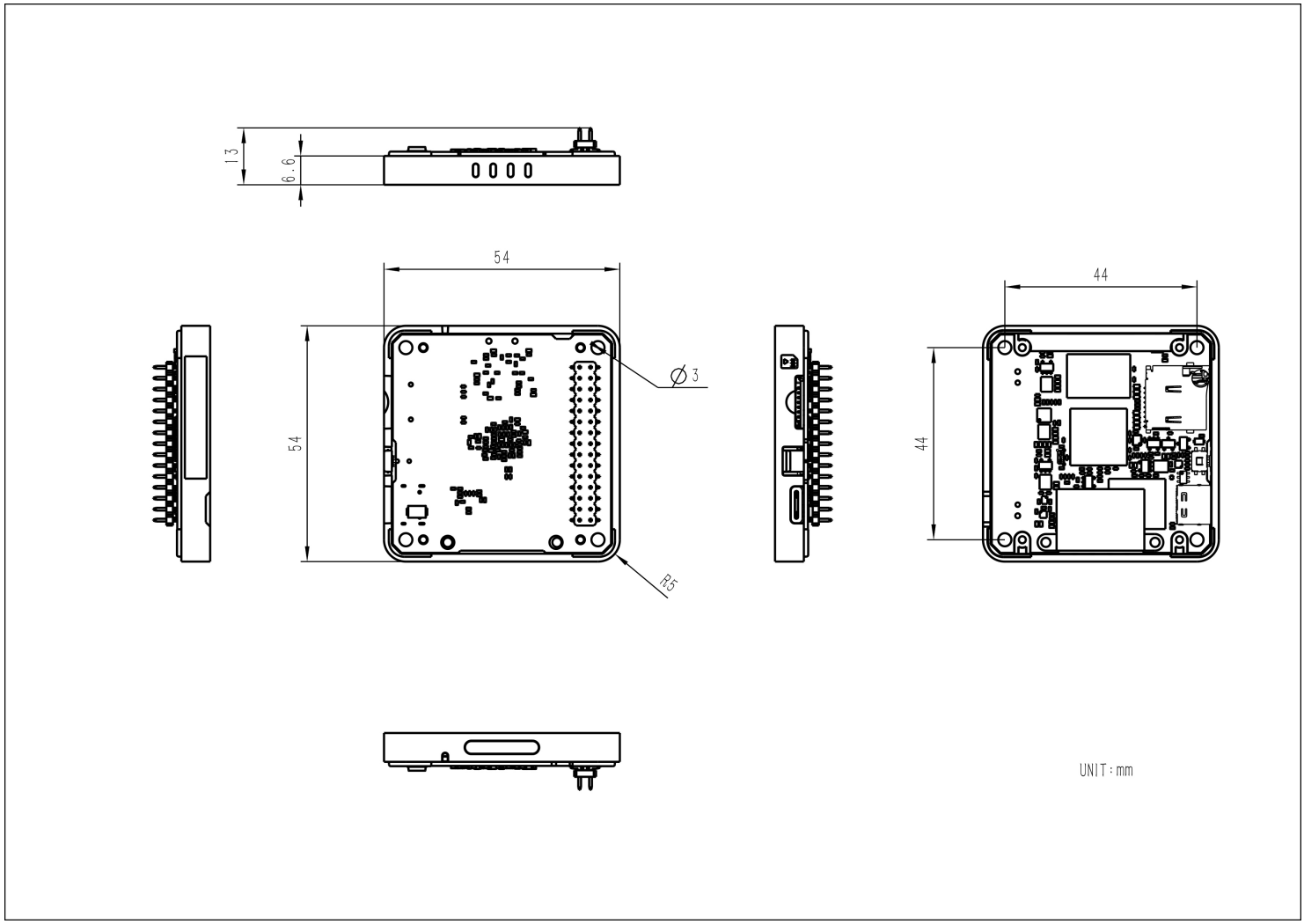
Taking `CoreS3` as an example, the first column (left green box) is the TX pin for serial communication, where users can choose one out of four options as needed (from top to bottom, the pins are G18, G7, G14, and G10). The default is set to IO18. To switch to a different pin, cut the connection on the solder pad (at the red line) — it's recommended to use a blade for this — and then connect to one of the three remaining pins below. The second column (right green box) is for RX pin selection, and, as with the TX pin, it also allows a choice of one out of four options.

allows a choice of one out of four options.

Schematics

- Module LLM Schematic Download

Model Size



Indicator Light

- LLM Module working status indicator:
  - Red: Device is initializing
  - Green: Device initialized successfully
- LLM Module upgrade status indicator:
  - Blue flashing: Application package updating
  - Red:Application package update failed
  - Green: Application package updated successfully

> **Note on Model Replacement**
>
> The LLM Module supports models in a proprietary format requiring special processing to function properly. Therefore, existing models on the market cannot be used directly.

UIFlow

- LLM Module UIFlow2.0 Quick Start
- LLM Module UIFlow2.0 API

Arduino

- LLM Module Arduino Quick Start
- LLM Module Arduino Library
- LLM Module Arduino Library API

Firmware Update

- LLM Module Image Update
- LLM Module ADB Tools
- LLM Module Factory firmware

Development Framework

- StackFlow

Development Resources

- AX630C Databrief
- LLM Module JSON API Documentation
- LLM Module AX630C API UIFlow 1.0 Docs
- LLM Module AX620E MSP/Sample
- LLM Module Linux Kernel 4.19.125-head
- AXERA LLM Example
- AXERA CV Example
- LLM Module Large Model Compilation Guide

Video

- Module LLM product introduction and example showcase

Module_LLM_Video.mp4

AI Benchmark Comparison