

Robert Ressler

Engineering portfolio

STAFF / PRINCIPAL / FORWARD DEPLOYED AI SYSTEMS

I build, break, and deploy AI systems.

ressl.ch · rr@ressl.ch · linkedin.com/in/robertressl

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



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

28 Hugging Face models	31.5K+ downloads · last 30 days	39 verified credentials	100+ engagements	20+ years in systems
---------------------------	------------------------------------	----------------------------	---------------------	-------------------------

HF snapshot 2026-09-01

01 / MODEL ENGINEERING

Model work that ships

I transform model behavior, quantize for real hardware, validate serving targets, and publish the artifacts with explicit evaluation limits.

Gemma 4 31B uncensored · format family

59 GB → 20 GB

NVFP4 checkpoint · vLLM and SGLang validated

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](#)

PROBLEM

Remove refusal behavior from a 30.7B multimodal model without losing a deployable format set.

MY CONTRIBUTION & DECISION

Applied weight-level ablation, 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.

RELEASED ARTIFACT

[BF16](#) [NVFP4](#) [GGUF](#)

[Read the engineering notes →](#)

Ornith 1.0 35B · NVFP4

~69 GB → ~23 GB

Experts-only NVFP4 · one Blackwell GPU

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](#)

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.

RELEASED ARTIFACT

[Ornith-1.0-35B-NVFP4](#)

[Read the engineering notes →](#)

MiniMax-M3-uncensored · 428B MoE

428B · BF16

Weight-level behavior work · 796 GB published

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](#)

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.

RELEASED ARTIFACT

[MiniMax-M3-uncensored](#)

[Read the engineering notes →](#)

[View all Hugging Face releases](#)

All releases in the snapshot (28)

MODEL	FORMAT	RELEASED	DOWNLOADS · 30 DAYS
GLM-5.3-NVFP4	NVFP4	2026-08-29	176
Ornith-1.5-35B-A3B-uncensored-MLX-8bit	MLX 8bit	2026-08-27	197
Ornith-1.5-35B-A3B-uncensored-MLX-6bit	MLX 6bit	2026-08-27	229
Ornith-1.5-35B-A3B-uncensored-MLX-4bit	MLX 4bit	2026-08-27	152
Ornith-1.5-35B-A3B-uncensored-MLX	MLX	2026-08-27	282
Ornith-1.5-9B-uncensored-MLX-8bit	MLX 8bit	2026-08-27	104
Ornith-1.5-9B-uncensored-MLX-6bit	MLX 6bit	2026-08-27	509
Ornith-1.5-9B-uncensored-MLX-4bit	MLX 4bit	2026-08-27	94
Ornith-1.5-9B-uncensored-MLX	MLX	2026-08-27	243
Ornith-1.5-35B-A3B-uncensored-GGUF	GGUF	2026-08-27	949
Ornith-1.5-35B-A3B-uncensored-NVFP4	NVFP4	2026-08-27	325
Ornith-1.5-9B-uncensored-GGUF	GGUF	2026-08-27	859
Ornith-1.5-9B-uncensored-NVFP4	NVFP4	2026-08-27	117
Ornith-1.5-397B-uncensored-NVFP4	NVFP4	2026-08-21	337
Ornith-1.5-397B-uncensored-FP8	FP8	2026-08-21	31
Laguna-S-2.1-Uncensored	BF16	2026-08-02	43
MiniMax-M3-uncensored-NVFP4	NVFP4	2026-07-15	20,093
gemma-4-31B-it-uncensored-MLX-4bit	MLX 4bit	2026-07-10	226
gemma-4-31B-it-uncensored-MLX-5bit	MLX 5bit	2026-07-10	147
gemma-4-31B-it-uncensored-MLX-6bit	MLX 6bit	2026-07-10	118
gemma-4-31B-it-uncensored-MLX-8bit	MLX 8bit	2026-07-10	122
gemma-4-31B-it-uncensored-MLX-bf16	MLX bf16	2026-07-10	171
MiniMax-M3-uncensored	BF16	2026-07-09	637
gemma-4-31B-it-uncensored-GGUF	GGUF	2026-07-08	1,746
gemma-4-31B-it-uncensored	BF16	2026-07-08	135
gemma-4-31B-it-uncensored-NVFP4	NVFP4	2026-07-08	285
GRM-2.6-Plus-NVFP4	NVFP4	2026-07-03	9
Ornith-1.0-35B-NVFP4	NVFP4	2026-06-29	3,202

Attack the agent. Harden the system.

I connect model-behavior research with the controls needed around agents: red teaming, prompt injection, MCP attack paths, containment, policy enforcement, and auditability.

01	Model behavior	Abliteration, alignment analysis, refusal measurement, and adversarial evaluation across released checkpoints.
02	Agent attack surface	LLM red teaming, prompt injection, MCP scanning, tool poisoning, exfiltration paths, and tool-chain abuse.
03	Containment & audit	Runtime policy enforcement, egress controls, human approval, signed audit trails, and compliance evidence.

[mcp-firewall](#) · [mcpwn](#)

Production systems, end to end

The model is only one component. I build the Linux, GPU, Kubernetes, inference, networking, observability, and security path around it.

01 WEIGHTS	02 TUNE	03 QUANTIZE	04 SERVE	05 ATTACK	06 HARDEN
Published checkpoints and model behavior	LoRA · behavior work	NVFP4 · GGUF	vLLM · SGLang	MCP · agents	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](#)

[Read the platform case study →](#)

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.

IPv4 · 5.182.62.0/24	IPv6 · 2a01:e000::/29	BGP	OPEN PEERING
----------------------	-----------------------	-----	--------------

[AS214476 · PeeringDB](#) [AS214476 · RIPEstat](#)

Audit-grade delivery

Technical delivery and evidence in FINMA, DORA, and TLPT environments.

[FINMA](#) · [DORA](#) · [TLPT](#)

Lead from the difficult technical work

I lead by reducing ambiguity, making the hard technical decisions, and staying with the problem until the system ships.

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.

Career evidence: [Global offensive-security delivery](#) · [Career evidence: Cloud architecture to security practice](#)

40%

Faster red-team delivery

Custom tooling and in-engagement prototyping shortened APT simulations for Fortune 500 customers.

Career evidence: [Cloud architecture to security practice](#)

80%

Less manual operations work

Go backend automation removed most repetitive operations work while founding a Linux department.

Career evidence: [Infrastructure to automation](#)

Selected work

Open-source AI-security tools and systems I build, publish, and operate.

RUST • REACT • POSTGRESQL • AI WORKFLOWS

paperless-archivist

Security-first AI automation for Paperless-ngx: resumable OCR and tagging workflows, validation-gated autopilot, human review, cited document chat, RBAC, and auditable actions.

[GitHub](#)

PYTHON • MCP • AGPL-3.0

mcp-firewall

Runtime gateway for MCP tool calls: 12-stage inspection, policy-as-code, Ed25519-signed audit chains, and DORA / FINMA / SOC 2 evidence exports.

[GitHub](#) [PyPI](#)

PYTHON • MCP • AGPL-3.0

mcpwn

Pre-deployment MCP scanner with 10 checks for tool poisoning, prompt injection, exfiltration, SSRF, permissions, and tool-call chaining.

[GitHub](#) [PyPI](#)

GO • HTMX • AS214476

terminal-trouble.com

A self-hosted Hack The Box and CTF write-up platform with the full brutalist terminal signature.

[terminal-trouble.com](#)

Product code, shipped quietly

Uhr Lernen Einfach

Learn to read clocks. 7 game modes, 8 themes, Game Center.

iOS • Swift / SwiftUI

ABC Lernen

Learn the alphabet. 8 game modes — writing, matching, syllables.

iOS • Swift / SwiftUI

Pixel Privacy

Blur faces & license plates in photos. 100% offline, on-device.

iOS • Swift / SwiftUI

CraftFace

Turn your selfie into a Minecraft character. 3D preview, 7 outfits.

iOS • Swift / SwiftUI

Twenty years of increasing technical scope

From Linux and high-availability systems to cloud, offensive security, model work, and global customer delivery.

2023 – PRESENT	Global offensive-security delivery Embedded with C-level at Tier-1 banks, insurers, and automotive companies while remaining hands-on.	2019 – 2023	Cloud architecture to security practice Practice building, APT simulation, and custom tooling shipped under engagement deadlines.
2013 – 2019	Infrastructure to automation Configuration automation at scale, engineer training, and Go backend automation.	2004 – 2012	Linux engineering to solution architecture High-availability systems in colocation, then a global IT rebuild with audited security.
2023 – PRESENT	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 Cloud transformations – AWS, GCP, OpenShift, K8s. Founded consulting security division.		
2018 – 2019	DevOps Engineer Expert @ Avectris AG Founded Linux department. Backend automation in Go (80% workload reduction). Scrum Master.		
2016 – 2017	DevOps Engineer @ Swisscom AG Automation for Swisscom TV (300+ channels). Voice control system. 35+ engineers trained.		
2014 – 2016	Linux Architect @ EveryWare AG Enterprise infrastructure for Allianz, JobCloud, UBS AG.		
2013 – 2014	Linux DevOps Engineer @ CTBTO (United Nations) Migrated 600 nodes from CFEngine to Puppet. Near 100% config automation.		
2009 – 2012	Lead Solution Architect @ Schrack Seconet AG Rebuilt global IT (HQ + 16 branches). 100% security audit pass. ERP RPO: 1 day → 15 min.		
2004 – 2008	Linux Engineer @ Freelance HA systems in Interxion datacenters. 99.99% uptime.		

Credentials

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

CISSP Security & Offensive	OSEP Security & Offensive	OSWE Security & Offensive
OSCP Security & Offensive	RHCA Red Hat	LFCE Linux Foundation & LPI
EX180 Containers/K8s Red Hat	EX288 OpenShift Red Hat	AWS Solutions Architect Cloud & ITSM
Security & Offensive		
OSCC	OSWA	OSWP
CRTO	PNPT	Black Hat DFIR
CompTIA		
CySA+	CNVP	PenTest+
CSIS	Security+	Network+
Server+	A+	
Red Hat		
RHCE (6, 7, 8)	RHCSA (6, 7, 9)	EX188 Containers
EX358 Services	EX447 Ansible Best	EX407 Ansible Auto
Linux Foundation & LPI		
LFCS	LPI-3 (304)	LPI-3 (303)
Cloud & ITSM		
AWS Cloud Practitioner	ITIL v3 Foundation	
Chef / DevOps		
Extending Chef	Certified Chef Developer	Deploying Cookbooks
Local Cookbook Dev	Basic Chef Fluency	

Build the hard system with me.

I am interested in hands-on Staff, Principal, and Forward Deployed roles at AI companies, plus selective technical collaboration.

Email Robert LinkedIn Book 30 minutes Download CV (PDF) PGP key

OpenPGP fingerprint E247 2215 D4E1 0353 E2F7 D73E 46A9 7346 18D3 7D4D