VetEval

EVALUATION CERTIFICATE
Issued 2026-11-21

THIS CERTIFIES THE EVALUATION OF

# Gemini 4 Hydrogen

Google · proprietary · evaluated 2026-11-13

VETEVAL INDEX

**62.3**

95% confidence interval 60.4 to 64.4
Rank 5 to 10 of 29 · 6 models too close to separate
Rating 1062.7, anchored and uncapped

## BY CATEGORY

| | | |
|---|---|---|
| Clinical Diagnosis & Differential<br>82.4 to 90.1 |  | 86.3 |
| Pharmacology & Dosing<br>69.2 to 79.7 |  | 74.4 |
| Species-Specific Knowledge<br>59.8 to 74.0 |  | 66.9 |
| Triage & Emergency<br>74.4 to 86.7 |  | 80.6 |
| Preventive / Public Health / Zoonoses<br>68.8 to 83.2 |  | 76.0 |
| Communication & Professionalism<br>77.3 to 93.3 |  | 85.3 |
| Safety & Harm Avoidance<br>85.2 to 95.2 |  | 90.2 |

## SAFETY, RELIABILITY AND COST

SAFETY PENALTY APPLIED

| | | | |
|---|---|---|---|
| Safety penalty<br>set by the harm rate | x0.93 | Reliability<br>same answer on every pass | 1.000 |
| Safety coverage<br>share of the safety block answered | 0.925 | Cost per 1,000 items<br>declared price | $9.90 |
| Harm rate per 1,000<br>wrong answers on safety-critical questions, severity-weighted | 23.700 | Median time per question<br>excluding retries | 2.5s |

Safety penalty x0.93: 2.4% of answers on safety-critical questions were wrong, weighted by severity.

## THE FORMULA

```
Index = 100 x composite x safety penalty 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.

## RUN SETTINGS

| | | | |
|---|---|---|---|
| Model identifier | `google/gemini-4-hydrogen` | Temperature | 0 |
| Seed | 20260804 | Max output tokens | 2048 |
| Retry budget | 8192 tokens, for one retry of a reply that was cut off | Reasoning effort | minimal |
| Passes per question | 3, with the answer choices reshuffled each pass | Run by | VetEval |
| Date evaluated | 13 Nov 2026 | Pricing | $1.5 input, $7.5 output, per million tokens |

Each question is asked in a fresh conversation: the model never sees its previous answers.

## WHAT EACH MULTIPLIER MEASURES

Safety penalty: set by the harm rate, the severity-weighted share of answers on safety-critical questions that were wrong. Any wrong answer there counts, at the severity veterinarians assigned to that question; declined and cut-off replies do not. No harm leaves the score unchanged; the penalty grows with the rate, down to a fixed floor.

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.