从端侧推理到智能体

Web AI 的今天和明天

—— Web 进化论 2025 年度大会 · 杭州

关于我

前端架构师

Web GDE

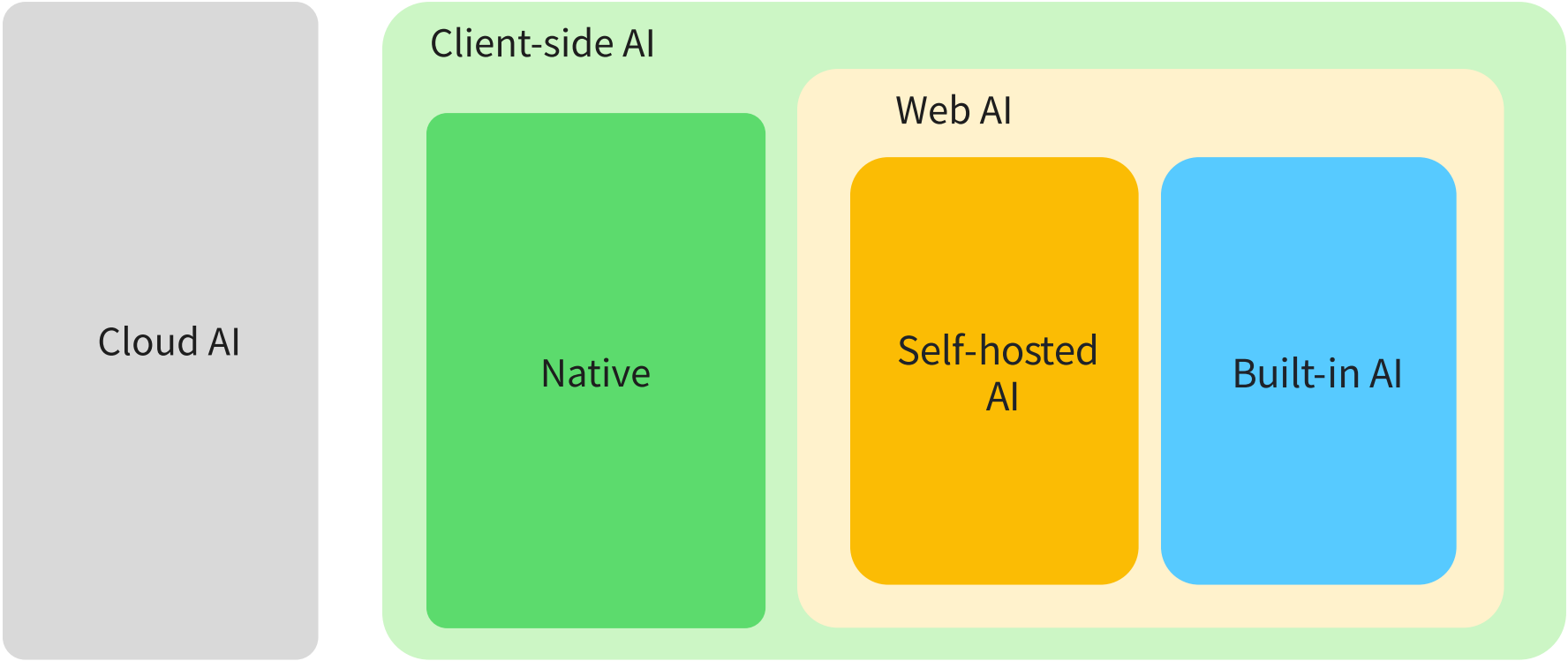
掘金社区优秀创作者

钱俊颖 (Jax)

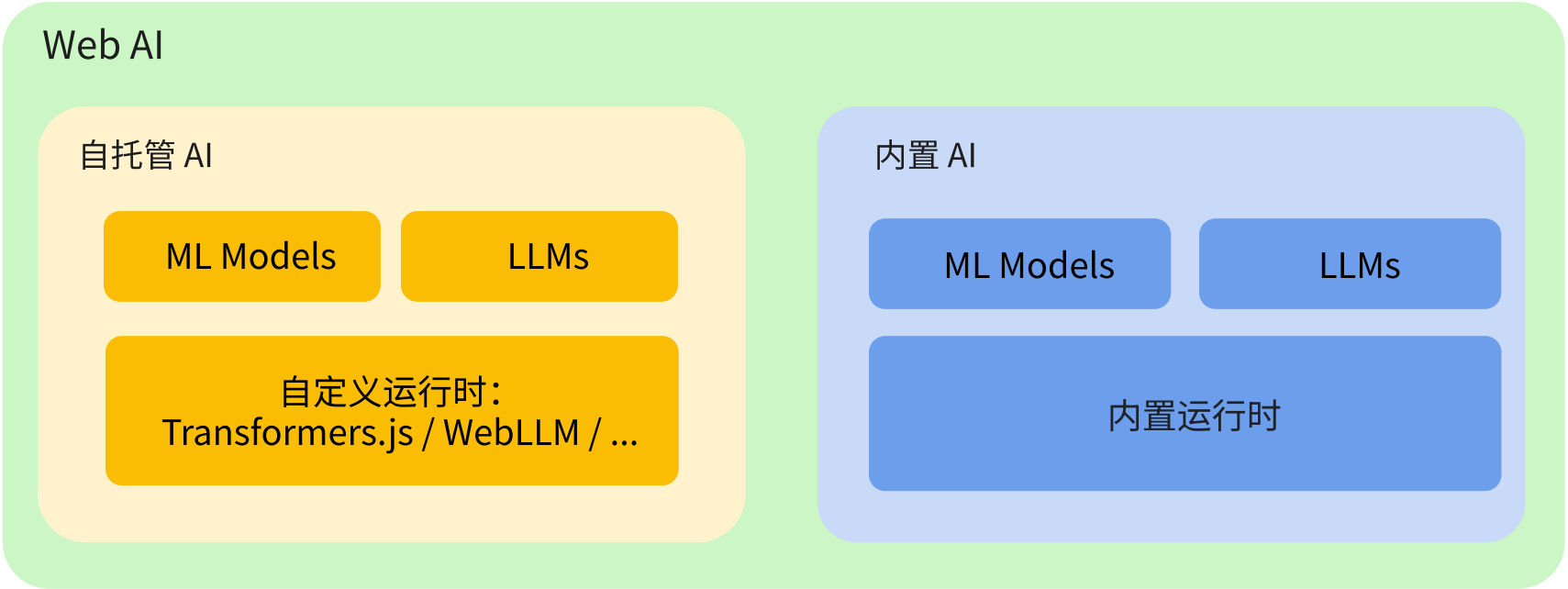
目录

目录

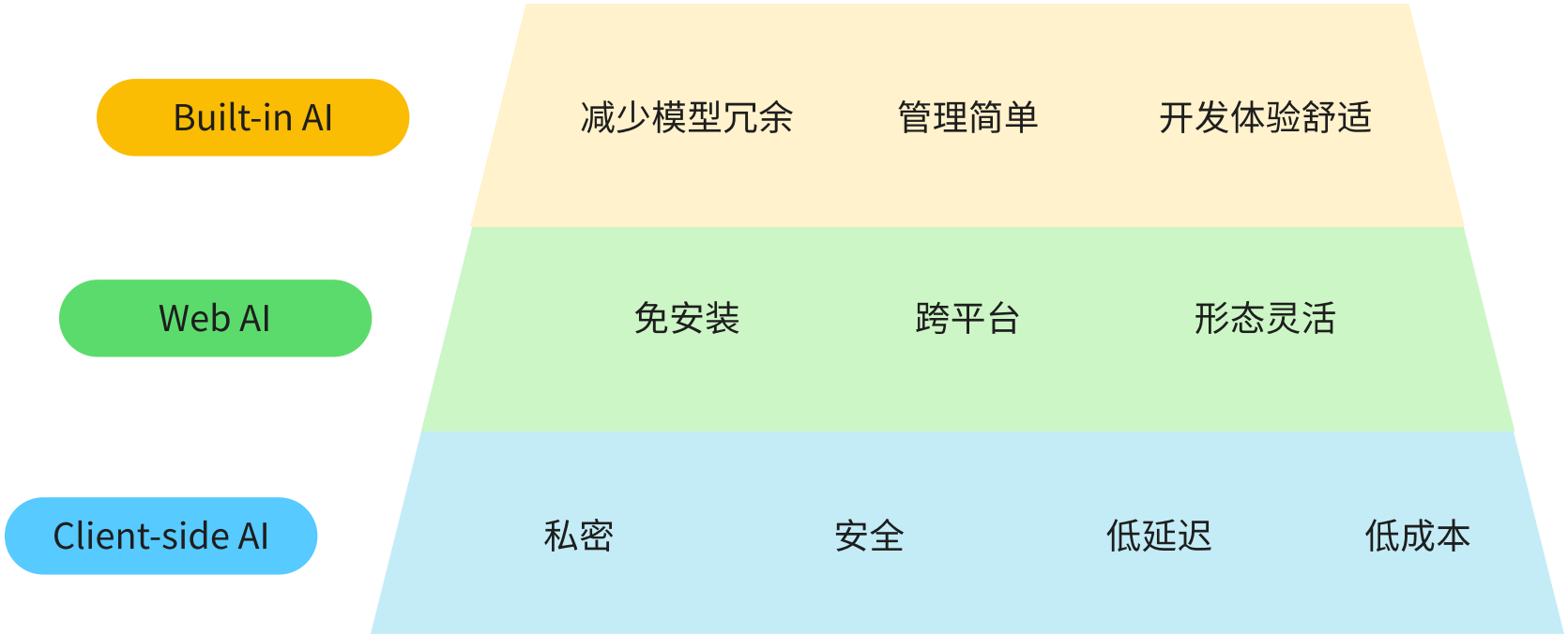
当我们聊起 Web AI，我们在聊什么？



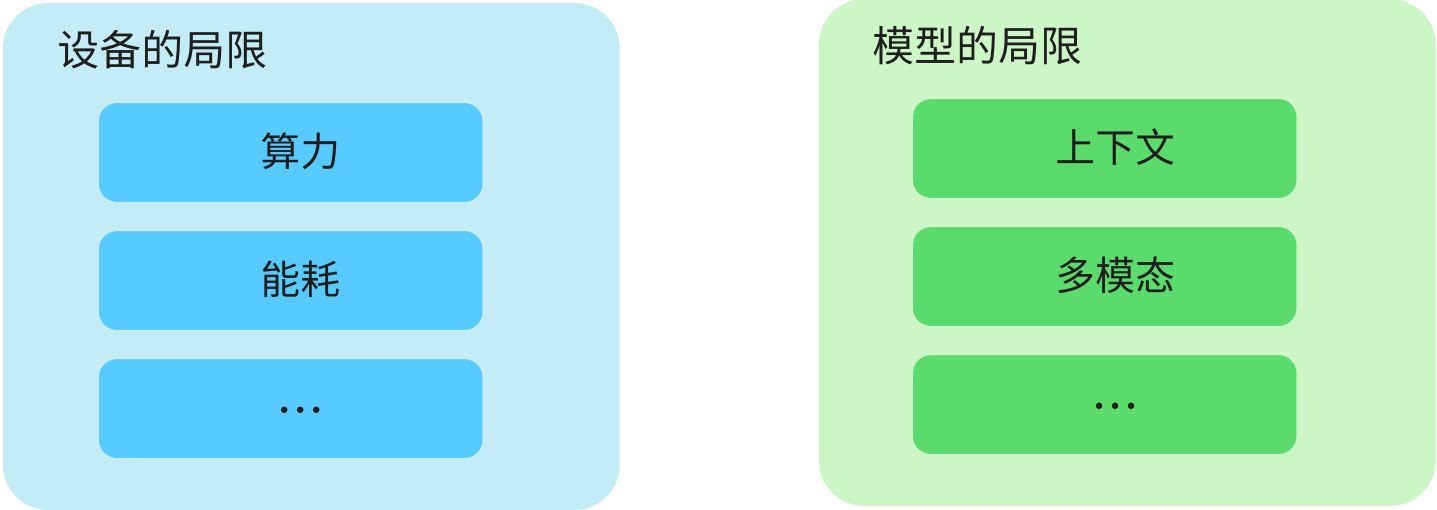
Web AI 分类


Web AI
自托管 AI
ML Models
LLMs
自定义运行时：
Transformers.js / WebLLM / ...
内置 AI
ML Models
LLMs
内置运行时

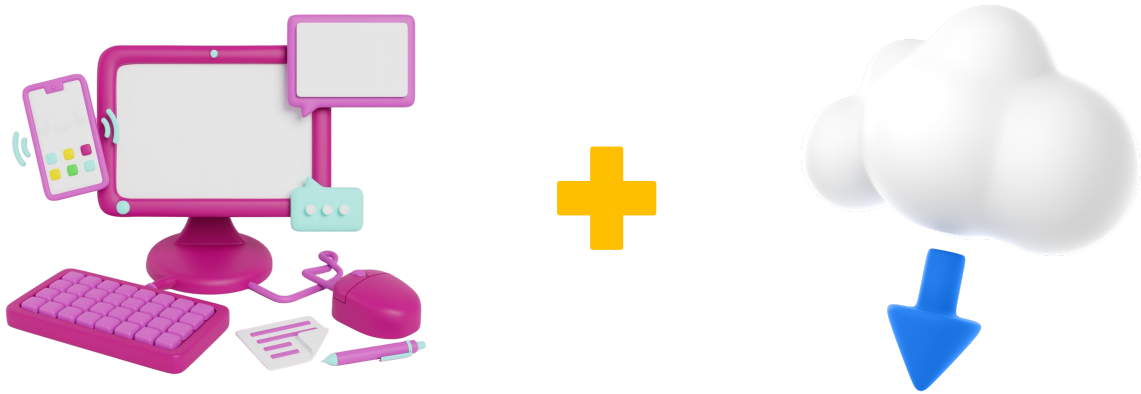
优势金字塔


Built-in AI
减少模型冗余
管理简单
开发体验舒适
Web AI
免安装
跨平台
形态灵活
Client-side AI
私密
安全
低延迟
低成本

局限 —— 端侧非银弹



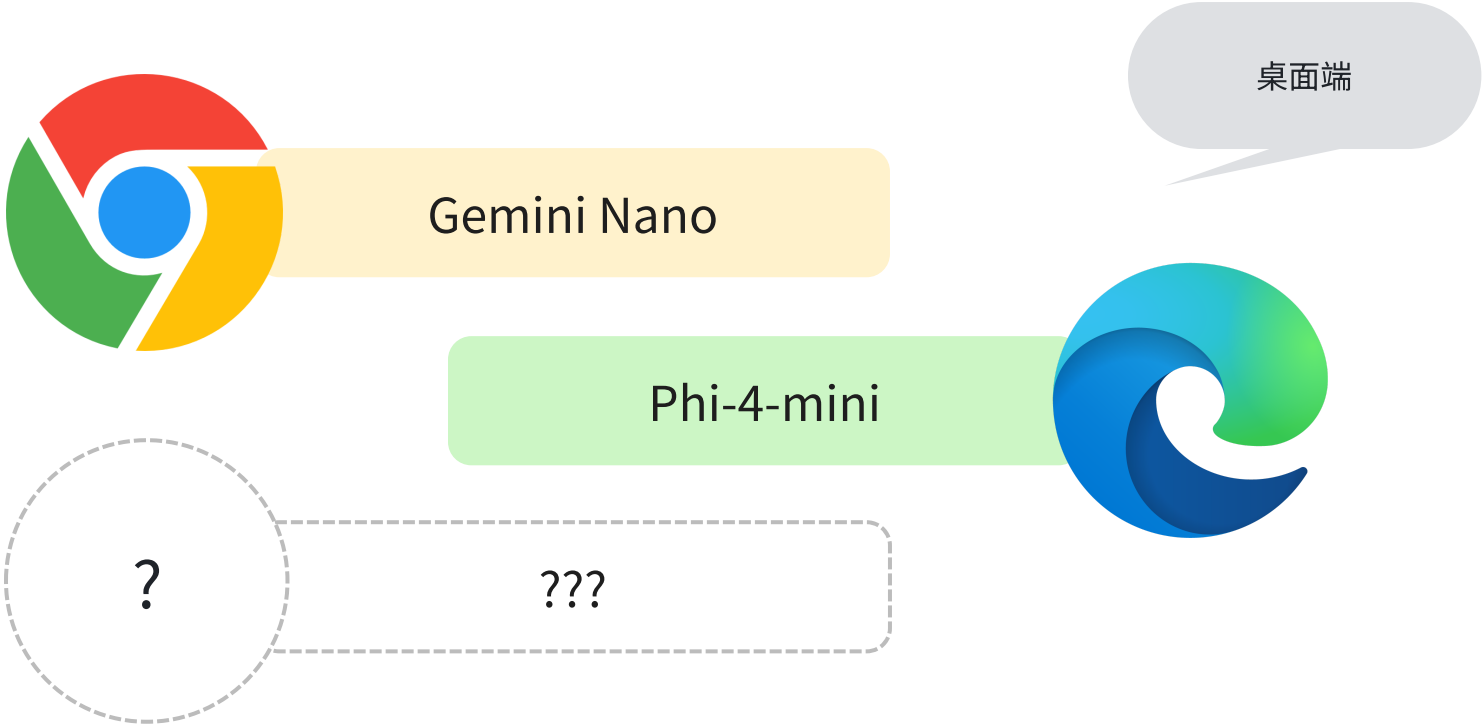
Hybrid 模式



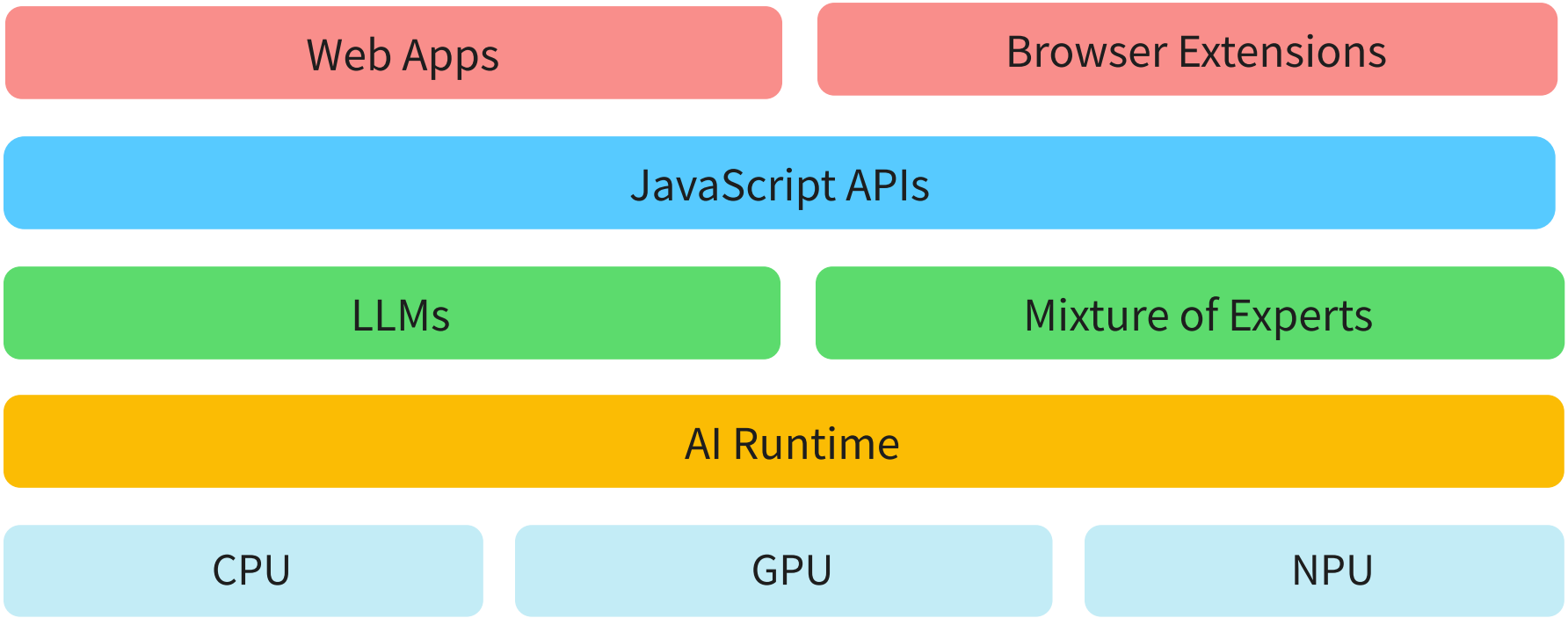
“此端基础，彼端就不基础”

目录

浏览器支持



Built-in AI


Web Apps
Browser Extensions
JavaScript APIs
LLMs
Mixture of Experts
AI Runtime
CPU
GPU
NPU

JavaScript APIs

Writer API

Summarizer API

Language Detector API

Proofreader API

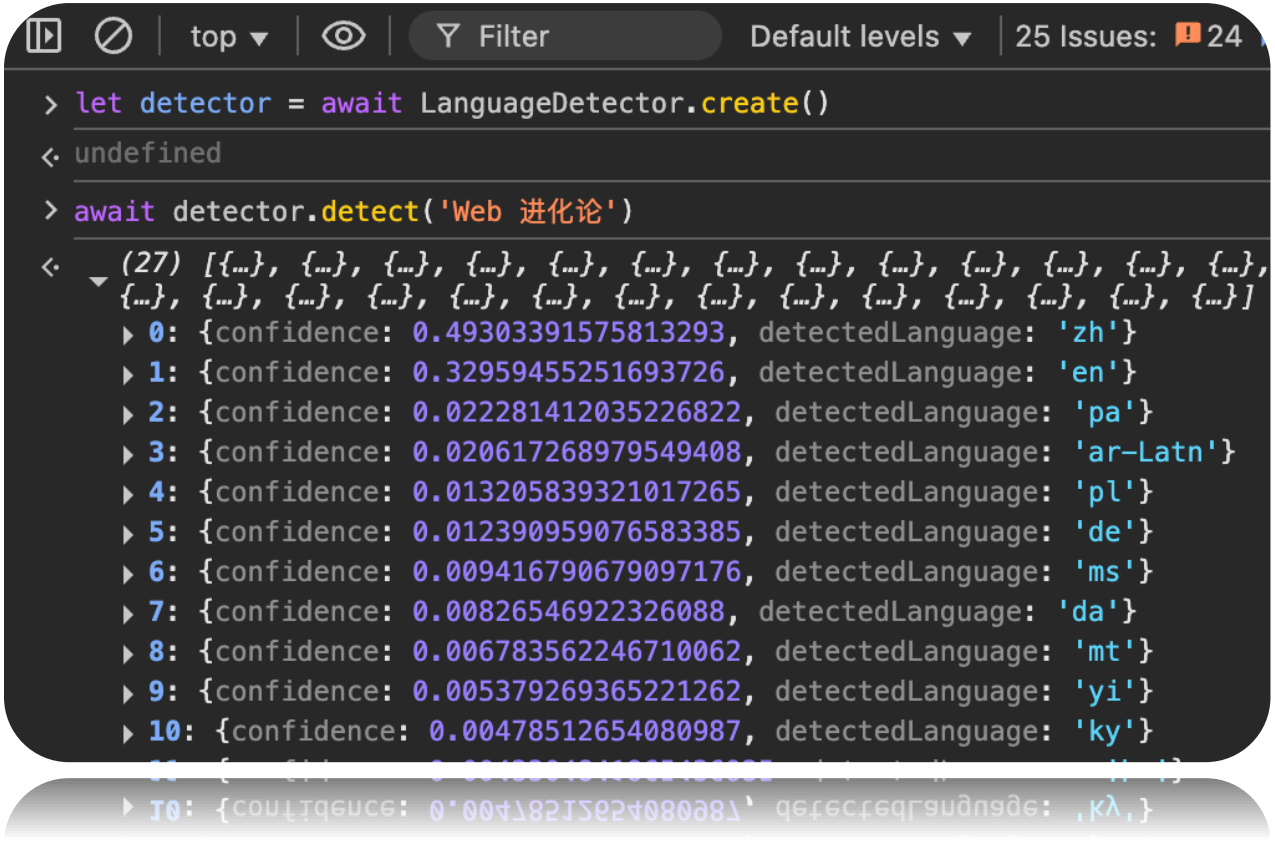
Rewriter API

Translator API

Prompt API

详情参见：https://developer.chrome.com/docs/ai/built-in-apis#api_status

Language Detector API



Translator API

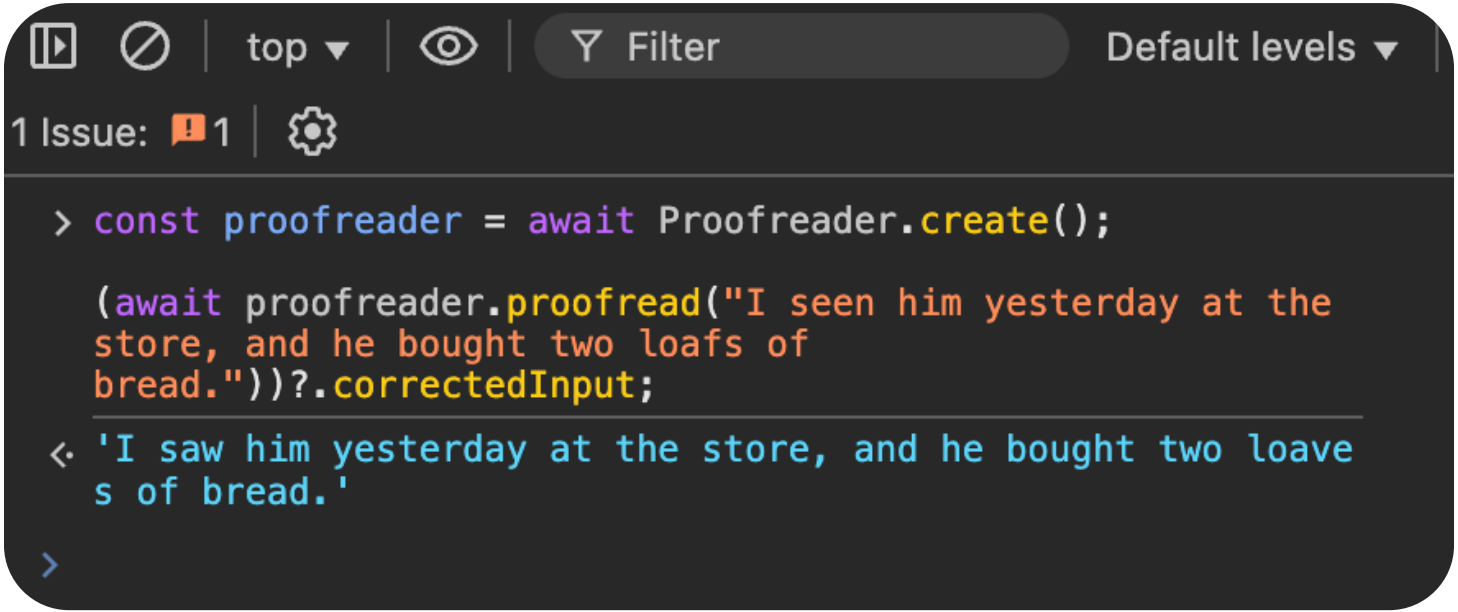
```
> let translator = await Translator.create({
      sourceLanguage: 'zh',
      targetLanguage: 'en'
  })

<· undefined

> await translator.translate('W3C Web中文兴趣组（Chinese Web Interest
  Group）组织本次活动，同时也为W3C AI Agent & the Web技术研讨会的中国区前瞻
  会议，旨在汇聚来自产业、科研机构、用户代表、开发者和标准化组织的相关利益方，共
  同探讨智能体（AI Agent） 在Web生态中的角色与发展，识别潜在的标准化方向，助力
  构建可持续、开放且互操作的Agentic Web。')

<· 'The W3C Web Chinese Interest Group (Chinese Web Interest Group) o
  rganizes this event, and also serves as W3C AI Agent & The The for
  ward-looking meeting of the Web Technology Seminar in China, aimin
  g to bring together relevant stakeholders from industries, scienti
  fic research institutions, user representatives, developers and st
  andardization organizations to discuss the agents (AI Agent) Role
  and development in the web ecology, identify potential standardiza
  tion directions, and help build a sustainable, open and interopera
  ble Agentic Web.'
```

Proofreader API



Writing Assistance API

```
> const summarizer = await Summarizer.create({
    type: "tldr",
    length: "short",
  });

< undefined

> await summarizer.summarize(str)

< 'The Chinese Web Interest Group is holding a meeting to di
  scuss how *AI agents* will shape the future of the web, co
  vering topics like construction, ecosystem integration, co
  mmunication, use cases, and standardization.'
```

Prompt API

Prompt Engineering

Multimodal

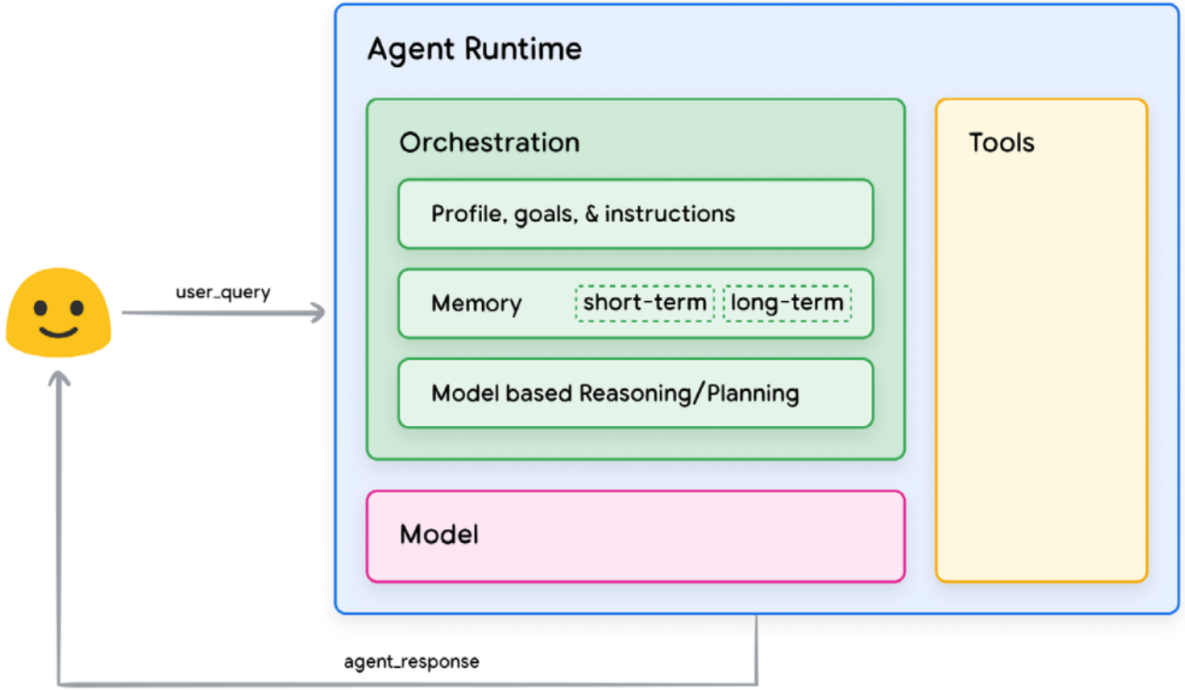
Structured Output

Tool Use

详情参见：https://github.com/webmachinelearning/prompt-api

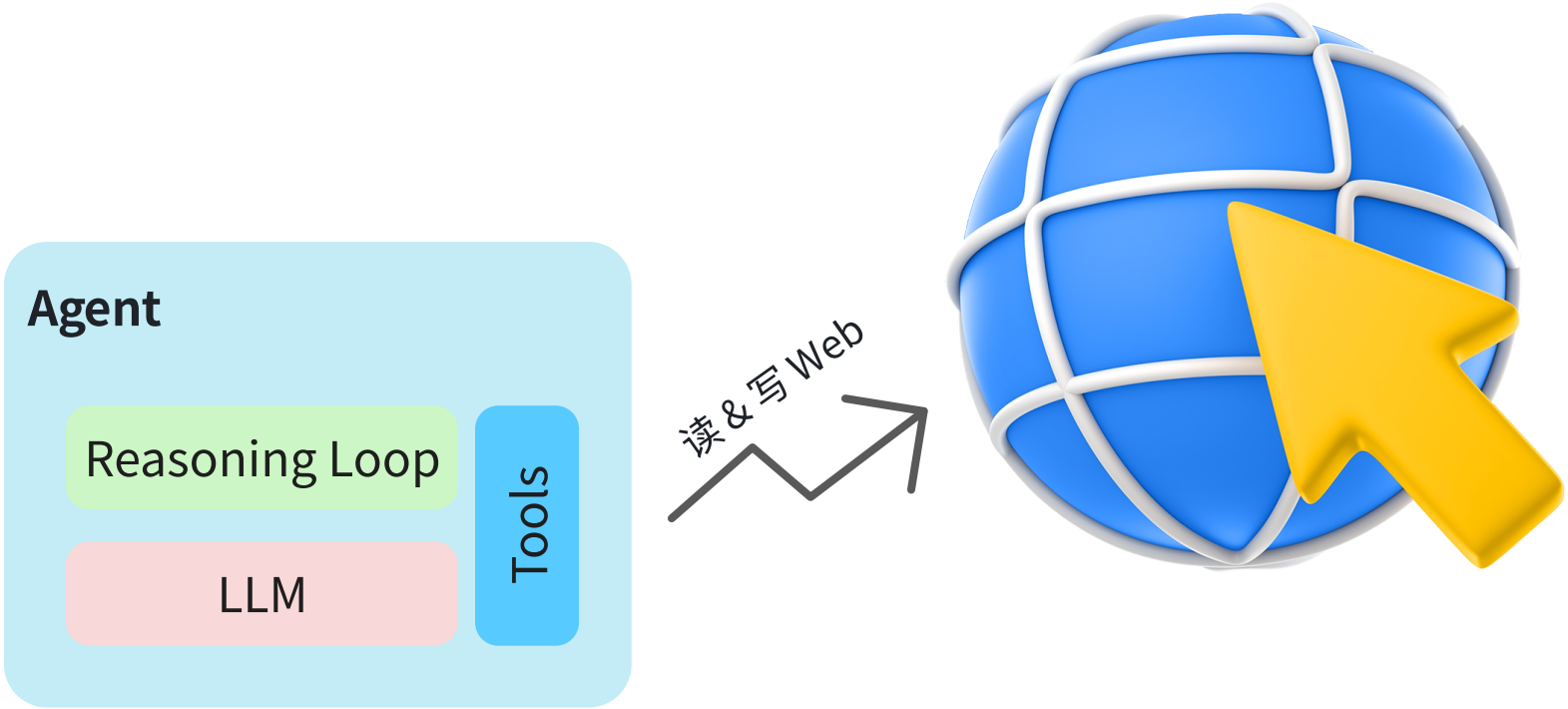
目录

Agent 的结构

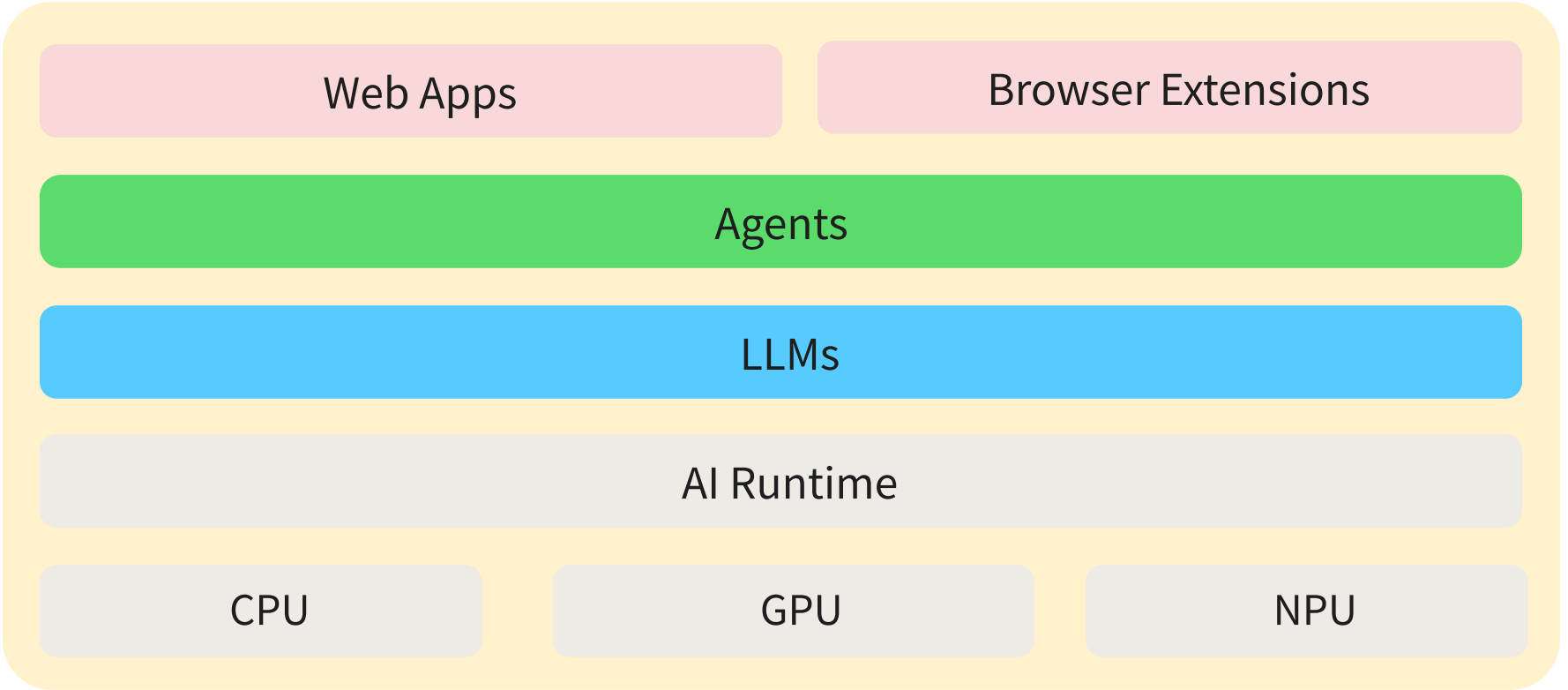


图片来源：https://www.kaggle.com/whitepaper-agents

Web Agent



Agent In Web AI



Web 端实现 AI Agent

Vanilla

- Self-hosted LLMs
- LLM Runtime
- Orchestration
- Tools

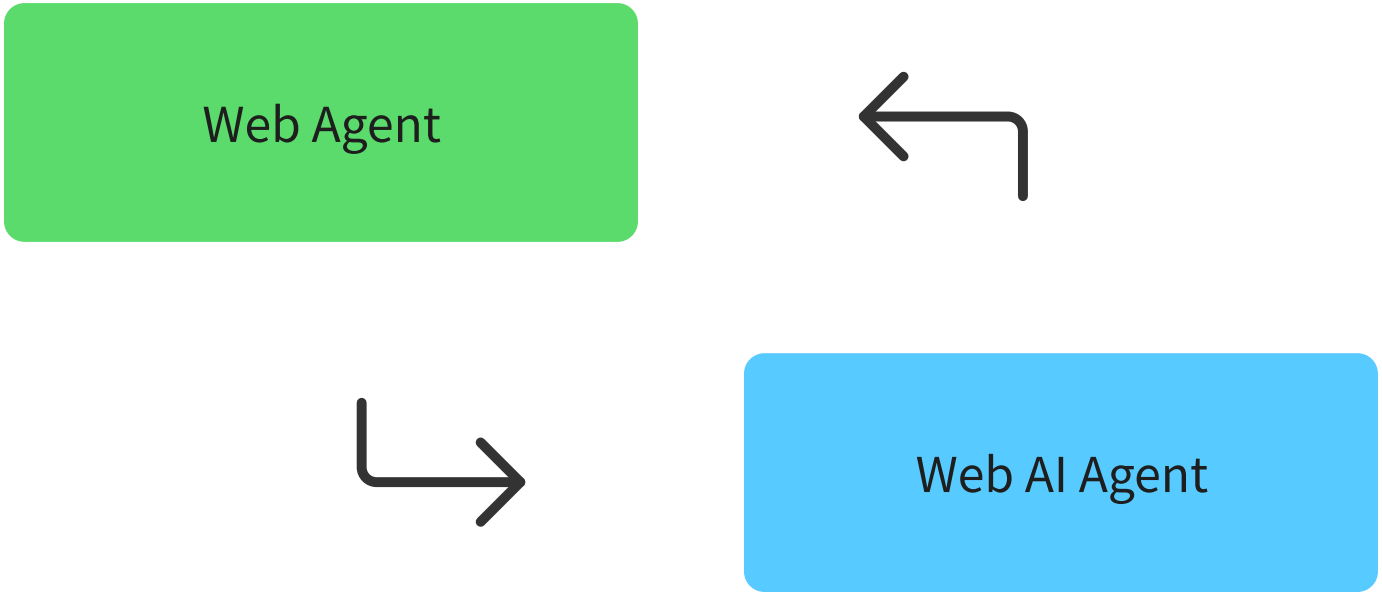
Built-in AI

- Orchestration
- Tools

LangGragh.js

- Cloud AI
- Orchestration
- Tools

互为工具



目录

假如你是秦始皇……

为什么要有标准?

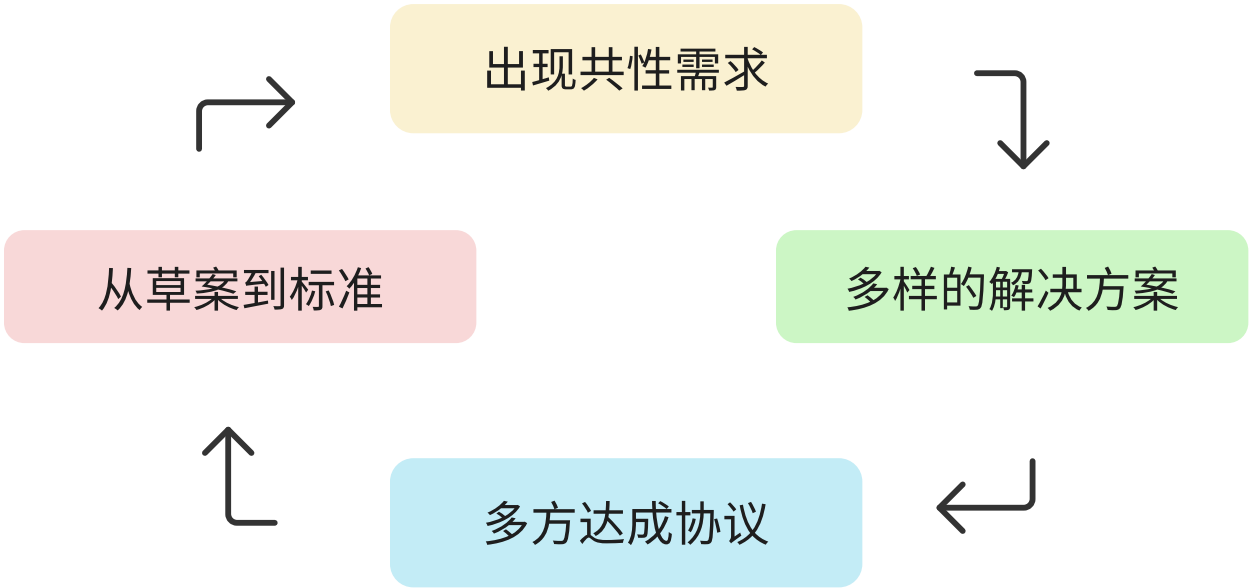
解决通用需求 / 痛点

统一规范

保持开放生态

提炼最佳实践

标准是怎么来的?



第一步

广义 Web AI / 端侧 Web AI

读写 Web 的 Agent / 运行在 Web 里的 Agent

……

正在推进的标准

https://webmachinelearning.github.io/

Community Group Drafts

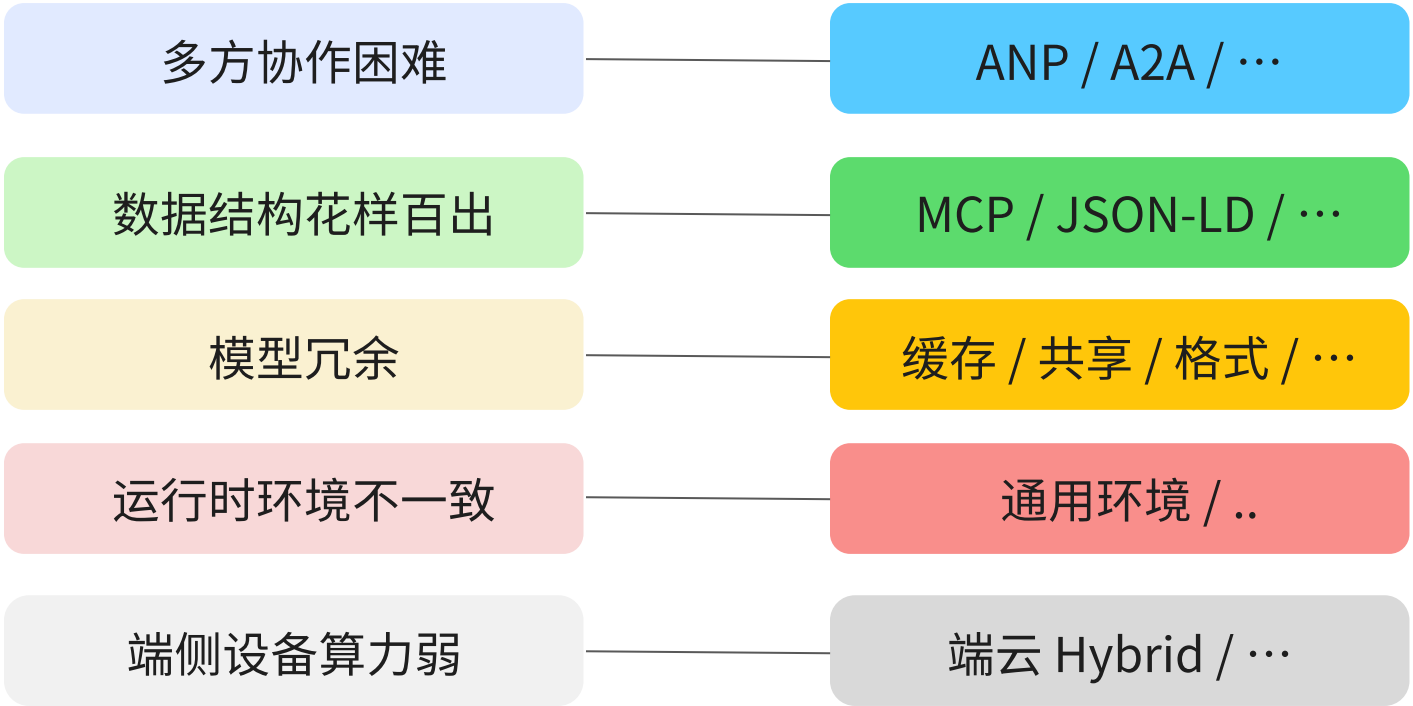
- Translator & Language Detector API
- Writing Assistance APIs

Community Group Explainers

- Proofreader API
- Prompt API (?)

Web AI 基础设施

痛点 ⬅ ➡ 标准


多方协作困难
ANP / A2A / ⋯
数据结构花样百出
MCP / JSON-LD / ⋯
模型冗余
缓存 / 共享 / 格式 / ⋯
运行时环境不一致
通用环境 / ..
端侧设备算力弱
端云 Hybrid / ⋯

总结

聚焦范围

着眼于端侧 Web 环境的智能化，讨论 Web AI 的结构和价值。

Agent In Web

不同层级的 Agent，相互之间的关系，在 Web 端的实现方式。

Built-in AI 能力

对 JavaScript 能力的扩展，全面适应 AI 场景。

Web AI 标准

标准化进程的现状，可预见的需求和缺口

感谢倾听

如果你也对 Web AI 感兴趣，
热烈欢迎一起交流！