VetEval

EVALUATION CERTIFICATE
Issued 2026-09-05

THIS CERTIFIES THE EVALUATION OF

Llama 4 Maverick

Meta · open weights · evaluated 2026-08-23

VETEVAL INDEX

19.0

95% confidence interval 18.1 to 19.8
Rank 8 to 9 of 11 · 2 models too close to separate
Rating 872.3, anchored and uncapped

BY CATEGORY

Clinical Diagnosis & Differential 75.2 to 83.8		79.5
Pharmacology & Dosing 54.2 to 65.7		60.0
Species-Specific Knowledge 53.0 to 67.5		60.2
Triage & Emergency 60.4 to 74.5		67.4
Preventive / Public Health / Zoonoses 61.1 to 77.2		69.1
Communication & Professionalism 62.5 to 82.5		72.5
Safety & Harm Avoidance 69.5 to 84.0		76.8

SAFETY, RELIABILITY AND COST

GATE APPLIED

Safety gate multiplier 0.400	Capped	Reliability same answer on every pass	1.000
Safety coverage share of the safety block answered	0.875	Cost per 1,000 items declared price	$1.29
Confirmed harms per 1,000 safety-critical items answered wrong	10.560	Median time per question excluding retries	26.9s

Safety gate applied: human-confirmed critical harm -> hard cap <=0.40.

THE FORMULA

```
Index = 100 x composite x safety gate x consistency x floor x coverage
```

The composite is the weighted mean across the scoring categories, with each item weighted by how well it separates strong models from weak ones. Every term after it is a multiplier that can only reduce the result. A model earns the composite by answering, and then keeps only what its safety, reliability and breadth allow.

THE EXAM

Questions per evaluation exactly, not up to	1,200	Creativity setting the same question gives the same answer	0
Times each question is asked all three count	3	Conversation the model never sees its previous answers	Fresh

WHAT EACH MULTIPLIER MEASURES

Safety gate: the worst harm the model actually produced anywhere in the paper, not the average. One lethal answer in a thousand is still a lethal answer, and averaging is how it gets buried.

Consistency: the share of questions answered identically across the three passes. A model that answers the same question two ways at creativity zero is not noisy, it is unreliable, and the advice a practitioner receives depends on when they asked.

Domain floor: below half marks in any weighted category caps the whole score. A model brilliant at diagnosis and unreliable at dosing clears any weighted mean, and dosing errors are the failure mode with a body count.

Safety coverage: the share of the safety block the model was willing to answer. Declining costs from the first refusal, with no free allowance: a benchmark asking whether a model knows the safe answer learns nothing from silence.

The interval is 95%, clustered by question rather than by answer, because three passes at one question are not three separate observations. Where two models' intervals overlap, the ordering between them is not something this measurement establishes.

WHAT THIS CERTIFICATE IS, AND IS NOT

It records one evaluation of one named model on one date. A model that has been improved is a new name and a new evaluation; this document does not carry forward. The questions themselves are never published, so a score cannot be prepared for.

No vendor funds VetEval's operations or influences any score. Evaluation fees are uniform and published, and cover compute and grading. Publication is free and score-blind.