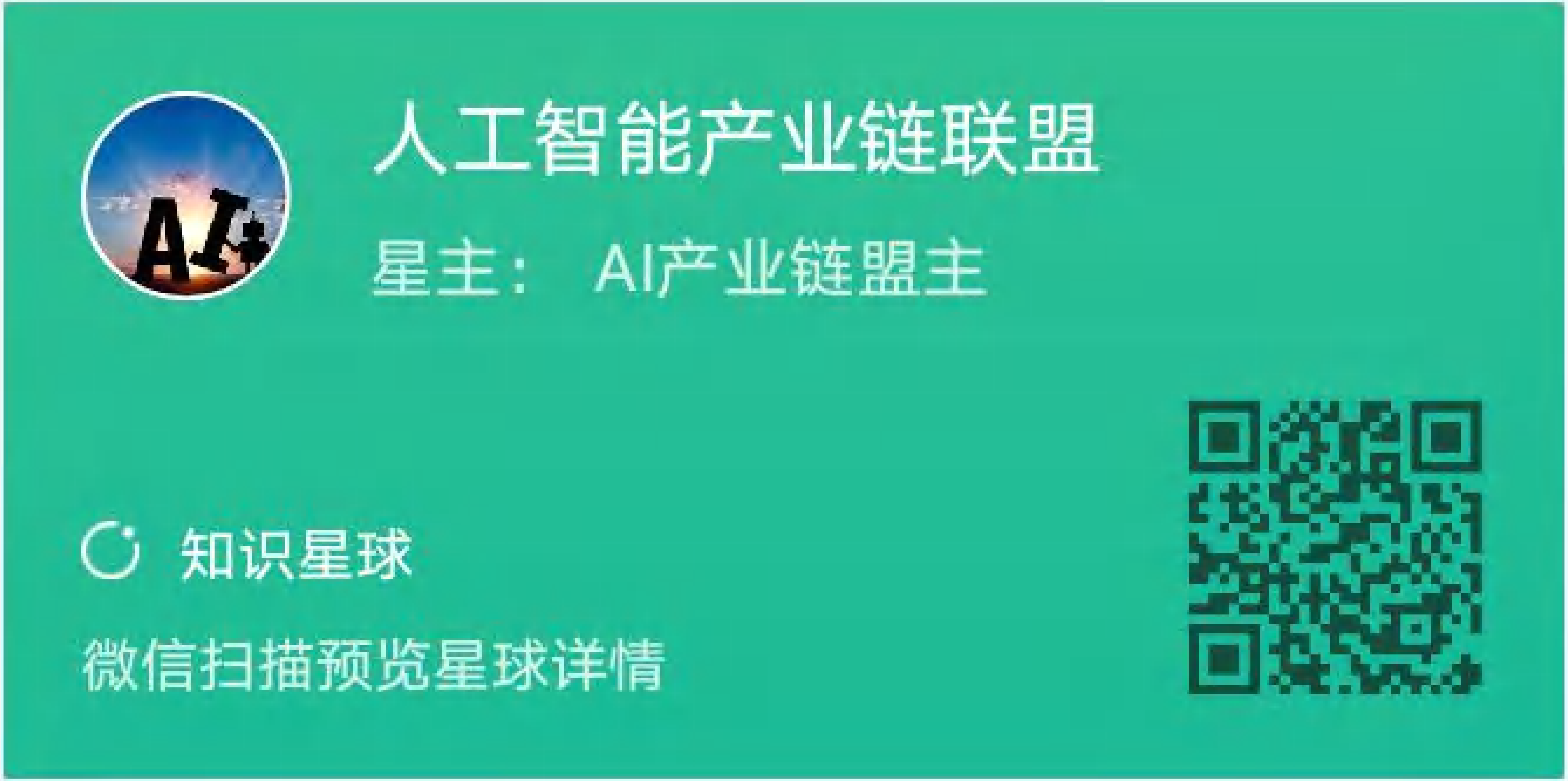
大语言模型的异构计算和加速

戴金权 (Jason Dai)
英特尔院士



人工智能产业链联盟

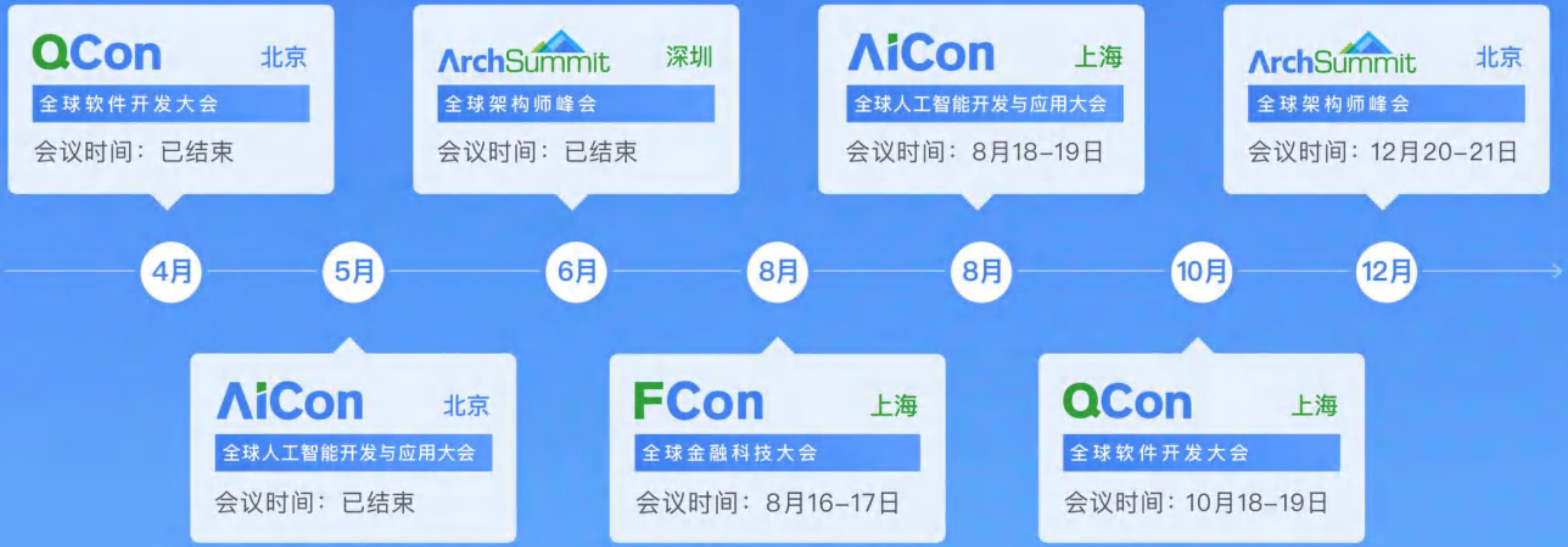
星主： AI产业链盟主

知识星球

微信扫描预览星球详情

极客邦科技 2024 年会议规划

促进软件开发及相关领域知识与创新的传播



4月

QCon 北京

全球软件开发大会

会议时间：已结束

5月

AiCon 北京

全球人工智能开发与应用大会

会议时间：已结束

6月

ArchSummit 深圳

全球架构师峰会

会议时间：已结束

8月

FCon 上海

全球金融科技大会

会议时间：8月16–17日

8月

AiCon 上海

全球人工智能开发与应用大会

会议时间：8月18–19日

10月

QCon 上海

全球软件开发大会

会议时间：10月18–19日

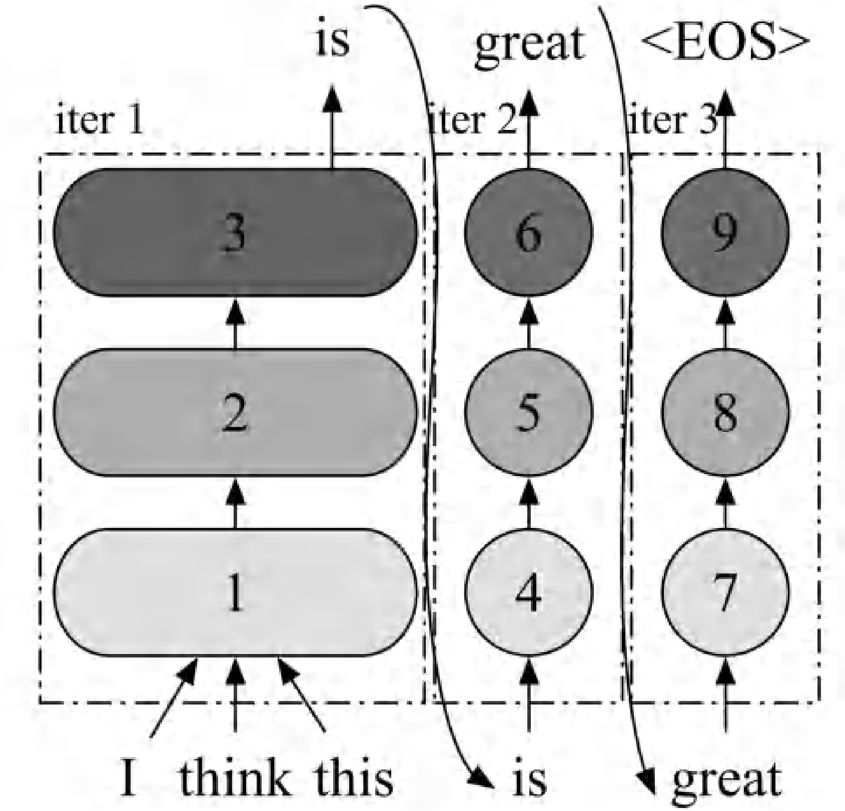
12月

ArchSummit 北京

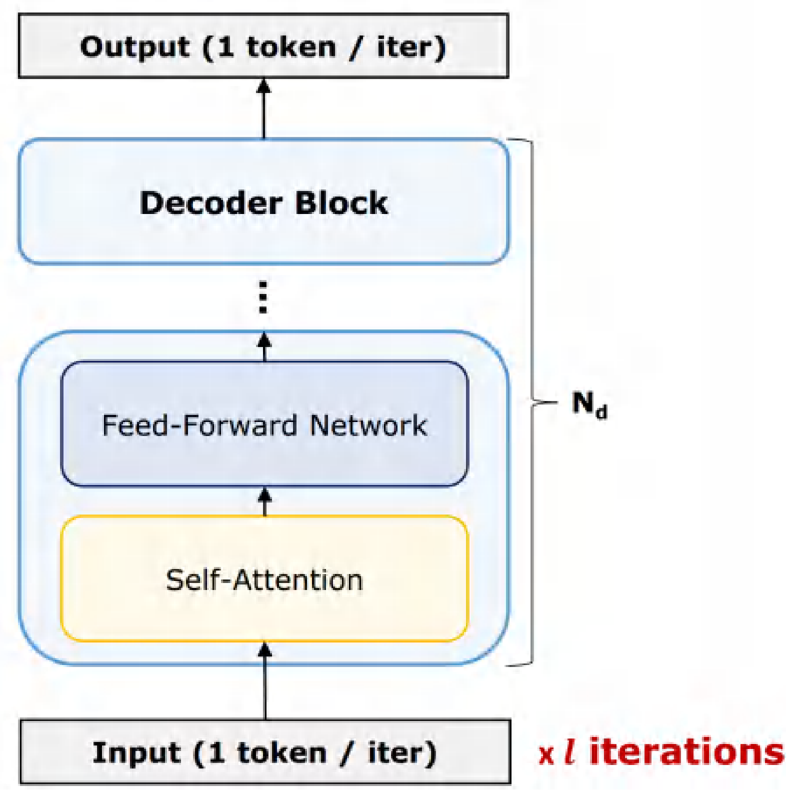
全球架构师峰会

会议时间：12月20–21日

自回归大语言模型(基于Transformer解码器架构)

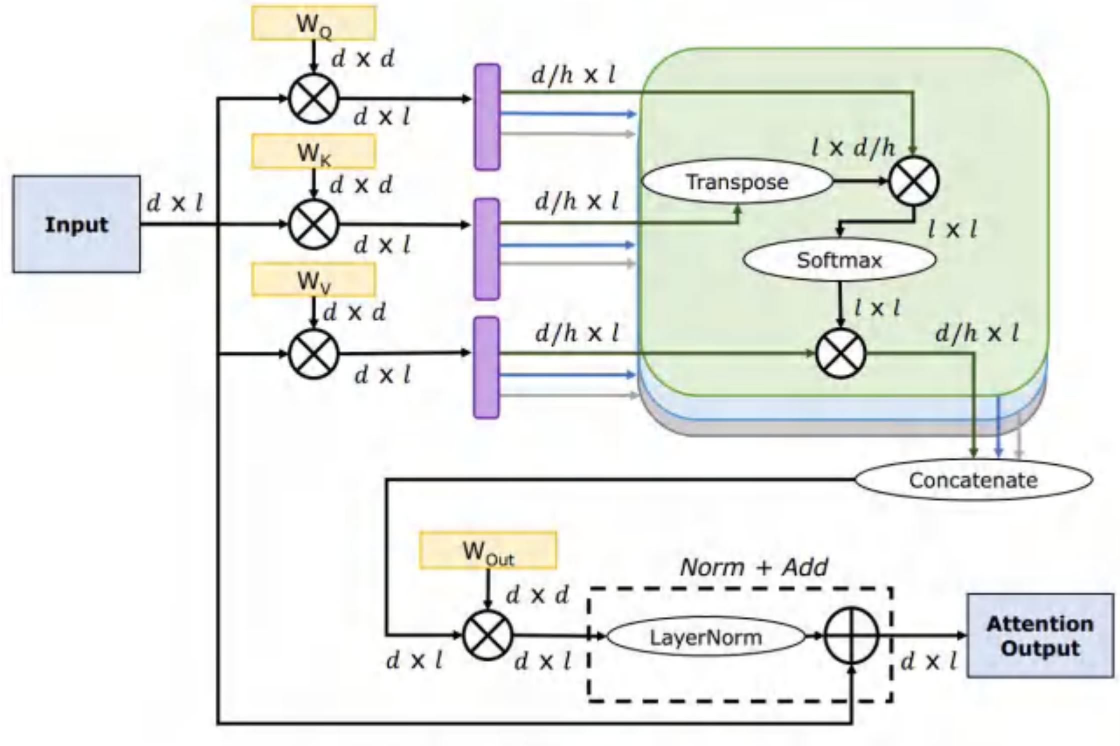


自回归大语言模型：预测下一个token

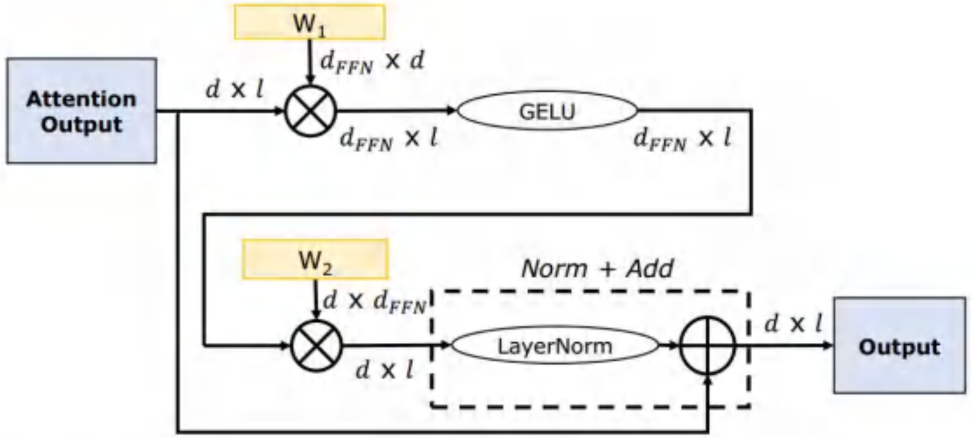


Transformer解码器架构

Transformer解码器架构



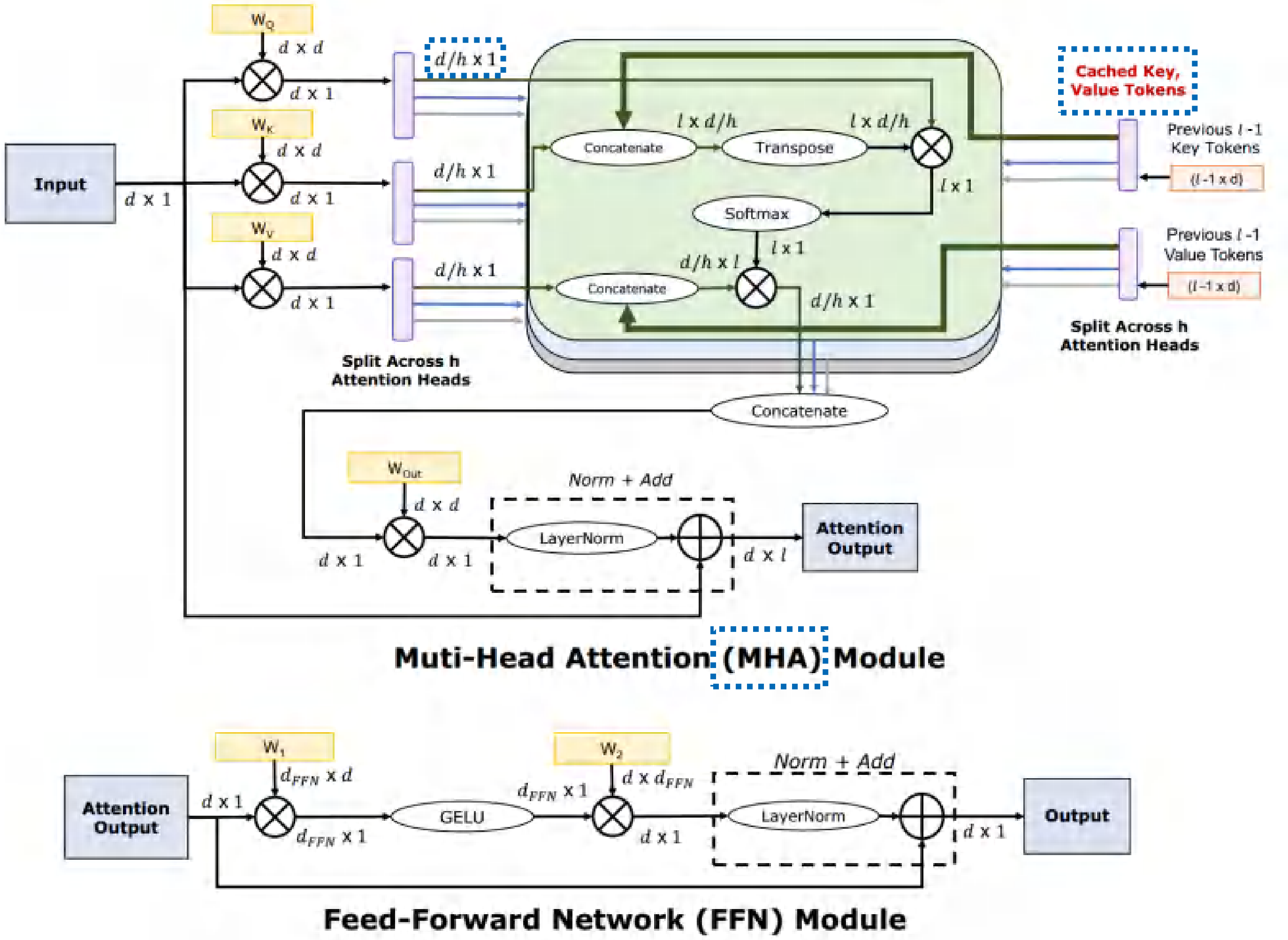
Muti-Head Attention (MHA) Module



Feed-Forward Network (FFN) Module

训练; 推理 (第一个token/Prefill)

Transformer解码器架构



Muti-Head Attention (MHA) Module

Feed-Forward Network (FFN) Module

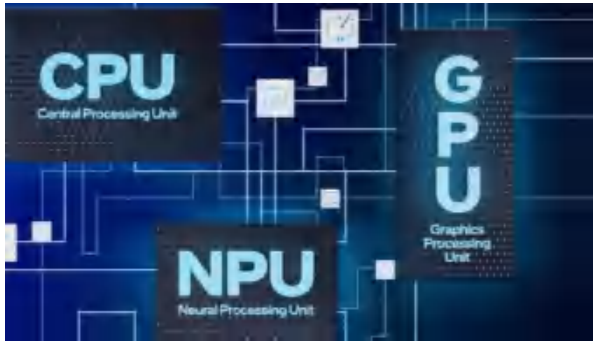
推理 (下一个token/Decode)

大语言模型推理和训练瓶颈

- 内存带宽
- 计算
- 显存大小
- 分布式计算 (互联)

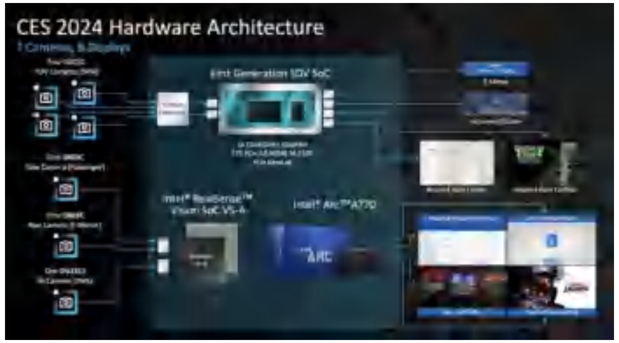
大模型的异构计算和加速

- XPU异构计算
  - CPU, GPU, NPU硬件加速



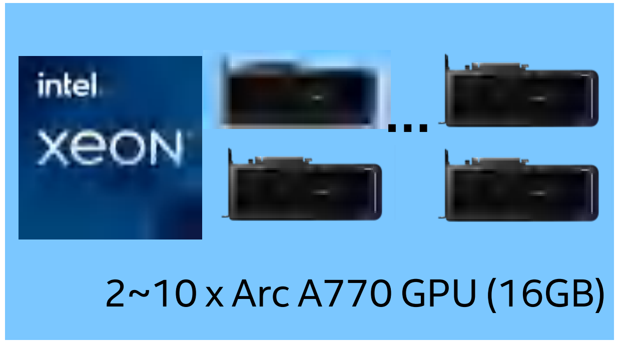
客户端

（Intel Core Ultra AI PC）


CES 2024 Hardware Architecture
Intel® RealSense™ Vision SoC VS-A
Intel® Arc™ A770
ARC

边缘端

（Intel AI座舱）



2~10 x Arc A770 GPU (16GB)

服务器

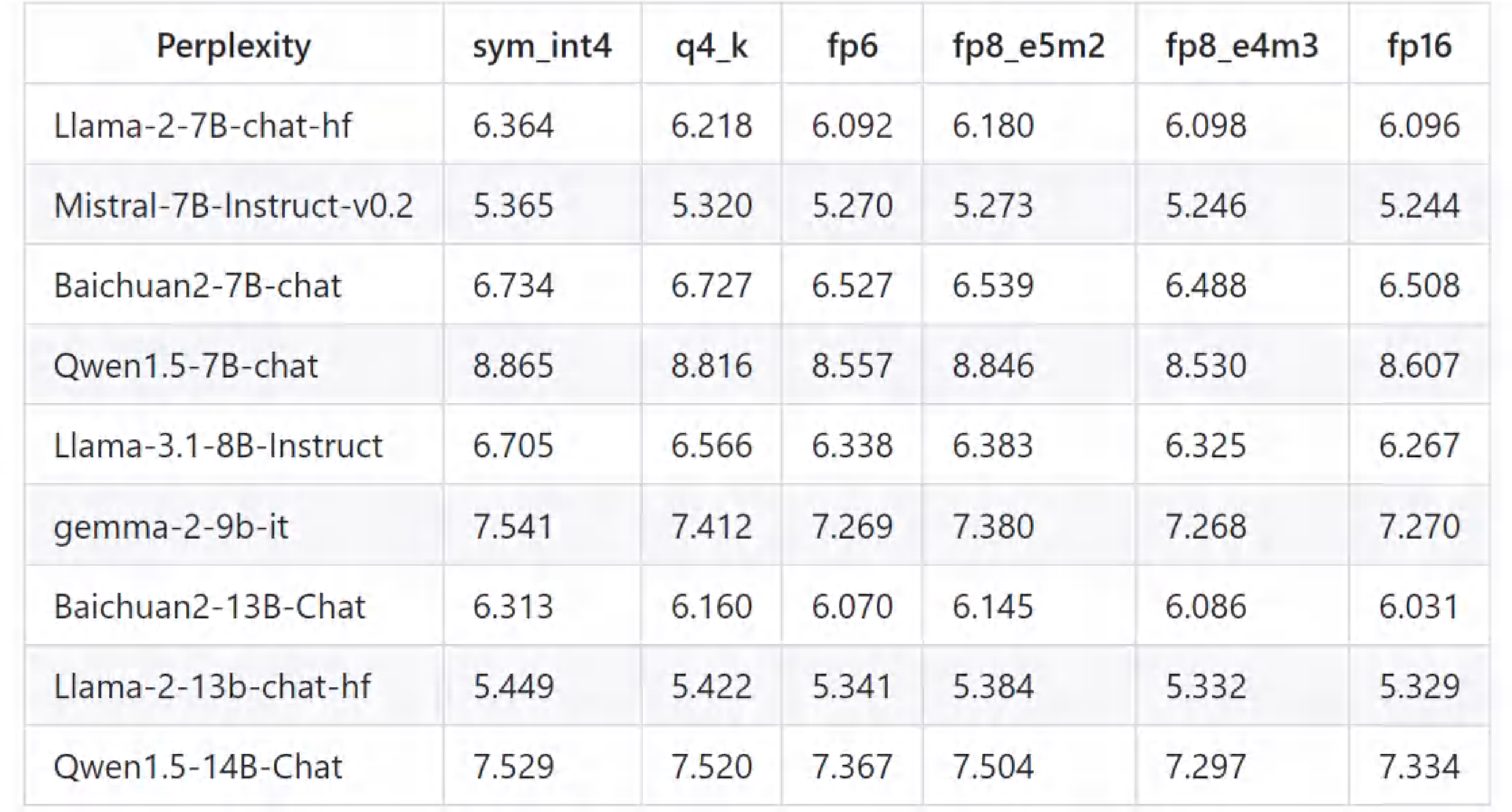
（ Intel Xeon+Intel Arc GPUs）

大模型的异构计算和加速

- 低比特计算
  - 模型量化/压缩 (WxAy)
  - 数据类型 (INTx, FPx)
  - 低比特算子
  - 显存(如kv cache) 使用量
  - 训练、微调 (如QLoRA)

低比特大模型的精度

困惑度 (Wikitext数据集)

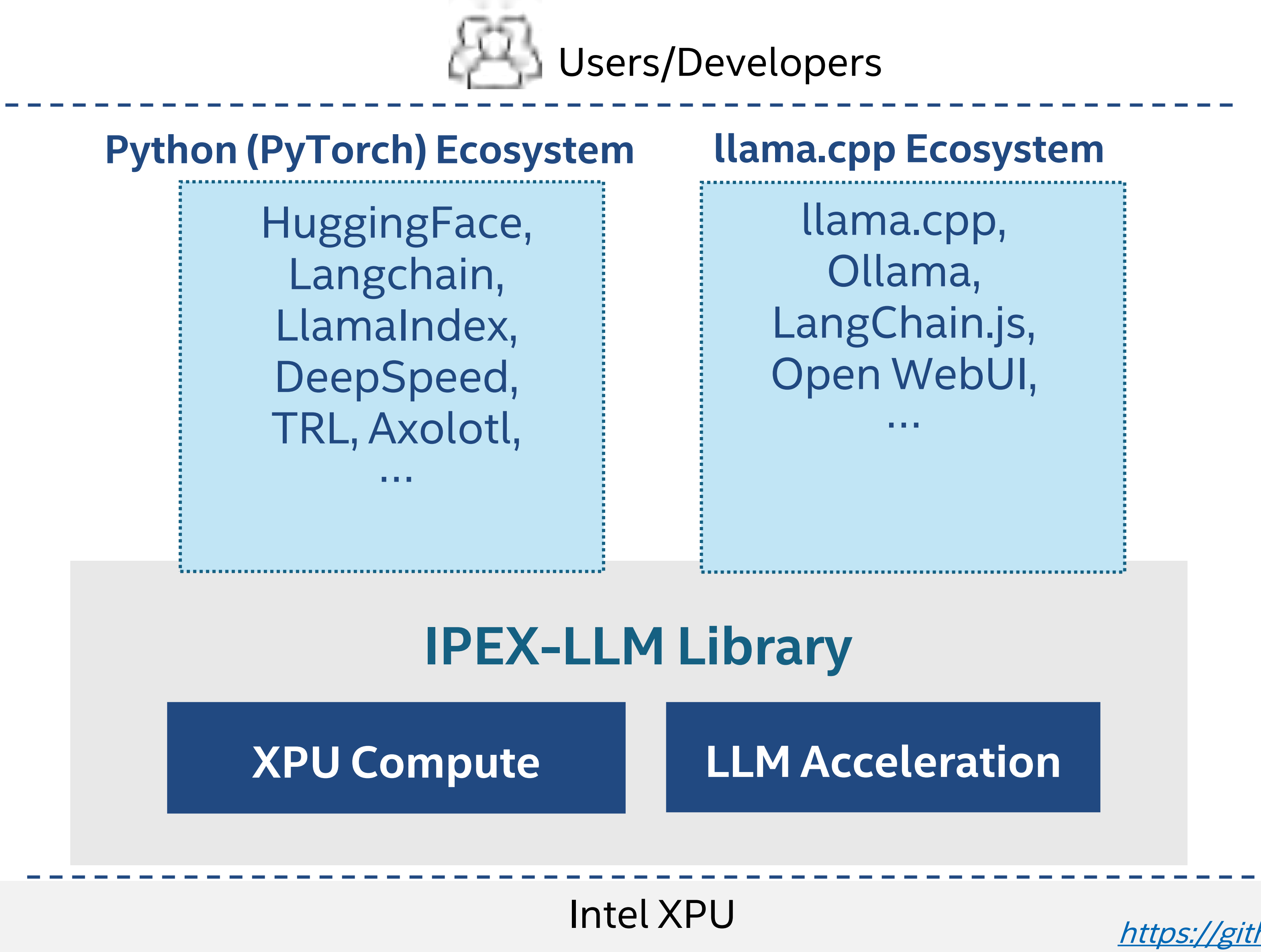


Perplexity	sym_int4	q4_k	fp6	fp8_e5m2	fp8_e4m3	fp16
Llama-2-7B-chat-hf	6.364	6.218	6.092	6.180	6.098	6.096
Mistral-7B-Instruct-v0.2	5.365	5.320	5.270	5.273	5.246	5.244
Baichuan2-7B-chat	6.734	6.727	6.527	6.539	6.488	6.508
Qwen1.5-7B-chat	8.865	8.816	8.557	8.846	8.530	8.607
Llama-3.1-8B-Instruct	6.705	6.566	6.338	6.383	6.325	6.267
gemma-2-9b-it	7.541	7.412	7.269	7.380	7.268	7.270
Baichuan2-13B-Chat	6.313	6.160	6.070	6.145	6.086	6.031
Llama-2-13b-chat-hf	5.449	5.422	5.341	5.384	5.332	5.329
Qwen1.5-14B-Chat	7.529	7.520	7.367	7.504	7.297	7.334

大模型的异构计算和加速

- 推理算法优化
  - Self-speculative decoding
  - KV Cache compression
  - Sliding window attention
  - Sparse attention
  - Flash attention/decoding
  - Continuous batching
  - Prefill/decoding disaggregation
  - …

IPEX-LLM：开源大模型XPU加速框架

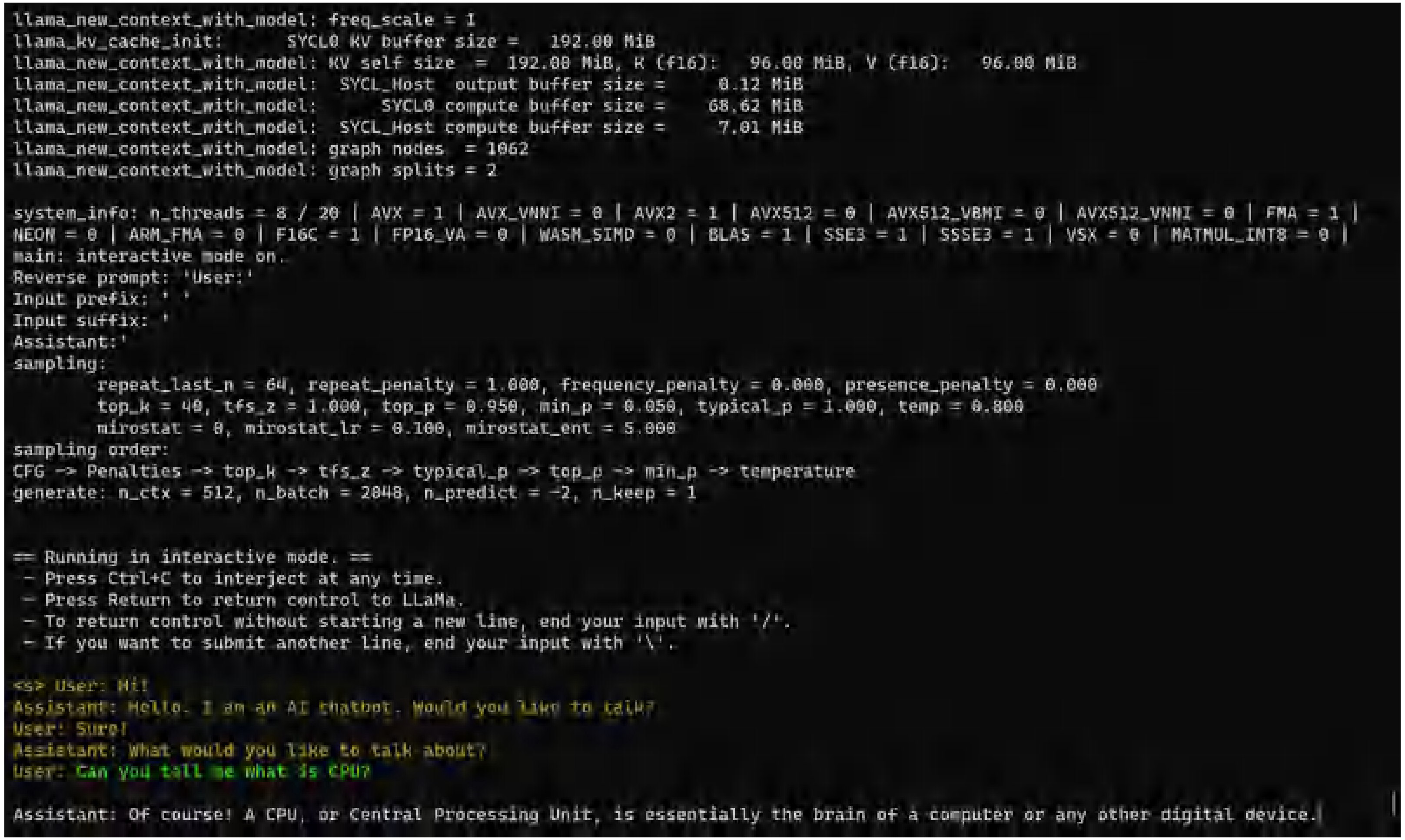


https://github.com/intel-analytics/ipex-llm/

英特尔 XPU 大模型加速体验

Intel UHD/Iris iGPU

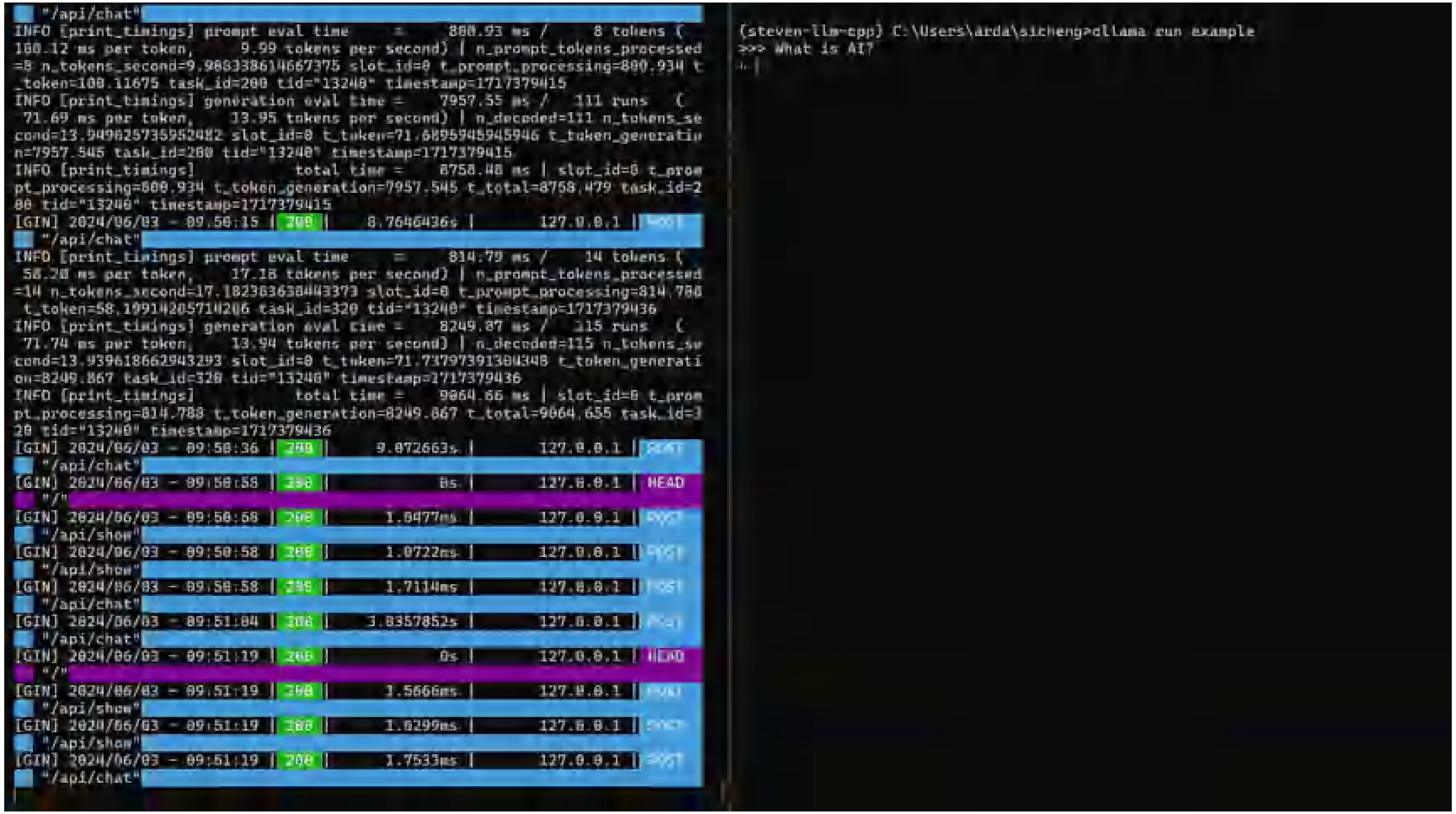
llama.cpp + IPEX-LLM (Phi-3-mini, Q4_0)



```
llama_new_context_with_model: freq_scale = 1
llama_kv_cache_init:      SYCL0 KV buffer size =   192.00 MiB
llama_new_context_with_model: KV self size  =  192.00 MiB, K (f16):   96.00 MiB, V (f16):   96.00 MiB
llama_new_context_with_model:  SYCL_Host  output buffer size =     0.12 MiB
llama_new_context_with_model:      SYCL0 compute buffer size =    68.62 MiB
llama_new_context_with_model:  SYCL_Host compute buffer size =     7.01 MiB
llama_new_context_with_model: graph nodes  = 1062
llama_new_context_with_model: graph splits = 2

system_info: n_threads = 8 / 20 | AVX = 1 | AVX_VNNI = 0 | AVX2 = 1 | AVX512 = 0 | AVX512_VBMI = 0 | AVX512_VNNI = 0 | FMA = 1 |
NEON = 0 | ARM_FMA = 0 | F16C = 1 | FP16_VA = 0 | WASM_SIMD = 0 | BLAS = 1 | SSE3 = 1 | SSSE3 = 1 | VSX = 0 | MATMUL_INT8 = 0 |
main: interactive mode on.
Reverse prompt: 'User:'
Input prefix: ' '
Input suffix: '
Assistant:'
sampling:
        repeat_last_n = 64, repeat_penalty = 1.000, frequency_penalty = 0.000, presence_penalty = 0.000
        top_k = 40, tfs_z = 1.000, top_p = 0.950, min_p = 0.050, typical_p = 1.000, temp = 0.800
        mirostat = 0, mirostat_lr = 0.100, mirostat_ent = 5.000
sampling order:
CFG -> Penalties -> top_k -> tfs_z -> typical_p -> top_p -> min_p -> temperature
generate: n_ctx = 512, n_batch = 2048, n_predict = -2, n_keep = 1


== Running in interactive mode. ==
 - Press Ctrl+C to interject at any time.
 - Press Return to return control to LLaMa.
 - To return control without starting a new line, end your input with '/'.
 - If you want to submit another line, end your input with '\'.

<s> User: Hi!
Assistant: Hello. I am an AI chatbot. Would you like to talk?
User: Sure!
Assistant: What would you like to talk about?
User: Can you tell me what is CPU?

Assistant: Of course! A CPU, or Central Processing Unit, is essentially the brain of a computer or any other digital device.
```

Intel Core Ultra AI PC

Ollama + IPEX-LLM (Mistral-7B, Q4_K_M)

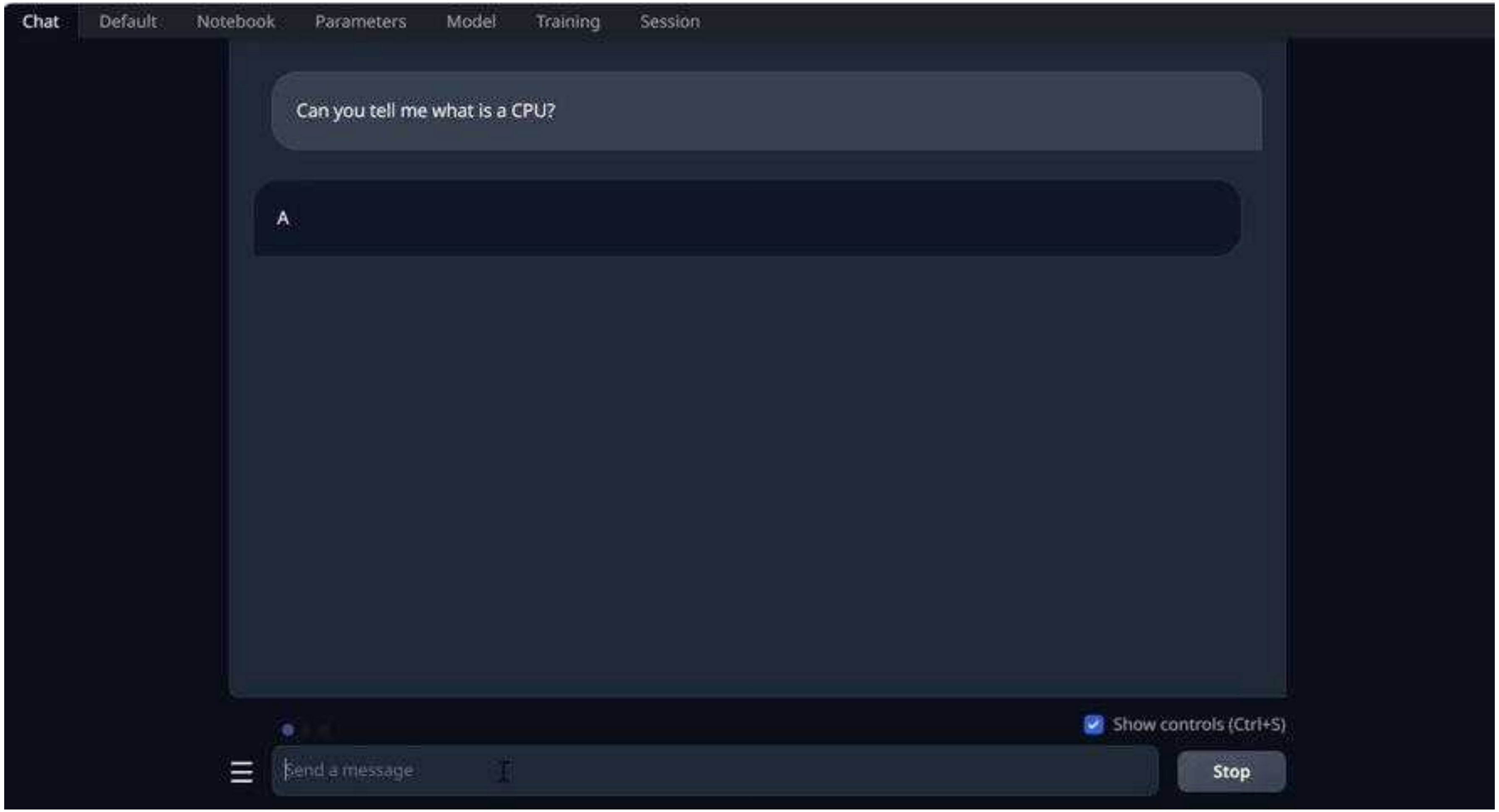


```
   "/api/chat"
INFO [print_timings] prompt eval time     =     800.93 ms /     8 tokens (
100.12 ms per token,     9.99 tokens per second) | n_prompt_tokens_processed
=8 n_tokens_second=9.988338614667375 slot_id=0 t_prompt_processing=800.934 t
_token=100.11675 task_id=200 tid="13240" timestamp=1717379415
INFO [print_timings] generation eval time =    7957.55 ms /   111 runs    (
 71.69 ms per token,    13.95 tokens per second) | n_decoded=111 n_tokens_se
cond=13.949625735952482 slot_id=0 t_token=71.6895945945946 t_token_generatio
n=7957.545 task_id=200 tid="13240" timestamp=1717379415
INFO [print_timings]             total time =    8758.48 ms | slot_id=0 t_prom
pt_processing=800.934 t_token_generation=7957.545 t_total=8758.479 task_id=2
00 tid="13240" timestamp=1717379415
[GIN] 2024/06/03 - 09:50:15 | 200 |    8.7646436s |       127.0.0.1 | POST
   "/api/chat"
INFO [print_timings] prompt eval time     =     814.79 ms /    14 tokens (
 58.20 ms per token,    17.18 tokens per second) | n_prompt_tokens_processed
=14 n_tokens_second=17.182363638443373 slot_id=0 t_prompt_processing=814.788
 t_token=58.19914285714286 task_id=320 tid="13240" timestamp=1717379436
INFO [print_timings] generation eval time =    8249.87 ms /   115 runs    (
 71.74 ms per token,    13.94 tokens per second) | n_decoded=115 n_tokens_se
cond=13.939618662943293 slot_id=0 t_token=71.73797391304348 t_token_generati
on=8249.867 task_id=320 tid="13240" timestamp=1717379436
INFO [print_timings]             total time =    9064.66 ms | slot_id=0 t_prom
pt_processing=814.788 t_token_generation=8249.867 t_total=9064.655 task_id=3
20 tid="13240" timestamp=1717379436
[GIN] 2024/06/03 - 09:50:36 | 200 |     9.072663s |       127.0.0.1 | POST
   "/api/chat"
[GIN] 2024/06/03 - 09:50:58 | 200 |            0s |       127.0.0.1 | HEAD
   "/"
[GIN] 2024/06/03 - 09:50:58 | 200 |     1.0477ms |       127.0.0.1 | POST
   "/api/show"
[GIN] 2024/06/03 - 09:50:58 | 200 |     1.0722ms |       127.0.0.1 | POST
   "/api/show"
[GIN] 2024/06/03 - 09:50:58 | 200 |     1.7114ms |       127.0.0.1 | POST
   "/api/chat"
[GIN] 2024/06/03 - 09:51:04 | 200 |   3.0357852s |       127.0.0.1 | POST
   "/api/chat"
[GIN] 2024/06/03 - 09:51:19 | 200 |            0s |       127.0.0.1 | HEAD
   "/"
[GIN] 2024/06/03 - 09:51:19 | 200 |     1.5666ms |       127.0.0.1 | POST
   "/api/show"
[GIN] 2024/06/03 - 09:51:19 | 200 |     1.0299ms |       127.0.0.1 | POST
   "/api/show"
[GIN] 2024/06/03 - 09:51:19 | 200 |     1.7533ms |       127.0.0.1 | POST
   "/api/chat"
```

```
(steven-llm-cpp) C:\Users\arda\sicheng>ollama run example
>>> What is AI?
```

Intel Arc A770 GPU

TextGeneration-WebUI + IPEX-LLM (Llama3-8B, FP8)

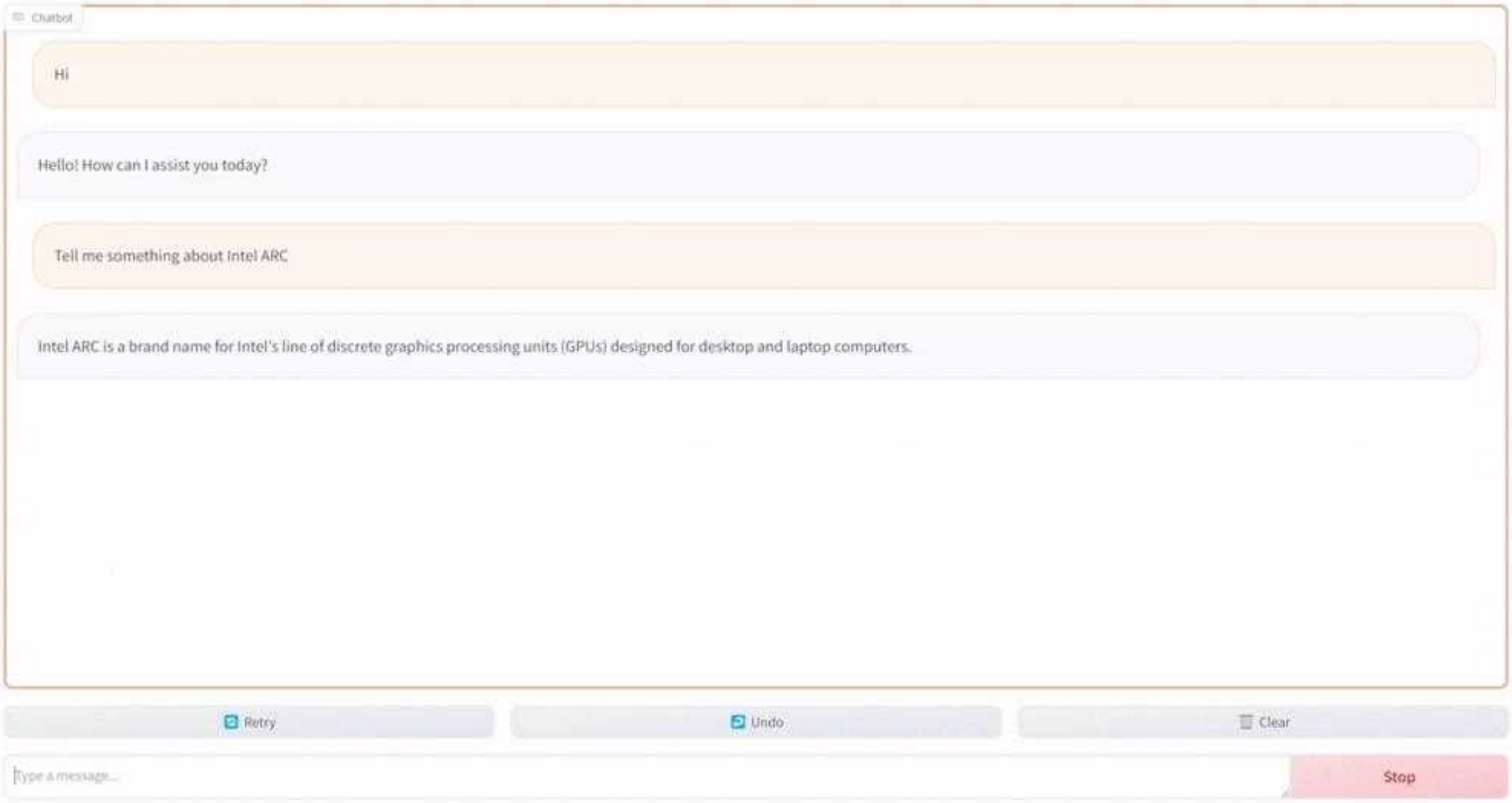


Can you tell me what is a CPU?

A

4 x Arc A770 GPU

FastChat + IPEX-LLM (QWen1.5-72B FP6)



Hi

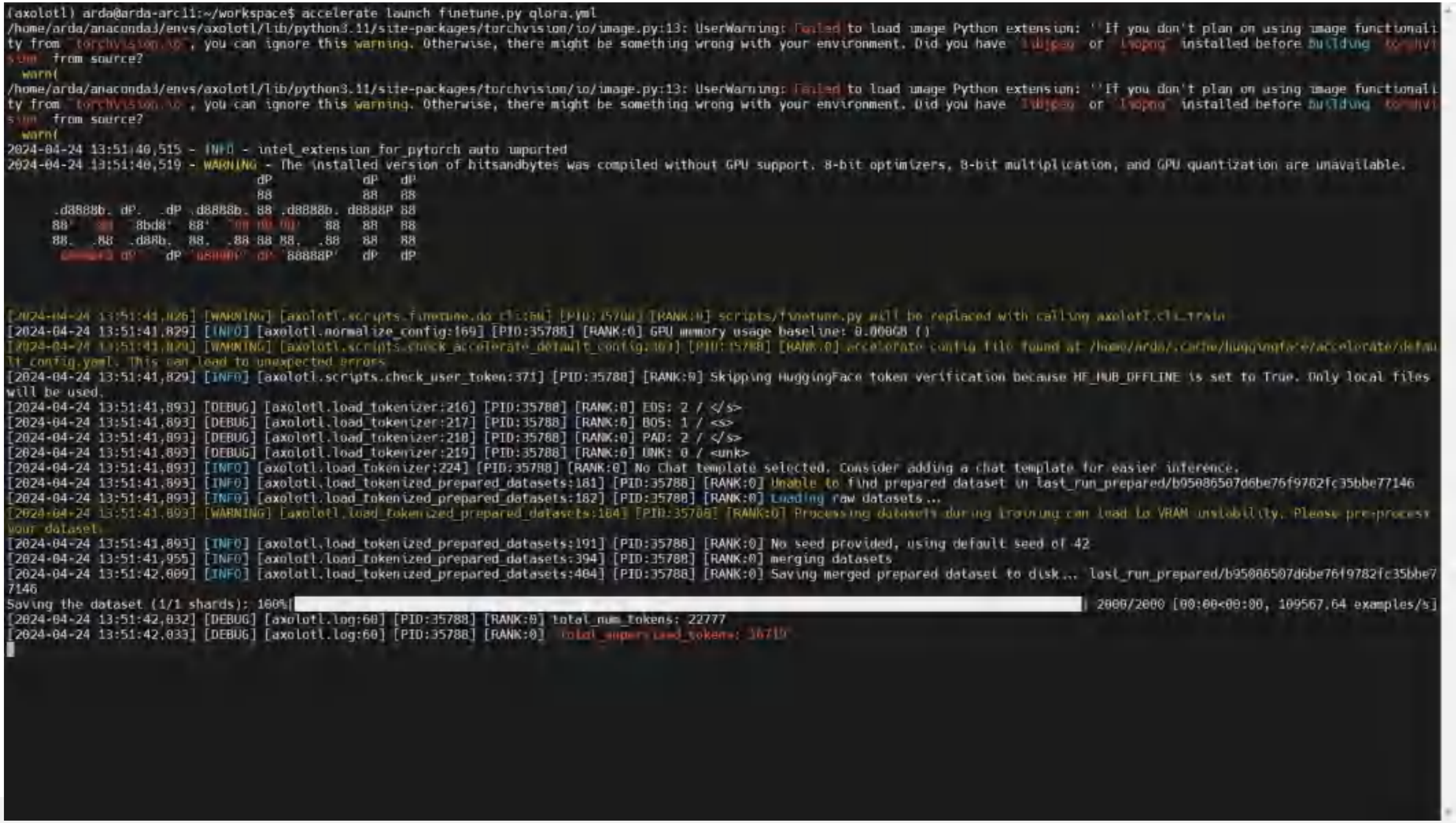
Hello! How can I assist you today?

Tell me something about Intel ARC

Intel ARC is a brand name for Intel's line of discrete graphics processing units (GPUs) designed for desktop and laptop computers.

LoRA/QLoRA on Xeon+Multi-Arc

支持 PEFT, TRL, Axolotl, Zero2/Zero3

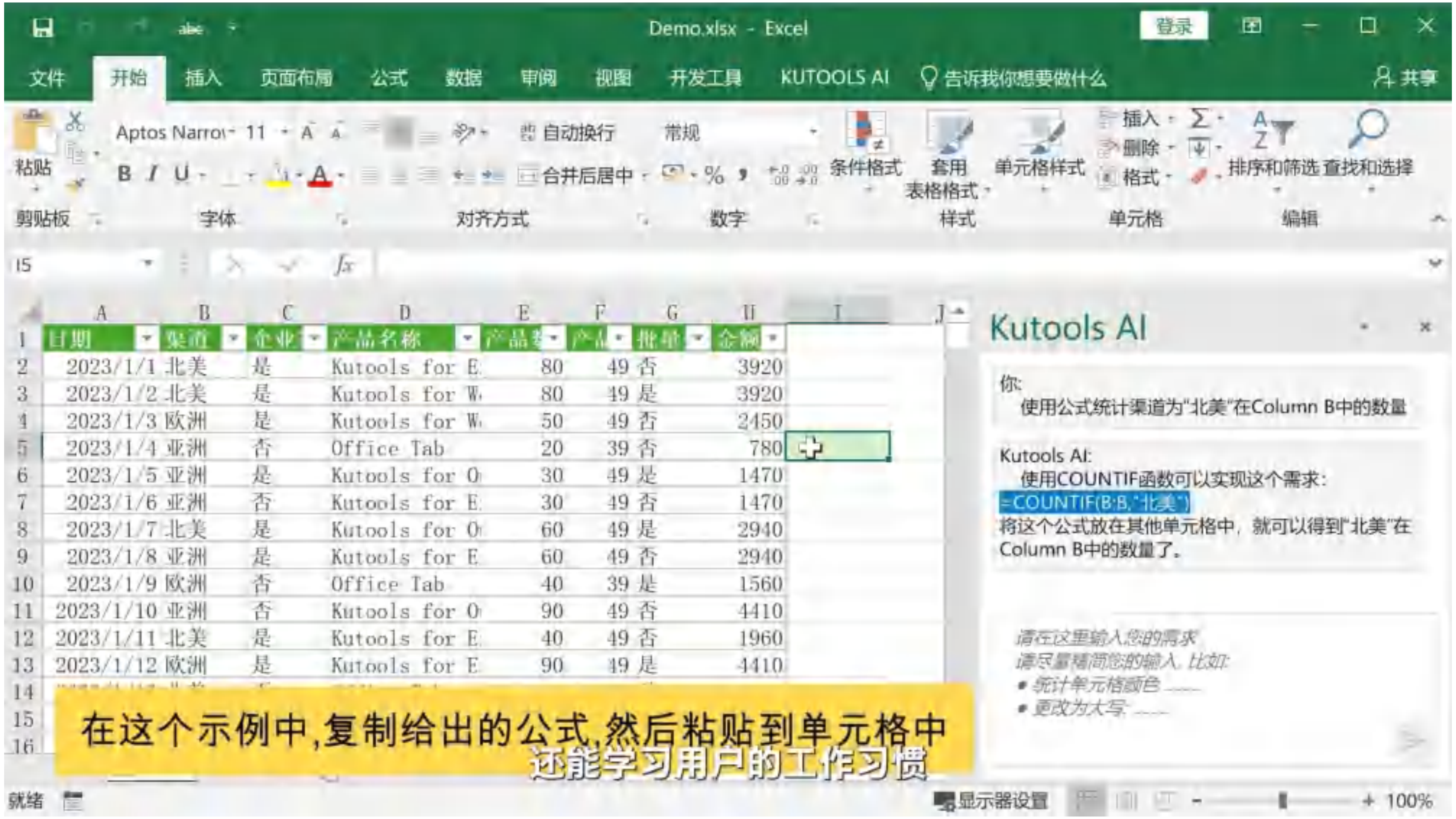


```
(axolotl) arda@arda-arc11:~/workspace$ accelerate launch finetune.py qlora.yml
/home/arda/anaconda3/envs/axolotl/lib/python3.11/site-packages/torchvision/io/image.py:13: UserWarning: Failed to load image Python extension: ''If you don't plan on using image functionali
ty from `torchvision.io`, you can ignore this warning. Otherwise, there might be something wrong with your environment. Did you have `libjpeg` or `libpng` installed before building `torchvi
sion` from source?
  warn(
/home/arda/anaconda3/envs/axolotl/lib/python3.11/site-packages/torchvision/io/image.py:13: UserWarning: Failed to load image Python extension: ''If you don't plan on using image functionali
ty from `torchvision.io`, you can ignore this warning. Otherwise, there might be something wrong with your environment. Did you have `libjpeg` or `libpng` installed before building `torchvi
sion` from source?
  warn(
2024-04-24 13:51:40,515 - INFO - intel_extension_for_pytorch auto imported
2024-04-24 13:51:40,519 - WARNING - The installed version of bitsandbytes was compiled without GPU support. 8-bit optimizers, 8-bit multiplication, and GPU quantization are unavailable.
                                 dP            dP   dP
                                 88            88   88
      .d8888b. dP.  .dP .d8888b. 88 .d8888b. d8888P 88
      88'  `88  `8bd8'  88'  `88 88 88'  `88   88   88
      88.  .88  .d88b.  88.  .88 88 88.  .88   88   88
      `88888P8 dP'  `dP `88888P' dP `88888P'   dP   dP



[2024-04-24 13:51:41,826] [WARNING] [axolotl.scripts.finetune.do_cli:60] [PID:35788] [RANK:0] scripts/finetune.py will be replaced with calling axolotl.cli.train
[2024-04-24 13:51:41,829] [INFO] [axolotl.normalize_config:169] [PID:35788] [RANK:0] GPU memory usage baseline: 0.000GB ()
[2024-04-24 13:51:41,829] [WARNING] [axolotl.scripts.check_accelerate_default_config:363] [PID:35788] [RANK:0] accelerate config file found at /home/arda/.cache/huggingface/accelerate/defau
lt_config.yaml. This can lead to unexpected errors
[2024-04-24 13:51:41,829] [INFO] [axolotl.scripts.check_user_token:371] [PID:35788] [RANK:0] Skipping HuggingFace token verification because HF_HUB_OFFLINE is set to True. Only local files
will be used.
[2024-04-24 13:51:41,893] [DEBUG] [axolotl.load_tokenizer:216] [PID:35788] [RANK:0] EOS: 2 / </s>
[2024-04-24 13:51:41,893] [DEBUG] [axolotl.load_tokenizer:217] [PID:35788] [RANK:0] BOS: 1 / <s>
[2024-04-24 13:51:41,893] [DEBUG] [axolotl.load_tokenizer:218] [PID:35788] [RANK:0] PAD: 2 / </s>
[2024-04-24 13:51:41,893] [DEBUG] [axolotl.load_tokenizer:219] [PID:35788] [RANK:0] UNK: 0 / <unk>
[2024-04-24 13:51:41,893] [INFO] [axolotl.load_tokenizer:224] [PID:35788] [RANK:0] No Chat template selected. Consider adding a chat template for easier inference.
[2024-04-24 13:51:41,893] [INFO] [axolotl.load_tokenized_prepared_datasets:181] [PID:35788] [RANK:0] Unable to find prepared dataset in last_run_prepared/b95086507d6be76f9782fc35bbe77146
[2024-04-24 13:51:41,893] [INFO] [axolotl.load_tokenized_prepared_datasets:182] [PID:35788] [RANK:0] Loading raw datasets...
[2024-04-24 13:51:41,893] [WARNING] [axolotl.load_tokenized_prepared_datasets:184] [PID:35788] [RANK:0] Processing datasets during training can lead to VRAM instability. Please pre-process
your dataset.
[2024-04-24 13:51:41,893] [INFO] [axolotl.load_tokenized_prepared_datasets:191] [PID:35788] [RANK:0] No seed provided, using default seed of 42
[2024-04-24 13:51:41,955] [INFO] [axolotl.load_tokenized_prepared_datasets:394] [PID:35788] [RANK:0] merging datasets
[2024-04-24 13:51:42,009] [INFO] [axolotl.load_tokenized_prepared_datasets:404] [PID:35788] [RANK:0] Saving merged prepared dataset to disk... last_run_prepared/b95086507d6be76f9782fc35bbe7
7146
Saving the dataset (1/1 shards): 100%|████████████████████████████████████████████████████| 2000/2000 [00:00<00:00, 109567.64 examples/s]
[2024-04-24 13:51:42,032] [DEBUG] [axolotl.log:60] [PID:35788] [RANK:0] total_num_tokens: 22777
[2024-04-24 13:51:42,033] [DEBUG] [axolotl.log:60] [PID:35788] [RANK:0] total_supervised_tokens: 16715
```

英特尔 XPU 大模型应用创新

Office助手

ExtendOffice展示



	A	B	C	D	E	F	G	H
1	日期	渠道	企业	产品名称	产品数	产品	批量	金额
2	2023/1/1	北美	是	Kutools for E	80	49	否	3920
3	2023/1/2	北美	是	Kutools for W	80	49	是	3920
4	2023/1/3	欧洲	是	Kutools for W	50	49	否	2450
5	2023/1/4	亚洲	否	Office Tab	20	39	否	780
6	2023/1/5	亚洲	是	Kutools for O	30	49	是	1470
7	2023/1/6	亚洲	否	Kutools for E	30	49	否	1470
8	2023/1/7	北美	是	Kutools for O	60	49	是	2940
9	2023/1/8	亚洲	是	Kutools for E	60	49	否	2940
10	2023/1/9	欧洲	否	Office Tab	40	39	是	1560
11	2023/1/10	亚洲	否	Kutools for O	90	49	否	4410
12	2023/1/11	北美	是	Kutools for E	40	49	否	1960
13	2023/1/12	欧洲	是	Kutools for E	90	49	是	4410

Kutools AI

你:
使用公式统计渠道为"北美"在Column B中的数量

Kutools AI:
使用COUNTIF函数可以实现这个需求:
=COUNTIF(B:B,"北美")
将这个公式放在其他单元格中，就可以得到"北美"在Column B中的数量了。

请在这里输入您的需求
请尽量精简您的输入, 比如:
- 统计单元格颜色 ......
- 更改为大写 ......

在这个示例中,复制给出的公式,然后粘贴到单元格中
还能学习用户的工作习惯

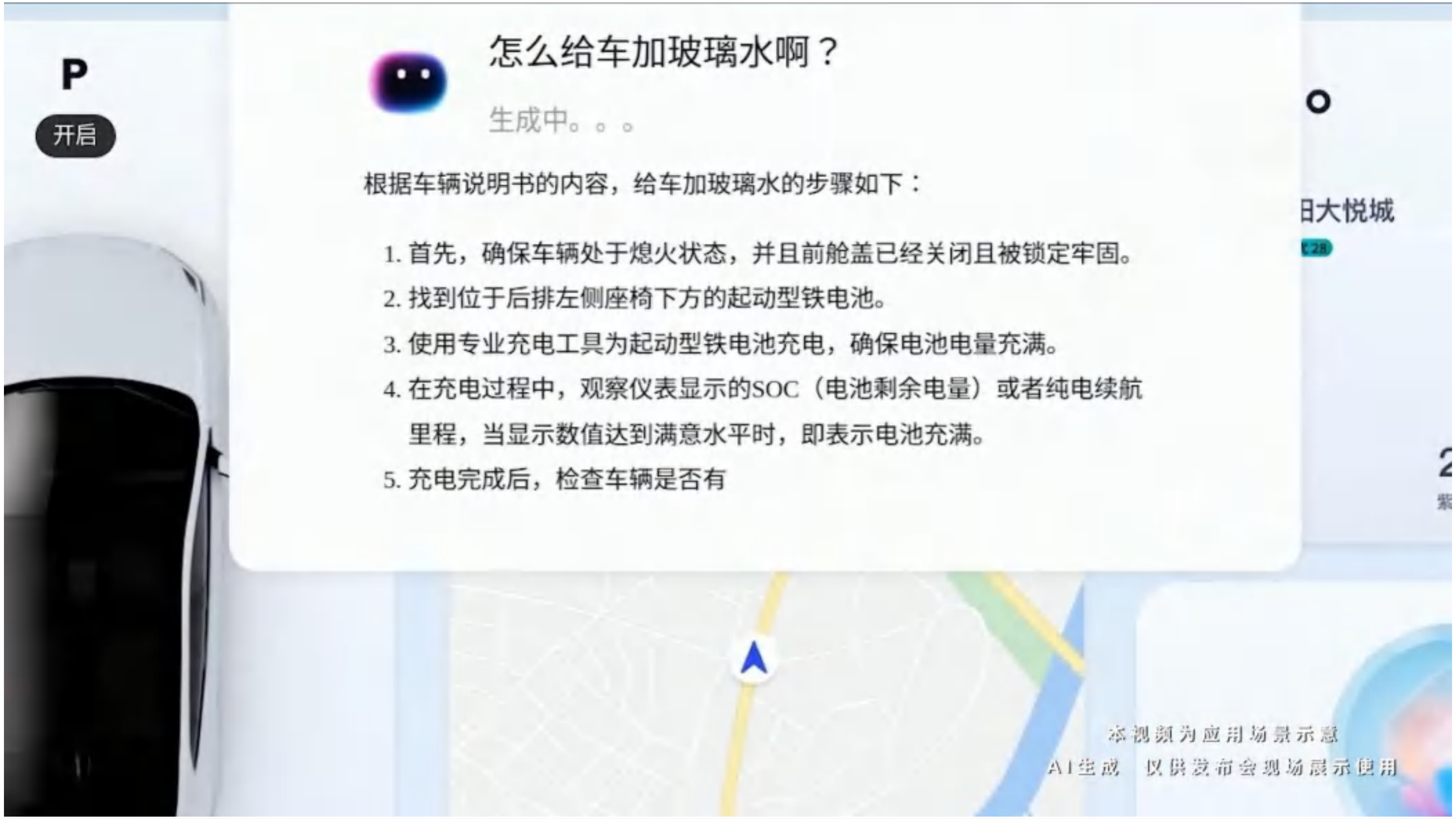
工业机器人代码生成

科东软件展示

转换的结果，直接控制机器人手臂进行操作

AI座舱-汽车助理

智谱AI展示



怎么给车加玻璃水啊？

生成中。。。

根据车辆说明书的内容，给车加玻璃水的步骤如下：

1. 首先，确保车辆处于熄火状态，并且前舱盖已经关闭且被锁定牢固。
2. 找到位于后排左侧座椅下方的起动型铁电池。
3. 使用专业充电工具为起动型铁电池充电，确保电池电量充满。
4. 在充电过程中，观察仪表显示的SOC（电池剩余电量）或者纯电续航里程，当显示数值达到满意水平时，即表示电池充满。
5. 充电完成后，检查车辆是否有

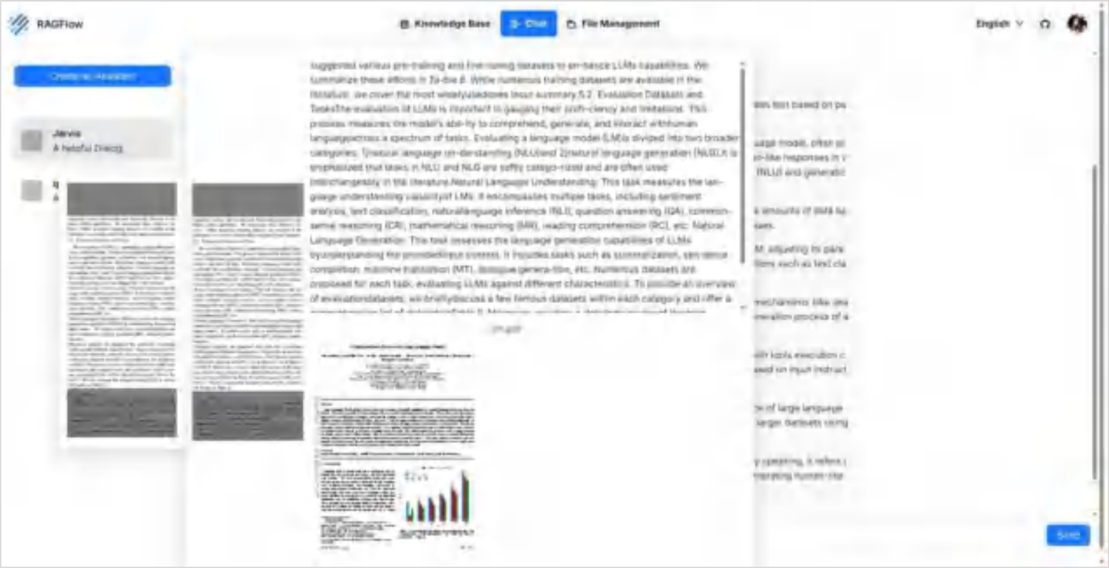
本视频为应用场景示意

AI生成　仅供发布会现场展示使用

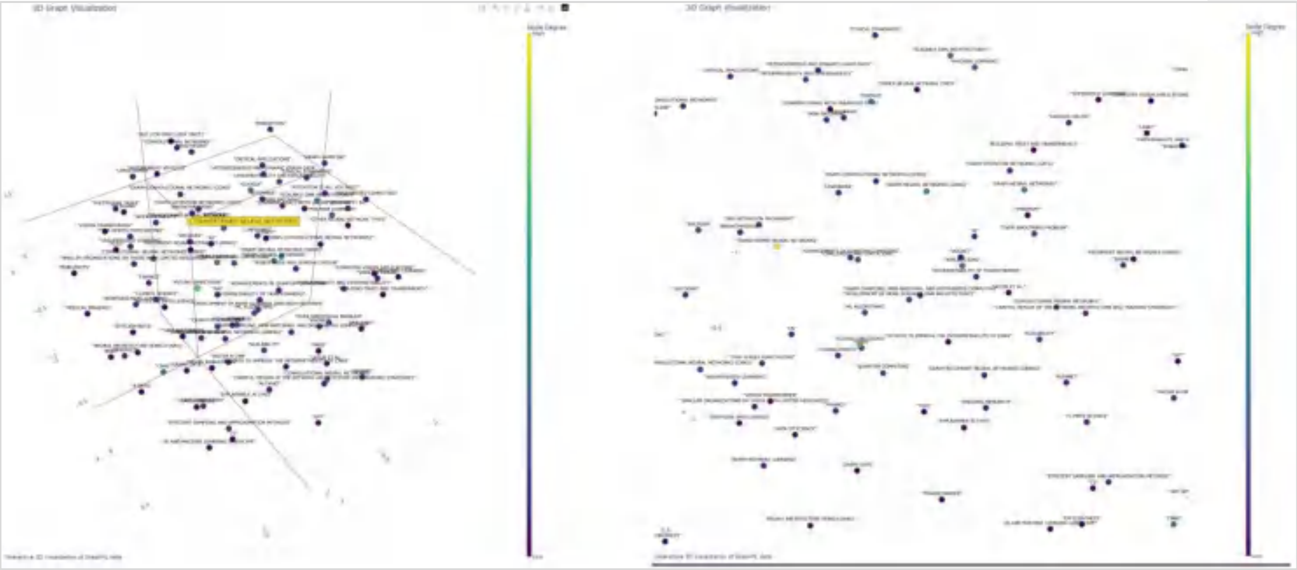
AI座舱-驾驶伴侣

百川智能展示

个人或企业本地RAG系统



在英特尔 XPU 上运行 RAGFlow
(https://github.com/intel-analytics/ipex-llm/blob/main/docs/mddocs/Quickstart/ragflow_quickstart.md)



在英特尔 XPU 上运行 GraphRAG
(https://github.com/intel-analytics/ipex-llm/blob/main/docs/mddocs/Quickstart/graphrag_quickstart.md)

AI人工智能产业链联盟

#每日为你摘取最重要的商业新闻#

更新 · 更快 · 更精彩

Zero

AI音乐创作人

水墨动漫联盟创始人

百脑共创联合创始人

人工智能产业链联盟创始人

中关村人才协会秘书长助理

河北北大企业家分会秘书长

墨玫星辰智能科技有限公司CEO

河北清华发展研究院智能机器人中心线上负责人

中关村人才协会数字体育与电子竞技专委会秘书长助理

主要业务:AI商业化答疑及课程应用场景探索，各类AI产品学习手册，答疑及课程

欢迎扫码交流

提供：学习手册/工具/资源链接/商业化案例/行业报告/行业最新资讯及动态

人工智能产业链联盟创始人

邀请你加入星球，一起学习

人工智能产业链联盟报告库

星主：人工智能产业链联盟创始人

每天仅需0.5元，即可拥有以下福利！
每周更新各类机构的最新研究成果。立志将人工智能产业链联盟打造成市面上最全的AI研究资料库，覆盖券商、产业公司、研究院所等...

Call to Actions

- 关注和试用 IPEX-LLM，并给我们反馈
  - *https://github.com/intel-analytics/ipex-llm/*

- 使用 IPEX-LLM 在 Intel XPU 平台开发大模型及其应用
  - 客户端-边缘-服务器（Intel Core Ultra AI PC、AI座舱、Xeon+Intel Arc GPUs）
  - 高效的大模型 XPU 加速的创新
  - 大模型应用场景的创新

谢谢!

Notices &Disclaimers

Performance varies by use, configuration and other factors. Learn more on the Performance Index site.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.