HUANG Xinyu

contact@x3huang.dev | Shanghai, China

TECHNICAL HIGHLIGHTS

- **AI Product Specialist**: 4.5 years' experience delivering intelligent speech SaaS solutions, owning complete lifecycle from backend architecture design to service deployment. Familiar with ML model operationalization fundamentals (feature engineering - model deployment - performance monitoring)
- **Cloud-Native Architect**: Expert in designing and developing microservices, familiar with best practice of DDD structure and multiple protocols, eg. HTTP, Webscoket, gRPC/Thrift protocols, RESTFul API, graphQL
- **High-Performance Systems Builder**: Developed intelligent speech gateway service supporting 1M+ QPS, implemented dynamic rate-limiting strategies & multi-layer caching system.
- **DevOps Practitioner**: Familiar with best CI&CD practices, designing workflows with various tools, eg. Jenkins, ArgoCD. Promoted test environment isolation, gray-scale deployment among teams.
- **Continuous Learning Advocate**: Proficient in Python/Golang, familiar with Cpp. Completed graduate-level coursework in ML/DL

EXPERIENCE

Freelancer May 2025 - Present

- **Building** my blog site on self-hosted VPS with: observability, using Prometheus, OTLP stack, Grafana; reliability, with k3s, helm charts; automated deployment, via github actions

Bytedance, Inc. Shanghai, China

Backend Software Engineer, Platform Infrastructure, Speech Team, Data Jan 2021 – Apr. 2025

- **Architected** intelligent speech gateway for LLM products, resolving gateway-to-model service challenges through dynamic rate-limiting strategies and Raft-based load balancing framework, achieving 99.99% voice session availability supporting 10B+ daily requests
- **Enhanced** VolcEngine's intelligent speech platform authentication flow, integrating JWT/STS/Volcano AKSK credential systems. Redesigned interactions of order system with Speech services, supporting model resource scaling from thousands to hundreds of thousands
- **Built** multimodal data analytics pipeline combining Flink real-time processing and Spark ETL, delivering statistics for pricing strategies and training data collection
- **Developed** education-focused LLM applications using proprietary workflow engine, creating duplex speech assessment pipelines integrating ASR and pronunciation detection models for "Open Language" and experimental Ed-tech products

San Diego Hunger Coalition (NGO) San Diego, US

Data Intern Jul 2019 – Aug 2019

- **Cleaned** 500k+ entries in Amazon RDS using the “star” schema
- **Automated** the generation of routine reports with Python scripts through Amazon RDS API

EDUCATION

UC San Diego, School of Global Policy and Strategy San Diego, US

Master of International Affairs Sept 2018 – Jun 2020

- Core courses: Quantitative Methods series, Statistical Learning series (ECE), Data Analysis (ECE)
- Capstone Essay: “Do People Divorce Because of a Policy? Impact of Limit Purchasing Orders in China on Divorce and Marriage Rates”
- Machine Learning course project: “On Automatic X-ray Images Generation from Radiology Reports”

Shanghai International Studies University Shanghai, China

Bachelor of Arts in Translation and Interpretation Sept 2014–June 2018

ADDITIONAL

My_wacig: Compiler of the Monkey Language in Go Oct 2022

- Individual project: a toy compiler https://github.com/XyLearningProgramming/my_wacig

黄昕宇

contact@x3huang.dev | 中国 上海

技术亮点

- **AI 应用开发**：4.5 年语音 AI 产品 SaaS 交付经验，完整参与后端服务架构设计到生产部署的全生命周期，熟悉机器学习模型服务化基础流程（特征工程-模型部署-效果监控）
- **云原生架构设计**：熟悉 DDD 领域驱动设计等常见软件设计范式、HTTP / Websocket / gRPC/Thrift 通信协议、RESTFul API 最佳实践，设计日均亿级调用微服务
- **高并发系统**：构建支持高并发的智能语音网关，设计动态限流规则与多级缓存策略，多种途径优化 MySQL 复杂查询
- **DevOps 实践者：**熟悉 CICD 流程和配套工具链，如 Jenkins,ArgoCD 等，编写并在团队内推广带有隔离环境测试、灰度发布等功能的自动化部署流程
- **持续学习者**：熟悉 Python/Golang/C++技术栈，系统学习过机器学习、深度学习研究生课程

工作经验

自由职业 2025 年 5 月 –

- **构建** VPS 上的个人网站，可靠性：k3s, helm charts 部署和管理多种服务；可观测性：prometheus, promtail, grafana 等实时搜集指标和日志；自动化部署：定制 github workflow 完成 CICD 流程

字节跳动 中国 上海

后端软件开发工程师，平台基建，语音团队，Data 2021 年 1 月 – 2025 年 4 月

- **制定**大模型产品的后端架构，处理从网关到大模型之间的服务架构问题，通过动态限流策略分发、基于 Raft 算法的语音会话负载均衡框架，实现语音服务 99.99%可用性，支持内外部用户和数万外部客户的每日数十亿请求
- **熟悉** AI 产品控制面功能，配合优化火山引擎智能语音平台和产品使用的权限验证流程，整合 JWT/STS/火山 AKSK 等认证方式；迭代火山订单系统和智能语音解决方案层的交互流程，支撑模型资源规模从千级到数十万级的跨越式增长
- **构建**多模态数据分析平台，通过 Flink+Spark 的实时计算与离线 ETL 相结合的方式，产出服务调用热力图、用户行为等数据看板，支持了模型定价决策与模型训练数据采集
- **设计**教育场景大模型应用解决方案，基于自研的流程编排引擎构建语音评估双工管线，整合 ASR 与发音检测等模型开发自适应智能评测体系，推动开言英语和其他实验性教育产品的开发

San Diego Hunger Coalition (NGO) 美国 圣地亚哥

数据实习生 2019 年 7 月 – 2019 年 8 月

- **自动化**生成常规报告和表单，通过编写与 Amazon RDS API 交互的 Python 脚本，自动化报告生成

教育背景

加州大学圣地亚哥分校，全球政策与战略学院 美国 圣地亚哥

国际关系硕士 2018 年 9 月 – 2020 年 6 月

- 核心课程：Quantitative Methods series, Statistical Learning series (ECE), Data Analysis (ECE)
- 毕业论文：“Do People Divorce Because of a Policy? Impact of Limit Purchasing Orders in China on Divorce and Marriage Rates”
- 机器学习课程项目：“On Automatic X-ray Images Generation from Radiology Reports”

上海外国语大学 中国 上海

英语文学学位（国际公务员班） 2014 年 9 月 – 2018 年 6 月

其他

My_wacig 2022 年 10 月

- 基于 Thorsten Ball 的 Monkey 动态语言，添加了新类型和语法的简单编译器：https://github.com/XyLearningProgramming/my_wacig