Inside OpenVet's Clinical Evaluation Framework for Veterinary AI

A detailed look at the benchmarking methodology used to measure clinical accuracy in the OpenVet AI hospital.


Adam Sager, Founder and CEO, OpenVet

MIAMI, FL, UNITED STATES, April 29, 2026 /EINPresswire.com/ -- FOR IMMEDIATE RELEASE

OpenVet today released a detailed explanation of the methodology it uses to evaluate clinical accuracy in veterinary artificial intelligence. The document describes the system's evaluation framework and the results of its foundational knowledge benchmark.

Reliable clinical AI requires evaluation methods that are as rigorous as the medical decisions the system is expected to support. Many AI systems appear convincing in conversation but fail under structured evaluation. OpenVet's benchmarking program was designed to measure whether an AI system meets the reliability threshold required for an AI hospital. In addition, OpenVet is issuing this statement to clarify certain aspects of a prior related press release.

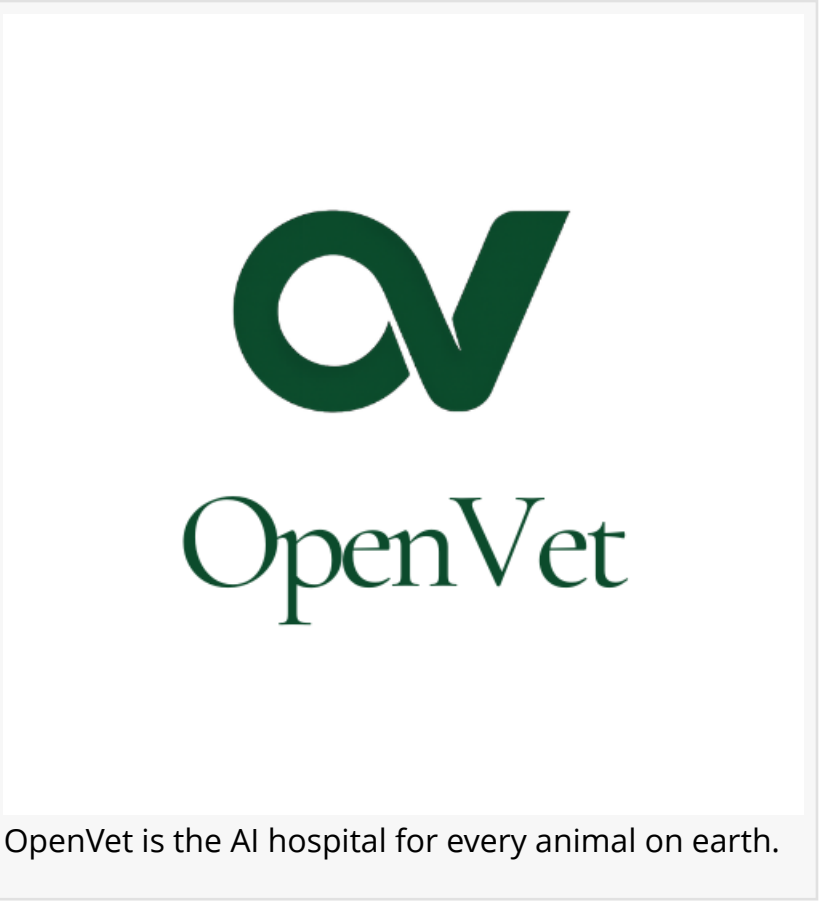
"In medicine, just providing answers is not enough," said Adam Sager, founder of OpenVet. "Clinical systems have to be measured rigorously before they can be trusted in real clinical environments. This benchmark is one part of how we approach that responsibility."

What we are building

OpenVet is the AI hospital for every animal on earth. Every case that passes through it becomes a permanent biological data point.

Every clinical interaction processed by the system becomes structured medical data. Because of this, accuracy is not simply a feature of the system. It is the foundation of the system. A wrong drug dose or a missed contraindication does not simply produce an incorrect answer. It produces incorrect clinical data that can influence future decisions.

For this reason, OpenVet evaluates accuracy continuously using multiple independent validation methods.



OpenVet is the AI hospital for every animal on earth.

Why veterinary medicine presents a harder problem than human medicine

In human medicine, most drugs are approved for a defined patient population supported by large clinical trials. Veterinary medicine often works differently.

As one example, under the Animal Medicinal Drug Use Clarification Act of 1994, veterinarians are permitted to prescribe drugs in an extralabel manner when no approved alternative exists. In practice, many drugs used in veterinary medicine were originally developed for human use or for another species.

Drug metabolism also differs significantly across species. Cats, for example, lack the hepatic glucuronidation pathway required to metabolize acetaminophen safely. A dose tolerated by a dog can be fatal in a cat. Differences in metabolism, toxicology, and therapeutic windows exist across species and sometimes across breeds.

A clinical system that does not reason at the level of species specific biology cannot be considered a safe decision support tool.

Our layered evaluation framework

OpenVet evaluates the system using four independent benchmarking layers. Each layer measures a different dimension of clinical performance.

Layer 1: Knowledge floor evaluation

Does the system correctly answer general clinical questions across species and veterinary domains?

Layer 2: Retrieval validation
When the system cites a source, is the source real, relevant, and does it actually support the answer?

Layer 3: Expert panel review
Do practicing veterinarians judge the answers to be clinically correct, appropriately qualified, and safe? This approach is consistent with published clinical AI evaluation methodologies in which independent physician panels score AI generated clinical responses across multiple domains of care.

Layer 4: Golden dataset evaluation
Does the system perform consistently over time against curated cases with known correct answers?

Passing one layer does not guarantee success at the next layer. The goal of the framework is to measure different types of clinical risk. This document focuses specifically on the methodology used for Layer 1, the foundational knowledge benchmark.

Why this matters

Many AI systems can appear convincing in conversation but fail under structured evaluation. In medicine, that difference matters. In veterinary medicine, where species, metabolism, dosing, and contraindications can shift case by case, the standard must be higher.

An AI hospital only becomes possible if the underlying system is measured rigorously, continuously, and across multiple layers of clinical risk. OpenVet's evaluation framework was designed to establish that level of reliability.

Clarification Regarding Prior OpenVet Press Release

To evaluate baseline clinical knowledge, OpenVet created an independent benchmarking dataset designed to reflect the breadth of veterinary medicine across species and domains. The benchmark was designed to reflect the scope and difficulty of veterinary licensing examinations used to define minimum professional competency. The benchmark was constructed independently of proprietary licensing examinations and without access to proprietary examination materials. Rather than using any proprietary examination content, OpenVet generated comparable benchmark questions derived from publicly available veterinary literature and academic sources. OpenVet did not take the NAVLE examination, did not take any NAVLE self assessments, and did not achieve a score on the NAVLE. Any prior language suggesting otherwise was imprecise and does not accurately reflect OpenVet's activities. OpenVet also

confirms that it did not access, use, rely upon, or incorporate any materials owned, administered, or distributed by the International Council for Veterinary Assessment, including NAVLE questions, NAVLE self assessments, or ICVA preparation materials, in developing, training, testing, or benchmarking its AI system.

OpenVet's benchmarking involved the use of publicly available academic and clinical sources and original AI-generated questions designed to assess general veterinary knowledge. OpenVet respects ICVA's role in safeguarding the integrity of veterinary licensure and issues this clarification to ensure accuracy and transparency in the marketplace.

Question generation and validation

An automated pipeline produced candidate multiple choice questions derived from publicly available veterinary literature.

Each candidate question underwent a validation stage before it could be included in the benchmark dataset.

Questions were required to meet four independent criteria:

• clinical correctness of the answer
• correct species context
• traceability to published veterinary evidence
• appropriate domain classification

If a candidate question failed any criterion, it was rejected and regenerated before evaluation.

This validation step ensured that the benchmark dataset itself met clinical and methodological quality standards before the system was tested against it.

This filtering stage is separate from the accuracy scoring stage described below.

Benchmark administration

The final benchmark consisted of 600 questions administered in three sequential tests of 200 questions each.

The dataset was stratified across species and domains including companion animals, livestock, and exotic species. Some questions contained visual clinical material such as radiographs.

To reduce the risk of memorization, new question sets were generated for each evaluation cycle. Questions used during development were not reused during final testing.
Accuracy scoring was based solely on whether the system selected the correct multiple choice

answer.

The validation criteria described earlier were applied to dataset construction and were not used as part of the scoring rule.

Iterative improvement process
The benchmark and evaluation process evolved through several development cycles as errors were identified and corrected.

Errors observed during testing were categorized into four types:

• retrieval failures where correct evidence existed but was not retrieved
• species context errors where evidence from the wrong species was used
• reasoning failures where correct evidence was retrieved but incorrectly synthesized
• knowledge gaps where relevant literature was absent from the corpus

Each failure type required different technical improvements including search architecture adjustments, species context enforcement, reasoning improvements, and expansion of the knowledge corpus.

After each development cycle, new benchmark questions were generated so the system could not learn or memorize specific questions.

Benchmark result

On the final evaluation, the system answered all 600 questions correctly (on OpenVet's independently constructed benchmark, not on the NAVLE examination). All questions were answered correctly on the first attempt during the final evaluation run.

Observed accuracy was therefore 600 correct answers out of 600 questions. Because any benchmark measures performance on a finite sample of questions, the result is reported with a statistical confidence interval.

Using an exact binomial method, the 95 percent confidence interval for the system's true accuracy is estimated to be between approximately 99.5 percent and 100 percent. This means that if the evaluation were repeated many times using comparable question sets, the true probability of answering a question correctly would be expected to be at least about 99.5 percent with 95 percent confidence.

What this result means

The benchmark establishes that the system can reliably answer general veterinary knowledge questions at the level expected of baseline professional competency.

This benchmark measures foundational knowledge. It does not by itself measure the full range of clinical reasoning, case management, or real world judgment required in practice. Benchmark performance alone is not sufficient for clinical deployment. OpenVet is designed as a decision support system used alongside veterinary judgment, not as a replacement for a licensed veterinarian. Establishing a reliable knowledge floor is one step in building a system that can be trusted in more complex clinical settings.

What comes next

Veterinary medicine contains a large body of knowledge that is not fully captured in published literature. Clinical judgment often depends on experience, context, and incomplete information.

OpenVet's architecture is designed to address this challenge. Every answer generated by the system includes explicit evidence sources and confidence signals. When evidence is limited or uncertain, the system is designed to surface that uncertainty rather than conceal it.

Building an AI hospital requires more than generating convincing answers. Clinical systems must be measured, audited, and evaluated under conditions that reflect real medical risk.

OpenVet's evaluation framework was designed to establish that level of reliability before AI systems participate in veterinary clinical care.

Future releases will describe the methodology behind OpenVet's additional evaluation layers, including retrieval validation, expert review, and curated case testing.

Reference

Hurt RT et al.
The Use of an Artificial Intelligence Platform OpenEvidence to Augment Clinical Decision Making for Primary Care Physicians.
Journal of Primary Care and Community Health. 2025;16. PMC12033599.

About OpenVet

OpenVet is the AI hospital for every animal on earth. Every case that passes through it becomes a permanent biological data point. The system gives every veterinarian instant access to cited, species aware clinical intelligence at the point of care for any animal.

Media Contact
press@openvet.ai
openvet.ai

Team OpenVet
OpenVet.AI
email us here
Visit us on social media:
LinkedIn
Facebook
X

This press release can be viewed online at: https://www.einpresswire.com/article/909049681

EIN Presswire's priority is source transparency. We do not allow opaque clients, and our editors try to be careful about weeding out false and misleading content. As a user, if you see something we have missed, please do bring it to our attention. Your help is welcome. EIN Presswire, Everyone's Internet News Presswire™, tries to define some of the boundaries that are reasonable in today's world. Please see our Editorial Guidelines for more information.
© 1995-2026 Newsmatics Inc. All Right Reserved.