



Arise2030 本地 LLM 部署指南

2026 年 6 月

法律事项

本文档中呈现的信息属于格兰菲智能科技股份有限公司和/或其附属公司（以下统称为“格兰菲”）所有，且格兰菲保留不经通知随时对本文档信息或对本文档所述任何产品和服务做出修改的权利。格兰菲不担保信息、文本、图案、链接或本文档内所含其他项目的准确性或完整性。本文档内容仅供参考，格兰菲不对本文档所述产品的可销售性、所有权、不侵犯知识产权、准确性、完整性、稳定性或特定用途适用性做任何暗示担保、保证。格兰菲可不经通知随时对本文档或本文档所述产品或服务做出更改，但不承诺因此更新本文档。

非经格兰菲事先书面许可，任何实体或个人均不得获得本文档的副本，且无权以任何方式处理本文档，包括但不限于使用、复制、修改、合并、出版、发行、销售或传播本文档的部分或全部。在任何情况下，格兰菲均不对因本文档引起的、由本文档造成的、或与之相关的任何索赔、损害或其他责任负责。格兰菲保留自行决定随时更改、修改、添加或删除本文档的部分或全部的权利。

格兰菲对本文档享有最终解释权。

目录

法律事项.....	2
目录.....	3
修订记录.....	4
1 概述/引言.....	5
2 运行环境要求.....	5
2.1 操作系统.....	5
2.2 Python 环境.....	5
3 驱动安装与验证.....	5
3.1 安装 Arise2030 驱动.....	5
3.2 安装 AriseLLM.....	6
4 模型量化与运行.....	6
4.1 模型量化.....	6
4.2 模型运行.....	7
4.3 运行常见问题.....	8
4.3.1 找不到 GPU 设备.....	8
4.3.2 OpenCL 编译失败.....	8
4.3.3 动态库加载失败.....	9
4.3.4 长 Prompt 导致桌面卡顿.....	9
4.3.5 jinja2 版本错误.....	9
4.3.6 模型加载过程中被系统终止.....	9

修订记录

版本	日期	作者	描述
A0	2026/6/4	格兰菲	初次发布

1 概述/引言

本文旨说明在 Arise2030 显卡上部署与运行 LLM 步骤。

2 运行环境要求

2.1 操作系统

支持以下 Linux 发行版（或兼容系统）：

- KOS
- UOS
- Ubuntu

2.2 Python 环境

支持版本：

Python 3.8 ~ Python 3.11

推荐版本：

Python 3.10

3 驱动安装与验证

3.1 安装 Arise2030 驱动

安装流程请参考官网驱动安装文档《Linux 驱动安装指南》。

通过如下命令验证驱动是否安装成功：

```
lspci | grep -i vga
```

如果能够正确识别 Arise GPU，则说明驱动安装成功。

3.2 安装 AriseLLM

AriseLLM 是基于 Arise GPU 平台优化的大语言模型推理工具。安装 AriseLLM Wheel 包命令如下：

```
pip3 install gf_arise_fastllm_pytools-x.x.x-py3-none-manylinux2014_x86_64.whl \
-i https://pypi.tuna.tsinghua.edu.cn/simple
```

验证安装，执行：

```
ftllm -v
```

正常情况下将显示当前 AriseLLM 版本信息。

如果执行 `ftllm -v` 提示命令不存在，但通过 `pip3 list` 命令能看到安装成功，则需要将 Python 用户目录加入 PATH：

```
export PATH=$PATH:$HOME/.local/bin
```

永久生效命令如下：

```
echo 'export PATH=$PATH:$HOME/.local/bin' >> ~/.bashrc
source ~/.bashrc
```

4 模型量化与运行

4.1 模型量化

AriseLLM 支持将 HuggingFace 格式模型转换为量化模型，以降低显存占用并提升推理效率。可以通过 <https://huggingface.co/> 网站或者 <https://hf-mirror.com/> 下载 AriseLLM 支持的模型

当前版本支持以下模型：

模型系列	支持模型
Qwen3	0.6B、1.7B、4B
Qwen2.5	0.5B、1.5B、3B
DeepSeek-R1-Distill	1.5B

本章节以 DeepSeek-R1-Distill-Qwen-1.5B 模型为例，hf-mirror 地址为：<https://hf-mirror.com/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B> .可以使用如下命令下载模型

```
mkdir ~/models
cd ~/models
git clone https://hf-mirror.com/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B
```

模型量化命令如下：

```
ftlm export /path/to/hf-model \
-o /path/to/output \
--dtype int4g32
```

参数说明：

参数	说明
-o	输出目录
--dtype int4g32	INT4 Group32 量化格式

以 DeepSeek-R1-Distill-Qwen-1.5B 为例，量化模型的命令如下：

```
ftlm export \
~/models/DeepSeek-R1-Distill-Qwen-1.5B \
-o ~/models/DeepSeek-R1-Distill-Qwen-1.5B-int4g32 \
--dtype int4g32
```

量化完成后将在输出目录生成可部署模型文件。

4.2 模型运行

启动交互式对话

```
ftlm run /path/to/model
```

以 DeepSeek-R1-Distill-Qwen-1.5B 为示例：

```
ftlm run \
~/models/DeepSeek-R1-Distill-Qwen-1.5B-int4g32
```

启动成功后即可进入交互式对话界面。

常用交互命令

作用	命令
清空历史会话	clear
退出交互程序	stop

4.3 运行常见问题

4.3.1 找不到 GPU 设备

现象：

clinfo 可以识别设备

但 AriseLLM 无法发现 Device

解决方案：

修改配置文件：

```
sudo vim /etc/default/grub
```

将：

```
GRUB_CMDLINE_LINUX_SECURITY="security=kysec"
```

修改为：

```
GRUB_CMDLINE_LINUX_SECURITY="security="
```

更新配置：

```
sudo update-grub  
reboot
```

4.3.2 OpenCL 编译失败

错误示例：

```
clBuildProgram failed  
clErr: -3
```


安装依赖：

```
sudo apt-get install libtinfo5
```

4.3.3 动态库加载失败

设置动态库路径：

```
export LD_LIBRARY_PATH=/yourpath/fastllm_pytools:$LD_LIBRARY_PATH
```

4.3.4 长 Prompt 导致桌面卡顿

启用同步模式：

```
export FASTLLM_ENABLE_PROMPT_OCL_SYNC=1
```

然后重新运行模型。

4.3.5 jinja2 版本错误

错误示例：

```
AttributeError:  
module 'jinja2'  
has no attribute 'pass_eval_context'
```

升级 jinja2：

```
pip3 install jinja2==3.1.5
```

4.3.6 模型加载过程中被系统终止

错误示例：

```
Killed
```

原因：

系统内存或 Swap 空间不足。

查看 Swap 大小和文件位置：

```
free -h
```

创建 32GB Swap:

```
sudo swapoff /swapfile  
sudo fallocate -l 32G /swapfile  
sudo mkswap /swapfile  
sudo swapon /swapfile
```

格兰菲智能科技有限公司