上海
接受远程工作及搬迁

罗一阳（Nikolai Zakharov）

nikolai.zakharov92@gmail.com
+86 136 2105 3870 个人网站
LinkedIn GitHub

个人简介

高级应用机器学习与生成式AI工程师，拥有8年以上学术机器学习研究与产业经验。交付过大语言模型、多模态、推荐、语音、自动驾驶及三维生成系统，覆盖模型训练、评测、推理优化及机器学习基础设施；清华大学计算机硕士。

工作经历

AI创业公司 · 生成模型与ML平台　　机器学习工程师　　2026.05--2026.07 · 上海

- 参与牙冠生成模型研发，并构建配套的多GPU训练与Kubernetes/Kubeflow基础设施。

VK · 社交媒体平台　　机器学习工程师 / 应用研究　　2025.01--10；2026.02--04 · 莫斯科

- 构建LLM描述生成服务，提升音乐视频推荐相关性；设计三阶段管道（蒸馏训练元数据提取器、音频-文本相关性预测器）；部署item2item推荐质量仪表板用于团队评测。
- 使用正则弱标注与LLM生成数据构建剧集识别、标准化及元数据提取；相似视频推荐的TVT提升4.5--6.3%；利用内容模型信号过滤同集重复上传，重复结果率从60%降至7%。

Kandinsky Lab（Sber AI） · GenAI　　高级机器学习工程师　　2025.11 -- 2026.01 · 莫斯科

- 构建多LLM合成数据流程，使用LoRA微调Qwen用于用户提示词增强，并在A100集群上线可复现的多GPU图生视频评测。

蔚来 NIO · 智能电动车与自动驾驶　　算法工程师 / 自动驾驶　　2023.07 -- 2025.01 · 上海

- 训练端到端Transformer轨迹规划模型，融合地图与视觉上下文，使用分布式训练并以Hydra管理实验。
- 开发基于规则的C++实时事件检测与标注模块，接入开环/闭环评测，自动收集问题场景用于分析。

T-Bank · 数字银行　　机器学习工程师 / 语音技术　　2021.09 -- 2023.02 · 莫斯科

- 训练并上线Transformer语音转换系统（95%说话人相似度），接入月活1000万+的语音助手产品；使用Go实现流式推理与张量组件，并以TensorRT优化端侧与云端推理。

华为 · 全球ICT公司　　机器学习工程师 / 消费者业务　　2019.07 -- 2021.09 · 北京

- 构建支持自动更新的多语言Trie树NLU引擎，覆盖90%+助手流量；与华为俄罗斯团队合作，上线端侧语音克隆并参与其端侧训练框架开发；研究RNN-T ASR异构端侧推理；为视线估计模型实现ARM NEON intrinsic，ARM性能略优于OpenCV。

教育与科研

清华大学　　2016--2019；2022--2023　北京

- 计算机科学与技术硕士（机器学习），全额奖学金；2022--2023年在清华完成一年博士课程。JVCIR论文：
  Towards Controllable Image Descriptions with Semi-Supervised VAE [DOI]。
- 华为诺亚方舟实验室研究实习（2018--2019）：使用TensorFlow 1重写DARTS，生成可通过TFLite直接在端侧运行的静态图；参与基于GAN的人脸识别数据增强研究。

技术能力

模型与训练	PyTorch、Transformer、LLM微调与蒸馏、三维生成建模、合成数据、分布式/多GPU训练。
评测与推理	多模态、几何及开环/闭环评测、数据版本、实验管理、TensorRT、模型服务。
工程系统	Python、C++、Go、TensorRT、TFLite、ARM NEON、Kubernetes/Kubeflow、Kueue、Docker、CI/CD。
语言	俄语（母语）；英语与中文（流利）；法语与土耳其语（基础）。