魏雍
Github
weiyong1024@gmail.com
(+86) 173-7657-1024

教育背景

- **硕士 - 北京航空航天大学 - 电子与通信工程** 中国，北京
  保研; 研一上学期发核心期刊达毕业标准; 北京市优干 *2016 年 9 月 – 2019 年 1 月*
- **本科 - 北京航空航天大学 - 电子信息工程** 中国，北京
  沈元荣誉学院 (入学 top50/3000+); 北京市三好; GPA 3.7 *2012 年 9 月 – 2016 年 6 月*
- **高中 - 大连育明高中** 中国，大连
  CPhO（全国中学生物理竞赛）省一等奖获高考免试保送资格 *2009 年 9 月 – 2012 年 6 月*

编程竞赛

- **Google Code Jam Kickstart (谷歌 2017 全球校招赛)**
  全球 108^{th}, 中国 18^{th}, 前 5% ***计分板链接 (id - WeiYong1024)***

开源项目

- **Second-Me**
  项目 PMC 角色，上线一周获 6K+ Star
  提出并实现一种在个人计算机上基于个人数据训练大模型的解决方案，以及一种基于个人大模型的新型社交网络。

工作经历

- **心识宇宙（天使轮加入，国内 LLM 应用先驱创业团队）** 中国，杭州
  基础设施技术主管，领导 4 人团队，构建大模型应用的 AiInfra+DataInfra+DevOps 体系 *2022 年 6 月 - 至今*
  - **公司代表产品**
    - **MeBot**：启发型个人助理，Web 版于 2024.8 获 ProductHunt 周榜第二，App 版已上苹果 AppStore。
    - **Second Me**：开源个人大模型训练、部署、模型网络技术框架，于 2025.3 上线，第一周获 6K+ Star。
  - **公司技术设计**
    - **LPM(Large personal model)**：基于 Lora 与用户在线数据训练个性化 LLM，可用于生产环境。
    - **MindOS**：国内最早的 LLM Agent Platform，比字节 Coze 平台上线时间早 17 个月。
    - **UMM(Unified mind model)**：统一心识模型，比 OpenAI 首次提出 AIAgent 概念早 16 个月。
  - **Infra 技术职责**
    - **AiInfra ——打造 LLM 应用基础设施最佳实践**
      - **MLOps 平台**：设计并实施包含数据采集、数据增强、LLM 训练、LLM 部署的全链路自动化 MLOps 平台。结合开源工具 ClearML 的模型观测能力，用于支撑 LPM 实时训练并做生产部署。
      - **LLM 统一接入层**：前期使用 Java 自建基础服务，实现 OpenAI、AzureOpenAI、GCPVertexAI、Claude 等模型的统一接入、部署管理、多账号池化、探活、权重分配、成本统计与可视化等功能。后期迁移至 LiteLLM 作为提供上述功能的最佳实践。
      - **LLM 工程平台**：使用 Langfuse 标准化算法离线实验管理、Prompt 资产管理、资产版本管理、线上 LLM-Tracing 等功能。
    - **DevOps ——高效、安全生产**
      - **生产环境管理**：设计并搭建 Prod/Pre/Test 等多套生产环境，包含严格 CodeReview 流程的可伸缩 GitServer、CI/CD 流程、零信任设备管理等基础设施。涉及技术：Gitlab、Helm、Java、Jenkins、JumpServer、Kubernetes、Octant、Python、Rancher、Shell 等。
      - **系统可观测性**：选型、设计并构建公司后台系统使用的分布式配置、分布式限流、分布式调度、应用性能监控及日志监控内部生产等基础设施。涉及技术：Apollo、Datadog、Grafana、Loki、Nacos、Prometheus、Sentinel、Skywalking、xxl-job、腾讯云 CLS 等。
      - **云及三方服务管理**：公司内部用户权限管理，精细化管控用云、用三方服务的成本，设计、制作及维护成本大盘。涉及技术：Azure、GCP、CronJob、TencentCloud 等。
    - **DataInfra ——管理数据资产、提供实验工具、数据驱动决策**
      - **在离线数据体系**：设计、构建并维护包含 OLTP->CDC->OLAP->DataVisualization 的在离线数据流用于离线数据开发，包含消息队列-> 埋点服务的异步数据流用于用户行为数据分析等场景。支撑各类指标的大盘、运营 BI 报表等需求。涉及技术：MySQL、Clickhouse、Debezium、Superset、RocketMQ 等。

        - **A/BTest 工具体系**：选型、设计、构建并维护包含实验配置/结果分析平台（GrowthBook）+ 前端埋点（GoogleAnalytics）+ 后端埋点（Springboot 实现，包含工程和算法）的全链路实验工具体系。实现前端页面、后端算法的功能开关和 A/B 实验。涉及技术：GoogleAnalytics、GrowthBook、Java、NodeJS、Python 等。
        - **产品隐私合规**：构建一系列隐私合规标准，包括但不限于数据脱敏、内部培训等，使公司产品达到北美 USDP 和欧盟 GDPR 隐私合规标准。
    * **业务层研发——提供业务开发依赖的基础服务**
        - **中台基础服务**：设计并构建中台基础微服务，为业务系统提供后端加密存储、用户隐私数据库、Web 爬虫、离线数据舒仓管理等通用基础服务，并以 RestfulAPI、二方库 +RPC 接口等形式输出。
  - **支撑业务**
    * 主导移动端小程序万物总动员、Web 端产品 MindOS、App MeBot 及各种子系统的全链路压测、发布过程重保；准备并制定 toB 公有云/专有云/私有化产品交付 SOP；服务架构随业务跨国/跨云地域迁移等。
    * 在全公司横向推进 blameless postmortem 文化，组织安全生产周会。

- 阿里云 中国，杭州
  高级工程师、后端、*Java*、*Springboot*、阿里云 *2020 年 8 月 - 2022 年 5 月*
  - **流量调度中间件**: 在集团高可用团队负责阿里云内部流量调度中间件。基于阿里内部的 PandoraBoot（Springboot 加强版）提供集群上容器粒度秒级健康指标监控、3σ 离群点检测，依托阿里内部 RPC 框架 HSF(Dubbo 内部版) 的权重表机制实现微服务的分钟级流量调度与恢复链路。防单点故障引起的集群雪崩，也可用于 JIT 预热。任职期间推广产品纳管了阿里集团全部核心应用，支撑 40w+ 容器，作为核心安全组件护航 2020 双十一 60wQPS 流量洪峰。
  - **私有云管理平台**: 时年 P8 团队目标建设政务场景下的专有云管理系统。基于 Java 框架 Springboot 开发，该系统分应用运维、资源运营、总集管理三个子系统。我负责其中应用运维中心子系统的后端开发。应用运维中心底层构建云资源列表、监控巡检、用量统计、自定义接口巡检、SQL 巡检等原子能力，并在上层此构建了慢 SQL 统计分析、业务大盘、体检报告等专家报表。该系统目前在全国超过 6 个局点商业化输出，重保护航 2021 某省会城市小学入学等项目、2022 春节核酸检测等项目。
  - **政务云管轻量化交付**: 时年新 P9 团队目标交付 7 个地方政务私有云，早期由开发同学使用 Rainbond 界面化地将来自原不同团队的微服务组合部署在客户 K8S 集群上，配置繁琐导致交付周期长。为解决该问题我基于 docker-compose 搭建单机版轻量化部署交付方案，基于通用的环境初始化脚本与交付步骤文档，仅通过维护不同客户现场的环境变量以中心化管控，纳管来自各个老团队的 20+ 个微服务。编辑 WBS 和 SOP，组织外包同学培训交付流程，单个经过培训的外包可以在 2 日内交付/升级一个现场。基于该方案，团队技术产品簇的单月商业化输出能力从个位数据点上升到两位数。

- 小马智行**(L4 自动驾驶解决方案)** 中国，北京
  工程师、*Infra*、*C++*、*Python*、*Linux*、*Bazel*、严苛的代码质量标准 *2019 年 2 月 - 2020 年 3 月*
  - **行车录音工具链**: 无人车路测需要一位跟车工程师通过外接键盘描述记录路测 issue，为去掉跟车工程师构建行车录音工具链，以麦克风和物理按钮作为硬件方案。首先在 Bazel 项目中引入基于 BSD 软件许可的输入设备接口库 evdev，并基于该库和 Linux 的 ALSA 音频驱动编写守护进程，负责硬件检测、监听来自外部按钮的信号和车载系统进程的管道信息、触发与停止麦克风录音，在行车过程中将 issue 信息存储为工控机上的音频文件。在数据处理阶段，首先在 GCP 上搭建语音转文字服务，并用 Python 脚本将音频文件的内容提取出来，然后使用 Google 的 Protobuf 工具将原信息序列化汇入 QA 数据流。
  - **车载语音系统**: 原车载语音模块使用 Google 的 Pico TTS 库在行车过程中将硬编码的语音内容文本实时播放，从而导致语言单一和 TTS 过程重复消耗计算资源的问题。新的车载语音系统使用基于音频文件的工作流程，旨在减少车载工控机的计算量并支持语音 I18N。为此 1. 研发环节用 Python 编写内部 CLI 工具 `car_sound_utils`(以下简称 CLI)，供内部工程师创建和变更现有语料库、管理语音包版本：首先在 CLI 中基于 AWS 上用 CloudFormation、Lambda 函数计算搭建的整套 TTS 服务封装音频创建命令，然后在 CLI 中添加上传、下载语音文件和上传语音包到内部 storage server 的命令，以及缓存机制加速上传下载。2. 车载系统环节，主进程初始化语音模块时使用 Linux 的 ALSA 驱动将音频流加载进内存，并在语音模块接到播放请求消息时播放。

实习经历

- **爱彼迎**: 2018 年夏爱彼迎发力中国市场，作为大陆首批 8 个实习生，全栈开发国内版房东年度回顾页。
- **旷视科技**: 2018 年初，旷视和 VIVO 合作期间，参与 X21 机型人脸识别模组的研发，做基于 InceptionV3（一种 CNN）移动端模型搜索优化。

系统经验与技术栈

- **系统经验**：MLOps、LLM 应用、Prompt 工程、超大规模分布式系统、高可用、可观测性、自动驾驶等
- **技术栈**：Cursor、Java、Python、Kubernetes、Shell、C++、全球主流云、主流运维工具等