Robert Ressl

STAFF / PRINCIPAL AI SYSTEMS ENGINEER · ZURICH, CH · HTTPS://RESSL.CH

Hands-on engineer across model tuning, inference infrastructure, offensive security, and regulated production. I lead from the code, the architecture, and the customer room.

Email · LinkedIn · Book 30 minutes · PGP fingerprint: E247 2215 D4E1 0353 E2F7 D73E 46A9 7346 18D3 7D4D

SELECTED OUTCOMES

Latest release: Ornith-1.5-35B-A3B-uncensored — Uncensored release family · 2026-08-27 · 2.1K+ downloads · 30 days

Gemma 4 31B uncensored · format family

PROBLEM Remove refusal behavior from a 30.7B multimodal model without losing a deployable format set.
MY CONTRIBUTION & DECISION Applied weight-level abliteration, then carried the same model lineage through BF16, ModelOpt NVFP4, and a GGUF K-quant ladder. For multimodal NVFP4 serving, the vision tower, embedder, and lm_head stayed in BF16.
RESULT BF16 evaluation: 0/686 effective refusals across four datasets. The 20 GB NVFP4 build, reduced from 59 GB, was validated in vLLM and SGLang. GGUF: 12–31 GB with 0–1/100 hard refusals across quant levels.
Evidence: BF16 · NVFP4 · GGUF

Ornith 1.0 35B · NVFP4

PROBLEM Fit a 34.7B Qwen3.5-MoE reasoning checkpoint onto a single Blackwell workstation without quantizing every subsystem.
MY CONTRIBUTION & DECISION Quantized routed experts to NVFP4 with NVIDIA ModelOpt; kept attention QKV, shared experts, and the vision encoder at higher precision.
RESULT About 23 GB versus 69 GB BF16, validated to load and generate with vLLM on an RTX PRO 6000 Blackwell.
Evidence: Ornith-1.0-35B-NVFP4

MiniMax-M3-uncensored · 428B MoE

PROBLEM Study refusal behavior in a 428B multimodal MoE while preserving its reasoning and routing structure.
MY CONTRIBUTION & DECISION Modified attention o_proj and every residual-writing down_proj across all 60 layers; released the result as 796 GB BF16 in 59 shards.
RESULT 0/16 hard refusals on the published sample; multimodal behavior, reasoning, and MoE routing were smoke-tested, not fully benchmarked.
Evidence: MiniMax-M3-uncensored

28 published models: huggingface.co/ressl

IMPACT

11 Countries with embedded delivery — Built the global offensive-security practice and its customer-delivery model for Tier-1 banks, insurers, and automotive companies; recognized with Heart of Kyndryl 2024.

40% Faster red-team delivery — Custom tooling and in-engagement prototyping shortened APT simulations for Fortune 500 customers.

80% Less manual operations work — Go backend automation removed most repetitive operations work while founding a Linux department.

EXPERIENCE

2023 — now **Associate Director, Offensive Security Global — Kyndryl**

Embedded with C-Level at Tier-1 banks, insurers, automotive across 11 countries. 100+ on-site engagements. Custom tooling shipped under engagement deadlines. DORA / FINMA / TLPT.

2022 — 2023 **Senior Red Team Lead Engineer, Global — Kyndryl**

Forward-deployed APT simulation for Fortune 500. Custom tooling and rapid prototyping in-engagement — 40% efficiency gain in delivery.

2021 — 2022 **Associate Partner Security — Kyndryl**

Built consulting practice and GTM strategy after IBM spin-off. Pre-sales scoping and customer-embedded delivery model.

2019 — 2021 **Security Consultant & Cloud Architect — IBM**

2018 — 2019 **DevOps Engineer Expert — Avectris AG**

2016 — 2017 **DevOps Engineer — Swisscom AG**

2014 — 2016 **Linux Architect — EveryWare AG**

2013 — 2014 **Linux DevOps Engineer — CTBTO (United Nations)**

2009 — 2012 **Lead Solution Architect — Schrack Seconet AG**

2004 — 2008 **Linux Engineer — Freelance**

SYSTEMS AND PLATFORMS

WEIGHTS Published checkpoints and model behavior **TUNE** LoRA · behavior work **QUANTIZE** NVFP4 · GGUF **SERVE** vLLM · SGLang **ATTACK** MCP · agents **HARDEN** Policy · audit

AI inference platform I build and operate a GitOps-managed Talos/Kubernetes inference platform across NVIDIA and AMD GPUs, covering model preparation, backend readiness, streaming, and operational telemetry. The case study documents a vLLM-router implementation and its trade-offs from a specific point in that work. Private platform · public architecture notes

Own dual-stack network · AS214476 I operate AS214476 and originate my own IPv4 and IPv6 address space over BGP. Routing policy, dual-stack operations, Linux networking, reliability, and observability are part of the system I own end to end.

Architecture deep-dive: ressl.ch/blog/inference-platform

SECURITY PRACTICE

MODEL BEHAVIOR Abliteration, alignment analysis, refusal measurement, and adversarial evaluation across released checkpoints.

AGENT ATTACK SURFACE LLM red teaming, prompt injection, MCP scanning, tool poisoning, exfiltration paths, and tool-chain abuse.

CONTAINMENT & AUDIT Runtime policy enforcement, egress controls, human approval, signed audit trails, and compliance evidence.

KEY CREDENTIALS

39 credentials across offensive security, Linux, automation, containers, and cloud.

SECURITY & OFFENSIVE
CISSP · OSEP · OSWE · OSCP — 4 verified

RED HAT
RHCA · EX288 OpenShift · EX180 Containers/K8s — 3 verified

LINUX FOUNDATION & LPI
LFCE — 1 verified

CLOUD & ITSM
AWS Solutions Architect — 1 verified

VERIFY IT LIVE

→ Interactive terminal with the Evidence CLI behind every figure above

→ Inference-platform architecture article

→ Complete work portfolio (full-site print view)