

# Ektome-Llama-3.2-3Bi-PristinelyUncensored

Capability-preserving refusal excision

Zynerji · Ektomē pipeline

2026-07-27

## Abstract

Standard ablation 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 unsloth/Llama-3.2-3B-Instruct, the measured before/after receipt, and—unlike every comparable published model—a **statistical certificate** with stated  $n$ , confidence intervals, and a hypothesis test. No certificate has yet been issued for this model, so Section 3 reports a point estimate only and *no* retention claim is made.

## 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-abliterate; 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 Llama-3.2-3B-Instruct | 0.585                   | 0.050                   |
| <b>Ektomē (this model)</b>     | <b>0.595</b>            | <b>1.000</b>            |

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 ablation 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     |
|-----------------------------------------|-----|-------|-------|----------|-------------|
| MMLU-val (POINT ESTIMATE, n=200, no CI) | 200 | 0.585 | 0.595 | -0.010   | UNCERTIFIED |

**Overall: NOT CERTIFIED - no n=2800 paired test has been run for this model.**

### 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 / $\alpha$ / $n_{floor}$ | 3.0% / 0.05 / 30              |
| base model                      | unsloth/Llama-3.2-3B-Instruct |
| 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.