Shanghai
Open to remote / relocation

Nikolai Zakharov

nikolai.zakharov92@gmail.com
+86 136 2105 3870 Website
LinkedIn GitHub

Summary

Senior Applied ML / GenAI Engineer with 8+ years spanning academic ML research and industry. Delivered LLM, multi-modal, autonomous-driving, speech, and 3D generative systems across model training, evaluation, optimized inference, and infrastructure; Tsinghua M.S.

Professional Experience

AI Startup · Generative Models & ML Platform — **ML Engineer** — May 2026 – Jul 2026 · Shanghai

- Worked on dental-crown generation models and built supporting multi-GPU/Kubernetes infrastructure.

VK · Social media platform — **ML Engineer / Applied Research** — Jan–Oct 2025; Feb–Apr 2026 · Moscow

- Built **LLM captioning** and 3-stage distillation for music-video recommendation; deployed an **item2item quality dashboard** for team evaluation.
- Built **TV-series detection and metadata extraction** from regex-bootstrapped, LLM-generated data; lifted **TVT by 4.5–6.3%** and used content-model signals to filter same-episode uploads, cutting duplicates from **60% to 7%**.

Kandinsky Lab (Sber AI) · GenAI — **Senior ML Engineer** — Nov 2025 – Jan 2026 · Moscow

- Fine-tuned Qwen with LoRA for prompt enhancement; built multi-LLM synthetic data and reproducible multi-GPU image-to-video evaluation on A100s.

NIO · Smart EV / autonomous driving — **Algorithm Engineer, Autonomous Driving** — Jul 2023 – Jan 2025 · Shanghai

- Trained an end-to-end **transformer planner** over map/visual context with distributed training and Hydra experiments.
- Built **C++ rule-based event tagging** for open/closed-loop evaluation and automatic collection of failure cases.

T-Bank · Digital bank — **ML Engineer, Voice Kit** — Sep 2021 – Feb 2023 · Moscow

- Trained/deployed voice conversion (**95% speaker similarity**) for a **10M+ MAU** assistant; built Go streaming/tensor components and **TensorRT** edge/cloud inference.

Huawei · Global ICT company — **ML Engineer, CBG** — Jul 2019 – Sep 2021 · Beijing

- Built auto-updating trie NLU (**90%+** traffic); shipped on-device voice cloning and contributed to its on-device training framework with Huawei Russia; prototyped heterogeneous on-device RNN-T ASR; wrote **ARM NEON** gaze kernels slightly faster than OpenCV on ARM.

Education and Research

Tsinghua University — 2016 – 2019; 2022 – 2023 · Beijing

- **M.S. Computer Science (ML)**, full scholarship; one year of PhD coursework (2022–2023); JVCIR paper: Towards Controllable Image Descriptions with Semi-Supervised VAE [DOI].
- **Huawei Noah's Ark Lab, Research Intern (2018–2019):** reimplemented DARTS in TensorFlow 1 with static graphs runnable on devices via TFLite; researched GAN-based face-recognition augmentation.

Skills

Models	PyTorch, transformers, LLM fine-tuning/distillation, 3D generative modeling, synthetic data, distributed/multi-GPU training.
Evaluation	Multimodal, geometric, and open/closed-loop evaluation, dataset versioning, experiment management.
Engineering	Python, C++, Go, TensorRT, TFLite, ARM NEON, model serving, Kubernetes/Kubeflow, Kueue, Docker, CI/CD.
Languages	Russian (native); English and Mandarin Chinese (fluent); French and Turkish (basic).