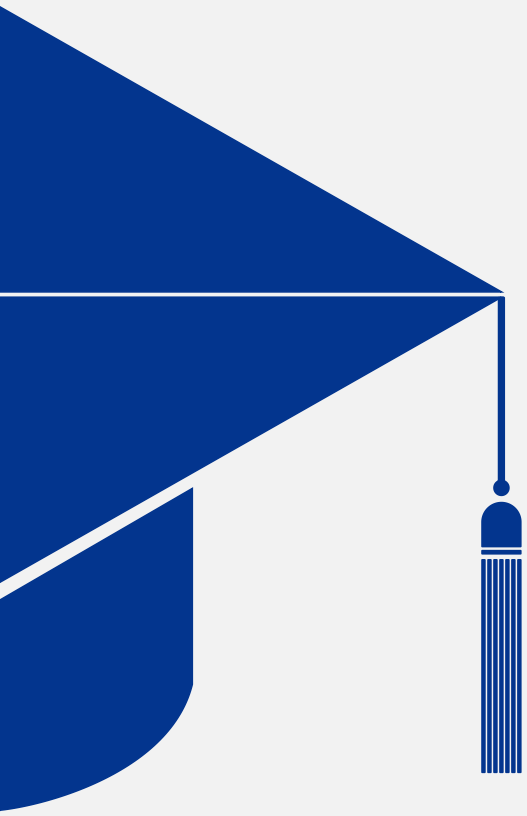


中国科学技术大学
University of Science and Technology of China

论文调研 - AIOS方向

马斌　2025.7.16

红专并进 理实交融

中国科学技术大学
University of Science and Technology of China

目录
CONTENTS

论文目录

沿革关系

侧重点分析

调研总结

近期进度汇报

论文目录

论文来源：Arxiv，dblp　　检索关键词：“AIOS”

- [1] GE Y, REN Y, HUA W, 等. LLM as OS, Agents as Apps: Envisioning AIOS, Agents and the AIOS-Agent Ecosystem[EB/OL]. arXiv, 2023[2025-07-11]. http://arxiv.org/abs/2312.03815.
- [2] XU S, LI Z, MEI K, 等. AIOS Compiler: LLM as Interpreter for Natural Language Programming and Flow Programming of AI Agents[EB/OL]. arXiv, 2024[2025-07-16]. http://arxiv.org/abs/2405.06907.
- [3] RAMA B, MEI K, ZHANG Y. Cerebrum (AIOS SDK): A Platform for Agent Development, Deployment, Distribution, and Discovery[EB/OL]. arXiv, 2025[2025-07-15]. http://arxiv.org/abs/2503.11444.
- [4] SHI Z, MEI K, JIN M, 等. From Commands to Prompts: LLM-based Semantic File System for AIOS[EB/OL]. arXiv, 2025[2025-07-16]. http://arxiv.org/abs/2410.11843.
- [5] ZHANG X, ZHANG Y. Planet as a Brain: Towards Internet of AgentSites based on AIOS Server[EB/OL]. arXiv, 2025[2025-07-15]. http://arxiv.org/abs/2504.14411.
- [6] MEI K, ZHU X, XU W, 等. AIOS: LLM Agent Operating System[EB/OL]. arXiv, 2025[2025-07-12]. http://arxiv.org/abs/2403.16971.
- [7] MEI K, ZHU X, GAO H, 等. LiteCUA: Computer as MCP Server for Computer-Use Agent on AIOS[EB/OL]. arXiv, 2025[2025-07-15]. http://arxiv.org/abs/2505.18829.

2023　2024　2025

方便起见，下面均用“[序号]+短标题+发表时间”表示。

沿革关系

[1] LLM as OS 2023.12

首次提出AIOS的**整体框架**

[2] AIOS Compiler 2024.5

尝试开发**LLM编译器**及代码表示，实现自然语言编程（NLProg）

初尝试后，确定了NLProg的可行性，继续后续深入研究

[3] Cerebrum 2025.3

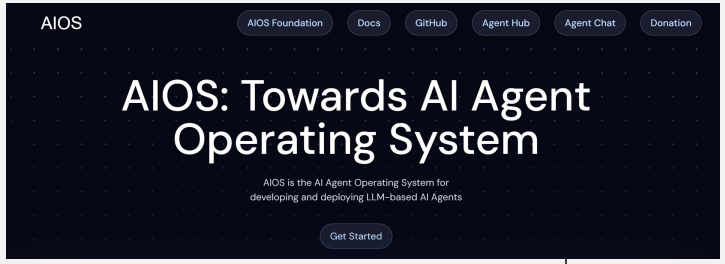
AIOS SDK,
(LLM, memory, storage, tool management)

[4] From Commands to Prompts 2025.3

File System,
基于LLM语义文件系统(LSFS)

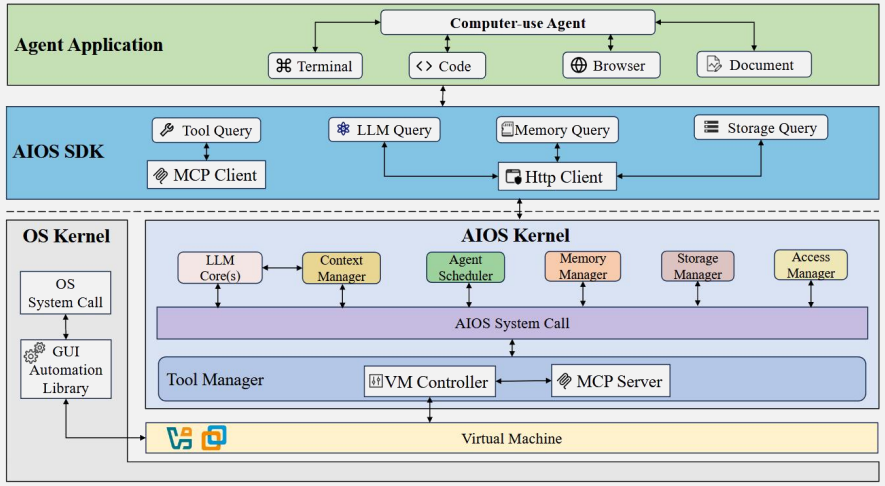
[5] Planet as a Brain 2025.5

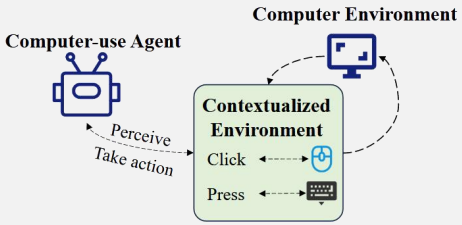
AIOS Server,
利用MCP和JSON-RPC,
首次部署AIOS-IoA



[6] AIOS 2025.5

总结先前工作，
产出**AIOS1.0**平台demo，
修正原先架构
(AIOS取代OS
-> AIOS基于OS)

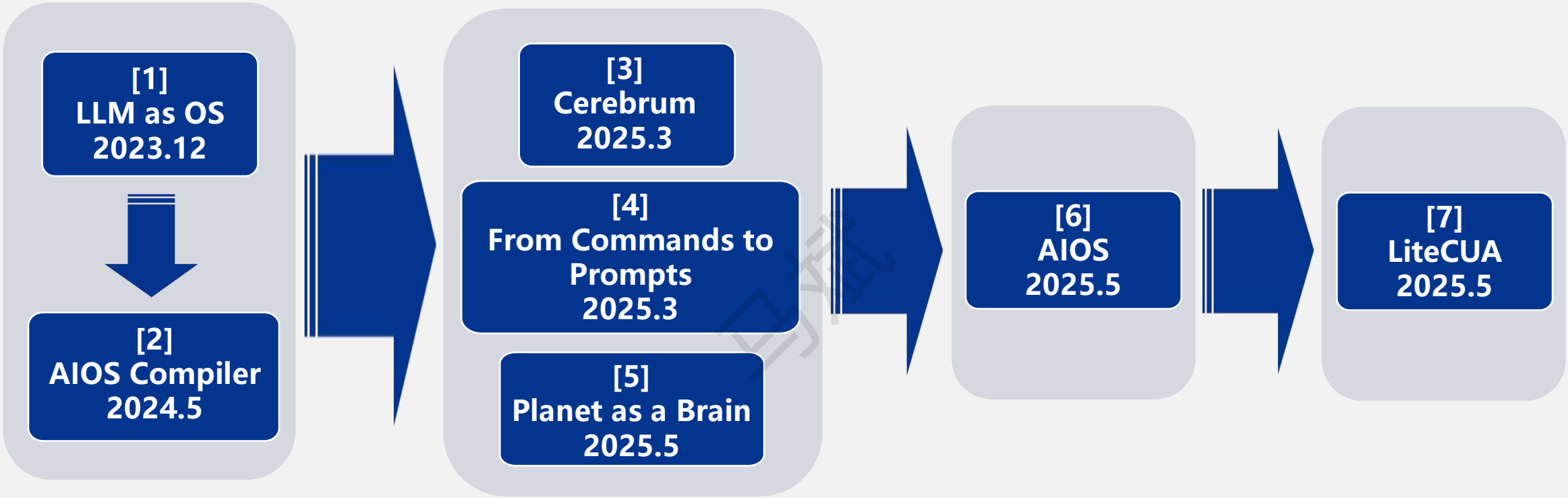




[7] LiteCUA 2025.5

应用场景其一，
强调CUA的优势
(human操作计算机
-> CUA操作计算机)

侧重点分析



初步架构想法
->可行性

实践想法

根据实践
修正架构

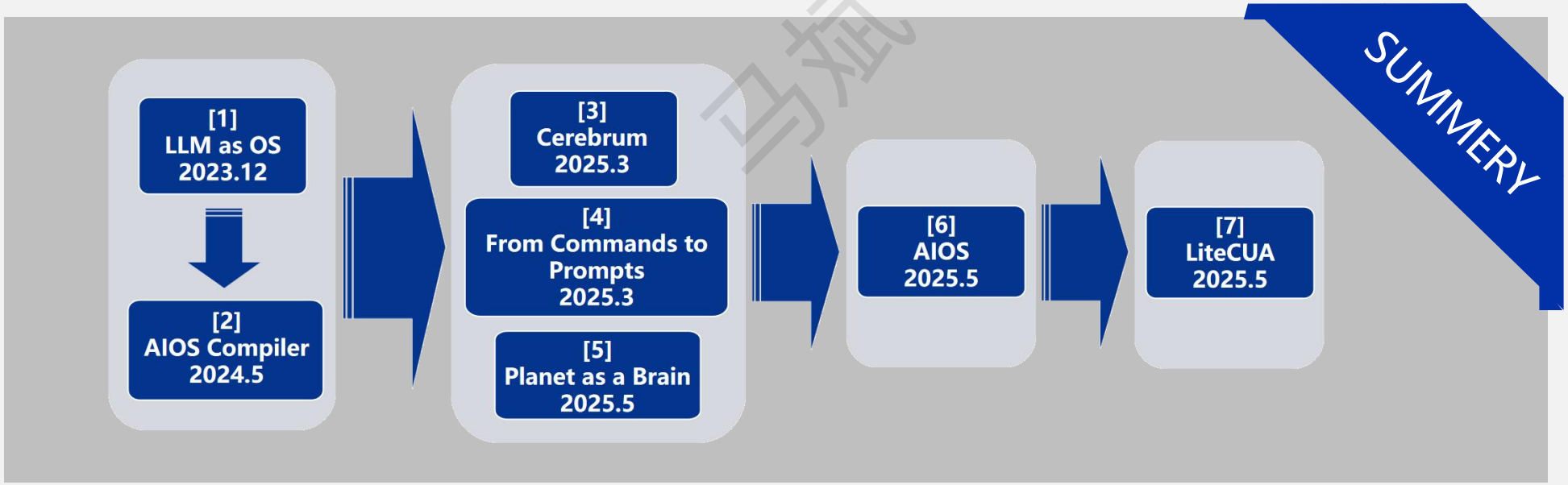
根据修正后的架构
探索应用场景

调研总结

调研的7篇AIOS相关论文研究有明显的沿革关系和技术演进路径，可集中对比分析。

体现了AIOS从架构猜想走向实际应用，并在实践中逐步修正并完善架构。

其中，研究目的围绕创新软件开发和人机交互的新范式，利用LLM推动技术民主化进行



近期进度汇报

- 论文调研：

  [1] [2] [3] [4] [5] [6] [7] （其中 [2] [3] [4] [5] [7] 仅精读了 Abstract 部分）

- 远程连接到服务器，成功本地部署QWEN-0.5B，运行实验(1) (2) (3) (4) 。

  每组实验分别在CPU和GPU上运行一次，得到对应“运行时间”和“数据传输时间”。

cpu_gpu_inference_comparison.py --- 1

测试模型：Qwen/Qwen2.5-0.5B-Instruct
测试数据：wikitext/wikitext-2-raw-v1

```
CPU推理时间：90.5342秒，数据传输时间：0.0000秒
GPU推理时间：5.5268秒，数据传输时间：0.0008秒
```

cpu_gpu_inference_comparison.py --- 3 (换dataset)

测试模型：Qwen/Qwen2.5-0.5B-Instruct
测试数据：squad_v2

```
CPU推理时间：97.0445秒，数据传输时间：0.0000秒
GPU推理时间：5.6994秒，数据传输时间：0.0005秒
```

cpu_gpu_inference_comparison.py --- 2

测试模型：Qwen/Qwen2.5-0.5B-Instruct
测试数据：wikitext/wikitext-2-raw-v1

```
CPU推理时间：95.0074秒，数据传输时间：0.0000秒
GPU推理时间：5.7042秒，数据传输时间：0.0005秒
```

cpu_gpu_inference_comparison.py --- 4 (换model)

测试模型：gpt2 (0.1B)
测试数据：squad_v2

```
CPU推理时间：24.3989秒，数据传输时间：0.0000秒
GPU推理时间：2.1689秒，数据传输时间：0.0002秒
```

中国科学技术大学
University of Science and Technology of China

THANKS

马斌 2025.7.16

红专并进 理实交融