居住地：莫斯科（可 relocate 至中国/新加坡/全球）

罗一阳（Nikolai Zakharov）

+86 13621053870
n.d.zakharov@outlook.com
LinkedIn
yiyang92.github.io

个人简介

高级机器学习工程师（多模态/LLM），8 年以上生产级 AI 经验。专注**音频大模型**、**多模态系统**、**推理优化**（vLLM、TensorRT）。交付：音乐推荐描述服务（VK）、RAG 智能体（+20%，蔚来）、实时语音合成（-88%，Tinkoff）。清华硕士，十年中国科技生态。精通俄语、英语、中文。

工作经历

VK AI / 应用研究组 — **高级机器学习工程师** — **2026.01 – 至今**

- 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**作为推荐团队核心指标。

VK Video / 视频推荐组 — **高级机器学习工程师** — **2025.01 – 2025.12**

- 架构推荐系统的 **NER+ 元数据提取管道**（LLM+ 规则），提升候选生成质量，TVT 提升 **25%**，重复结果率 **60% 降至 7%**。

蔚来汽车 / 自动驾驶大模型组 — **高级机器学习工程师** — **2023.03 – 2025.01**

- 架构 **AI 智能体系统**（金点子平台）：RAG 检索 +LLM 评估（+**20%**）；构建 RAG-based LLM 助手；部署 C++ 实时 AD 事件检测，支持车载评估。

Tinkoff 银行 / 语音技术组 — **机器学习工程师** — **2021.09 – 2023.02**

- 训练并部署**语音转换系统**（Transformer，95% 相似度），服务 1000 万 + 月活；优化 **HiFiGAN 声码器**（GPU 算子融合，RTF 降低 **88%**）；架构 **TensorRT** 边缘部署。

华为 / 小艺语音助手 — **机器学习工程师** — **2019.07 – 2021.09**

- 构建**多语言 Trie 树 NLU 引擎**，覆盖 **90%+** 语音助手流量；开发云端与端侧**语音克隆系统**。

其他经历

Sber AI / 分布式系统组 — **研究合作** — **2025.12 – 2026.01**

- 构建 Kandinsky 扩散模型的**提示词优化系统**；设计**分布式 LLM 评估基础设施**，支持 GPU 集群编排。
- 设计 LLM benchmark 分布式评估基础设施：PostgreSQL 队列系统与 GPU 集群编排。

华为诺亚方舟实验室 — **研究实习生** — **2018 – 2019**

- 开发 **AutoML 系统**（基于 DARTS），边缘设备精度提升 15%；基于 GAN 的 identity-preserving 人脸识别数据增强。

教育背景

清华大学 — **北京，中国**

- **硕士**（2016 – 2019），计算机科学与技术（机器学习），全额奖学金。论文：《基于深度学习的多样性图像描述生成》。
- **博士**（2022 – 2023），课程完成，返回产业界。

技术专长

- **LLM 系统**：RAG 管道、智能体框架、模型蒸馏、推理优化（vLLM、TensorRT、量化）
- **多模态**：音频-文本模型、音乐理解、描述生成、扩散模型
- **ML 基础设施**：分布式训练（FSDP、DeepSpeed）、Kubernetes、KServe、GPU 集群
- **编程**：Python（PyTorch、HuggingFace）、C++（实时推理）、Go