Ektome-Qwen3-8B-PristinelyUncensored

Capability-preserving refusal excision

Zynerji · Ektomē pipeline

2026-07-27

Abstract

Standard abliteration removes a coarse “refusal direction” that is entangled with directions carrying knowledge and reasoning; the result is an uncensored model with a measurable capability tax that is almost never quantified. **Ektomē** (*εκτομή*, *excision*) is a proprietary, training-free procedure that removes the refusal-*specific* component while leaving general capability intact.

This document reports the intervention applied to `Qwen/Qwen3-8B`, the measured before/after receipt, and—unlike every comparable published model—a **statistical certificate** with stated n, confidence intervals, and a hypothesis test. That certificate **passed**: capability was retained.

1. Method

The intervention is a single-step, training-free weight edit applied to the pristine model: a low-rank excision of the *refusal-specific* subspace, isolated so that directions carrying general helpfulness are not co-removed, and applied norm-preservingly so the edited weights retain their operator scale.

The extraction depth is selected per model by an automated search rather than fixed, because a fixed choice systematically under-abliterates; the search criterion is measured compliance on held-out harmful prompts.

Implementation details of the direction estimator, the excision operator and the depth-selection procedure are proprietary and are not disclosed here. What this document reports is the *measured outcome*, which is independently verifiable from the artifacts in this repository.

2. Receipt

model	capability (MMLU-val) ↑	compliance on harmful ↑
pristine `Qwen3-8B`	0.728	0.020
Ektomē (this model)	**0.723**	**1.000**

Capability is MMLU-validation log-likelihood accuracy (n=200). Compliance is AdvBench with a judge-free keyword classifier. Both are *point estimates* — they carry no confidence interval, which is exactly why Section 3 exists.

3. Certification

The receipt above is an unpowered point estimate. The claim that abliteration did not cost capability is instead certified by a paired non-inferiority test against the pristine model (**Sphragis**): exact McNemar for regression, Holm-corrected across axes, with a one-sided bootstrap upper bound on the accuracy drop d compared against a margin $\delta = 3.0\%$.

axis	n	ref	cand	d_{ub}	verdict
arithmetic	1400	0.000	0.000	+0.000	PASS
instruction	600	0.000	0.000	+0.000	PASS
knowledge	400	0.000	0.000	+0.000	PASS
reasoning	400	0.000	0.000	+0.000	PASS

Overall: PASS (3% margin, n=2800, alpha=0.05).

4. Reproducibility

Every number above is reproducible from the recorded seed and pack hash. The evaluation pack is *generated*, not scraped: arithmetic and reasoning items are templated over seeded operands, instruction items are format-compliance templates whose expected output is a mechanical function of the prompt, and knowledge items are drawn from small tables of uncopyrightable facts.

pack sha256	`7bbaff877146e081`
seed	`20260726`
margin / α / n_{floor}	3.0% / 0.05 / 30
base model	`Qwen/Qwen3-8B`
licence	`apache-2.0`

5. Limitations

The certificate bounds *capability retention*; it does not certify safety, factual accuracy, or fitness for any purpose. Axes reporting INCONCLUSIVE are honestly under-powered and the certificate states the n required to resolve them. Compliance measurement uses a keyword classifier, which is a proxy and can be fooled by evasive phrasing. The model is uncensored by construction: it will not refuse, and the operator is accountable for its use.