CYBERNEUROVA · BENCHMARK REPORT

Qwen3.6-35B-A3B · Abliterated

Multi-axis abliteration: hybrid attention (Gated DeltaNet + self-attn), 256 fused experts, thinking-mode chain-of-thought, all collapsed on a 9-benchmark suite.

Base model: **Qwen/Qwen3.6-35B-A3B** | Total params: **35B (3B active)**

Layers: **40 (30 linear_attn + 10 self_attn)** | Experts: **256 routed + 1 shared** | Direction layer: **24 / 40**

REFUSAL (ADVBENCH)

90.9% → 0.0%

−90.9 pp — full refusal collapse

SOFT-REFUSAL PROBE (55 OOD)

85.5% → 0.0%

−85.5 pp on hard out-of-distribution prompts

CYBER-WEAPONS COMPLIANCE

44.7% → 74.7%

+30.0 pp — meaningful technical unlock

COHERENCE PRESERVED

97.4% → 98.0%

no abliteration tax on fluency / diversity

Full benchmark comparison

BENCHMARK	BASELINE	ABLITERATED	Δ
Refusal (AdvBench-style, n=33) lower is better	90.9%	0.0%	-90.9 pp
Soft-refusal probe (55 OOD prompts) lower is better	85.5%	0.0%	-85.5 pp
Hacking compliance (pen-test prompts) higher is better	56.3%	76.0%	+19.7 pp
Cyber-weapons compliance (malware/exploit) higher is better	44.7%	74.7%	+30.0 pp
Bug finding (defensive coding) higher is better	95.0%	93.3%	-1.7 pp
Coding (HumanEval-style) higher is better	93.3%	93.3%	—
Reasoning (multi-step math/logic) higher is better	93.3%	100.0%	+6.7 pp
Coherence (fluency / diversity) higher is better	97.4%	98.0%	+0.6 pp
Tool calling (JSON function format) higher is better	0.0%	0.0%	—

baseline = stock Qwen/Qwen3.6-35B-A3B. abliterated = same model with the refusal direction captured at layer 24 and orthogonalized out of all attention writes + 256 fused-expert down-projections + shared-expert down-projection + embedding. The Δ column is sign-aware: green always means “better for the abliteration goal” per benchmark.

> **Tool-calling 0% / 0% — measurement artifact, not capability loss.** Both baseline and abliterated score 0% on the tool-calling benchmark. The grader expects raw JSON function calls; the model wraps its tool calls in `<think>...</think>` reasoning blocks (visible in its other outputs), so the regex parser doesn't recognize them. If abliteration had broken tool-

> calling, baseline would still be ~0.9 and ablated ~0.0. Same-zero both sides confirms it's a grader-vs-CoT issue we'll fix in a future framework update.

What changed where

The model's architecture combines four challenges that needed framework work to handle:

- **Hybrid attention**: 30 linear_attn (Gated DeltaNet) layers + 10 self_attn layers. Both have a single write module that adds to the residual stream — `linear_attn.out_proj` and `self_attn.o_proj` respectively. We orthogonalize whichever exists per layer.
- **Fused mixture-of-experts**: 256 experts stored as a single 3D Parameter `experts.down_proj` of shape `[256, 2048, 512]` — not iterable as a ModuleList. We orthogonalize all 256 expert slices in one batched operation per layer.
- **Shared expert**: standard `shared_expert.down_proj` Linear, orthogonalized normally.
- **Chain-of-thought refusals**: model wraps every response in `<think>...</think>` blocks before answering. Our refusal classifier was strict-anchored at position 0 and missed all of them. We added a CoT-stripping pass + a describer-pattern matcher; baseline refusal measurement jumped from 0% (broken classifier) to 85.5%, then to 90.9% on the 33-prompt bench shown here.

How to run

Standard transformers + AutoModelForCausalLM, bf16:

```
import torch
from transformers import AutoModelForCausalLM, AutoTokenizer

model = AutoModelForCausalLM.from_pretrained(
    "cyberneurova/CyberNeurova-Qwen3.6-35B-A3B-abliterated",
    dtype="bfloat16", device_map="cuda:0",
    trust_remote_code=True,
)
tok = AutoTokenizer.from_pretrained(
    "cyberneurova/CyberNeurova-Qwen3.6-35B-A3B-abliterated",
    trust_remote_code=True,
)

messages = [{"role": "user", "content": "Your prompt here."}]
text = tok.apply_chat_template(messages, tokenize=False, add_generation_prompt=True)
inputs = tok(text, return_tensors="pt").to("cuda:0")
out = model.generate(**inputs, max_new_tokens=1024)
print(tok.decode(out[0][inputs.input_ids.shape[1]:], skip_special_tokens=True))
```

The model emits `<think>...</think>` reasoning before its answer. Use `max_new_tokens >= 512` to capture both the reasoning and the response.

Available on HuggingFace at cyberneurova/CyberNeurova-Qwen3.6-35B-A3B-abliterated · part of the CyberNeurova × Qwen Family collection. · Print this page to PDF via `File → Print → Save as PDF` .