Yong Wei
Make changes

Email : weiyong1024@gmail.com
Mobile : (+86) 173-7657-1024

Education

- **Master - EE - Beihang University** — Beijing, China
  Exempt Entry Exam; Finish courses and research publishing mission during the 1^{st} semester — *Sep. 2016 – Jan. 2019*
- **Bachelor - EE - Beihang University** — Beijing, China
  Exempt NCEE; Honors College(top 50 among 3000+ freshmen); GPA 3.7 — *Sept. 2012 – Jun. 2016*

Competitive programming

- **Google Code Jam Kickstart 2017**
  Top 5%*, rank* 108^{th} *globally.* ***scoreboard, id:WeiYong1024***

Career

- **Mindverse** — Hangzhou, PRC
  TechLead - Infrastructure. leading team — *Jun. 2022 - Now*
  - **Scope1 - System & DevOps**: [1]Successfully implemented company's microservices containerization, cluster orchestration and multiple environments setup for R&D team. [2]Managed cloud resources, built multi-cloud architecture in multiple locations, and optimized cloud consts. [3]Promoted best engineering practices such as Master-Based Development, CodeReview tools and mechanisms. [4]Developed and maintained internal tools with HA such as GitServer, CICD pipelines and load testing systems. [5]Designed and constructed GDPR compliant products and services.
  - **Scope2 - AiInfra**: [1]Assisted algorithm engineers in conducting various inference experiments with acceleration frameworks such as DeepSpeed, FastTransformer. Supported GPT scheduling, pooling, and other requirements. [2]Wrapped OpenAI API call layer, created a pool consisting of OpenAI/AzureOpenAI accounts, and enabled algorithm engineers to flexibly invoke GPT inference capabilities based on their needs.
  - **Scope3 - Data & BI**: Built online-offline data pipelines from scratch, and implemented multiple dashboards for the company's system water level monitoring, business metrics monitoring, BI data analysis, cost monitoring, and other purposes.
- **Aliyun** — Hangzhou, PRC
  Senior SDE - Middleware, backend, leading team — *Aug. 2020 - May. 2022*
  - **Team1 - Alibaba middleware team**: Took over and developed the group's traffic scheduling middleware, which managed all core applications within Alibaba Group, supported over 400,000 containers, and provided cluster-level, second-level traffic scheduling at the individual container dimension, outlying point discovery, minute-level traffic isolation and recovery. This middleware played a crucial role in ensuing the smooth running of the 2020 Double 11 shopping festival.
  - **Team2 - Private cloud**: Participated in the development of a private cloud operation and maintenance system for the government and was fully responsible for the backend of the application operation center subsystem(which accounted for 1/3 for the completed system). Was fully responsible for the product delivery to six customers across the country and worked with outsourced teams to complete it within two days. Succeeded in support in various projects including the 2021 Elementary School Project and the 2022 Spring Festival Nucleic Acid Project.
- **Pony.ai** — Beijing, PRC
  SDE - Infrastructure — *Feb. 2019 - Mar. 2020*
  - **Project1 - Onboard in-vehicle voice module**: Designed and implemented the voice module in Self Driving Vehicles' onboard system, developed and maintained tools that enabled engineers to customize in-vehicle voice.
  - **Project2 - Voice logging workflow**: In the context of reducing the number of on-board personnel in SDV to a single operator, designed and implemented the entire tool set of driver recording, storage,ASR, and issue distribution.

Internship

- **Airbnb**: In the summer of 2018, Airbnb expanded into the Chinese market, and as one of the first 8 interns in the Mainland China office, developed the domestic version of the annual host review page with full-stack tech.
- **Megvii(Face++)**: During the cooperation between Megvii and VIVO in early 2018, participated in the development of the facial recognition module for the X21 model and conducted mobile model search optimization based on InceptionV3(a type of CNN) for mobile devices.

System experience & Tech stack

- **System experience**: LLM, Cloud computing, HA, DevOps, Self driving vehicles
- **Tech stack**: Java、Python、Kubernetes and so on

魏雍
变革与自我变革

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

教育背景

- **硕士 - 北航 - 电子与通信工程** 中国，北京
  保研; 研一上发核心期刊达毕业标准; 北京市优干 *2016 年 1 月 – 2019 年 1 月*
- **本科 - 北航 - 电子信息工程** 中国，北京
  物理竞赛保送; 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*

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)***

工作经历

- **心识宇宙** 中国，杭州
  TechLead - 基础设施、带 *4* 人团队 *2022 年 6 月 - 至今*
  - **Scope1 - System & DevOps**: 从零到一实现：[1] 公司微服务容器化、上集群编排、产研多套环境搭建；[2] 云资源管理，搭建多地、多云架构，优化用云成本；[3] 工程最佳实践：主推主干开发、CodeReview 工具与机制；[4] 内部工具及其高可用：如 GitServer、CICD 流水线、压测系统等；[5]GDPR 标准产品隐私合规建设。等等。
  - **Scope2 - AiInfra**: [1] 协助算法同学进行多种推理加速框架（如 DeepSpeed、FastTransformer 等）的推理实验，支持 GPU 调度、池化等需求；[2] 封装 OpenAI 的 API 调用层，混合不同模型权限 OpenAI 账号池、AzureOpenAI 账号池，让算法团队用好 GPT 的模型能力。等等。
  - **Scope3 - Data & BI**: 从零到一搭建在离线数据 Pipeline，并实现多种大盘，用于公司系统水位监控、业务指标监控、业务 BI 数据分析、成本监控等。
- **阿里云** 中国，杭州
  高级工程师 - 中间件、*web* 后端、带虚线团队 *2020 年 8 月 - 2022 年 5 月*
  - **所在团队 1 - 集团中间件**: 接手并开发集团流量调度中间件。上线功能纳管了阿里集团全部核心应用，支撑 40W+ 容器，提供集群单容器维度秒级流量监控、离群点发现、分钟级流量调度与恢复链路。护航 2020 年双十一双峰平稳进行。
  - **所在团队 2 - 专有云**: 参与开发一款 2G 的基于阿里云专有云底座上的运维管理系统——智能云管，全权负责云上应用运维中心子系统（占完整系统的 1/3）的后端设计实现。全权负责产品对全国 6 个局点的交付，实现纯外包团队 2 日交付一个现场。重保郑州 2021 小学入学项目、2022 春节核酸项目等。
- **小马智行** 中国，北京
  工程师 - 基础设施 *2019 年 2 月 - 2020 年 3 月*
  - **代表项目 1 - 车载系统语音模块**: 设计并实现无人车载系统中的语音模块，开发并维护让工程师定制语音的工具。
  - **代表项目 2 - 行车录音分发 Issue 工具链**: 在当年希望无人车日常随车人数从 2 降到 1 的大目标下，设计并实现车内司机录音、存储、ASR、Issue 分发的整条链路。

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，全栈开发了当年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。

行业经历与技术栈

- **系统经验**：大模型、云计算、高可用、DevOps、自动驾驶
- **技术栈**：Java、Python、Kubernetes 等