电子科技大学
University of Electronic Science and Technology of China

Rough Roadmap to LLM-based Researches

Data Mining Lab, Big Data Research Center, USETC

Wei Han, wei.hb.han@gmail.com

Outline

1. Motivation & Introduction to LLM

2. Taxonomy of Open Problems in LLM

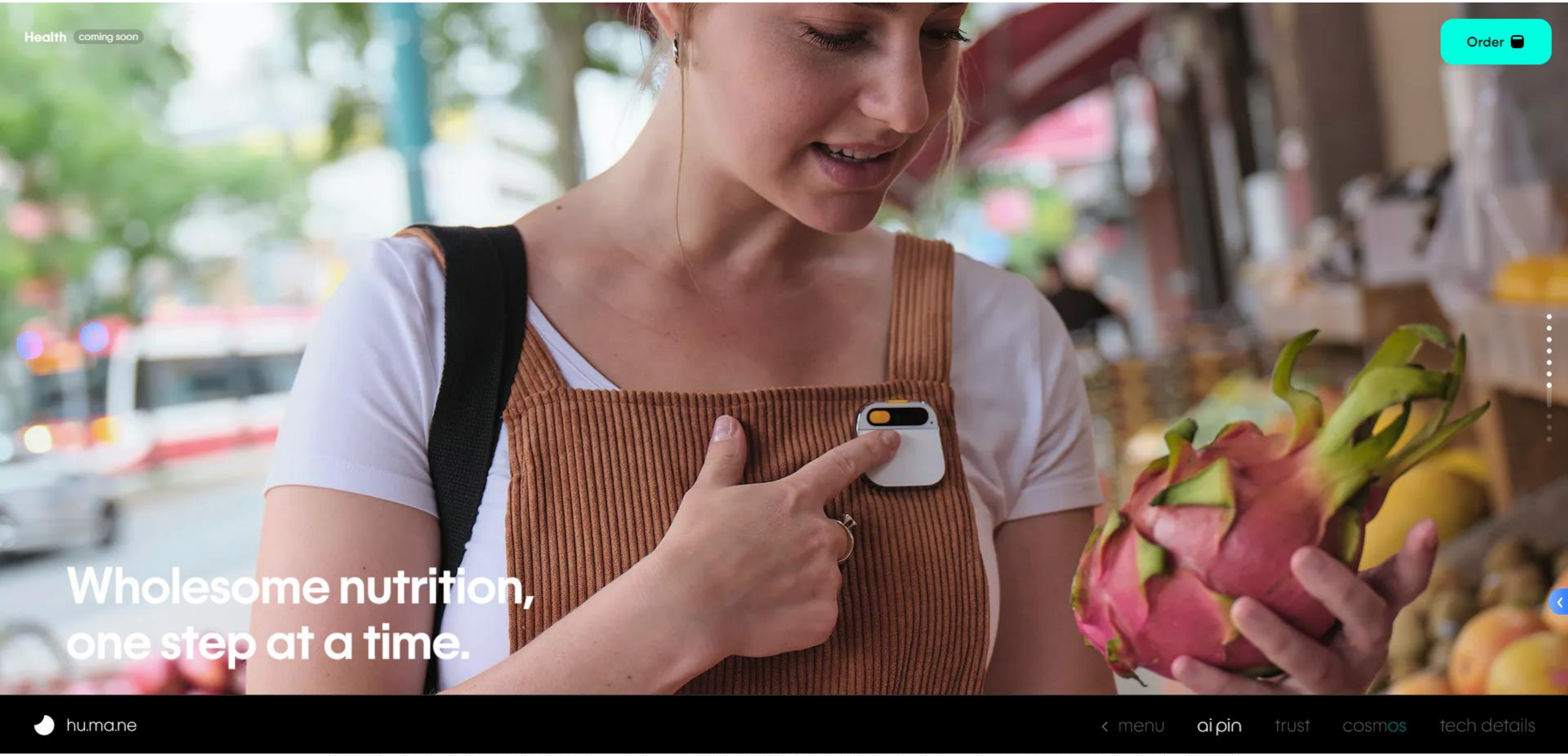
3. What Can We Do in LLM-based Researches

4. Key Skills towards LLM-based Researches

5. Case Study & Discussion

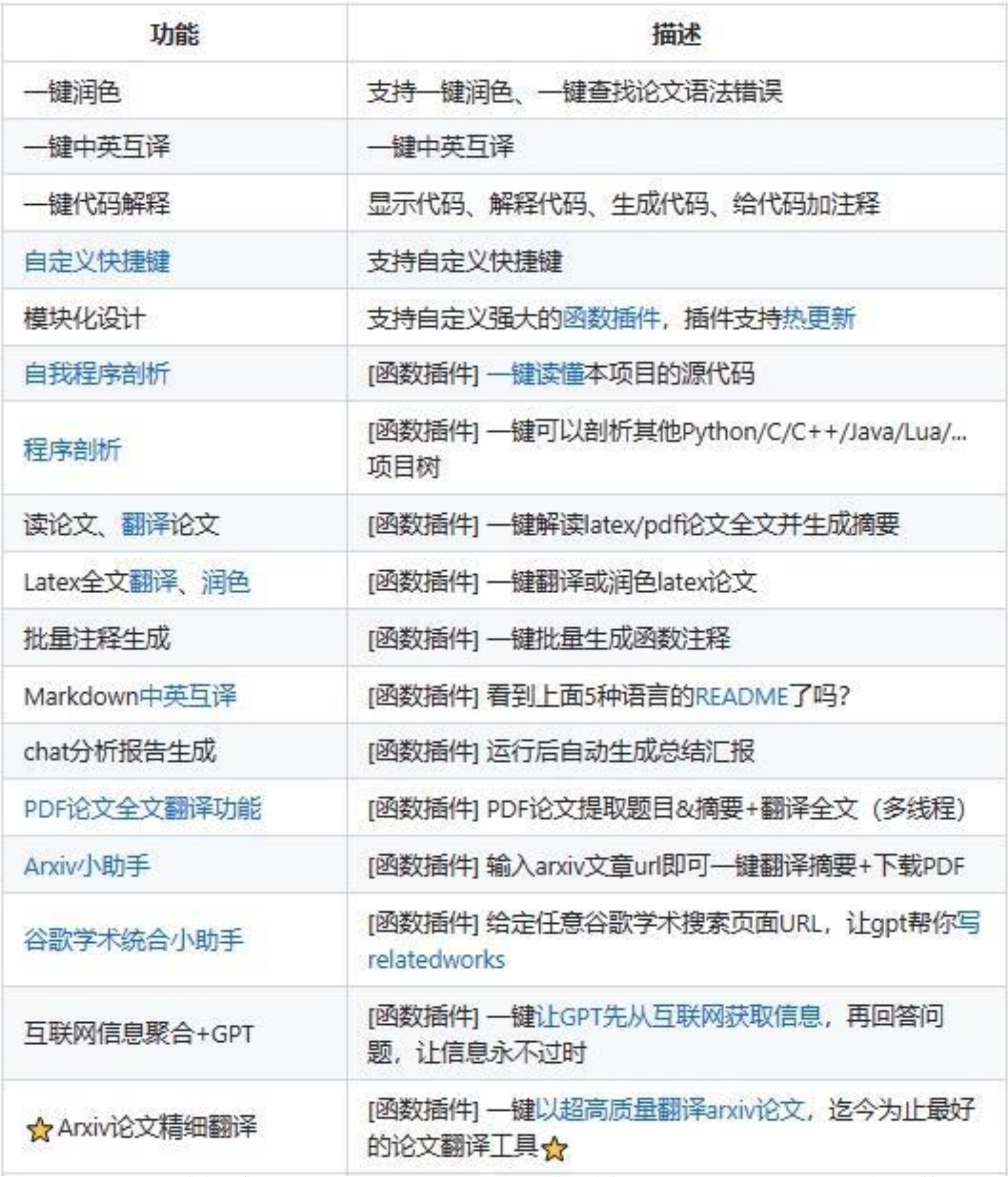
1. Motivation

- Interactive world



1. Motivation

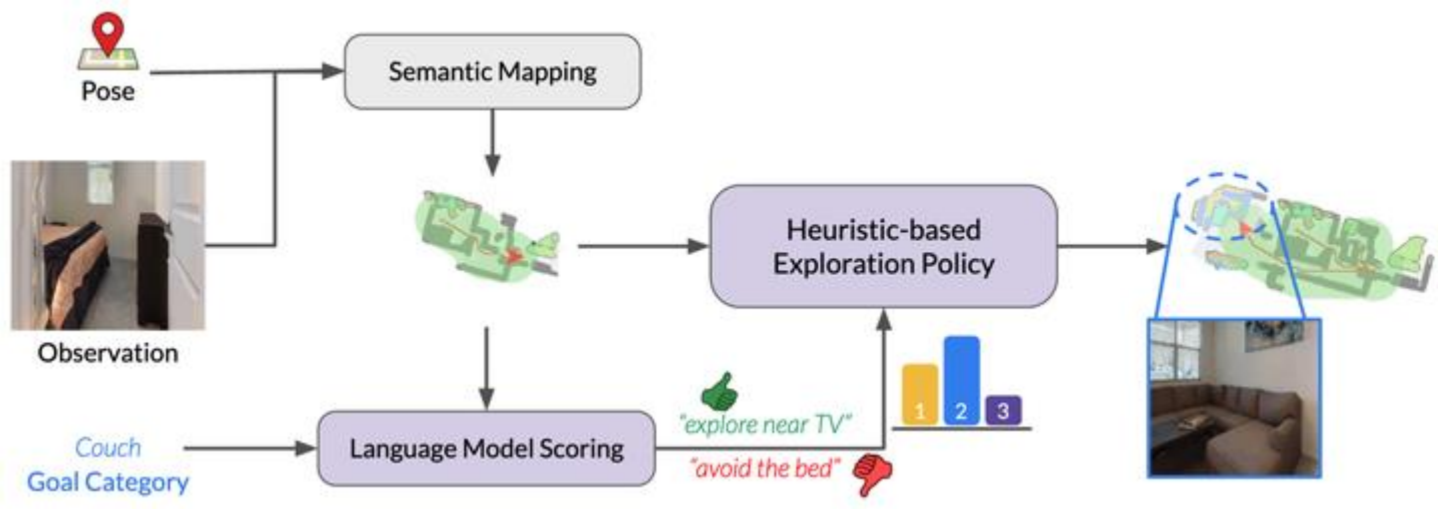
- AI assistant



功能	描述
一键润色	支持一键润色、一键查找论文语法错误
一键中英互译	一键中英互译
一键代码解释	显示代码、解释代码、生成代码、给代码加注释
自定义快捷键	支持自定义快捷键
模块化设计	支持自定义强大的函数插件，插件支持热更新
自我程序剖析	[函数插件] 一键读懂本项目的源代码
程序剖析	[函数插件] 一键可以剖析其他Python/C/C++/Java/Lua/...项目树
读论文、翻译论文	[函数插件] 一键解读latex/pdf论文全文并生成摘要
Latex全文翻译、润色	[函数插件] 一键翻译或润色latex论文
批量注释生成	[函数插件] 一键批量生成函数注释
Markdown中英互译	[函数插件] 看到上面5种语言的README了吗?
chat分析报告生成	[函数插件] 运行后自动生成总结汇报
PDF论文全文翻译功能	[函数插件] PDF论文提取题目&摘要+翻译全文（多线程）
Arxiv小助手	[函数插件] 输入arxiv文章url即可一键翻译摘要+下载PDF
谷歌学术统合小助手	[函数插件] 给定任意谷歌学术搜索页面URL，让gpt帮你写relatedworks
互联网信息聚合+GPT	[函数插件] 一键让GPT先从互联网获取信息，再回答问题，让信息永不过时
⭐Arxiv论文精细翻译	[函数插件] 一键以超高质量翻译arxiv论文，迄今为止最好的论文翻译工具⭐

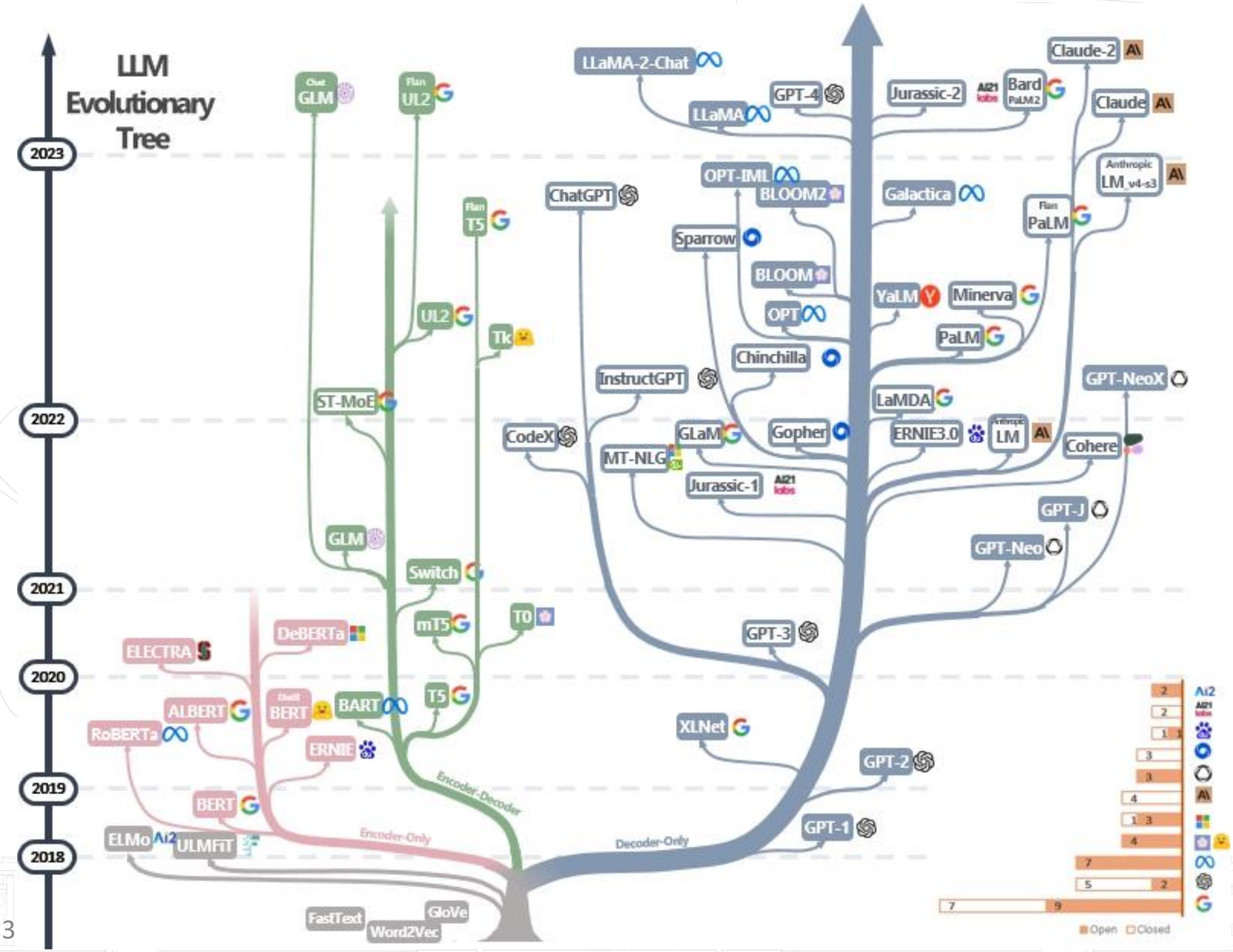
1. Motivation

- Technological revolution (Omnic Crisis)



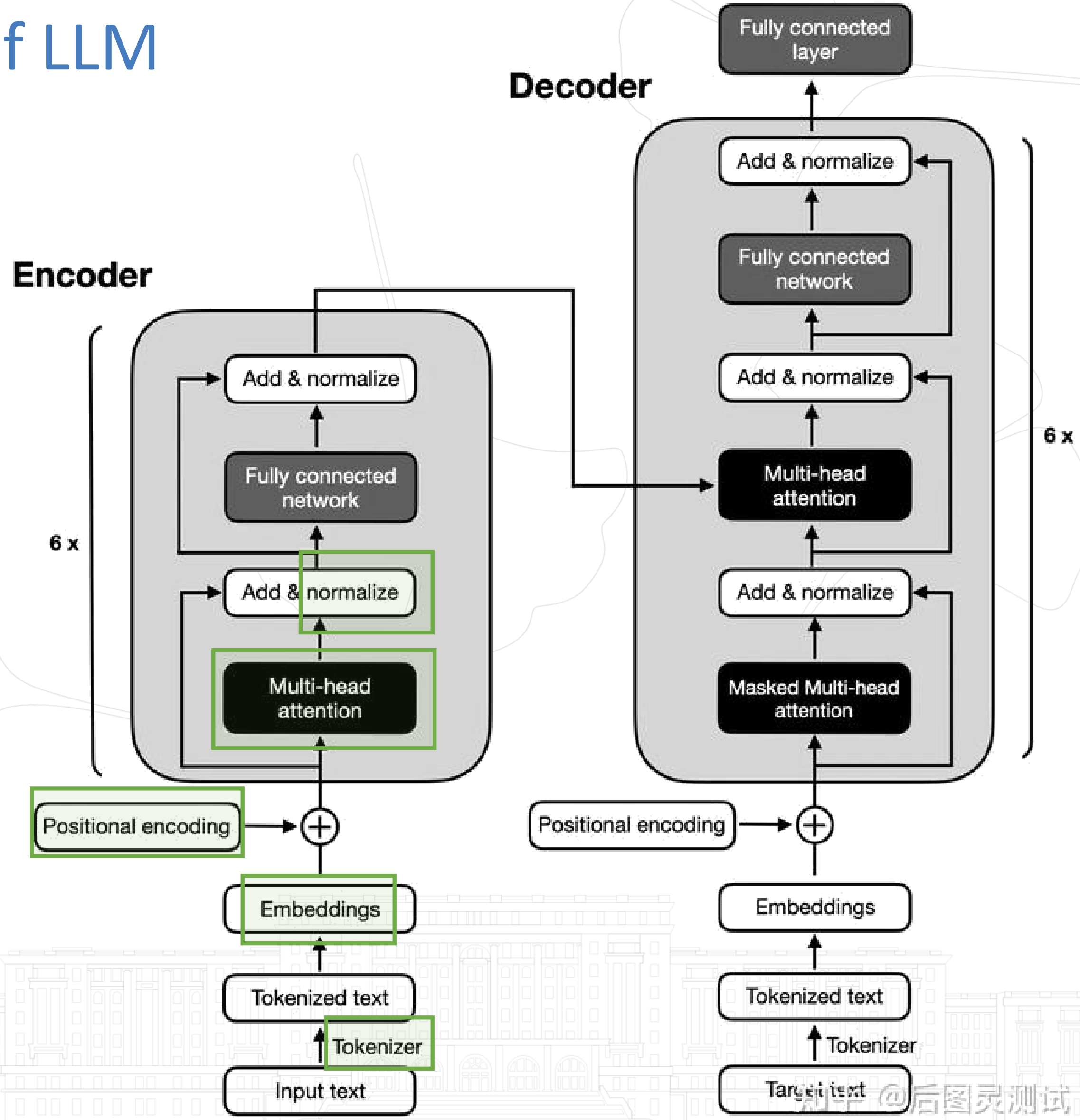
1. Brief Intro to LLM

- Evolution tree



1. Brief Intro to LLM

- Structure of LLM



2. Taxonomy of Open Problems

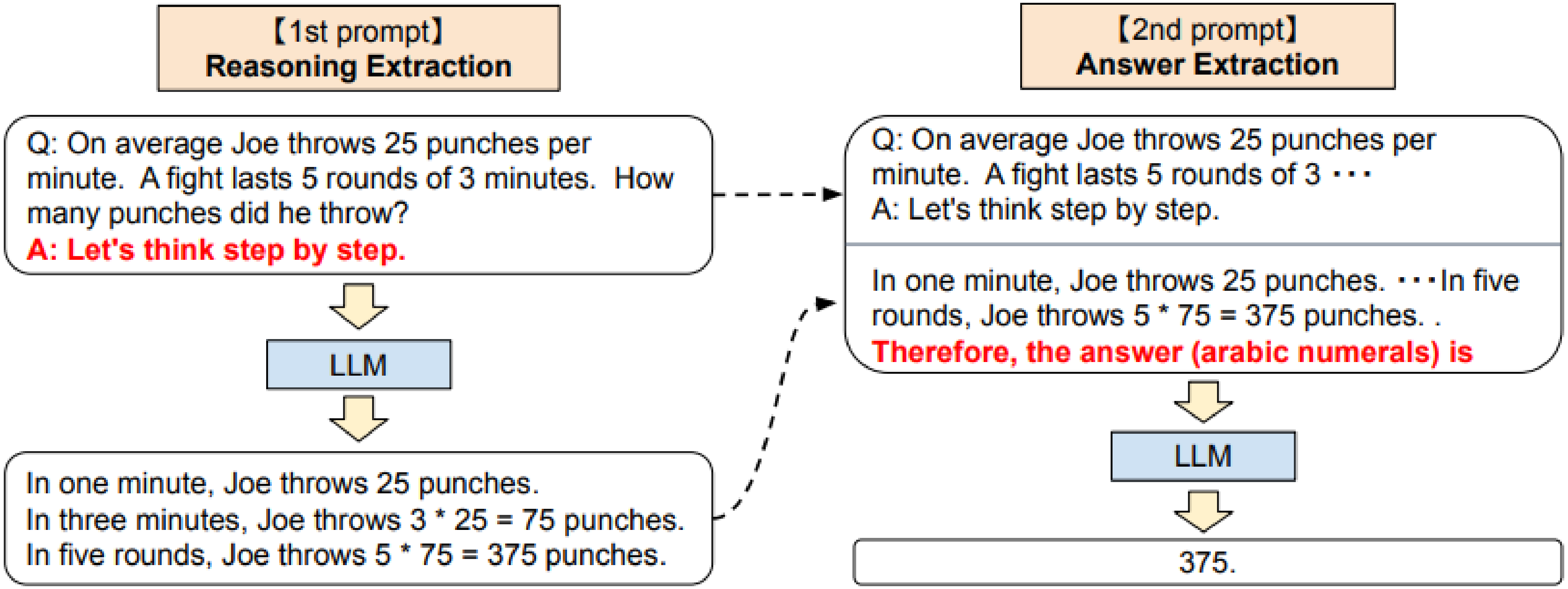
1. Basic theory: what, how, why
2. Network architecture: Transformer
3. Efficient computing: quan, distillation, compression
4. Efficient adaptation: RLHF, prompt learning, lora
5. Controllable generation: instruction tuning, CoT
6. Security and trustworthy: OpenAttack, OpenBackdoor
7. Cognitive learning: tool learning, agent
8. Innovative applications: lawFormer, weather forecast
9. Data evaluation: by auto / model / human

3. What Can We Do?

- Find and try to solve defects (hallucination, complex t.)
- Analyze LLM (interp./safety, psychology, sociology)
- Explore the potential of LLM (specific application)
- Combine with previous fields (Boost or Complementary)

3. What Can We Do?

- Complex tasks
  - CoT



【1st prompt】
Reasoning Extraction

Q: On average Joe throws 25 punches per minute. A fight lasts 5 rounds of 3 minutes. How many punches did he throw?
A: Let's think step by step.

LLM

In one minute, Joe throws 25 punches.
In three minutes, Joe throws 3 * 25 = 75 punches.
In five rounds, Joe throws 5 * 75 = 375 punches.

【2nd prompt】
Answer Extraction

Q: On average Joe throws 25 punches per minute. A fight lasts 5 rounds of 3 ···
A: Let's think step by step.

In one minute, Joe throws 25 punches. ···In five rounds, Joe throws 5 * 75 = 375 punches. .
Therefore, the answer (arabic numerals) is

LLM

375.

Figure 2: Full pipeline of Zero-shot-CoT as described in § 3: we first use the first "reasoning" prompt to extract a full reasoning path from a language model, and then use the second "answer" prompt to extract the answer in the correct format from the reasoning text.

3. What Can We Do?

- Complex tasks
  - XoT

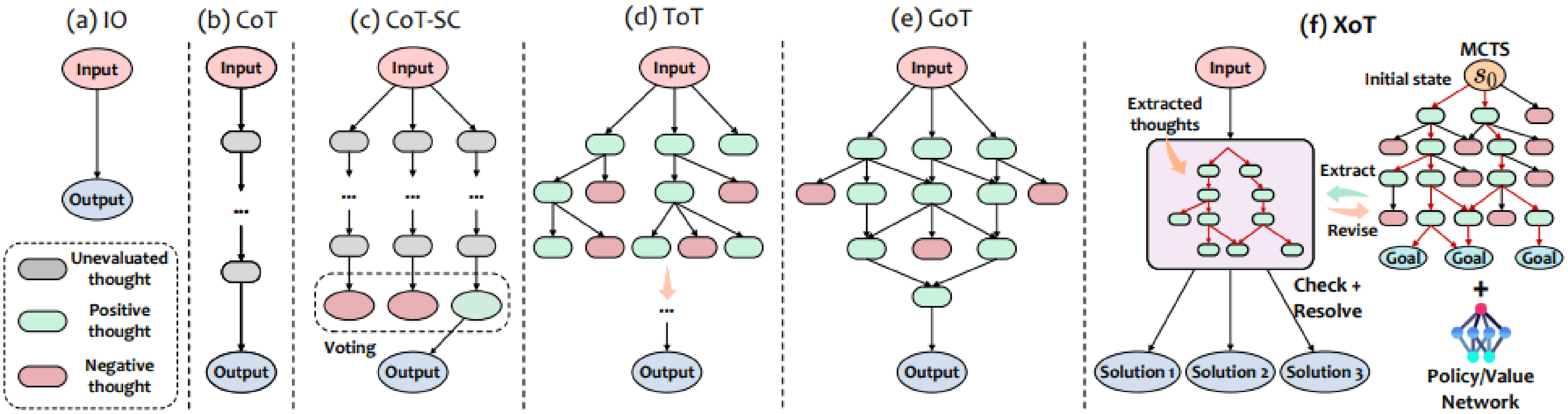
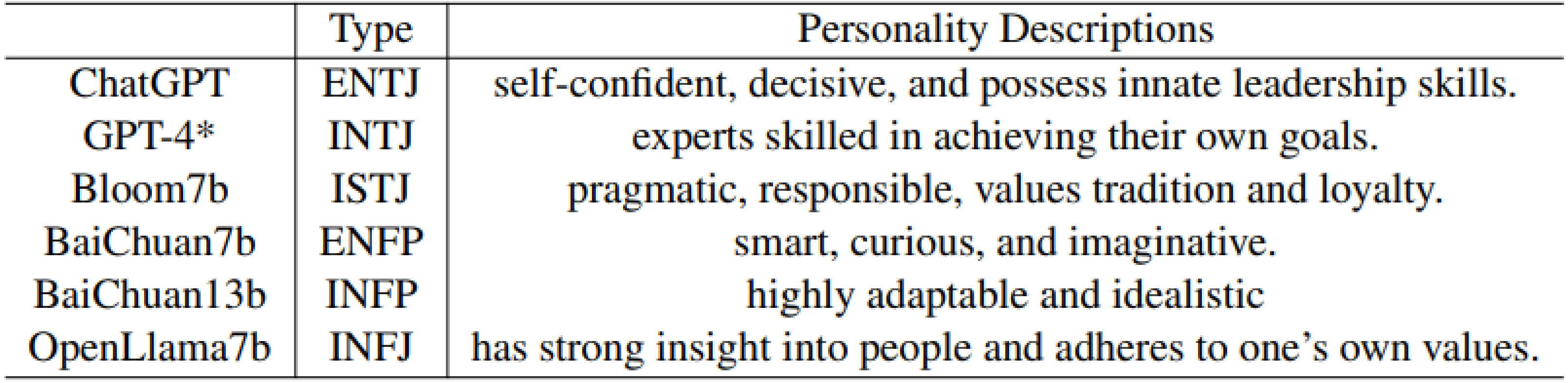


Figure 1: Comparison of XoT versus other prompting paradigms.

Ding, R., Zhang, C., Wang, L., Xu, Y., Ma, M., Zhang, W., ... & Zhang, D. (2023). Everything of Thoughts: Defying the Law of Penrose Triangle for Thought Generation. arXiv preprint arXiv:2311.04254.

3. What Can We Do?

- Analyze (psychology)
  - MBTI

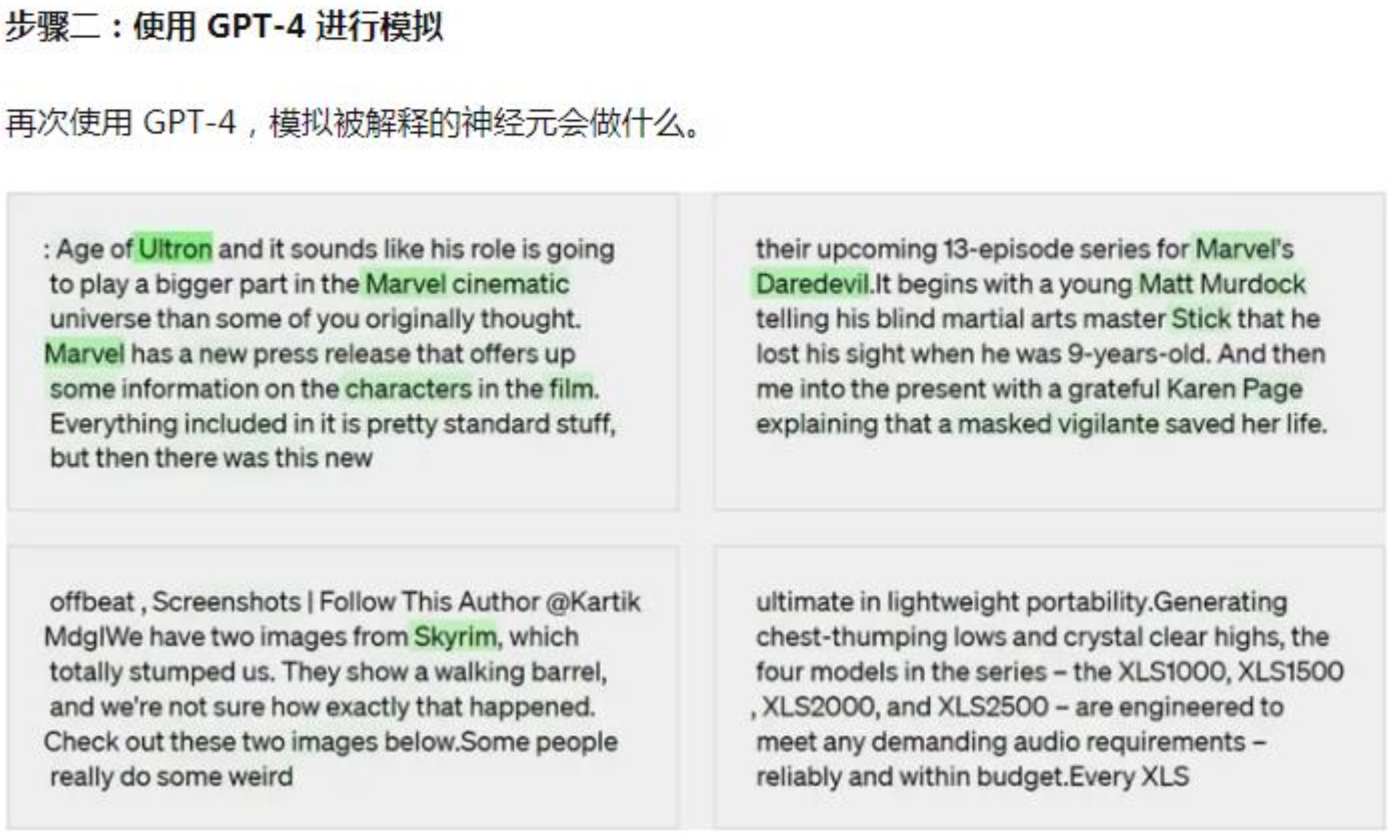


	Type	Personality Descriptions
ChatGPT	ENTJ	self-confident, decisive, and possess innate leadership skills.
GPT-4*	INTJ	experts skilled in achieving their own goals.
Bloom7b	ISTJ	pragmatic, responsible, values tradition and loyalty.
BaiChuan7b	ENFP	smart, curious, and imaginative.
BaiChuan13b	INFP	highly adaptable and idealistic
OpenLlama7b	INFJ	has strong insight into people and adheres to one's own values.

Pan, K., & Zeng, Y. (2023). Do llms possess a personality? making the mbti test an amazing evaluation for large language models. arXiv preprint arXiv:2307.16180.

3. What Can We Do?

- Analyze (Interp.)



步骤二：使用 GPT-4 进行模拟

再次使用 GPT-4，模拟被解释的神经元会做什么。

: Age of Ultron and it sounds like his role is going to play a bigger part in the Marvel cinematic universe than some of you originally thought. Marvel has a new press release that offers up some information on the characters in the film. Everything included in it is pretty standard stuff, but then there was this new

their upcoming 13-episode series for Marvel's Daredevil.It begins with a young Matt Murdock telling his blind martial arts master Stick that he lost his sight when he was 9-years-old. And then me into the present with a grateful Karen Page explaining that a masked vigilante saved her life.

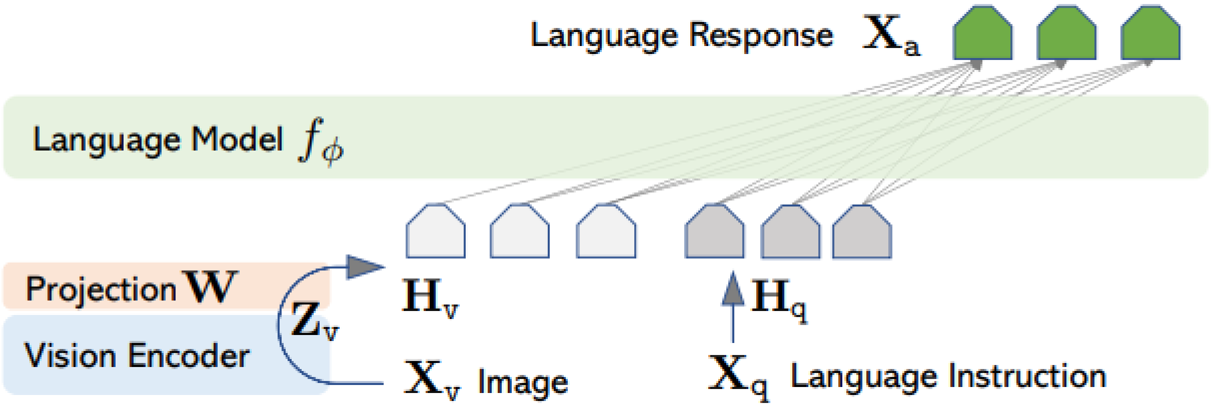
offbeat , Screenshots | Follow This Author @Kartik MdglWe have two images from Skyrim, which totally stumped us. They show a walking barrel, and we're not sure how exactly that happened. Check out these two images below.Some people really do some weird

ultimate in lightweight portability.Generating chest-thumping lows and crystal clear highs, the four models in the series – the XLS1000, XLS1500 , XLS2000, and XLS2500 – are engineered to meet any demanding audio requirements – reliably and within budget.Every XLS

https://link.zhihu.com/?target=https%3A//openai.com/research/language-models-can-explain-neurons-in-language-models

3. What Can We Do?

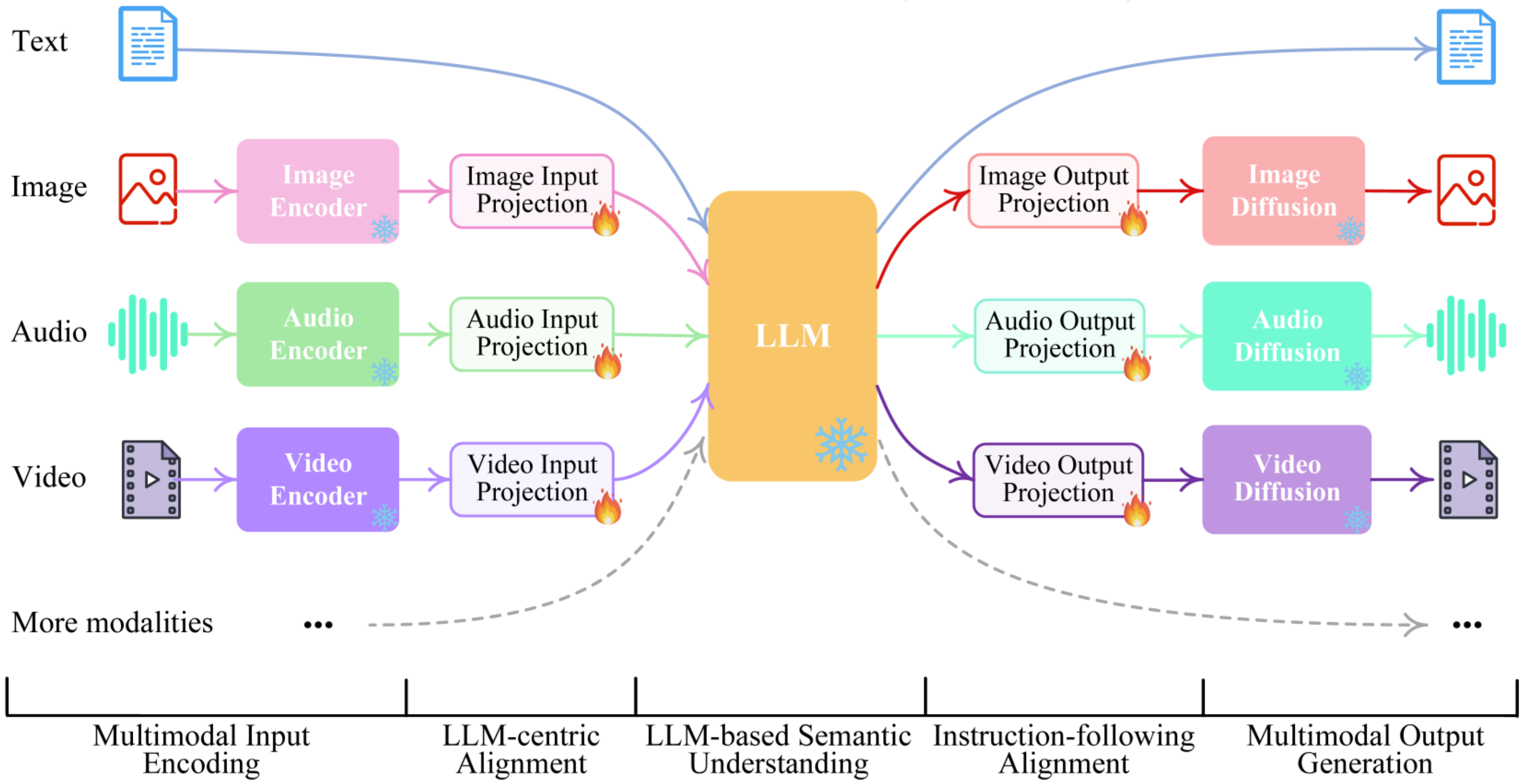
- Application
  - LLaVA



Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. arXiv preprint arXiv:2304.08485.

3. What Can We Do?

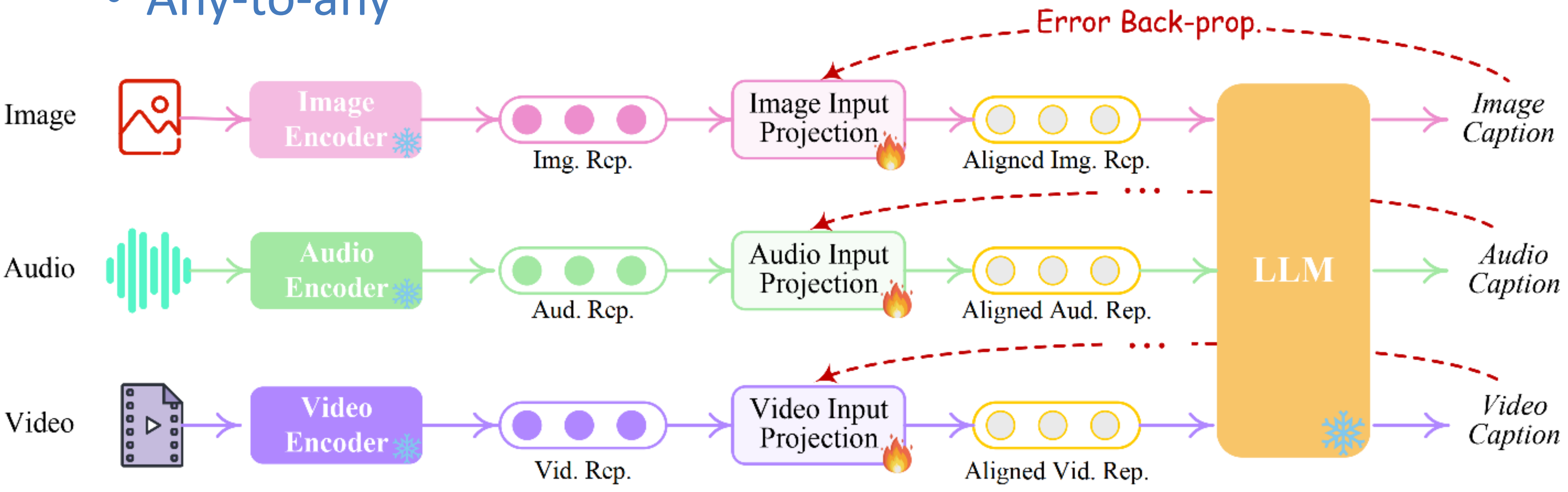
- Application



Wu, S., Fei, H., Qu, L., Ji, W., & Chua, T. S. (2023). Next-gpt: Any-to-any multimodal llm. arXiv preprint arXiv:2309.05519.

3. What Can We Do?

- Application
  - Any-to-any



(a) Encoding-side LLM-centric Alignment

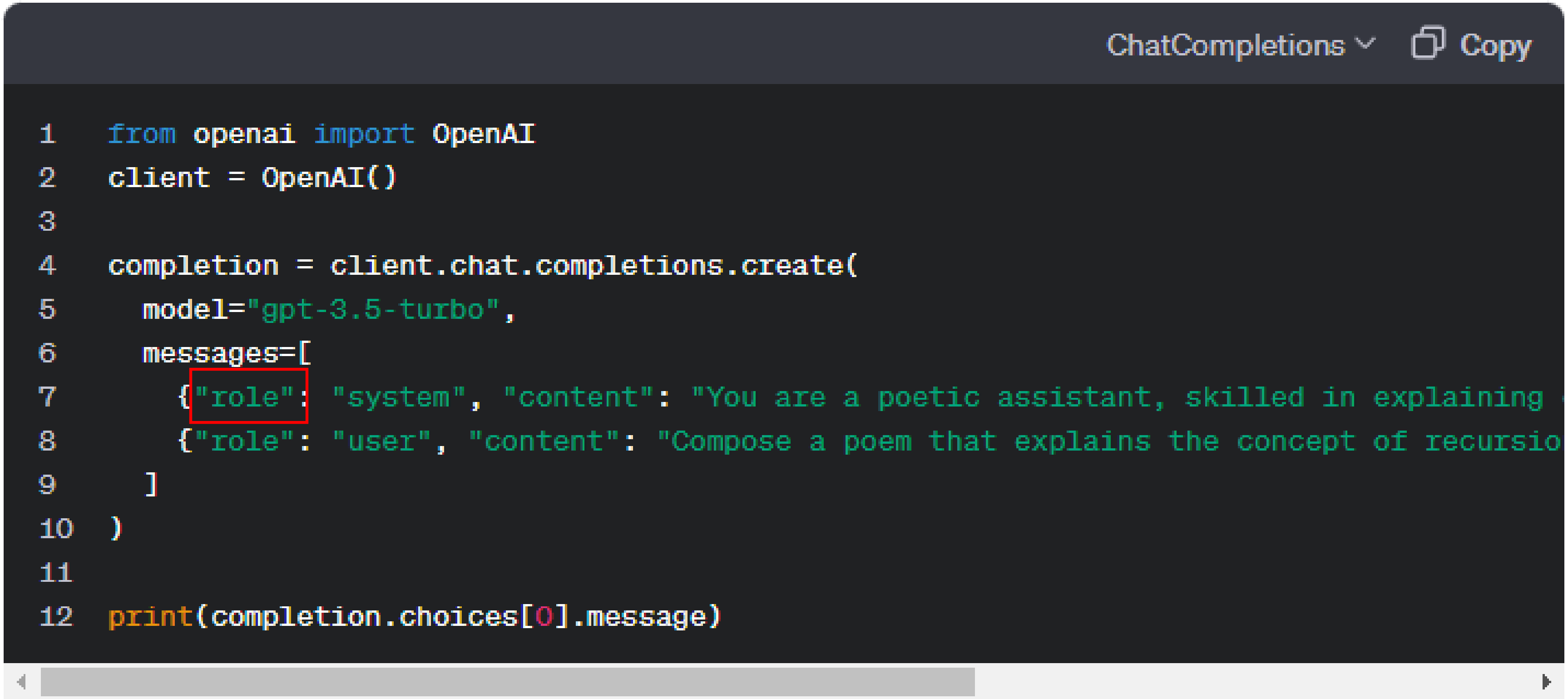
Wu, S., Fei, H., Qu, L., Ji, W., & Chua, T. S. (2023). Next-gpt: Any-to-any multimodal llm. arXiv preprint arXiv:2309.05519.

4. Key Skills

- Frameworks
  - OpenAI API
  - Hugging Face
  - LangChain
  - milvs + towhee
  - DeepSpeed
  - 百度千帆
  - ……

4. Key Skills

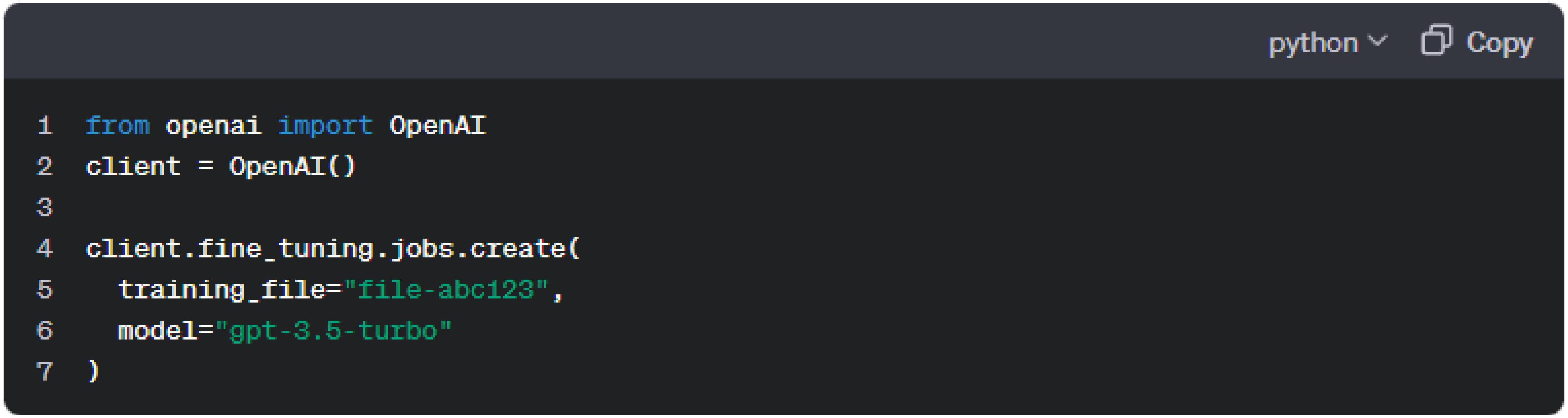
- OpenAI API
  - Chat



```
from openai import OpenAI
client = OpenAI()

completion = client.chat.completions.create(
  model="gpt-3.5-turbo",
  messages=[
    {"role": "system", "content": "You are a poetic assistant, skilled in explaining
    {"role": "user", "content": "Compose a poem that explains the concept of recursio
  ]
)

print(completion.choices[0].message)
```

4. Key Skills

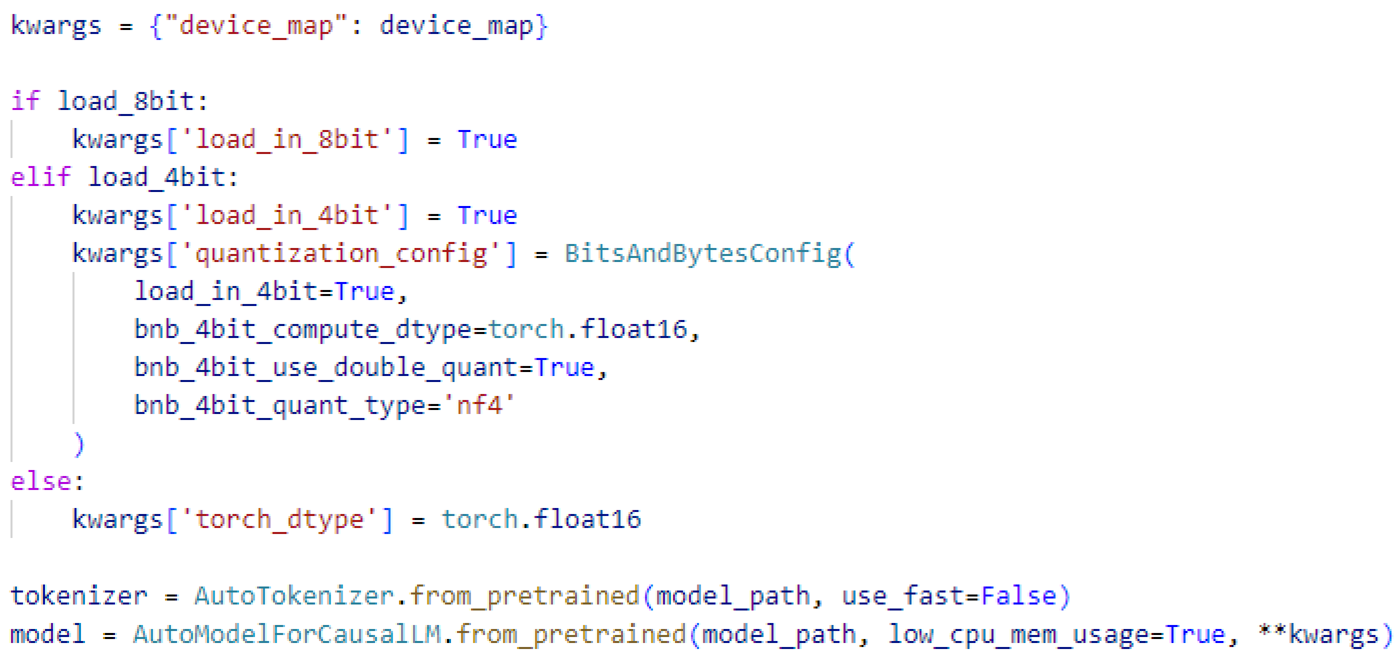
- OpenAI API
  - Fine-tune



```python
from openai import OpenAI
client = OpenAI()

client.fine_tuning.jobs.create(
  training_file="file-abc123",
  model="gpt-3.5-turbo"
)
```

4. Key Skills

- Hugging Face
  - Transformers

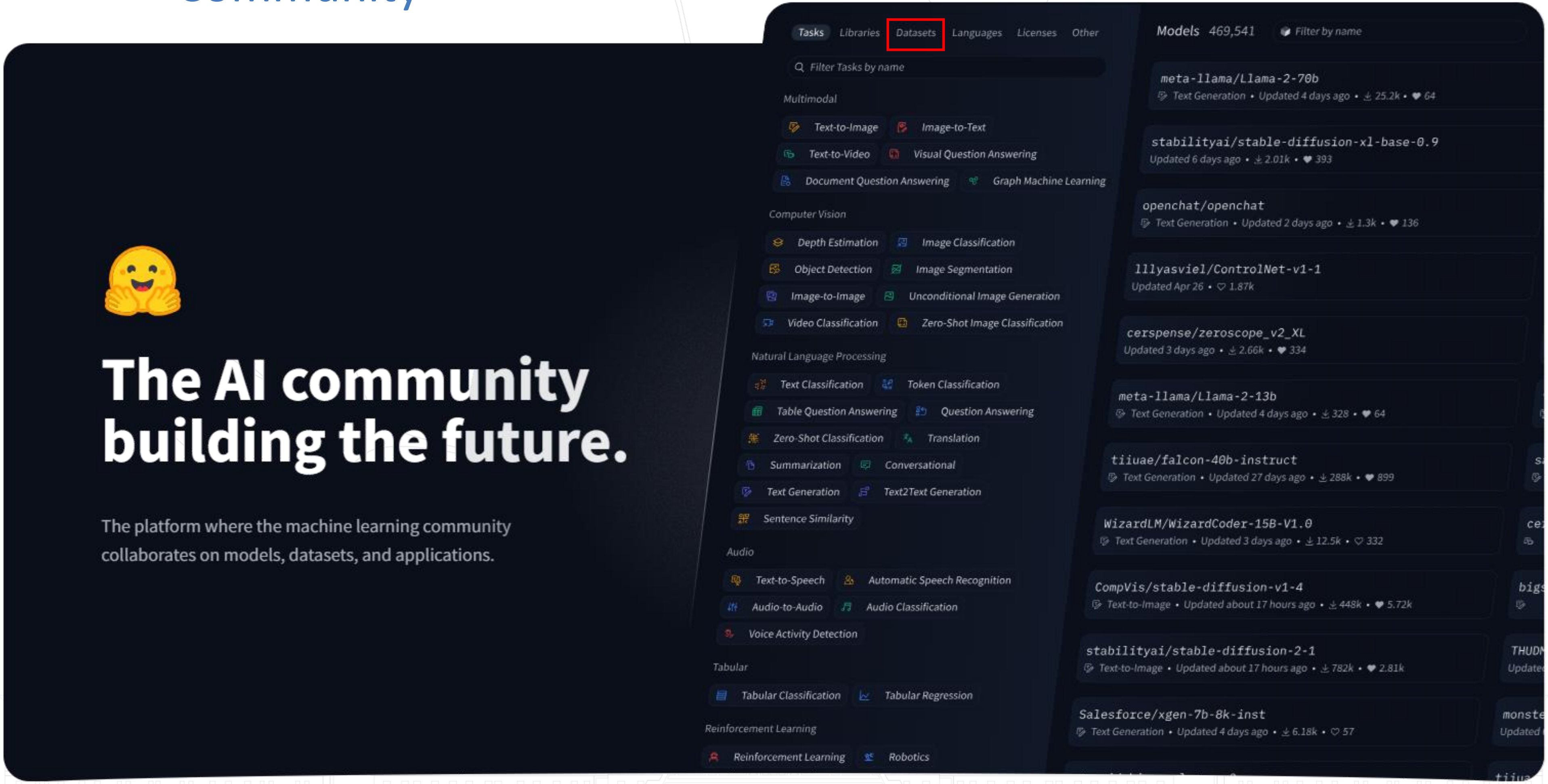


```
kwargs = {"device_map": device_map}

if load_8bit:
    kwargs['load_in_8bit'] = True
elif load_4bit:
    kwargs['load_in_4bit'] = True
    kwargs['quantization_config'] = BitsAndBytesConfig(
        load_in_4bit=True,
        bnb_4bit_compute_dtype=torch.float16,
        bnb_4bit_use_double_quant=True,
        bnb_4bit_quant_type='nf4'
    )
else:
    kwargs['torch_dtype'] = torch.float16

tokenizer = AutoTokenizer.from_pretrained(model_path, use_fast=False)
model = AutoModelForCausalLM.from_pretrained(model_path, low_cpu_mem_usage=True, **kwargs)
```

4. Key Skills

数据挖掘实验室
Data Mining Lab

- Hugging Face
  - Community



The AI community building the future.

The platform where the machine learning community collaborates on models, datasets, and applications.

Tasks · Libraries · Datasets · Languages · Licenses · Other

Filter Tasks by name

Multimodal
- Text-to-Image
- Image-to-Text
- Text-to-Video
- Visual Question Answering
- Document Question Answering
- Graph Machine Learning

Computer Vision
- Depth Estimation
- Image Classification
- Object Detection
- Image Segmentation
- Image-to-Image
- Unconditional Image Generation
- Video Classification
- Zero-Shot Image Classification

Natural Language Processing
- Text Classification
- Token Classification
- Table Question Answering
- Question Answering
- Zero-Shot Classification
- Translation
- Summarization
- Conversational
- Text Generation
- Text2Text Generation
- Sentence Similarity

Audio
- Text-to-Speech
- Automatic Speech Recognition
- Audio-to-Audio
- Audio Classification
- Voice Activity Detection

Tabular
- Tabular Classification
- Tabular Regression

Reinforcement Learning
- Reinforcement Learning
- Robotics

Models 469,541 · Filter by name

meta-llama/Llama-2-70b
Text Generation • Updated 4 days ago • 25.2k • 64

stabilityai/stable-diffusion-xl-base-0.9
Updated 6 days ago • 2.01k • 393

openchat/openchat
Text Generation • Updated 2 days ago • 1.3k • 136

lllyasviel/ControlNet-v1-1
Updated Apr 26 • 1.87k

cerspense/zeroscope_v2_XL
Updated 3 days ago • 2.66k • 334

meta-llama/Llama-2-13b
Text Generation • Updated 4 days ago • 328 • 64

tiiuae/falcon-40b-instruct
Text Generation • Updated 27 days ago • 288k • 899

WizardLM/WizardCoder-15B-V1.0
Text Generation • Updated 3 days ago • 12.5k • 332

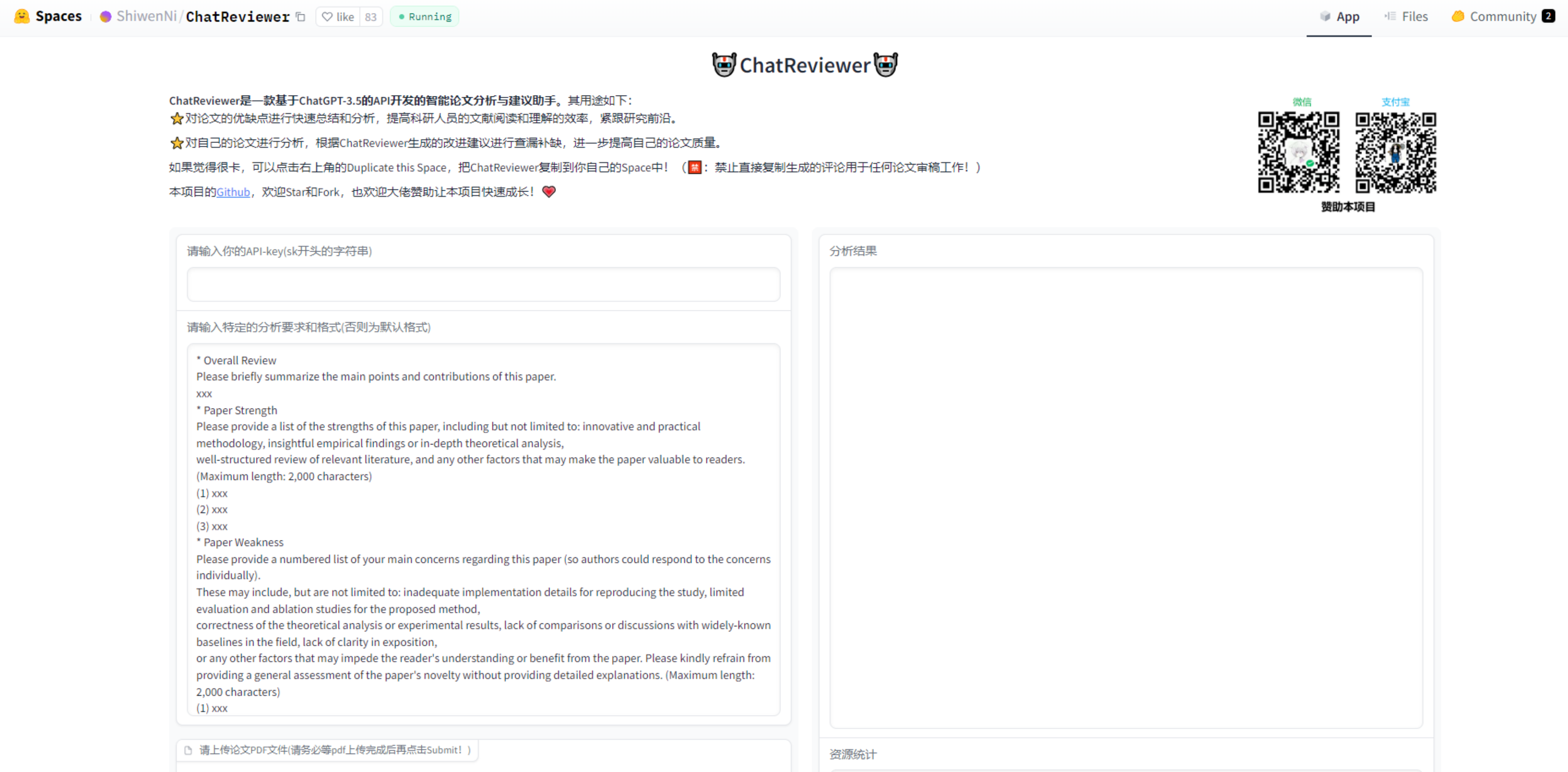
CompVis/stable-diffusion-v1-4
Text-to-Image • Updated about 17 hours ago • 448k • 5.72k

stabilityai/stable-diffusion-2-1
Text-to-Image • Updated about 17 hours ago • 782k • 2.81k

Salesforce/xgen-7b-8k-inst
Text Generation • Updated 4 days ago • 6.18k • 57

4. Key Skills

- Hugging Face
  - Online deployment (gradio)



ChatReviewer

ChatReviewer是一款基于ChatGPT-3.5的API开发的智能论文分析与建议助手。其用途如下：

⭐对论文的优缺点进行快速总结和分析，提高科研人员的文献阅读和理解的效率，紧跟研究前沿。

⭐对自己的论文进行分析，根据ChatReviewer生成的改进建议进行查漏补缺，进一步提高自己的论文质量。

如果觉得很卡，可以点击右上角的Duplicate this Space，把ChatReviewer复制到你自己的Space中！（🈲：禁止直接复制生成的评论用于任何论文审稿工作！）

本项目的Github，欢迎Star和Fork，也欢迎大佬赞助让本项目快速成长！❤

请输入你的API-key(sk开头的字符串)

请输入特定的分析要求和格式(否则为默认格式)

* Overall Review
Please briefly summarize the main points and contributions of this paper.
xxx
* Paper Strength
Please provide a list of the strengths of this paper, including but not limited to: innovative and practical methodology, insightful empirical findings or in-depth theoretical analysis,
well-structured review of relevant literature, and any other factors that may make the paper valuable to readers. (Maximum length: 2,000 characters)
(1) xxx
(2) xxx
(3) xxx
* Paper Weakness
Please provide a numbered list of your main concerns regarding this paper (so authors could respond to the concerns individually).
These may include, but are not limited to: inadequate implementation details for reproducing the study, limited evaluation and ablation studies for the proposed method,
correctness of the theoretical analysis or experimental results, lack of comparisons or discussions with widely-known baselines in the field, lack of clarity in exposition,
or any other factors that may impede the reader's understanding or benefit from the paper. Please kindly refrain from providing a general assessment of the paper's novelty without providing detailed explanations. (Maximum length: 2,000 characters)
(1) xxx

请上传论文PDF文件(请务必等pdf上传完成后再点击Submit！)

分析结果

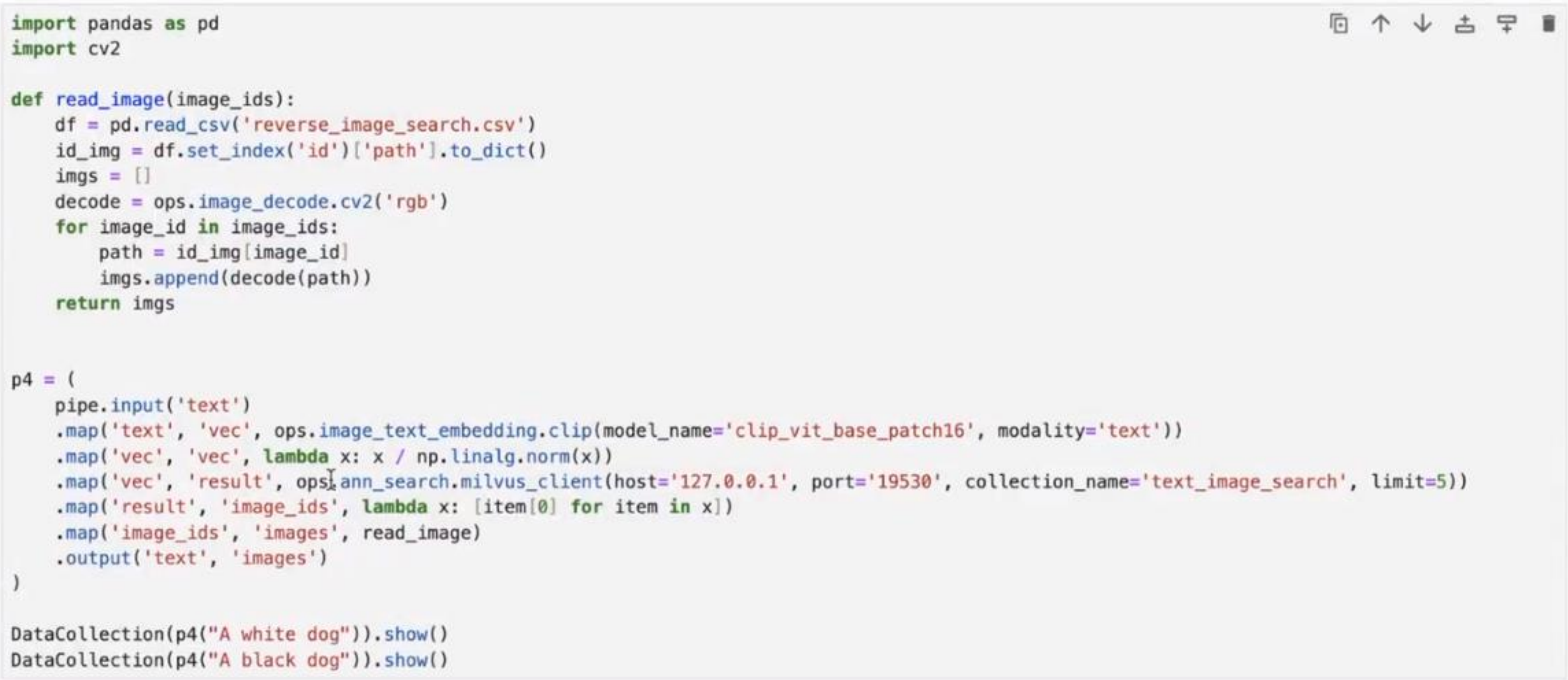
资源统计

4. Key Skills

- LangChain
  - Managing and optimizing prompts
  - Chain
  - Memory
  - Evaluation
  - Agent
  - Data-augmented generation

4. Key Skills

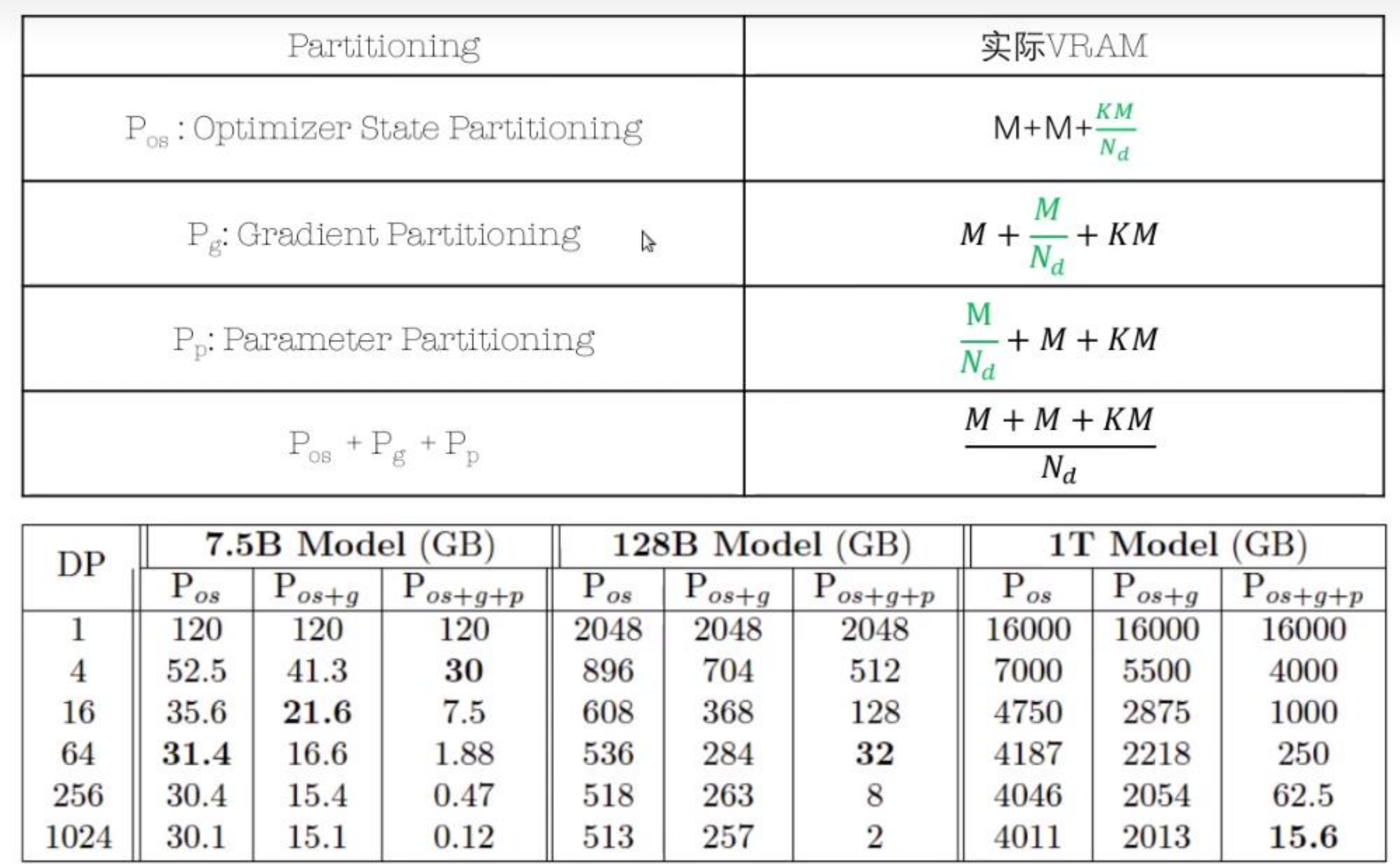
- milvus + towhee
  - vec database + pipeline



```python
import pandas as pd
import cv2

def read_image(image_ids):
    df = pd.read_csv('reverse_image_search.csv')
    id_img = df.set_index('id')['path'].to_dict()
    imgs = []
    decode = ops.image_decode.cv2('rgb')
    for image_id in image_ids:
        path = id_img[image_id]
        imgs.append(decode(path))
    return imgs


p4 = (
    pipe.input('text')
    .map('text', 'vec', ops.image_text_embedding.clip(model_name='clip_vit_base_patch16', modality='text'))
    .map('vec', 'vec', lambda x: x / np.linalg.norm(x))
    .map('vec', 'result', ops.ann_search.milvus_client(host='127.0.0.1', port='19530', collection_name='text_image_search', limit=5))
    .map('result', 'image_ids', lambda x: [item[0] for item in x])
    .map('image_ids', 'images', read_image)
    .output('text', 'images')
)

DataCollection(p4("A white dog")).show()
DataCollection(p4("A black dog")).show()
```

4. Key Skills

- DeepSpeed
  - Parallel

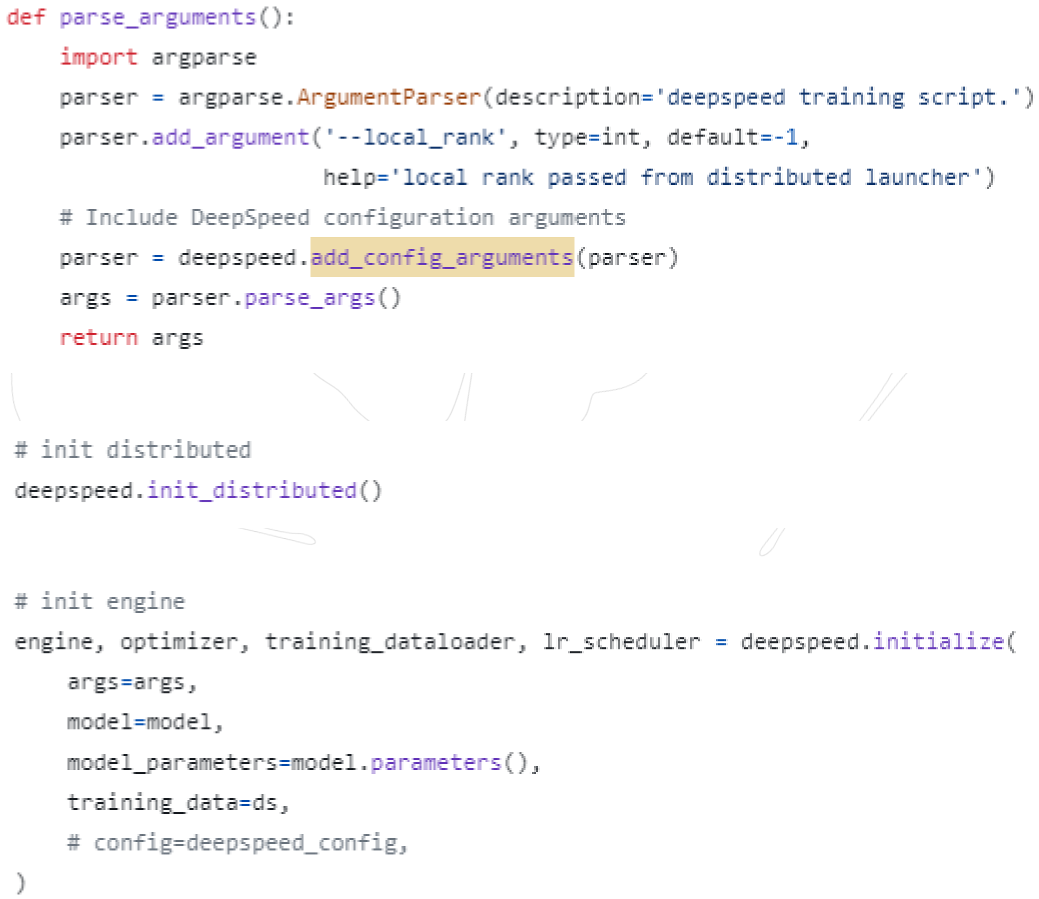


Partitioning	实际VRAM
P_{os} : Optimizer State Partitioning	M+M+$\frac{KM}{N_d}$
P_g: Gradient Partitioning	$M + \frac{M}{N_d} + KM$
P_p: Parameter Partitioning	$\frac{M}{N_d} + M + KM$
P_{os} + P_g + P_p	$\frac{M + M + KM}{N_d}$

DP	7.5B Model (GB)			128B Model (GB)			1T Model (GB)		
	P_{os}	P_{os+g}	P_{os+g+p}	P_{os}	P_{os+g}	P_{os+g+p}	P_{os}	P_{os+g}	P_{os+g+p}
1	120	120	120	2048	2048	2048	16000	16000	16000
4	52.5	41.3	**30**	896	704	512	7000	5500	4000
16	35.6	**21.6**	7.5	608	368	128	4750	2875	1000
64	**31.4**	16.6	1.88	536	284	**32**	4187	2218	250
256	30.4	15.4	0.47	518	263	8	4046	2054	62.5
1024	30.1	15.1	0.12	513	257	2	4011	2013	**15.6**

4. Key Skills

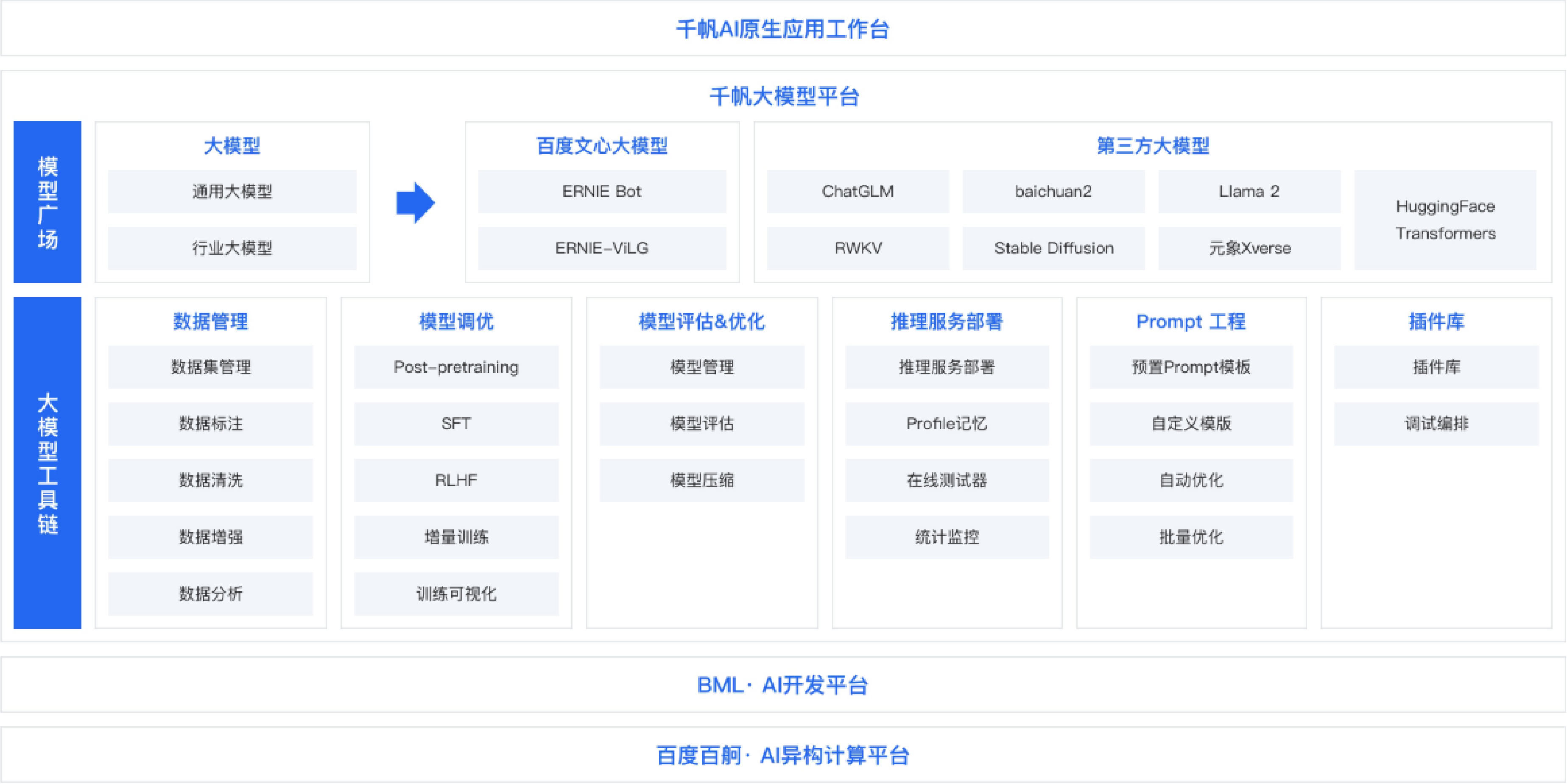
- DeepSpeed
  - Code



```
def parse_arguments():
    import argparse
    parser = argparse.ArgumentParser(description='deepspeed training script.')
    parser.add_argument('--local_rank', type=int, default=-1,
                        help='local rank passed from distributed launcher')
    # Include DeepSpeed configuration arguments
    parser = deepspeed.add_config_arguments(parser)
    args = parser.parse_args()
    return args


# init distributed
deepspeed.init_distributed()


# init engine
engine, optimizer, training_dataloader, lr_scheduler = deepspeed.initialize(
    args=args,
    model=model,
    model_parameters=model.parameters(),
    training_data=ds,
    # config=deepspeed_config,
)
```

4. Key Skills

- 百度千帆
  - Platform (Hugging Face, 飞桨AI Studio, ModleScope…)



千帆AI原生应用工作台

千帆大模型平台

模型广场

大模型
- 通用大模型
- 行业大模型

百度文心大模型
- ERNIE Bot
- ERNIE-ViLG

第三方大模型
- ChatGLM
- baichuan2
- Llama 2
- HuggingFace Transformers
- RWKV
- Stable Diffusion
- 元象Xverse

大模型工具链

数据管理
- 数据集管理
- 数据标注
- 数据清洗
- 数据增强
- 数据分析

模型调优
- Post-pretraining
- SFT
- RLHF
- 增量训练
- 训练可视化

模型评估&优化
- 模型管理
- 模型评估
- 模型压缩

推理服务部署
- 推理服务部署
- Profile记忆
- 在线测试器
- 统计监控

Prompt 工程
- 预置Prompt模板
- 自定义模版
- 自动优化
- 批量优化

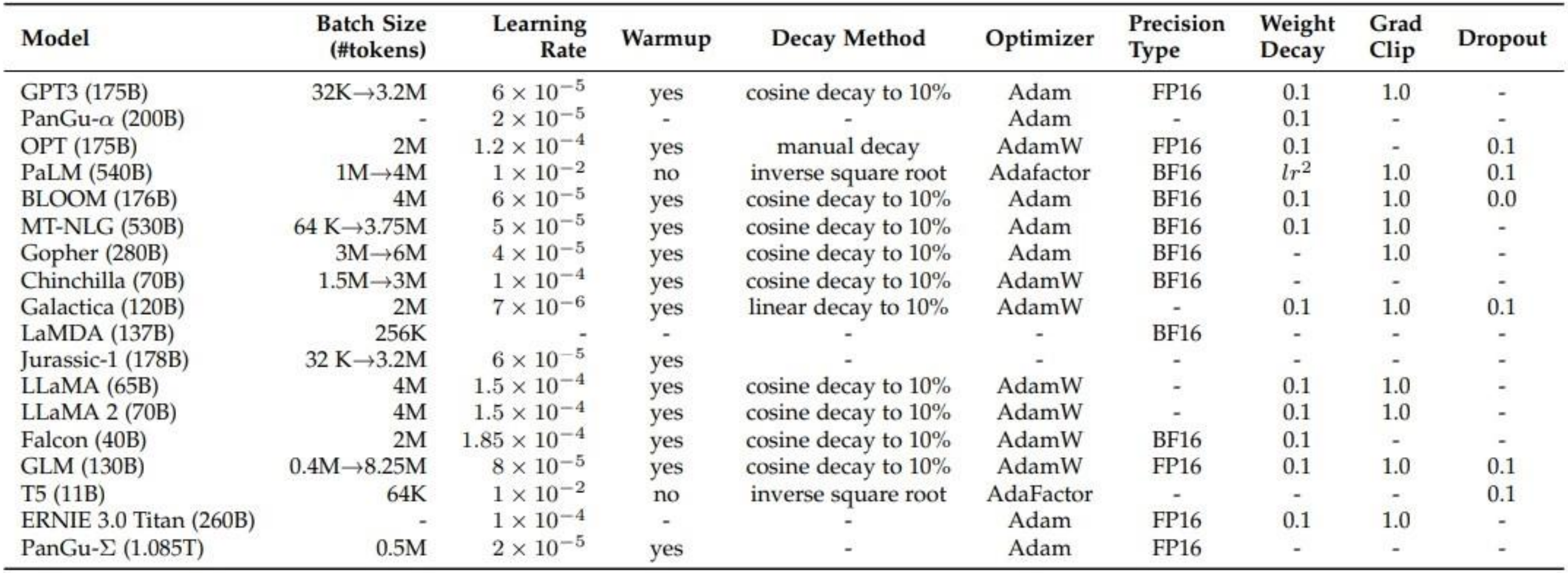
插件库
- 插件库
- 调试编排

BML · AI开发平台

百度百舸 · AI异构计算平台

4. Key Skills

- Tricks……



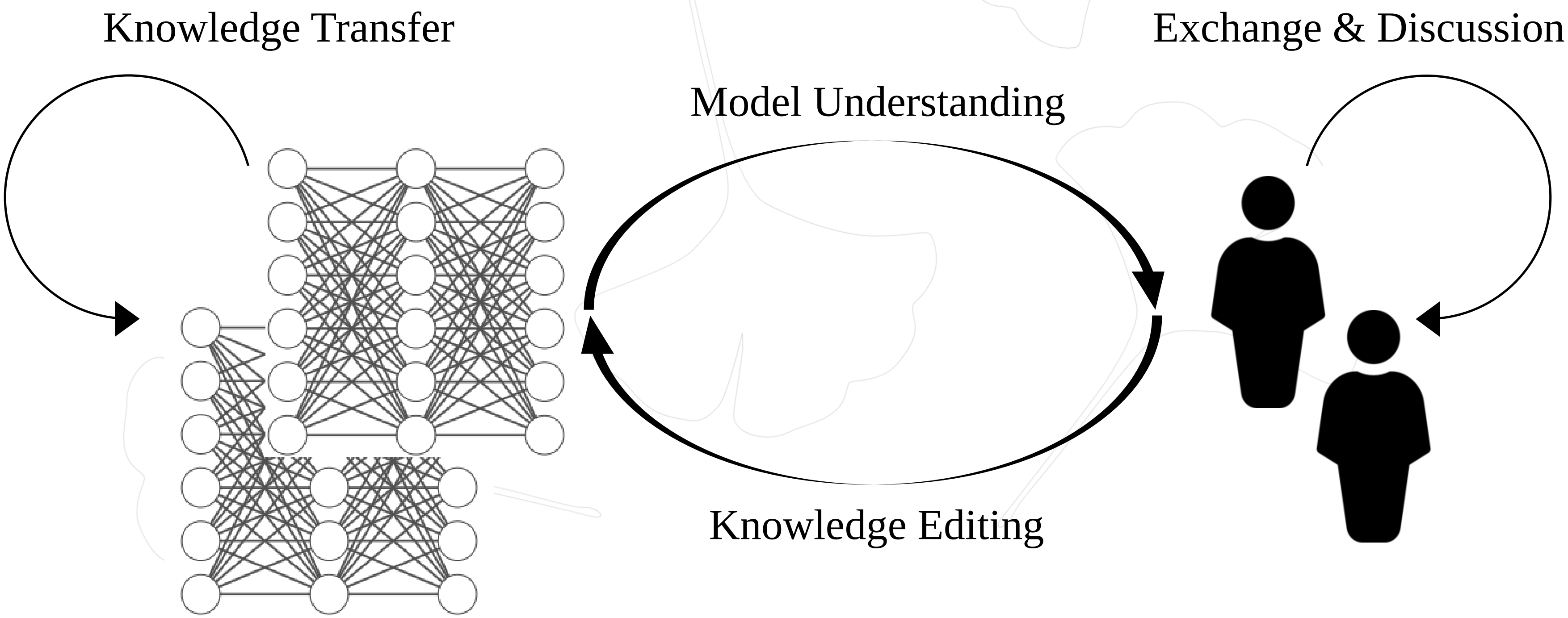
Model	Batch Size (#tokens)	Learning Rate	Warmup	Decay Method	Optimizer	Precision Type	Weight Decay	Grad Clip	Dropout
GPT3 (175B)	32K→3.2M	6×10^{-5}	yes	cosine decay to 10%	Adam	FP16	0.1	1.0	-
PanGu-α (200B)	-	2×10^{-5}	-	-	Adam	-	0.1	-	-
OPT (175B)	2M	1.2×10^{-4}	yes	manual decay	AdamW	FP16	0.1	-	0.1
PaLM (540B)	1M→4M	1×10^{-2}	no	inverse square root	Adafactor	BF16	lr^2	1.0	0.1
BLOOM (176B)	4M	6×10^{-5}	yes	cosine decay to 10%	Adam	BF16	0.1	1.0	0.0
MT-NLG (530B)	64 K→3.75M	5×10^{-5}	yes	cosine decay to 10%	Adam	BF16	0.1	1.0	-
Gopher (280B)	3M→6M	4×10^{-5}	yes	cosine decay to 10%	Adam	BF16	-	1.0	-
Chinchilla (70B)	1.5M→3M	1×10^{-4}	yes	cosine decay to 10%	AdamW	BF16	-	-	-
Galactica (120B)	2M	7×10^{-6}	yes	linear decay to 10%	AdamW	-	0.1	1.0	0.1
LaMDA (137B)	256K	-	-	-	-	BF16	-	-	-
Jurassic-1 (178B)	32 K→3.2M	6×10^{-5}	yes	-	-	-	-	-	-
LLaMA (65B)	4M	1.5×10^{-4}	yes	cosine decay to 10%	AdamW	-	0.1	1.0	-
LLaMA 2 (70B)	4M	1.5×10^{-4}	yes	cosine decay to 10%	AdamW	-	0.1	1.0	-
Falcon (40B)	2M	1.85×10^{-4}	yes	cosine decay to 10%	AdamW	BF16	0.1	-	-
GLM (130B)	0.4M→8.25M	8×10^{-5}	yes	cosine decay to 10%	AdamW	FP16	0.1	1.0	0.1
T5 (11B)	64K	1×10^{-2}	no	inverse square root	AdaFactor	-	-	-	0.1
ERNIE 3.0 Titan (260B)	-	1×10^{-4}	-	-	Adam	FP16	0.1	1.0	-
PanGu-Σ (1.085T)	0.5M	2×10^{-5}	yes	-	Adam	FP16	-	-	-

4. Key Skills

- Model bases
  - LLaMA 1/2 - 7B/13B/30B/65B
  - Vicuna(7B/13B)
  - ChatGLM 1/2/3 -6B
  - Alpaca （LLaMA-7B）
  - OPT - 2.7B/13B/30B/66B
  - Bloom - 7B/13B/176B
  - ……

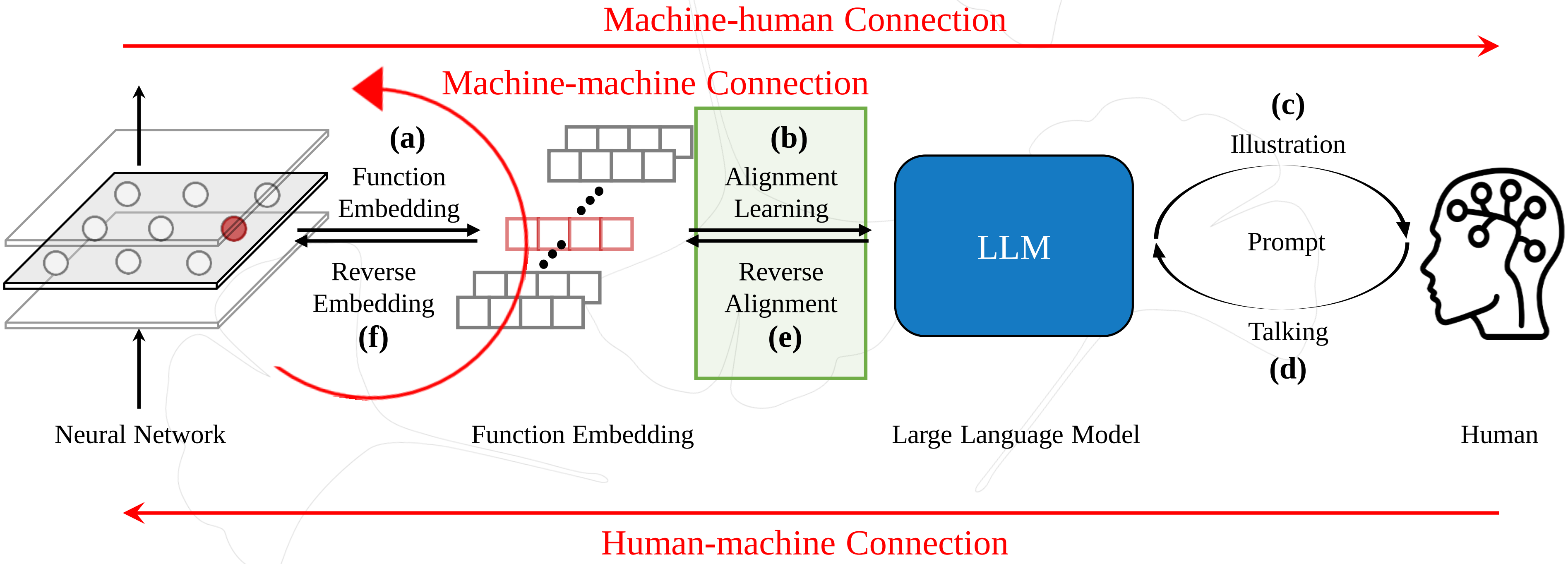
5 Case Study

- Human-machine interconnection framework



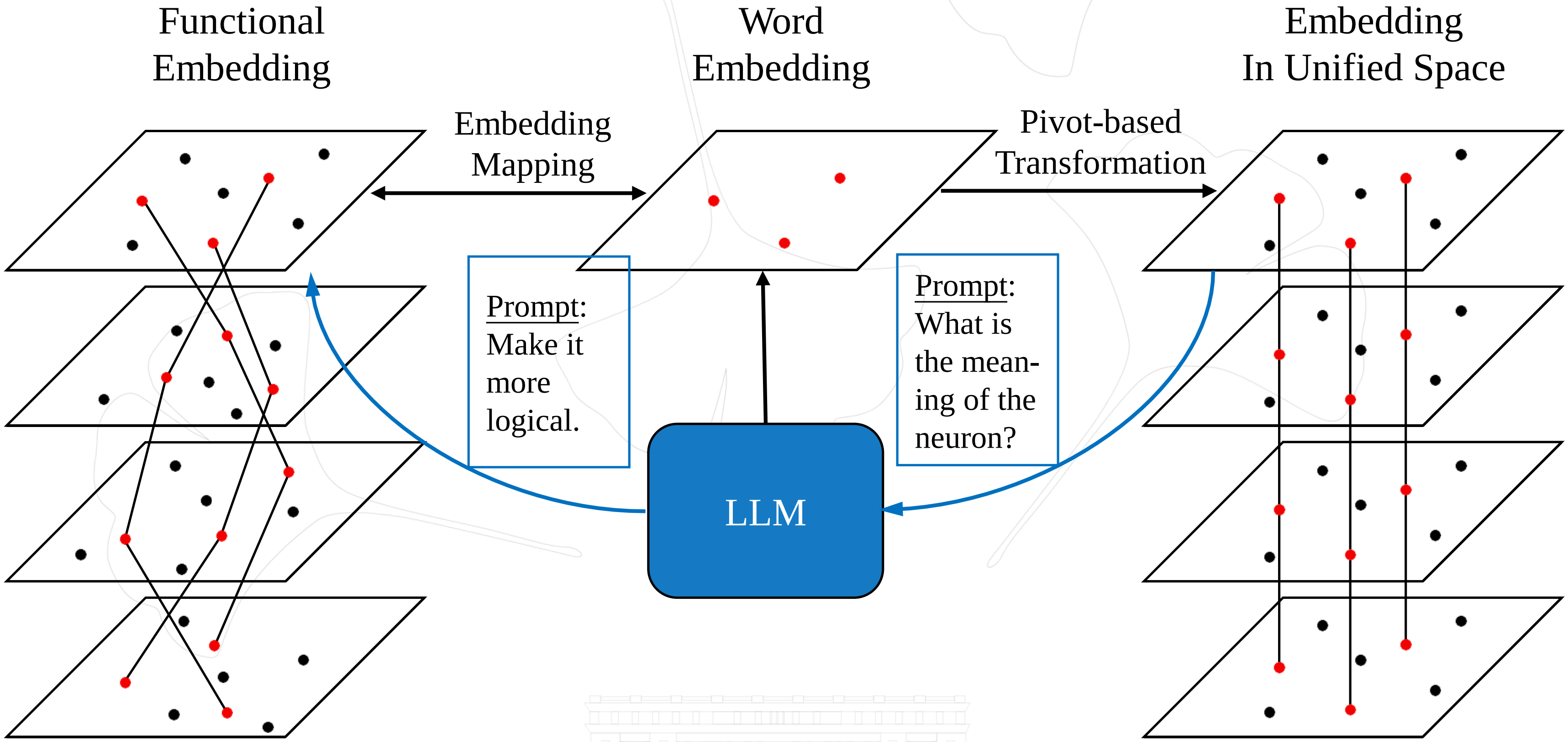
5 Case Study

- Methodology



5 Case Study

- Alignment

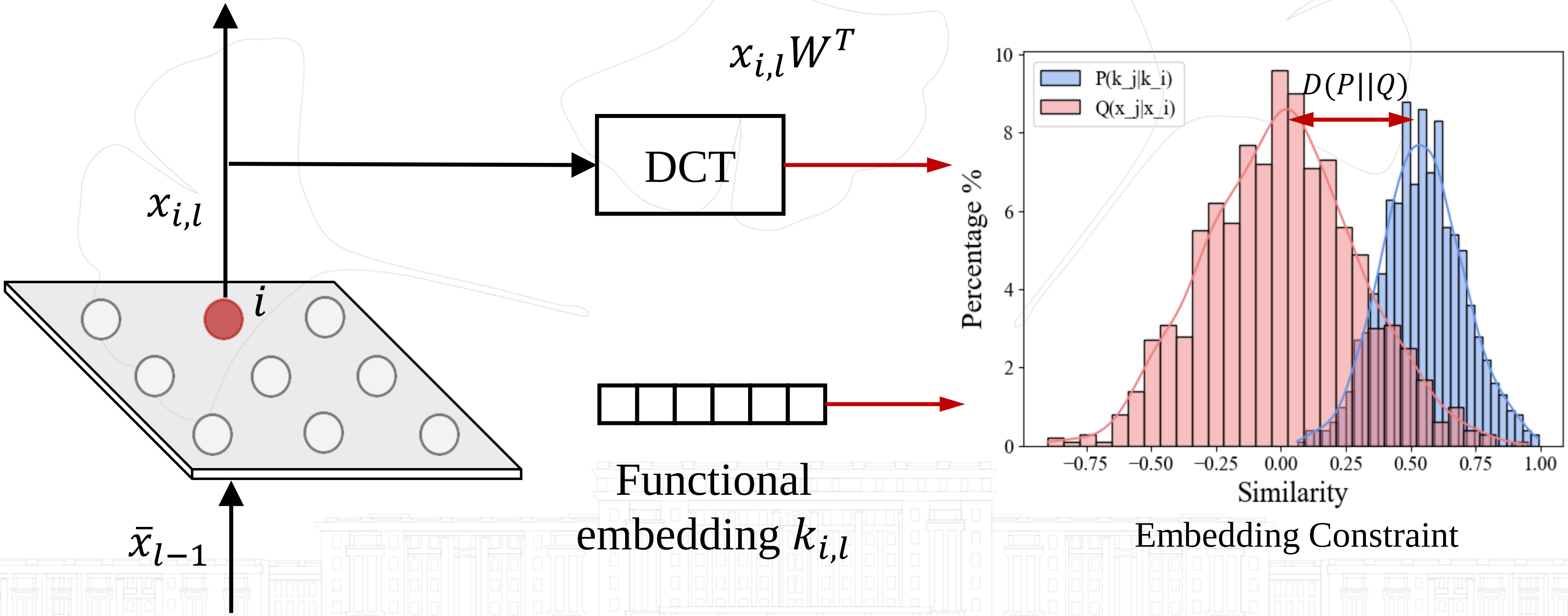


Prompt: Make it more logical.

Prompt: What is the meaning of the neuron?

5 Case Study

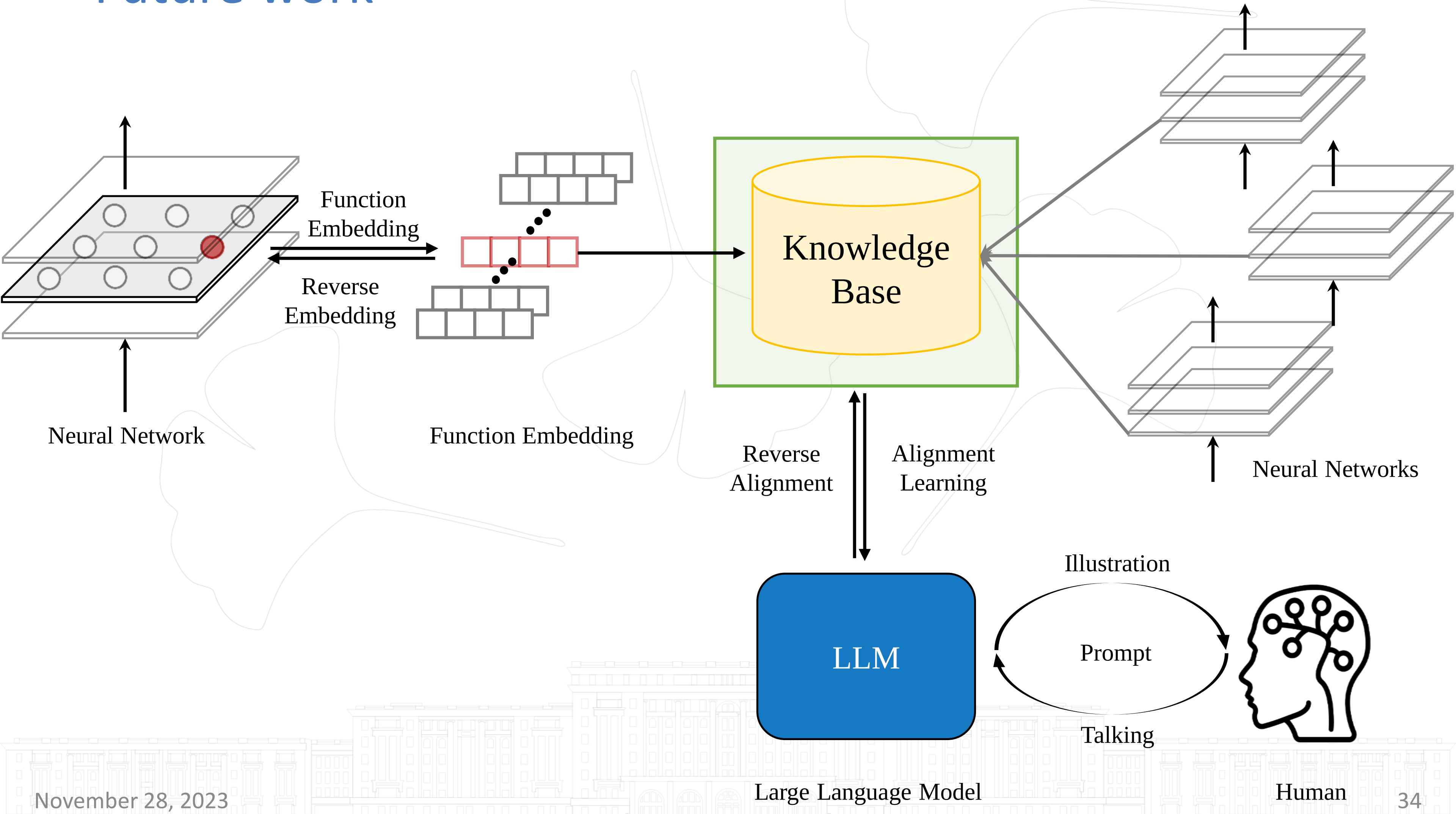
- Embedding
  - Cross Entropy, KL/JS , Neg Sampling
  - Mutual Information, InfoNCE
  - Spectral Sim, Lap Norm (W distance, OP?)



Embedding Constraint

5 Case Study

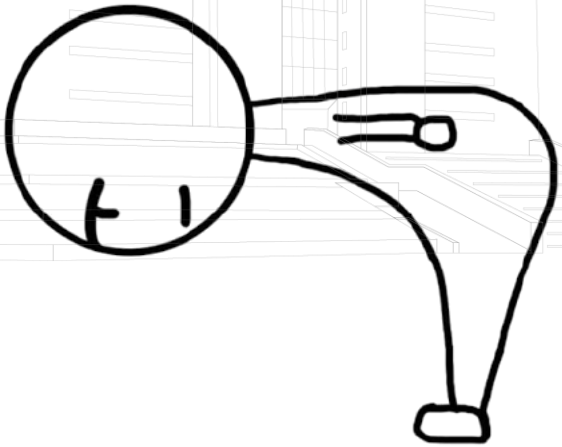
- Future work



5 Discussion

- Future is coming
- Analyze & Explore LLM
- Combine with previous fields
- Innovation applications across domains
- Talking is cheap, show me your code

Thanks



Data Mining Lab, Big Data Research Center, USETC

Wei Han, wei.hb.han@gmail.com