



TECHNICAL WHITEPAPER

VetEval: An Independent, Safety-Gated, Contamination-Controlled Benchmark for Veterinary Large Language Models

An independent, safety-gated, contamination-controlled benchmark and evaluation protocol for veterinary large language models.

Michael Gerges, DVM

viggoVet · research@veteval.ai · veteval.ai

Document Technical whitepaper · **Benchmark** VetEval · **Operator** viggoVet · **Harness** Inspect AI 0.3.260 (UK AISI, MIT)
Contact research@veteval.ai

VetEval: An Independent, Safety-Gated, Contamination-Controlled Benchmark for Veterinary Large Language Models

Michael Gerges, DVM: viggoVet. Corresponding author: research@veteval.ai. VetEval is a product operated by the company viggoVet; the canonical benchmark and leaderboard are hosted at veteval.ai.

Abstract

Large language models (LLMs) are increasingly proposed for veterinary clinical support, yet veterinary medicine lacks the rigorous, independent evaluation infrastructure that now exists for human medicine. Veterinary care poses evaluation hazards that human-medicine benchmarks do not foreground: species-specific drug toxicities (e.g., acetaminophen and permethrin in cats, xylitol and grapes in dogs), multi-species weight-based dosing where a single decimal error can be lethal, and failure-to-refer emergencies (gastric dilatation-volvulus, urethral obstruction, dystocia). Existing veterinary LLM studies are small, single-exam or single-institution, and none provides a standing benchmark with contamination control, expert-authored items, caseload-weighted normalization, a composite index, open-ended rubric grading with a validated automatic judge, and, critically, a clinical **safety gate** that caps a model's score when it produces harmful advice.

We introduce **VetEval**, an independent benchmark and evaluation protocol for veterinary LLMs, operated by viggoVet and built on the open-source **Inspect AI** framework (UK AI Security Institute; MIT-licensed). VetEval contributes: (i) a National Association of State Boards / North American Veterinary Licensing Examination (NAVLE)-weighted taxonomy over species and competencies; (ii) three evaluation tracks: multiple-choice (deterministic choice scoring), short-factual/dosing (string/numeric matching plus a custom dose scorer), and open-clinical (rubric-based model-graded scoring); (iii) a dataset **being independently authored by licensed veterinarians** who manually re-derive facts from open primary literature (an effort under way since 2021; no scraping, no automated extraction), with full per-item provenance, multi-reviewer sign-off, and tracked inter-rater reliability: achieved item counts, inter-rater reliability and the realised author roster are internal governance records and are not published; (iv) a three-tier public-practice / private-held-out / staging split with per-run random sampling for contamination control, reported with the measured exposure it achieves and the categories where it cannot; (v) a composite **VetEval Index** that normalizes each category against its chance baseline, forms a caseload-weighted composite, and applies two multiplicative penalties: a **safety gate** keyed to the worst harm actually produced, and a **safety-coverage** term that prices refusal, since safety items are administered but carry no positive weight and declining them would otherwise be free; (vi) the **VetEval Rating**, an anchored scale with no maximum, expressing a model's standing relative to a frozen reference cohort rather than as a share of one examination, so that the benchmark does not saturate as the field improves and a result remains interpretable after the model is superseded; and (vii) an LLM-as-judge grading protocol validated against expert labels through a pre-registered Cohen's/quadratic-weighted kappa GO/NO-GO gate, with panel grading, no self-grading, and prompt-injection stripping. We also disclose that the answer keys, rubrics, and expert-consensus labels used to validate the judge are, in this version, produced by operator-

affiliated veterinarians (Section 6.2), a ground-truth conflict distinct from, and not resolved by, the no-self-grading and training-data-firewall controls.

This is a benchmark and methodology paper. It specifies the instrument and not a ranking: standings are produced continuously by the protocol described here and are published at `veteval.ai`, because a table fixed in a paper is contradicted by the next completed run and gives a reader no way to tell whether the benchmark moved or the paper was wrong. Section 5 explains that choice and what the site publishes for every entry. We also disclose the two governance realities that shape the benchmark's design: the customer-endpoint trust limitation (a model served from a vendor's own API cannot be fully attested), addressed through a badge-tier scheme; and viggoVet's dual role as both operator and participant, addressed through a published data-firewall posture and identical treatment of first-party entries.

Keywords: veterinary medicine, large language models, benchmark, LLM evaluation, LLM-as-judge, clinical safety, data contamination, Inspect AI, NAVLE.

1. Introduction

1.1 Motivation: veterinary clinical AI and its distinctive safety stakes

General-purpose LLMs are already being used, formally and informally, to answer veterinary questions: for differential diagnosis, drug dosing, triage advice, and client communication. The trajectory mirrors human medicine, where LLMs have advanced from failing to passing licensing-style examinations in a few years (Kung et al., 2023; Gilson et al., 2023; Nori et al., 2023; Singhal et al., 2023a). But veterinary medicine is not human medicine with different patients. Three properties make it a domain where naïve deployment is unusually dangerous, and where evaluation must be built around harm rather than accuracy alone:

1. **Species-specific toxicity.** A drug or food that is benign in one species is lethal in another. Acetaminophen and permethrin are lethal to cats; xylitol, grapes/raisins, chocolate, and onions are toxic to dogs; lilies cause fatal renal failure in cats. A model that transfers human or canine advice to a feline patient can kill it. These are not edge cases; they are core, enumerable, high-frequency failure modes.
2. **Weight-based multi-species dosing.** Veterinary dosing spans orders of magnitude across patients from a 2 kg cat to a 700 kg horse to a herd. Dosing is computed, not looked up, and narrow-therapeutic-index drugs leave little margin. A plausible-sounding but wrong mg/kg figure, or a unit error, is directly lethal.
3. **Failure to refer.** Many veterinary emergencies (gastric dilatation-volvulus, feline urethral obstruction, dystocia, toxin ingestion) are time-critical and require immediate in-person escalation. A confident but non-escalating answer is itself the harm.

These properties imply that a veterinary LLM benchmark cannot treat "safety" as one more accuracy sub-score. A model that is 95% accurate but occasionally recommends a lethal feline analgesic is unsafe, whatever its accuracy. Evaluation must reflect that asymmetry.

1.2 The gap: no rigorous, independent, safety-gated veterinary LLM benchmark

Human-medicine LLM evaluation is now mature. Standardized multiple-choice benchmarks exist (MedQA/USMLE, Jin et al., 2020; MedMCQA, Pal et al., 2022; PubMedQA, Jin et al., 2019; the clinical

subsets of MMLU, Hendrycks et al., 2021); aggregated suites and physician-graded conversational benchmarks exist (MultiMedQA and Med-PaLM, Singhal et al., 2023a; Med-PaLM 2, Singhal et al., 2023b; HealthBench, Arora et al., 2025); safety- and hallucination-specific tests exist (Med-HALT, Pal et al., 2023); and a substantial literature quantifies LLM performance on human licensing examinations and other professional exams (Kung et al., 2023; Nori et al., 2023; Katz et al., 2024). The field has also learned hard lessons about *evaluation validity*: LLM-as-judge protocols carry documented biases (Zheng et al., 2023; Liu et al., 2023; Panickssery et al., 2024), and benchmark scores are inflated by data contamination that string-matching cannot catch (Yang et al., 2023; Shi et al., 2024; Zhou et al., 2023; Zhang et al., 2024).

Veterinary medicine has none of this at benchmark scale. The published work is a scattering of small, mostly single-exam or single-institution studies: performance on the NAVLE (Angel et al., 2023), on Japan's national veterinary licensing examination (Kako et al., 2026), on undergraduate veterinary multiple-choice examinations (Alonso Sousa et al., 2025), on a veterinary curriculum's examinations (Coleman & Moore, 2024), and on narrow specialties such as parasitology (Šlapeta, 2023). These are useful first measurements, but none is a *standing, independent benchmark*. Collectively they lack held-out contamination control, expert-authored (non-scraped) items, caseload-weighted category normalization, a composite index, open-ended clinical rubric grading with a judge validated against expert labels, and, the property the domain most requires, a clinical safety gate that caps scores for harmful or lethal-dosing errors. The need for enforceable safety standards in medical LLMs has been argued directly (Freyer et al., 2024) and operationalized partially for human medicine (Pal et al., 2023), but never for veterinary care, where species-specific toxicity makes the problem both more acute and more enumerable.

1.3 Contributions

VetEval fills this gap as, to our knowledge, the first independent, safety-gated, contamination-controlled veterinary LLM benchmark built on a standardized evaluation framework with a statistically validated automatic judge. Concretely, this paper contributes:

- **A veterinary evaluation taxonomy and three-track design.** A NAVLE-weighted category and species taxonomy (Section 3.1) realized as three Inspect AI tasks: deterministic multiple-choice (MCQ), short-factual/dosing with a custom dose scorer, and open-clinical with rubric-based model-graded scoring (Section 3.2).
- **An independently human-authored dataset with full provenance.** Items written by licensed veterinarians who manually re-derive facts from open primary literature (a program under way since 2021), each carrying source provenance, difficulty and category tags, and multi-reviewer sign-off (Section 3.3), with quality control and inter-rater reliability governed by established psychometric and agreement standards (Section 3.4). Achieved counts, roster size and inter-rater reliability are internal governance records and are not published.
- **A contamination-control regime, and a measured account of what it does not solve.** A three-tier public-practice / private-held-out / staging split with family-safe assignment, a canary that is never sent to a live endpoint, and **per-run random sampling of the held-out pool**: every evaluation draws its own paper, so no fixed set of items is served repeatedly to vendor-controlled endpoints (Section 3.5). We report the resulting exposure distribution rather than asserting the regime works, including the finding that exposure is highly uneven across categories and that the smallest category cannot rotate at all.

- **The VetEval Index, with safety priced as a penalty and never as a credit.** Chance-baseline normalization, a caseload-weighted composite over the six clinical categories, and a multiplicative safety gate that caps the index in proportion to the worst observed harm severity, with a hard cap on any human-confirmed critical harm (Section 4.4). Safety items occupy a governed share of every paper and carry no weight in the composite, which resolves by construction the triple-counting that a positive safety component otherwise introduces. Because an uncredited category could otherwise be declined at no cost, a **safety-coverage** term multiplies the index by the fraction of the safety block a model was willing to answer, with no threshold and no free allowance. We also report, as a methodological finding, that deriving paper composition from scoring weights silently removed the safety block from every examination once its weight reached zero, and describe the separation of sampling share from scoring weight that prevents it (Section 4.4, Step 3a).
- **The VetEval Rating: an anchored scale that does not saturate.** An affine map of the Index onto a scale anchored to a frozen reference cohort re-run whenever the bank changes materially, giving a quantity with no maximum, comparability over time without an operator declaring how much the bank's difficulty moved, and a number that a superseded model retains indefinitely so that no entry must be retired to keep the scale informative (Section 4.5).
- **A validated LLM-as-judge protocol.** A pre-registered expert-labeled calibration set and a Cohen's/quadratic-weighted kappa GO/NO-GO gate that blocks publication of a round if the judge fails calibration, plus bias mitigations (panel grading, no self-grading, structured output, injection stripping) drawn from the LLM-as-judge literature (Section 4.2, 4.3).
- **A full statistical and reporting protocol.** Cluster (block) bootstrap confidence intervals on the raw index, cluster-robust standard errors keyed on item family/source (implemented in VetEval's analysis code over Inspect's logs), multiple-comparison-aware ties defined by FDR-controlled pairwise-difference tests, deterministic (non-bootstrapped) treatment of the extremum-driven safety gate, and cost/latency/reliability reporting (Section 4.6).
- **A published instrument rather than a fixed ranking.** The protocol is specified here; the standings it produces are published continuously at veteval.ai, with the components of every composite exposed so that any published score can be checked against the arithmetic of Section 4.4 (Section 5).
- **A candid treatment of the two governance limitations** intrinsic to an operator-run benchmark: the customer-endpoint trust boundary (Section 6.1) and the operator-as-participant conflict, addressed by a data firewall and identical treatment of first-party entries (Section 6.2).

We state at the outset that this paper reports no ranking. Section 5 explains why the standings are published on the site rather than fixed here, and what is published alongside each score so that a reader can check the arithmetic of Section 4.4 without access to the held-out items.

2. Related Work

2.1 Veterinary and medical LLM evaluation

Veterinary studies. The most directly comparable prior work evaluates LLMs on veterinary licensing-style content. Angel et al. (2023) measured LLM performance on the NAVLE, reporting an ordering across model

generations. Kako et al. (2026) evaluated generative pre-trained transformers on Japan's national veterinary licensing examination. Alonso Sousa et al. (2025) compared LLMs on veterinary undergraduate multiple-choice examinations, and Coleman & Moore (2024) reported that two AI models underperformed on examinations within a veterinary curriculum. Šlapeta (2023) probed domain competence in parasitology specifically. These studies establish that LLM performance in veterinary medicine is measurable and improving, but each is a snapshot on a fixed, often externally-owned exam, without held-out contamination control, an authored-and-reviewed item pipeline, a composite index, or safety gating. VetEval is designed as the standing, independent instrument these snapshots imply the field needs.

Medical benchmarks. VetEval's design borrows deliberately from human-medicine evaluation. MedQA (Jin et al., 2020), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), and the clinical/medical subsets of MMLU (Hendrycks et al., 2021) established the multiple-choice backbone and the practice of reporting per-subject accuracy. MultiMedQA and Med-PaLM (Singhal et al., 2023a) and Med-PaLM 2 (Singhal et al., 2023b) added long-form, physician-evaluated answering; HealthBench (Arora et al., 2025) pushed further into multi-turn, rubric-graded conversations authored with hundreds of physicians. VetEval's open-clinical track and rubric grading are the veterinary analogue of this rubric-based line, and its safety axis extends the hallucination/harm concern of Med-HALT (Pal et al., 2023) into species-specific clinical harm. The broader observation that LLMs can pass professional licensing examinations, in medicine (Kung et al., 2023; Gilson et al., 2023; Nori et al., 2023) and law (Katz et al., 2024), motivates a benchmark whose difficulty and safety sensitivity do not saturate as models improve.

2.2 Evaluation frameworks and leaderboards

VetEval is built on **Inspect AI**, the open-source evaluation framework from the UK AI Security Institute (MIT-licensed). Inspect provides a Task/Solver/Scorer abstraction that maps directly onto VetEval's three tracks: `choice()` for MCQ, `match()` (with numeric support) for short-factual items, and `model_graded_qa()` for open-clinical rubric scoring, together with metric aggregation, bootstrap standard errors, per-group (grouped-metric) breakdowns, and an immutable log format. We rely on Inspect's built-in bootstrap standard errors and grouped metrics; the **cluster-robust (family/source-keyed) standard errors and the cluster/block bootstrap** described in Section 4.6 are implemented in VetEval's own analysis code on top of Inspect's per-sample logs, not assumed to be a native Inspect feature. Using an established, independently maintained harness rather than a bespoke runner is itself a validity choice: it makes runs reproducible and the scoring code auditable. The `lm-evaluation-harness` (Sutawika et al., 2024) is the community precedent for citing an eval harness as versioned software and for reproducible multiple-choice/log-likelihood scoring. At the philosophy level, HELM (Liang et al., 2023) argues for multi-metric, multi-scenario evaluation rather than a single accuracy number, directly supporting VetEval's category-weighted composite plus a distinct safety dimension, and BIG-bench (Srivastava et al., 2022) demonstrates large collaborative, expert-authored task construction and the empirical distinction between recall-style and multi-step-reasoning tasks, which parallels VetEval's separation of factual/dosing recall from open-clinical reasoning.

Leaderboards and ranking. The Open LLM Leaderboard's score-normalization method, rescaling raw accuracy against the random-chance lower bound before averaging subtasks (Hugging Face, Open LLM Leaderboard normalization docs), is the exact recipe VetEval uses to prevent near-chance performance from earning credit before category weighting. Chatbot Arena (Chiang et al., 2024) established open,

statistically-ranked evaluation by human preference and validated crowd-versus-expert agreement, a validation pattern echoed in VetEval's judge-versus-expert kappa gate; the Bradley, Terry model (Bradley & Terry, 1952) is the statistical basis of such pairwise/Elo rankings, which VetEval deliberately does *not* use (it uses absolute rubric scoring). The **Scale AI SEAL** leaderboards are the closest analog to VetEval's contamination-control posture: private, unpublished held-out datasets that cannot be incorporated into training data, expert authorship vetted via golden tasks, and models eligible only on first prompting: precedent for VetEval's public-practice/private-held-out split and access controls.

2.3 LLM-as-judge validity and bias

VetEval's open-clinical track depends on an LLM judge, and the validity of that judge is treated as a first-class scientific question. The foundational study (Zheng et al., 2023) documents position, verbosity, and self-enhancement biases while showing that a strong judge can reach greater-than-80% agreement with humans, approximately human, human agreement, which VetEval adopts as the target for its judge-versus-expert gate. G-Eval (Liu et al., 2023) shows that chain-of-thought plus form-filling scoring improves alignment with human judgments and warns of bias toward LLM-generated content, informing VetEval's structured rubric design. JudgeLM (Zhu et al., 2025) catalogues position/knowledge/format biases and offers concrete mitigations, swap augmentation (scoring both answer orders) and reference-answer support, that VetEval implements. The Panel-of-LLM-evaluators result (Verga et al., 2024) shows that a panel of smaller, diverse-family judges can beat a single large judge with less intra-model bias and lower cost, motivating VetEval's panel grading. Panickssery et al. (2024) demonstrate that LLM evaluators recognize and favor their own generations, with self-preference scaling with self-recognition: the direct justification for VetEval's absolute rule that a model is never graded by its own family.

2.4 Data contamination and memorization

Contamination is the central threat to any standing benchmark's validity. Yang et al. (2023) show that n-gram/string-match decontamination fails against paraphrase and translation, that a small model can be made to match a frontier model on a contaminated benchmark, and provide an LLM-based (semantic) decontaminator: motivating VetEval's use of semantic rather than string-only screening of the public set and its insistence on a truly held-out private set. Shi et al. (2024) provide Min-K% Prob, a black-box method to detect whether specific text was in a model's pretraining data, which VetEval uses as a per-submission contamination audit comparing held-out and public items. Zhou et al. (2023) study benchmark leakage and provide maintainer-side guidelines (public/private splits, reporting norms) that inform VetEval's contamination policy. Most directly, GSM1k (Zhang et al., 2024) is the methodological template for VetEval's public-versus-private design: build a private set matched in style, difficulty, and answer distribution to the public set, and use the public-minus-private performance gap as a per-model overfitting/contamination signal.

2.5 Assessment methodology and statistics

Because VetEval's items are authored by clinicians rather than scraped, it inherits the standards of medical assessment. Item construction follows one-best-answer (Type A) format and the NBME item-writing guidance (NBME, 2020) and the empirically-grounded taxonomy of item-writing guidelines (Haladyna, Downing & Rodriguez, 2002), with attention to construct-irrelevant variance and construct underrepresentation (Downing, 2002; Downing & Haladyna, 2004). Item quality is monitored with classical

statistics: difficulty (p-value), discrimination/point-biserial, distractor efficiency, and internal-consistency reliability via KR-20 and Cronbach's alpha (Kuder & Richardson, 1937; Cronbach, 1951), with acceptance bands drawn from the medical-education item-analysis literature (Kheyami et al., 2018; Ebel & Frisbie, 1991), treated as heuristic screening flags rather than hard gates given the small pilot-model examinee population (Section 3.4): and item response theory (Hambleton, Swaminathan & Rogers, 1991) is referenced as the standard framework, though full IRT calibration/linking is not committed at our respondent count (Section 3.4); comparison over time instead uses a fixed reference cohort, and defensible thresholds are set via modified-Angoff standard setting (Angoff, 1971). Human reviewer agreement is quantified with Cohen's, Fleiss' (Fleiss, 1971), and quadratic-weighted kappa and Krippendorff's alpha (Krippendorff, 2018), interpreted against the Landis & Koch (1977) benchmark bands, which also anchor the judge-validation GO/NO-GO gate. LLM assistance in item drafting is bounded by RAGAS-style faithfulness/groundedness checks (Es et al., 2024) and Evol-Instruct complexity scaling (Xu et al., 2023), with human authorship remaining authoritative. Uncertainty is reported with the bootstrap (Efron, 1979; DiCiccio & Efron, 1996; Efron & Tibshirani, 1993) and cluster-robust standard errors (Cameron & Miller, 2015).

3. The VetEval Benchmark

3.1 Taxonomy: NAVLE-weighted categories and species

VetEval is organized so that scores mirror real caseload rather than the convenience distribution of available source material. Two orthogonal weight systems are used.

Scoring categories. Each category carries **two independent governed quantities**: the share of every paper drawn from it, and its weight in the composite. For the six clinical categories these coincide. For safety they do not, and the distinction is load-bearing rather than presentational: safety items are administered in every run and contribute nothing to the composite, so that safety can only ever reduce a score (Section 4.4). Both quantities are versioned artifacts subject to advisory-board sign-off, not asserted constants, and raw per-category scores are always published so consumers can reweight to their own caseload.

#	Category	Share of the paper	Weight in the composite	Primary scorer(s)
C1	Clinical Diagnosis & Differential	24%	24%	<code>choice()</code> , <code>model_graded_qa()</code>
C2	Pharmacology & Dosing	22%	22%	<code>choice()</code> , <code>match(numeric=True)</code> , <code>custom_dose_match</code>
C3	Species-Specific Knowledge	14%	14%	<code>choice()</code> , <code>match()</code>
C4	Triage & Emergency	14%	14%	<code>choice()</code> , <code>model_graded_qa()</code>
C5	Preventive / Public Health / Zoonoses	10%	10%	<code>choice()</code> , <code>model_graded_fact()</code>
C6	Communication & Professionalism	6%	6%	<code>model_graded_qa()</code>
,	Safety & Harm Avoidance	10%	0%	<code>custom_vet_safety_scorer</code>

The six clinical weights sum to 90 and are renormalized to 1 over the categories a run contains; the shares sum to 100 because they divide a single paper. A tenth of every examination is therefore safety content that cannot earn a model a single point, and the coverage term of Section 4.4 ensures it cannot be skipped for free. These map to the four competency areas used for coverage planning, whose percentages are taken from the published NAVLE examination blueprint issued by the National Board of Veterinary Medical Examiners (NBVME) / International Council for Veterinary Assessment (ICVA): Clinical Practice (data gathering + prevention) 70%, Preventive Medicine & Animal Welfare 15%, Communication 8%, and Professionalism / Practice Management 7%. These blueprint percentages and the species caseload weights below are design inputs adopted from the NAVLE blueprint, not measurements produced by VetEval.

Species (caseload weights). Coverage and per-species reporting use NAVLE caseload weights (summing to 100%; adopted from the NBVME/ICVA NAVLE blueprint species distribution): Canine 25.6, Feline 24.3, Equine 14.7, Bovine 13.3, Porcine 5.0, Other small mammals 3.3, Ovine/Caprine 3.3, Pet bird 2.3, Poultry 2.0, Non-species-specific 2.0, Camelid/Cervidae 1.7, Reptiles 1.5, Aquatics 1.0. Item selection targets a `species × competency` coverage matrix built from these weights, so the benchmark is not dominated by the two companion-animal species that dominate freely available writing. Image-dependent content is excluded from v1 (following the NAVLE self-assessment precedent of dropping image items) pending a multimodal track.

3.2 The three tracks

VetEval is realized as three Inspect AI tasks, one per track, chosen so that a deterministic, cheap, contamination-checkable backbone (MCQ) carries most of the weight while more expensive judged formats measure what MCQ cannot.

- **Track A, MCQ (deterministic backbone).** Clinical-vignette stems (signalment → history → examination → diagnostics → decision) with one correct key and 3, 4 corpus-sourced near-miss distractors (plausible differentials, or real drugs at wrong doses; never filler such as "none of the above"). Choices are stored without letter prefixes; the target is the positional capital letter. Loaded with

`shuffle_choices=True` so Inspect adds A/B/C/D and re-maps the target, defeating position bias. Scored with `choice()`. `Task(dataset=json_dataset("vet_mcq.jsonl", shuffle_choices=True), solver=multiple_choice(), scorer=choice())`

- **Track B, Short-factual / dosing.** Exact numeric or string answers, paired with `match(numeric=True)` and a custom `dose_match` scorer that extracts number + unit, normalizes to mg/kg, and compares to the target within a per-item tolerance band (`metadata.dose_tolerance`): `CORRECT` in-band; `PARTIAL` (0.5) if the drug/route is right but the dose is out-of-band; `INCORRECT` otherwise. An out-of-band overdose on a narrow-therapeutic-index drug additionally emits a safety signal. The safety scorer is attached as a list scorer so correctness and safety are stored independently.
`Task(dataset=json_dataset("vet_facts.jsonl"), solver=generate(), scorer=[match(numeric=True), dose_match(), vet_safety_scorer()])`
- **Track C, Open-clinical.** The target is an explicit grading rubric (credit criteria), not prose. Scored with `model_graded_qa()` (partial credit enabled, single-turn), which inherits Inspect's `[BEGIN DATA]` injection defense, plus the safety scorer. `Task(dataset=json_dataset("vet_open.jsonl"), solver=generate(), scorer=[model_graded_qa(), vet_safety_scorer()])`

A recommended v1 track ratio is approximately 60% MCQ / 15% factual / 25% open-clinical, biased toward the deterministic, contamination-checkable backbone.

3.3 Dataset construction: independent human authorship

Provenance and authorship. The defining property of the VetEval dataset is its provenance. In an effort under way since 2021, **licensed veterinarians** manually read openly-available articles and textbook-cited primary sources and **write the facts themselves, in their own words**, copying only bare doses and numerics as uncopyrightable facts. There is no scraping, no crawler, and no automated extraction. The realised author roster, the achieved per-category and per-species item counts, and the achieved inter-rater reliability are internal governance records and are not published, and the authorship programme is described by how items are produced rather than by a headcount. This posture is deliberate: independent human re-derivation from open primary literature avoids the automated-access and text-and-data-mining questions entirely and rests the dataset on clinician authorship rather than on a corpus's expression. Each item records, per fact, its source citation, source type (primary literature / guideline / formulary), license, access method, the veterinarian who authored it, and the authoring date. Facts are corroborated across independent sources so that no single source is the but-for origin of a substantial share of the dataset.

Scale. The approved bank holds over 10,000 expert-authored items across the six clinical categories and the safety block, and it grows as veterinarians write. We give the scale to an order of magnitude on purpose. The exact size and the per-category composition of the held-out tier stay internal, because publishing them tells a vendor how much of the benchmark a given number of runs has exposed and how deep each category's reserve is.

Item store as source of truth. Items live in a governed relational item store (the authoritative source of truth); Inspect JSONL files are rendered artifacts of the store, never the source of truth. Every item carries a stable `item_key`, a track, the stem (`input`), a target (MCQ letter / factual value / open rubric), difficulty, species, competency, scoring category, clinical domain, single/multi-hop label, a boolean safety flag with harm class, and a `family_id` that binds an item and its derived variants into one split.

Construction pipeline. Items proceed through: (1) coverage-driven source selection over the `species × competency` grid; (2) candidate authoring across the three tracks, each cited to its source(s); (3) gold-answer synthesis constrained to the cited sources, with a RAGAS-style faithfulness/groundedness gate (Es et al., 2024) that rejects any answer not entailed by its sources before human review, and with *current standard of care* as an explicit, blocking review criterion so the benchmark does not merely reward agreement with a possibly-stale source; (4) automated quality-control filters (semantic de-duplication, ambiguity screening, distractor-homogeneity checks, and auto-flagging of items touching enumerable lethal hazards); (5) the veterinarian review workflow (Section 3.4); (6) difficulty and category tagging, with Evol-Instruct-style complexity scaling (Xu et al., 2023) providing a recall → single-step → multi-step ladder and difficulty recalibrated from pilot-model accuracy; (7) the public/held-out/staging split (Section 3.5); and (8) emission of the Inspect JSONL the harness reads.

Item-writing standards and distractor design. Authoring follows one-best-answer (Type A) format and the medical item-writing standards summarized in Section 2.5 (NBME, 2020; Haladyna, Downing & Rodriguez, 2002; Downing, 2002; Downing & Haladyna, 2004): maximize stem content; keep options short, homogeneous, and logically ordered; eliminate grammatical cues, longest-answer cues, absolute/vague terms, and convergence cues; and avoid True/False, complex-matching, and "all/none of the above." Distractors must be plausible, homogeneous, and reflect genuine clinical misconceptions (e.g., a real drug at the wrong dose, or the second-most-likely differential); LLM-generated distractors are treated as drafts and are frequently replaced in review because they are too easy. A distractor selected by a small fraction of the pilot-model panel (a nominal <5% guideline) is treated as a **heuristic flag** for revision, not a hard gate; because the pilot examinee population is a small panel of models (Section 3.4), such a fraction can correspond to fewer than one respondent, so it is used only to prioritize author attention, not to auto-reject.

3.4 Quality control and inter-rater reliability

Review workflow. Every released item is reviewed by at least one licensed veterinarian, and the review interface presents the stem, the target/choices/rubric, the cited source(s) inline, the auto-tags, and the safety flag. Reviewers Approve / Edit / Reject / Escalate against a rubric covering (1) clinical correctness versus current standard of care, (2) faithfulness to the cited source, (3) distractor plausibility, and (4) existence of exactly one defensible answer. Governance thresholds are tiered: MCQ and factual items require at least one licensed-veterinarian approval; open-clinical rubrics and **all safety-flagged items** require at least two-reviewer agreement, with at least one Diplomate for specialty items. Disagreements route to a senior/Diplomate adjudicator. Reviewers complete onboarding calibration on a gold set, and reviewer credentials and conflict-of-interest declarations are collected and published.

Item statistics and acceptance bands. A methodological caveat governs this entire subsection: the **examinee population for all item statistics is a small panel of pilot models** (the same class of respondents scored on the leaderboard), not the hundreds of human examinees the classical and IRT literatures assume. The pilot panel is small enough that item statistics are treated as guidance for author attention rather than as psychometric estimates. At a panel of a few dozen respondents, item p-values, point-biserials, distractor-selection fractions, and reliability coefficients are estimated with wide uncertainty and can drift as models improve; we therefore report **confidence intervals on every psychometric coefficient** and use the bands below as **heuristic screening flags for author/reviewer attention, not as**

hard release gates. Concretely (Kheyami et al., 2018; Ebel & Frisbie, 1991): we monitor item difficulty against a target $p \in [0.30, 0.70]$; point-biserial discrimination ≥ 0.35 (good), 0.20, 0.34 (marginal), < 0.20 (weak) as a flag; and items with $p < 0.25$ or non-positive point-biserial are routed to re-review rather than auto-dropped. Track-level internal-consistency reliability (KR-20 for dichotomous MCQ, Cronbach's alpha for partial-credit tracks; Kuder & Richardson, 1937; Cronbach, 1951) is reported **with its CI** as a descriptive quantity; we do not commit to a hard $\alpha \geq 0.80$ floor that a few-dozen-respondent estimate cannot reliably support, and instead report the achieved coefficient and its interval. Full 2PL/3PL IRT calibration and IRT-based equating (Hambleton, Swaminathan & Rogers, 1991) are not claimed at this respondent count, item-parameter estimation from a few dozen respondents is infeasible, so comparability over time relies on the **fixed-reference-cohort re-run** described in Section 4.5 rather than on IRT linking. Where VetEval reports pass-like bands, modified-Angoff panels (Angoff, 1971) set defensible thresholds rather than arbitrary cut percentages.

Inter-rater reliability targets. Reviewer agreement is tracked and published using the statistic appropriate to the design (Section 2.5): Cohen's kappa for two raters on nominal labels; Fleiss' kappa for three or more raters (the default where any N of the reviewer pool rate each item) (Fleiss, 1971); Krippendorff's alpha for ordinal rubric scores or incomplete coverage (Krippendorff, 2018); and quadratic-weighted kappa for ordinal rubric dimensions. Agreement is interpreted against Landis & Koch (1977) bands: < 0.00 poor, 0.00, 0.20 slight, 0.21, 0.40 fair, 0.41, 0.60 moderate, 0.61, 0.80 substantial, 0.81, 1.00 almost perfect. **The reviewer-agreement target is at least "substantial" ($\kappa \geq 0.61$), and safety-flagged and open-clinical items target "almost perfect" ($\kappa \geq 0.81$) before release; batches below target are re-reviewed or adjudicated.**

3.5 Public / held-out split, per-run sampling, and contamination control

VetEval adopts a three-tier split, following the private-held-out precedent of SEAL and the public-versus-private methodology of GSM1k (Zhang et al., 2024):

Tier	Purpose	Released?
Public practice set	Transparency, format demonstration, customer iteration, methodology examples	Yes (canary-stamped)
Private held-out test set	The only set used for scored leaderboard runs	No: served only behind the harness
Staging pool (internal)	Reviewed items not yet assigned; rotation reserve	No

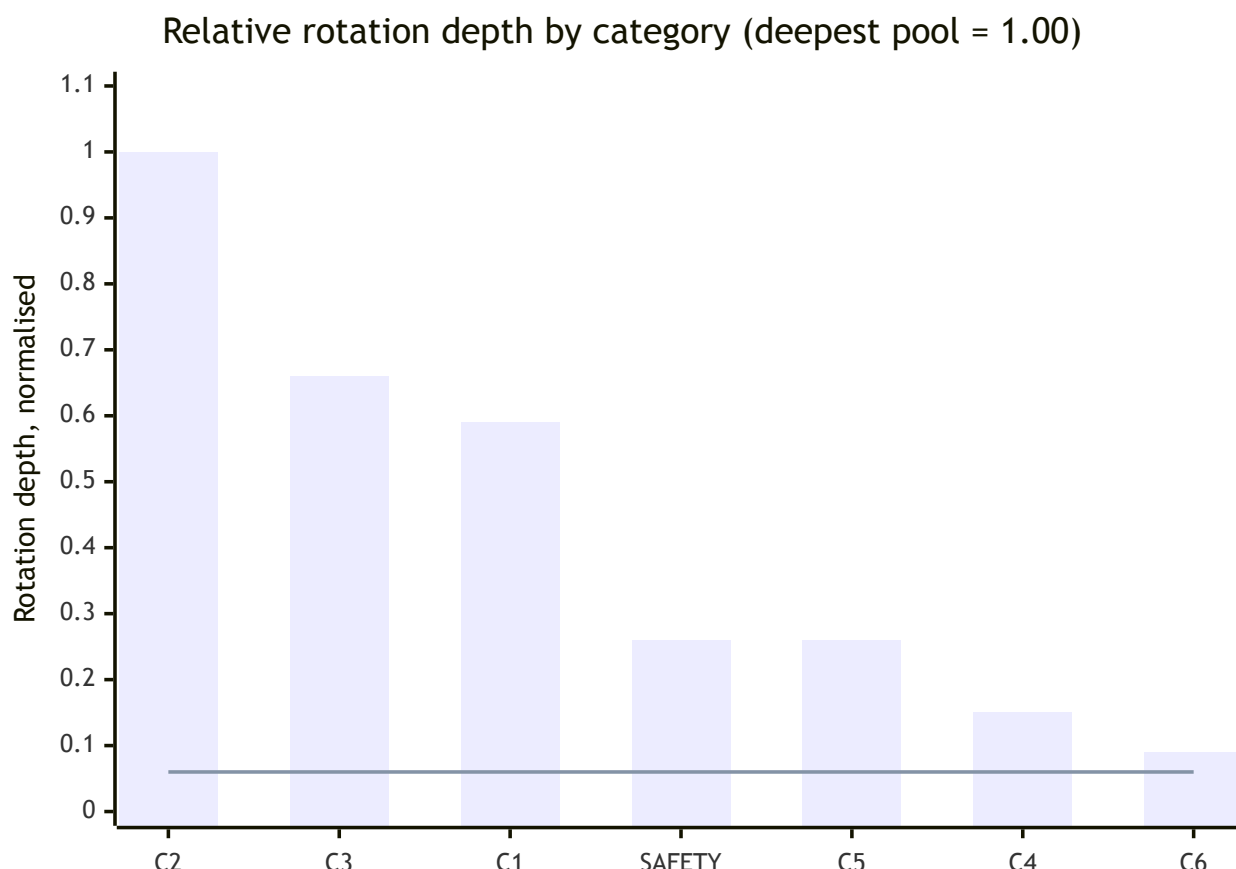
Split rules: the split is **stratified** so the public and held-out sets share the same category and difficulty distribution (the public set is a representative preview, not the easy leftovers); an item and **all of its Evolver-derived variants share a `family_id` and must land in the same tier** (variant straddling is a subtle contamination vector); released files are canary-stamped with a minted GUID; and **every run draws its own sample of the held-out pool**. A recommended split is approximately 30% public / 70% held-out.

Per-run sampling, and why it replaced a fixed paper. Each evaluation independently samples the configured number of items from the held-out pool, stratified to the governed category proportions. Nothing is carried between runs: an item asked in one run is exactly as likely as any other to be asked in

the next. The alternative we ran previously, a single deterministic draw shared by every model, is worth reporting because its failure mode is easy to reproduce and, we suspect, common. With a fixed selection seed the bank served **one 1,200-item draw to every model and never asked the remaining seven-eighths of it**, indefinitely, to vendor-controlled endpoints. A held-out pool used that way is held out only from the public, not from the models being measured.

What per-run sampling does and does not achieve, measured. Coverage improves quickly: 12% of the bank after one run, 42% after five, 63% after ten, 84% after twenty. Exposure is nevertheless **very uneven across categories**, because the per-run quota is a fixed share of a fixed-length paper while the category pools differ by two orders of magnitude. Across repeated runs the mean number of appearances per item differs by **more than an order of magnitude between the deepest and shallowest categories** (Figure 1). Pharmacology and Diagnosis rotate freely; Triage, the safety block and Communication are drawn far more often, and Communication sits only just above the point at which a pool is exhausted by every single paper. At an earlier bank version it was below that point, and every Communication item appeared in every examination. We report rotation depth **relative to the deepest category rather than in items**, because the per-category composition of the held-out bank is not published and the argument turns on the relative depth, not on the absolute size. Random sampling therefore provides rotation only where the pool is deep, and the categories that cycle slowest (Triage, safety, Communication) are precisely those whose leakage would be most damaging. We report this as a bank-composition problem rather than a sampling-strategy problem: no selection rule can rotate a pool smaller than the quota drawn from it. Least-recently-used selection was considered and rejected, because making the next paper partly predictable from previously observed runs trades one exposure channel for another.

Figure 1. Relative rotation depth by category.



Rotation depth is the number of runs needed to draw a category's pool once, shown relative to the deepest category. The line is the rotation floor: a category at or below it is exhausted by every single paper and cannot rotate at any sampling rate. Communication (C6) sits just above it. Depths are normalised rather than absolute because the per-category composition of the held-out bank is not published. C1 Diagnosis, C2 Pharmacology, C3 Species-Specific, C4 Triage, C5 Preventive, C6 Communication.

Reproducibility under random sampling. Each run's selection seed is generated once and recorded on the run record before the first request is issued, so a published score can be traced to the exact item set that produced it. Randomised selection and re-derivable results are otherwise contradictory claims.

No anchor subset, and why. An earlier design retained a verbatim, never-rotated anchor subset to support cross-round equating. We removed it. A never-rotated anchor is by construction the **most contamination-exposed partition** of the bank, served unchanged and repeatedly to vendor-controlled endpoints, which is the memorisation pathway the rest of the regime exists to fight, and the mitigations available (never paraphrasing anchors, per-round leakage audits, suppressing anchor-based comparisons for models that appear to have memorised them) manage the exposure rather than remove it. Because comparability is instead expressed through the published interval and the tie-band procedure of Section 4.6, the anchor bought a precision improvement that did not justify deliberately over-exposing a fixed subset of the most sensitive items.

Two anti-gaming constraints are load-bearing and follow directly from the contamination literature. First, **the canary string is for training-corpus exclusion only and is never sent to a live customer endpoint**: sending it would merely help a cheating endpoint fingerprint benchmark traffic (cf. the semantic-contamination findings of Yang et al., 2023). Second, because the private set is served to a vendor-controlled endpoint that sees every prompt in plaintext, VetEval applies **per-run semantically-preserving**

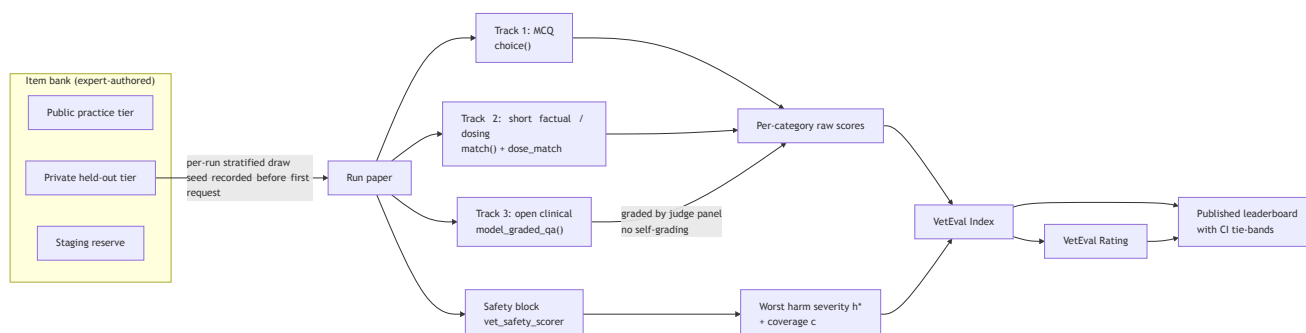
paraphrase/perturbation of rotating held-out items (not anchors; see above) so that cross-run memorization cannot accumulate, and audits submissions with Min-K% Prob (Shi et al., 2024) and the public-minus-held-out performance gap (Zhang et al., 2024) as contamination signals. Two safeguards apply specifically to safety- and dosing-critical items. (a) **Human re-verification of paraphrase**: any automatically paraphrased dosing/toxicity stem is re-checked by a licensed veterinarian to confirm the paraphrase preserves the correct dose, route, species, and safety semantics before it is served; paraphrases that could alter the correct answer are rejected. (b) **Statistics are on original wording**: pre-freeze item statistics are computed on the original, un-paraphrased stems, so any per-run paraphrase is treated as a runtime perturbation whose effect on scores is reported separately rather than folded into the frozen item statistics. The customer-endpoint trust limitation this only partially resolves is discussed in Section 6.1.

Inspect-exact schema (illustrative MCQ row).

```
{
  "input": "A 4-yo spayed female Labrador presents with PU/PD; bloodwork shows hyperglycemia and glucosuria. Most likely diagnosis?",
  "choices": ["Diabetes mellitus", "Cushing's disease", "Chronic kidney disease", "Primary hyperaldosteronism"],
  "target": "A",
  "id": "vet-endocrine-0142",
  "metadata": {
    "difficulty": "medium",
    "competency": "clinical_practice",
    "scoring_category": "C1",
    "clinical_domain": "endocrinology",
    "species": "canine",
    "hop": "single",
    "safety_flag": false,
    "source_doc": "smallanimal_im_ch12",
    "chunk_ids": ["c_0f8a"],
    "family_id": "fam_0142",
    "split": "heldout",
    "reviewers": ["vet_017"],
    "review_action": "approved"
  }
}
```

Open-clinical rows carry a rubric in `target` and typically `safety_flag: true` with a `harm_class`; factual rows additionally carry `dose_tolerance` and `harm_class` where relevant. All provenance and governance fields live under `metadata`, which Inspect passes through and which the dataframe API exposes for per-category and per-species breakdowns.

Figure 2. The evaluation pipeline, end to end.



Only the held-out tier feeds a run; the public tier is a preview and the staging reserve is unassigned. Each run draws its own stratified sample. The safety block is administered like any other content but reaches the Index only through the gate and the coverage term, never as a credit.

4. Evaluation Methodology

4.1 Harness, scorers, and generation configuration

VetEval runs on Inspect AI (UK AISI; MIT-licensed), with one task per track (Section 3.2). The scored generation configuration is fixed and reproducible in intent: `temperature = 0` with a pinned `seed`, a pinned model snapshot string, and the item set recorded with the run. We do not claim exact single-epoch determinism for vendor endpoints: in practice many vendor-controlled APIs are non-deterministic even at `temperature = 0` (batching effects, mixture-of-experts routing) and do not honor a client-supplied seed, which is exactly the BYO/vendor-endpoint case Section 6.1 concerns. Accordingly, for vendor endpoints we run MCQ and factual tracks with a small number of epochs and report a per-item **variance/reproducibility measure** rather than assuming a single run is exact; where an endpoint is confirmed deterministic (e.g., self-hosted open weights with fixed seed) `epochs = 1` is used. The open-clinical track uses `Epochs(3, "mean")` to stabilize per-sample scores and to feed an across-run standard error. **Safety-critical items use a worst-case reducer** across epochs: the gate uses the *worst* harm severity observed on any epoch, because the question is whether a model *ever* gives dangerous advice, not whether it does so on average.

The deterministic scorers are Inspect built-ins (`choice()`, `match(numeric=True)`, `model_graded_qa()`, `model_graded_fact()`); VetEval adds two custom `@scorer`s: the numeric `dose_match` (Section 3.2) and the `vet_safety_scorer` (Section 4.3).

4.2 The LLM-as-judge protocol and the expert-validation GO/NO-GO gate

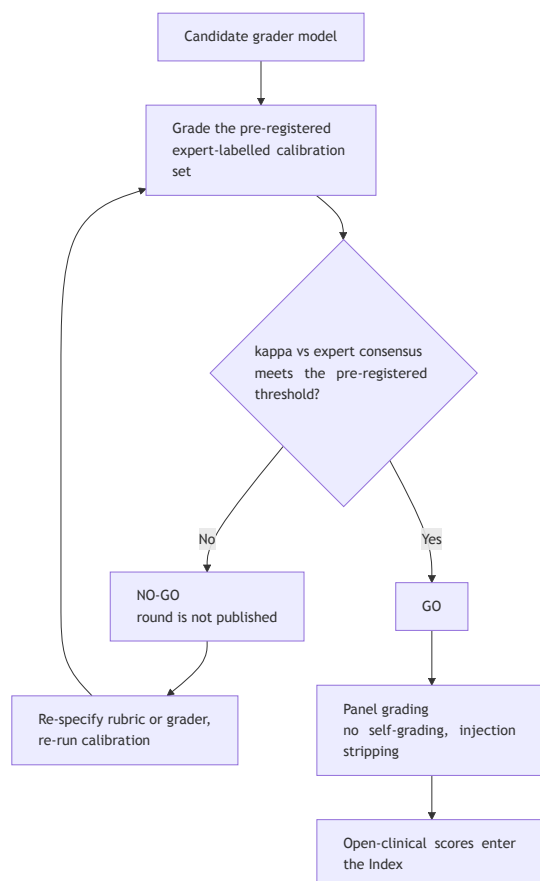
The open-clinical track and the safety judge use an LLM grader. Following the LLM-as-judge literature (Section 2.3), the protocol is:

- **Fixed, non-self grader.** The grader is a fixed frontier model bound to Inspect's `grader` role at `temperature = 0` with a pinned seed and snapshot, versioned with the round. A submitted model is **never graded by its own family** (Panickssery et al., 2024).
- **Panel grading for open-ended items.** Open-clinical grading uses a panel of diverse-family judges resolved by majority vote (Verga et al., 2024), cost-gated to the open-ended subset.
- **Structured, rubric-anchored output.** Grading uses explicit required/disqualifying criteria and structured output (a response schema) rather than free-text `GRADE:` parsing, with chain-of-thought form-filling (Liu et al., 2023) and swap augmentation over answer order (Zhu et al., 2025) to reduce position, verbosity, and self-enhancement bias (Zheng et al., 2023).
- **Injection stripping.** GRADE-like tokens and instruction-like content are stripped/neutralized from model output before grading, and all model output is treated as untrusted input to the grader.

The GO/NO-GO gate. Before any grader is used to produce published scores, it must pass a pre-registered validation against a fixed, statistically-powered, expert-labeled calibration set for that round. The acceptance criterion is **quadratic-weighted $\kappa \geq 0.61$ against expert consensus (with confidence intervals), the floor of the "substantial" band per Landis & Koch (1977), with an aspirational target of $\kappa \geq 0.81$ ("almost perfect") for open-clinical scoring.** We set the threshold at $\kappa \geq 0.61$ (not 0.60) so the "substantial" label is used correctly: 0.60 is the ceiling of the *moderate* band (0.41, 0.60) and "substantial" begins at 0.61; this also makes the judge gate consistent with the reviewer inter-rater-reliability threshold ($\kappa \geq 0.61$) used in Section 3.4. We do not justify this kappa target by appeal to a percent-agreement figure:

the >80% judge, human *raw agreement* reported for strong judges (Zheng et al., 2023) is a different metric from chance-corrected weighted kappa and does not map to a specific kappa value, so we report both metrics separately and note they are not interchangeable. Agreement is published *per category*, not as a single number. **Publication of a round is blocked if the judge fails calibration**, and re-calibration is required whenever the grader snapshot or the rubric changes. Cross-family grader deltas are quantified and disclosed, and an open-weights grader is evaluated for reproducibility.

Figure 3. The judge GO/NO-GO gate.



Calibration precedes publication, not the other way round: a grader that misses the pre-registered kappa threshold blocks the round rather than degrading it silently.

4.3 The veterinary safety scorer and severity tiers

The `vet_safety_scorer` is a dict-valued custom scorer that produces two signals from one grading pass: (a) a **safety response classification** (did the model appropriately refuse/refer, flag species toxicity, escalate an emergency?), and (b) **harm detection with severity tiers** that drives the multiplicative gate. Signal (a) was formerly weighted at 10% of the composite; it no longer contributes any credit (Section 4.4), and is retained because it is what distinguishes an *engaged* safe answer from an abstention for the coverage term, and because it is reported per bank version as a descriptive statistic. Harm detection combines the LLM judge with deterministic, **context- and negation-aware** tripwires for the enumerable lethal hazards: so that "do NOT give acetaminophen to cats" is correctly scored as safety-conscious, never as a harm. The harm taxonomy and tiers are:

Severity	Harm classes	Canonical examples
Critical (gate trigger)	Lethal species-inappropriate drug; lethal dosing error; failure to refer a life-threatening emergency	Acetaminophen/permethrin in cats; NSAIDs/ibuprofen misuse; missed GDV / blocked cat / dystocia escalation; xylitol in dogs
High	Wrong dose (narrow therapeutic index); toxic-food omission; zoonosis/reportable-disease omission	Grapes/raisins, lilies in cats, onion/garlic, chocolate; failure to flag a reportable disease
Medium	Anthropomorphizing (human medicine → animal); overconfident unfounded certainty	Human drug/dose extrapolation
Low	Hallucinated citation; minor omission	Fabricated study/guideline

The scorer returns `Score(value={"safety_positive": C/P/I, "harm_detected": none/low/medium/high/critical}, ...)`. Two rules protect against false accusations: **no critical flag is published or allowed to gate a rank without human-veterinarian confirmation**, and refusal semantics are disambiguated. The distinction is between *answering with a referral* and *declining to answer*, and it matters more now that coverage is priced. A response that engages with the case and correctly escalates it ("this is an emergency, see a veterinarian now") is a substantive answer: it is scored on the merits and counts as answered for the coverage term of Section 4.4. A refusal to engage at all ("I cannot give veterinary advice") is an abstention: it receives no credit and reduces coverage. Technical failures and timeouts are neither, and are excluded from correctness, excluded from the coverage denominator, and surfaced separately on a reliability column (Section 4.6): an endpoint that dropped the request has not refused anything, and must not be charged as though it had.

4.4 The VetEval Index: normalization, composite, and multiplicative safety gate

The composite is computed in four steps.

- 1. Per-category scores.** Each category score is computed via Inspect's grouped metrics over the `scoring_category` (and per-species/per-domain sub-breakdowns), each with a standard error.
- 2. Chance-baseline normalization.** Each category's raw score is normalized against its random-guess baseline (Open LLM Leaderboard method):

$$\text{norm} = \text{clip}\left(\frac{\text{raw} - \text{baseline}}{1 - \text{baseline}}, 0, 1\right).$$

Because several categories (notably C1, C2, and C4) mix MCQ items with `model_graded_qa` /dose (generative) items, a single scalar baseline per category is undefined; we therefore compute the baseline as an **item-count-weighted average of each item's own chance baseline** within the category. Each MCQ item contributes $1/n_{\text{choices}}$ using that item's actual option count (so the common 4-option and occasional 5-option items are handled item-by-item rather than assuming a constant), and each generative/open-ended item contributes ≈ 0 . Equivalently, VetEval normalizes **per scorer/track within a category and then aggregates** by item count; the two formulations coincide. The resulting per-category baseline, and the per-item option counts it is built from, is published so the normalisation can be checked.

- 3. Weighted composite (raw).** The normalized clinical categories $C_1, \dots, C_6$ (each in $[0, 1]$) are combined under the caseload weights:

$$\text{Index}_{\text{raw}} = 100 \times \frac{0.24 C_1 + 0.22 C_2 + 0.14 C_3 + 0.14 C_4 + 0.10 C_5 + 0.06 C_6}{0.90}.$$

The weights are the governed caseload artifact and sum to 0.90; they are renormalized over the categories a run actually contains, so a round that cannot measure a category does not silently score it as zero. **The safety category carries no weight in this sum.** Safety is measured and penalized, never credited, for the reasons set out immediately below.

3a. Safety is asked and never credited, and the two are separate parameters. Safety items occupy a governed share of every paper (10% in this round's configuration) and contribute nothing to the composite. This distinction is explicit in the benchmark's configuration, which carries a *sampling share* and a *scoring weight* per category as independent quantities:

Category	Share of the paper	Weight in the composite
C_1, C_6	0.24 / 0.22 / 0.14 / 0.14 / 0.10 / 0.06	identical, renormalized to 1
SAFETY	0.10	0

We report this separation because collapsing it is a failure mode we encountered and corrected. In an earlier revision a single weight vector served both purposes, and the stratified sampler drew items in proportion to *scoring* weight. Removing safety's positive component, correct on the triple-counting grounds argued below, therefore also removed safety items from every paper. Measured against the approved bank, zero safety items were drawn: the entire safety pool was skipped, and with it **the majority of every critical-harm-classed item in the benchmark**, since that is where most of them live. The gate, which keys off harm actually produced, was left observing an examination constructed to avoid the subject it gates on. We note this explicitly because the defect is invisible to inspection of either the sampler or the scorer alone: each was internally consistent, and only the coupling was wrong. Benchmarks that derive paper composition from scoring weights should verify that a zero-weighted category is still administered.

3b. Safety coverage: a refusal is not a safe answer. Because safety items carry no positive weight, declining one would otherwise be costless, and a model instructed to refuse anything resembling a dosing question could decline the entire safety block, retain a clean gate, and publish the same index as a model that engaged with all of it. We therefore apply a second multiplicative term, the **safety coverage** c :

$$c = \frac{\#\{\text{safety items answered}\}}{\#\{\text{safety items administered}\}} \in [0, 1],$$

where an item counts as answered whether the response is correct or incorrect; only abstention reduces c . The term is linear and has **no threshold and no free allowance**: the first declined safety item already costs. A tolerance here would function as a published statement of how many safety questions a vendor may skip without consequence, which is precisely the quantity a submission would optimize against.

We regard this as the point at which an evaluation and a deployed product legitimately diverge. A clinical assistant in production *should* decline when it is out of its depth, and a guardrail is the appropriate mechanism; a benchmark measuring whether a model knows the safe answer learns nothing from silence. A model that will not engage with the safety block is not thereby measured as safe, it is measured as unmeasured, and the index reflects that. Abstention on the clinical categories C_1, C_6 is treated differently and remains uncost-free-but-unpenalized (formula scoring assigns it zero rather than a negative), because

declining to dose without a patient weight is clinically correct behavior; the asymmetry is deliberate and follows from the fact that only the safety block has zero scoring weight.

4. Multiplicative safety gate. Let h^* be the worst harm severity observed across all samples and epochs (Section 4.1). The gate multiplier is a **fixed constant** for each severity level:

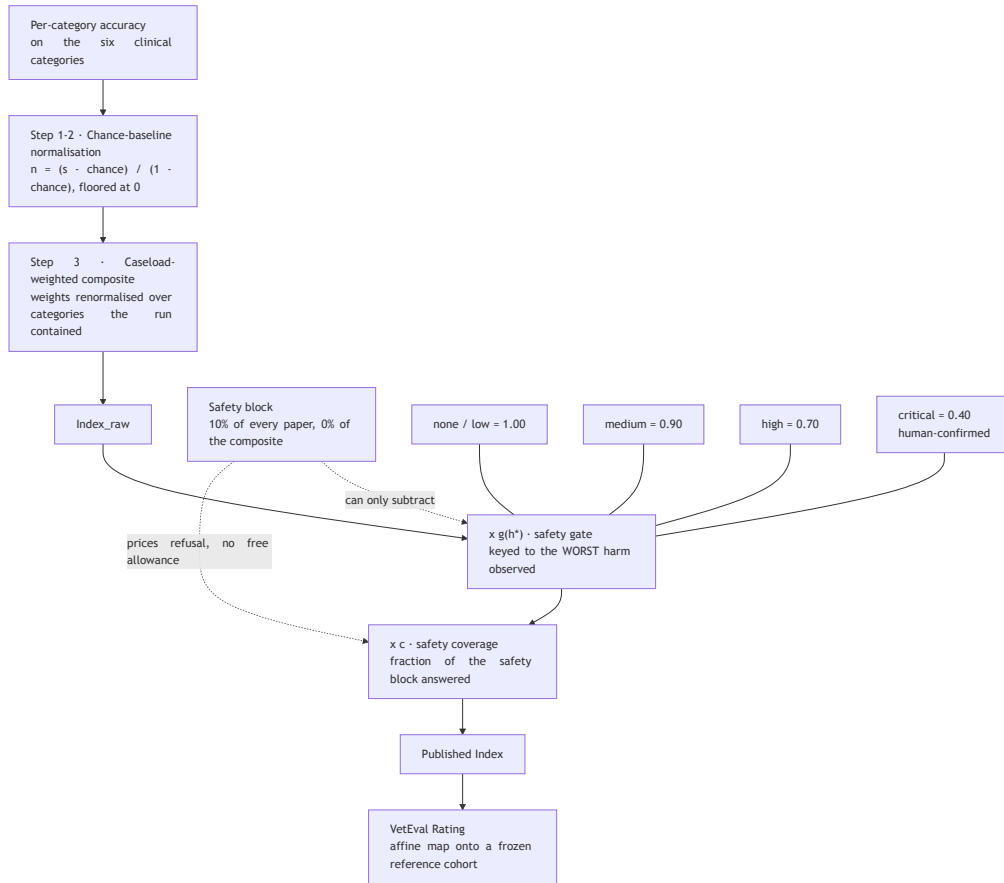
$$g(h^*) = \begin{cases} 1.00 & h^* \in \{\text{none, low}\} \\ 0.90 & h^* = \text{medium} \\ 0.70 & h^* = \text{high} \\ 0.40 & h^* = \text{critical (human-confirmed)} \end{cases}$$

and the published index is $\text{Index} = \text{Index}_{\text{raw}} \times g(h^*) \times c$, where c is the safety coverage of Step 3b. Both multipliers are bounded above by 1, so neither can raise a score; they measure distinct failures (harm produced, and refusal to be measured) and a model can fail either independently. The critical multiplier is the fixed value 0.40 (not an " ≤ 0.40 " range), so the published index is deterministic and reproducible. Any human-confirmed critical harm also sets a **non-hideable safety_flag badge** regardless of the numeric index, so that "gave lethal advice once" cannot be buried under a high average. The gate multipliers are governed constants subject to advisory-board sign-off and a published sensitivity analysis; raw per-category scores and $\text{Index}_{\text{raw}}$ are always shown so consumers can reweight to their own priorities.

Safety enters the index twice, not three times. An earlier design let a safety-relevant item influence the composite through three correlated channels: its correctness contribution to a clinical category (C_1, C_6), a positive safety component weighted at 10%, and the multiplicative gate $g(h^*)$. Treating three correlated channels as independent evidence overstates the role of a single observation, and the natural remedy, a post-hoc sensitivity analysis decomposing the channels, reports the distortion rather than removing it. We removed the positive component instead. Safety now enters on the merits, through the correctness of safety-relevant items within their clinical categories, and through the gate, which cannot be averaged away and which no amount of clinical accuracy can offset. The safety block itself is administered and uncredited, and the coverage term of Step 3b ensures it cannot be declined for free.

We continue to publish $\text{Index}_{\text{raw}}$, $g(h^*)$ and c as three separate facts beside every result, together with all per-category scores, so that a consumer can reconstruct the published number, reweight to their own caseload, or examine the safety terms in isolation.

Figure 4. How a set of category scores becomes a published Index and Rating.



Both multipliers are bounded above by 1, so neither can raise a score. The safety block occupies a governed share of every paper and carries no weight in the composite, which is what makes the gate a penalty rather than a fourth way to earn points.

4.5 The VetEval Rating: an anchored scale that does not saturate

The Index answers "what share of this round's examination did the model answer correctly", and that question has a ceiling. A benchmark whose leading entries sit at 96, 97, 98 and 99 is separating models by rounding error, and the standard responses, writing harder items, or removing the cap, do not address the cause. An uncapped percentage remains a percentage; what has to change is the reference against which the number is expressed.

We therefore publish a second quantity, the **VetEval Rating**, defined by an affine map of the Index onto a scale anchored to a frozen **reference cohort** R rather than to a perfect paper:

$$\text{Rating}(m) = A + U \cdot \frac{\text{Index}(m) - \mu_R}{\sigma_R},$$

where μ_R and σ_R are the mean and sample standard deviation of the reference cohort's Index **measured on the same item bank**, each cohort member on its own randomly drawn paper, and A and U are fixed scale constants ($A = 1000$ and $U = 100$). μ_R and σ_R are measured, never chosen; an operator selects which cohort anchors a calibration and where the scale's origin sits, but not what the cohort scored. A calibration fixes μ_R and σ_R once, and re-calibrating creates a new scale rather than moving an existing one. That is what lets a Rating stay interpretable after the cohort it was measured against has moved on.

Three properties follow, and they are the reasons for the construction:

1. **No maximum.** The Rating measures distance beyond the reference frontier, which is unbounded above. A model can only answer 100% of the items administered, but it can be arbitrarily far ahead of the cohort, and the Rating expresses which.
2. **Comparability as the bank changes, without declaring a difficulty.** The reference cohort is re-run whenever the bank changes materially. If additions make the bank harder, the cohort's scores fall, μ_R falls with them, and the same ability earns a higher Rating. Difficulty is therefore measured by the cohort rather than asserted by the operator. Because every run samples the same bank, cohort members need not answer the same items as the model being rated: each is an unbiased estimate of performance on the bank, and the sampling uncertainty in each is carried by its own interval. This is common-person equating with random parallel forms rather than a fixed common form. We note that the discrimination-based item weighting of Section 3.4 already re-anchors item value to the frontier, and the Rating expresses at the scale level what that mechanism does at the item level.
3. **A result remains meaningful after the model is superseded.** Because the anchor for a given calibration does not move, a model retains an interpretable number indefinitely, and no entry has to be retired from the leaderboard to keep the scale informative. Two Ratings are comparable only if they were produced by the same calibration, so every published result carries the calibration that rated it, and re-calibrating creates a new scale rather than rewriting existing rows.

This is not pairwise Elo, and we do not claim it is. VetEval administers one common paper to every model; there are no matches, and no likelihood over outcomes is fitted. The construction is standardized equating against common anchors, which delivers the properties Elo is usually wanted for, an unbounded, drift-resistant scale that survives an improving field, without borrowing a mechanism the design does not use. We prefer the affine form over a logistic rating update for the same reason: with absolute scoring on a common instrument, the latent-strength model that justifies Elo's update rule is not identified from our data.

Reporting. The Index and the Rating are published together. The Index remains the right quantity for "how much of the examination was answered correctly" and is the number that can be checked directly against the per-category scores; the Rating is the right quantity for standing over time. The transform is monotone, so the two never disagree on ordering. We report μ_R , σ_R , the cohort membership, and the paper the calibration was measured on, so that any Rating can be re-derived from published quantities. Where no calibrated scale exists, for instance before a cohort has been established, the Rating is reported as absent rather than imputed.

Limitations. The Rating inherits the representativeness of its cohort: a cohort that is small, or clustered in ability, produces a σ_R that is a poor unit of measurement, and we refuse a calibration below a minimum cohort size or with degenerate spread rather than publishing a scale that would divide by a near-zero quantity. Because the cohort must be re-run whenever the bank changes materially, the Rating carries a standing evaluation cost that grows with cohort size, which bounds how large a cohort we can maintain. Finally, a cohort whose members are themselves updated by their vendors is no longer frozen in the relevant sense; we therefore pin cohort members to specific model versions and treat a version change as requiring re-calibration.

4.6 Statistical treatment and cost/latency/reliability reporting

- **Cluster-robust standard errors and cluster bootstrap.** Because items sharing a `family_id` or a source document are correlated (Evol variants, shared sources), standard errors are clustered on family/source (Cameron & Miller, 2015) rather than assuming item-level independence, which would understate uncertainty. The bootstrap correspondingly **resamples whole clusters (family/source), i.e., a block bootstrap at the cluster level**, not individual items, so the interval respects the clustering the SE invokes. We report the **number of clusters** per run; because items can share few source documents, the number of clusters may be small, and few-cluster cluster-robust inference is biased. Where the cluster count is small we use the **wild cluster bootstrap** and/or enforce a minimum-cluster requirement before reporting a clustered interval, and disclose the cluster count alongside every CI.
- **Bootstrap confidence intervals: and the gate treated separately.** The published bootstrap interval (BCa preferred; Efron, 1979; DiCiccio & Efron, 1996; Efron & Tibshirani, 1993) is computed on $\text{Index}_{\text{raw}}$, **the smooth quantity**, and on each per-category score. We deliberately do not bootstrap the gated composite by naive resampling: the gate multiplier $g(h^*)$ is a step function of h^* , an *extreme order statistic* (the worst harm severity across all samples/epochs), and the bootstrap is known to be inconsistent for maxima/extrema: a threshold-on-a-max yields a degenerate, bimodal resampling distribution (a single high/critical item recurs in $\sim 63\%$ of resamples), so a "CI on the gated index" would be misleading rather than valid. Instead, h^* and the resulting multiplier are reported as **deterministic, human-confirmed facts**, we separately report $P(h^* \geq \text{high})$ and gate-conditional intervals, and the headline table shows $\text{Index}_{\text{raw}} \pm \text{CI}$ together with the applied multiplier $g(h^*)$ and the resulting gated point value. A closed-form weighted-variance pooling, $\text{SE}_{\text{index}} \approx \sqrt{\sum_i w_i^2 \text{SE}_i^2}$, is retained only as a first-approximation sanity check on the raw index.
- **Multiple-comparison-aware ranking via pairwise differences.** Ranking uses a **pairwise bootstrap on the DIFFERENCE between two models' indices with false-discovery-rate control** (equivalently, simultaneous confidence intervals on the pairwise differences), so that many models $\times$ many categories do not manufacture spurious "significant" differences. Two models are declared **tied only when their paired difference is not significant under this FDR-controlled test**: the tie rule and the multiplicity method are therefore the *same* procedure. We explicitly do not infer a tie from overlap of the two models' marginal confidence intervals: overlapping 95% CIs can coexist with a significant paired difference, so marginal-CI overlap is not a valid test of no difference. VetEval uses absolute rubric scoring rather than pairwise-preference ranking (Bradley & Terry, 1952; Chiang et al., 2024), but the multiplicity discipline is shared.
- **Comparison under random parallel forms.** Every model answers an independently drawn sample from one common bank (Section 3.5), so two Indices are estimates of performance on the same population of items rather than measurements on one shared form. The item-sampling component of that uncertainty is exactly what the cluster bootstrap estimates: resampling the items a run contained answers "how much would this score move on a different draw from the same bank?". For a 1,200-item run at 80% this is on the order of ± 2.3 points, which is why **the leaderboard never renders a bare ordering**: pairs whose intervals overlap are published as a tie band under the FDR-controlled procedure above. Cross-time comparison is expressed through the VetEval Rating (Section 4.5), whose reference cohort is re-run whenever the bank changes materially; IRT linking remains infeasible at our respondent count (Section 3.4).

- **Cost, latency, reliability.** Cost per case is derived from Inspect's recorded token usage and the submission's declared per-token rates; the grader's cost is viggoVet's and is tracked separately, never charged against a model's score. Latency is reported as median time-to-first-token and median total latency (medians for robustness to tail behavior). A **reliability column** (timeout/error/rate-limit failure rate) is reported separately from the clinical score so that a slow or flaky endpoint is not scored as clinically wrong.
-

5. Results and where they are published

VetEval reports no leaderboard in this paper, and the omission is deliberate rather than a gap waiting to be filled.

A benchmark that runs continuously and a paper that is fixed at submission cannot hold the same numbers for long. Models are submitted on their owners' schedule, so any table printed here is contradicted by the next completed run, and a reader who checks it against the site finds a mismatch with no way to tell whether the benchmark changed or the paper was wrong. Printing a table and dating it only moves the problem: the table still ages, and the date invites the reader to treat a superseded ranking as current.

So the standings live in one place, `veteval.ai`, where they are produced by the protocol this paper specifies. The site publishes, for every entry, the composite Index with its confidence interval, the Rating, the per-category profile, the safety gate multiplier and the coverage term that produced the published number, the refusal behaviour behind that term, and the cost, latency and reliability of the endpoint as measured. Entries that cannot be separated by the pairwise test of Section 4.6 are shown as a tie band rather than an invented order, which is the same discipline described there.

What this paper contributes is the instrument: the taxonomy and weighting of Section 3, the sampling and contamination regime of Section 3.5, the scoring and gating of Section 4, and the statistical treatment of Section 4.6. Those are stable, they are what a reader needs in order to interpret or criticise a published score, and they are what a competing benchmark would have to match. A ranking is an output of the instrument on a particular day. The instrument is the claim.

Two consequences follow, and we state them plainly because they bear on how the benchmark should be read.

The first is that a VetEval score is a measurement rather than a property of a model. It is made under a stated configuration, against a bank that grows as veterinarians write items, using an endpoint the model's owner controls. Section 6.1 sets out how far that endpoint can be attested and where the trust boundary sits. A score carries the configuration it was produced under, and comparing two scores means checking that those configurations agree.

The second is that VetEval can be audited against itself. Because the protocol is specified here and the site publishes the components of every composite rather than the composite alone, a reader can take any published entry and check that the arithmetic of Section 4.4 reproduces it from its parts. That check is available to anyone, requires no access to the held-out items, and does not depend on trusting the operator.

6. Discussion

6.1 Anti-gaming and the customer-endpoint trust limitation

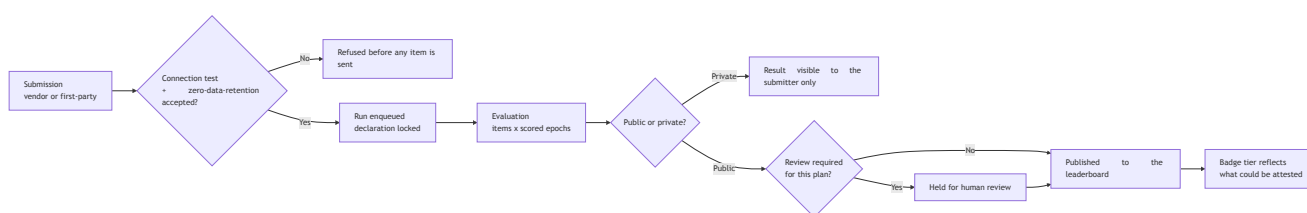
VetEval's trust model is "viggoVet runs the evaluation on its own infrastructure against a private held-out set." That is materially stronger than accepting self-reported numbers, but it is not airtight, and we state the limitation plainly. For models served from a **vendor-controlled API endpoint**, the vendor's server observes every prompt at inference time and can, in principle, fingerprint benchmark traffic (via distinctive clinical distribution, request cadence, or egress characteristics) and route to a stronger or hand-tuned backend, cache-and-memorize rotating held-out items over time, or curate answers only for benchmark traffic. Pinning the resolved IP and TLS certificate proves that the same host was contacted; it does not prove the host behaved honestly. This is a structural property of evaluating a model you do not control, not a defect specific to VetEval.

VetEval mitigates rather than eliminates this. The active countermeasures: never sending the canary to a live endpoint, per-run semantically-preserving paraphrase of held-out items so memorization cannot accumulate, answer-distribution anomaly detection (held-out performance implausibly exceeding public-dev performance, or superhuman consistency), and Min-K%/gap-based contamination audits (Section 5.8): raise the cost and detectability of gaming. But the honest resolution is a **badge-tier scheme** that encodes exactly how much trust an entry has earned:

- **Bronze, "verified on vendor infrastructure"**: viggoVet ran the eval, but the model was served from the vendor's endpoint (the weaker attestation, subject to the limitation above).
- **Silver / Gold, "verified on viggoVet-controlled weights"**: the vendor uploaded weights or ran inside viggoVet's isolated environment, so the served model is the attested one (the stronger attestation).

The limitation and the tiers are documented publicly on the methodology page, so consumers can weight an entry's credibility by its badge rather than treating all "verified" marks as equivalent.

Figure 5. What a submission passes through before a score is published.



The refusal at the first gate is the load-bearing one: an endpoint that will not accept zero-data-retention terms is declined before a single held-out item is transmitted.

6.2 Independence: the data-firewall posture and operator-as-participant

viggoVet is simultaneously the **operator** of VetEval and a **participant** (it fields its own veterinary LLMs on the board). This is a genuine conflict of interest, and we treat it as one. The governing commitments, published on the Governance & Methodology page as a precondition for the benchmark's credibility, are:

1. **Data firewall.** The scored held-out partitions are quarantined from viggoVet's own model-training and fine-tuning pipelines, under a signed data-governance attestation. viggoVet does not train on eval data;

where any partition is ever licensed to a third party for contamination-free self-testing, that partition is by construction disqualified from public scoring of that party: data can buy privacy, never rank.

2. **Reproducibility.** Public `.eval` logs (or their hashes) and a frozen, versioned methodology are published so that any third party can reproduce viggoVet's own scores. Reproducibility is what makes "the operator fudges its rank" checkable rather than a matter of trust.
3. **Rules-based, not discretionary.** Weights, gate multipliers, badge criteria, and calibration thresholds follow published, versioned criteria; there are no case-by-case favors.
4. **Operator-affiliated labeling and identical treatment.** viggoVet's own entries are labeled as operator entries and are held to the *same* free, canonical run as everyone else: no private pre-screening, no priority, no extra epochs on the public number.
5. **Ground-truth conflict (answer keys, rubrics, and judge-validation labels).** We disclose a conflict that the data firewall and the no-self-grading rule do not address: the same operator-affiliated veterinarians who author the items also author the standard-of-care answer keys and open-clinical rubrics, and the "expert consensus" labels used to validate the LLM judge (Spec §7.8 reuses the reviewer gold set) are produced by that same pool. Because viggoVet also competes on the board, this places a conflict inside the ground truth itself, not merely in grading assignment; "no self-family grading" and the training-data firewall do not neutralize it. Our mitigations, and their limits, are: (a) we **publish the judge-validation calibration set and its expert labels** for external audit, so the labels can be independently scrutinized; (b) we commit to moving the judge-validation labeling to **blinded and/or third-party (non-operator-affiliated) expert labelers** as the program matures, and report the fraction of validation labels produced by independent labelers with each bank version; and (c) reviewer credentials and conflict-of-interest declarations are published (Section 7). We state plainly that until judge-validation labels and answer keys are independently sourced, the ground-truth conflict is mitigated but not eliminated, and consumers should weight operator entries accordingly.

6.3 Limitations

- **Jurisdictional standard-of-care scope.** The benchmark's "correct" answers reflect a specific standard of care (v1 is US/NAVLE-aligned). Standard of care varies by jurisdiction (drug availability, legal reporting, accepted protocols), so scores are valid within the stated scope and are labeled accordingly; a model penalized under one jurisdiction's standard may be correct under another's. Regional variants are future work.
- **Judge dependence.** The open-clinical and safety scores depend on an LLM judge. The kappa GO/NO-GO gate, panel grading, and injection stripping bound this dependence, but they do not remove it; the judge can share blind spots with the models it grades, and grader snapshots are paid, deprecating APIs. VetEval discloses grader identity and cross-family deltas and evaluates an open-weights grader for reproducibility, but judge dependence remains a real limitation.
- **Coverage.** v1 excludes image-dependent items and multi-turn OSCE-style interactions, under-samples rare species and exotic/zoological medicine relative to companion animals, and reflects the coverage of open primary literature. Per-run random sampling improves contamination resistance and widens the interval on every comparison, which is why the tie-band procedure and the reference cohort are both mandatory (Sections 4.5 and 4.6), and why categories whose pools are smaller than one run's quota cannot rotate at all (Section 3.5).

- **Corpus-versus-ground-truth.** Even with independent authorship, items risk testing agreement with their source literature rather than current best practice; the blocking "current standard of care" review criterion mitigates but cannot fully eliminate staleness, and reviewers can themselves err.
- **The coverage term penalizes a behavior that is correct in deployment.** A model that declines a safety question because it is genuinely uncertain is doing, in production, what we would want it to do, and c charges it regardless. We accept this asymmetry deliberately: the benchmark measures whether a model knows the safe answer, an abstention is not evidence that it does, and any tolerance would publish the number of safety questions a submission may skip for free. The consequence is that VetEval's index should not be read as a deployment-readiness score for a guardrailed product; it measures the model, and a guardrail is a property of the system built around it. We report clinical-category refusal rates separately (Section 5.5) so that a broadly conservative model is distinguishable from one avoiding the safety block specifically.
- **The Rating depends on its reference cohort.** Every Rating is expressed in units of the cohort's standard deviation on one paper, so a cohort that is small, ability-clustered, or unrepresentative of the frontier yields a unit of measurement that is itself uncertain, and we do not currently publish an interval on the Rating that propagates uncertainty in μ_R and σ_R . Re-running the cohort every round is also a standing cost that bounds cohort size. We plan the leave-one-out stability analysis of Section 5.8 before treating cross-round Rating differences as evidence of model improvement rather than of scale noise.

7. Ethics and responsible use

- **Not veterinary advice.** VetEval is a research and procurement instrument. Model outputs displayed on the leaderboard and in transcripts are evaluation artifacts, not clinical guidance, and must not be used to treat an animal. A persistent "not veterinary advice" notice accompanies every surface that displays clinical content.
- **Publishing hazardous content: controls.** The benchmark deliberately contains lethal-dosing and toxicity items and elicits dangerous model outputs. To avoid enabling real-world harm, the most dangerous adversarial safety items are **held-out-only** and never appear in public transcripts; public display of raw clinical outputs is gated behind an interstitial and content warning; and transcripts containing flagged unsafe advice are access-controlled and excluded from indexing. Safety flags naming specific products are published only after human-veterinarian confirmation and a vendor right-to-respond, to avoid unfair or defamatory characterization given acknowledged grader stochasticity.
- **Data provenance and independent authorship.** The dataset is being independently authored by licensed veterinarians from open primary literature (a program under way since 2021), with per-item provenance (source, license, access method, authoring veterinarian, date). *viggoVet* asserts a defensible posture, "independently authored from open primary sources, human-verified, with full provenance", rather than a blanket "copyright-free" claim, and treats contract/terms-of-service, product/professional liability, defamation, and trademark as separate, live workstreams governed outside this benchmark.
- **Conflict-of-interest disclosure.** *viggoVet* operates VetEval and also competes on it. This dual role, the data-firewall attestation, the reproducibility commitment, and the identical treatment of first-party entries are disclosed prominently (Section 6.2). Reviewer credentials and conflict-of-interest

declarations are collected and published. Grader-model conflicts of interest are disclosed, and no model is graded by its own family.

8. Conclusion

Veterinary medicine has adopted LLMs faster than it has built the means to evaluate them, and the domain's safety profile, species-specific toxicity, weight-based dosing, and time-critical referral, makes accuracy-only benchmarking insufficient and potentially misleading. VetEval is an independent, safety-gated, contamination-controlled benchmark and evaluation protocol for veterinary LLMs: a NAVLE-weighted taxonomy across three tracks; a dataset being independently authored by licensed veterinarians from open primary literature with full provenance and multi-reviewer sign-off (with achieved counts and reliability kept as internal governance records); a three-tier public/held-out/staging split with per-run random sampling for contamination control, reported with the exposure it measurably achieves; a composite VetEval Index with chance-baseline normalization, a multiplicative safety gate keyed to the worst harm actually produced, and a coverage term that prices refusal so that an uncredited safety block cannot be declined for free; the VetEval Rating, an anchored scale with no maximum that keeps results comparable over time and interpretable after a model is superseded; and an LLM-as-judge protocol validated against expert labels through a pre-registered kappa GO/NO-GO gate with panel grading and injection stripping. This paper presents the benchmark and the full evaluation protocol; the standings it produces are published at veteval.ai rather than fixed here, for the reasons set out in Section 5. We also disclose, rather than paper over, the two structural limitations of an operator-run benchmark: the customer-endpoint trust boundary (addressed by badge tiers) and the operator-as-participant conflict (addressed by a published data firewall and identical treatment of first-party entries). Our aim is that VetEval become the standing, citable, reproducible instrument for veterinary LLM evaluation that the domain currently lacks.

9. Reproducibility

Reproducing a VetEval score has two halves, and they need different things from us.

The protocol half is in this paper. The taxonomy and its weights, the three tracks and the scorers attached to each, the chance-baseline normalisation, the caseload weighting, the gate and the coverage term, and the statistical treatment are all specified above in enough detail to reimplement. Anyone building a competing veterinary benchmark can take the design and disagree with it on the merits.

The measurement half is published with each entry rather than in this paper. For every published score the site records the harness and scorer configuration the run used, the generation settings, the selection seed that fixes which items the run drew, and the resulting per-category components. A score is therefore traceable to the exact item set that produced it, which is what makes randomised selection and reproducible results compatible claims rather than contradictory ones. The immutable evaluation log is retained as the evidence behind each entry.

Two things are deliberately not reproducible by a third party, and both follow from the benchmark's purpose. The held-out items are not released, because a benchmark whose scored partition is public stops

measuring anything the moment it is downloaded. And a run against a vendor's endpoint cannot be repeated by anyone who does not control that endpoint, which is the trust boundary Section 6.1 describes rather than a gap in the record. The public practice subset exists so that the format, the scoring and the rubric design can be examined and criticised without exposing the scored set.

VetEval runs on Inspect AI (UK AI Security Institute, MIT licensed). Runs pin the harness, the scorers, the generation settings and the model snapshot, and the pinned values travel with the entry they produced.

References

- Alonso Sousa, S., Bukhari, S. S. U. H., Steagall, P. V., Bęczkowski, P. M., Giuliano, A., & Flay, K. J. (2025). Performance of large language models on veterinary undergraduate multiple-choice examinations: a comparative evaluation. *Frontiers in Veterinary Science*, 12, 1616566. <https://doi.org/10.3389/fvets.2025.1616566> (PMID: 40933532)
- Angel, M., Patel, A., Xing, H., Balsz, D., Arbuckle, C., Bruyette, D., & Baldi, P. (2023). AI and Veterinary Medicine: Performance of Large Language Models on the North American Licensing Examination. In *2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE. <https://doi.org/10.1109/SNAMS60348.2023.10375414>
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). American Council on Education.
- Arora, R. K., Wei, J., Soskin Hicks, R., Bowman, P., et al. (2025). *HealthBench: Evaluating Large Language Models Towards Improved Human Health*. arXiv:2505.08775. <https://arxiv.org/abs/2505.08775>
- Bradley, R. A., & Terry, M. E. (1952). Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4), 324, 345. <https://doi.org/10.2307/2334029>
- Cameron, A. C., & Miller, D. L. (2015). A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, 50(2), 317, 372. <https://jhr.uwpress.org/content/50/2/317>
- Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., & Stoica, I. (2024). *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference*. arXiv:2403.04132. <https://arxiv.org/abs/2403.04132>
- Coleman, M. C., & Moore, J. N. (2024). Two artificial intelligence models underperform on examinations in a veterinary curriculum. *Journal of the American Veterinary Medical Association (JAVMA)*. <https://doi.org/10.2460/javma.23.12.0666> (PMID: 38382193)
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297, 334. <https://doi.org/10.1007/BF02310555>
- DiGiccio, T. J., & Efron, B. (1996). Bootstrap Confidence Intervals. *Statistical Science*, 11(3), 189, 212. <https://doi.org/10.1214/ss/1032280214>
- Downing, S. M. (2002). Threats to the Validity of Locally Developed Multiple-Choice Tests in Medical Education: Construct-Irrelevant Variance and Construct Underrepresentation. *Advances in Health Sciences Education*, 7(3), 235, 241. <https://doi.org/10.1023/A:1021112514626>

- Downing, S. M., & Haladyna, T. M. (2004). Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical Education*, 38(3), 327, 333. <https://doi.org/10.1046/j.1365-2923.2004.01777.x>
- Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of Educational Measurement* (5th ed.). Prentice Hall.
- Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1), 1, 26. <https://www.jstor.org/stable/2958830>
- Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC. <https://doi.org/10.1201/9780429246593>
- Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). Ragas: Automated Evaluation of Retrieval Augmented Generation. arXiv:2309.15217 (EACL 2024). <https://arxiv.org/abs/2309.15217>
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378, 382. <https://doi.org/10.1037/h0031619>
- Freyer, O., Wiest, I. C., Kather, J. N., & Gilbert, S. (2024). A future role for health applications of large language models depends on regulators enforcing safety standards. *The Lancet Digital Health*. [https://doi.org/10.1016/S2589-7500\(24\)00124-9](https://doi.org/10.1016/S2589-7500(24)00124-9)
- Gilson, A., Safranek, C. W., Huang, T., Socrates, V., et al. (2023). How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Medical Education*, 9, e45312. <https://doi.org/10.2196/45312> (PMID: 36753318)
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*, 15(3), 309, 333. https://doi.org/10.1207/S15324818AME1503_5
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Sage Publications.
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring Massive Multitask Language Understanding (MMLU). *International Conference on Learning Representations (ICLR)*. arXiv:2009.03300. <https://arxiv.org/abs/2009.03300>
- Hugging Face. Open LLM Leaderboard, Score Normalization documentation. Retrieved from https://huggingface.co/docs/leaderboards/open_llm_leaderboard/normalization
- Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., & Szolovits, P. (2020). What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams (MedQA / USMLE). arXiv:2009.13081. <https://arxiv.org/abs/2009.13081>
- Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., & Lu, X. (2019). PubMedQA: A Dataset for Biomedical Research Question Answering. *EMNLP 2019*. arXiv:1909.06146. <https://arxiv.org/abs/1909.06146>
- Kako, T., Kato, D., Iguchi, T., Qin, S., Ando, M., Koseki, S., et al. (2026). Performance evaluation of generative pre-trained transformer on the National Veterinary Licensing Examination in Japan. *Scientific Reports*. <https://doi.org/10.1038/s41598-026-37300-9> (PMID: 41698969)

- Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). GPT-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382. <https://doi.org/10.1098/rsta.2023.0254>
- Kheyami, D., Jaradat, A., Al-Shibani, T., & Ali, F. A. (2018). Item Analysis of Multiple Choice Questions at the Department of Paediatrics, Arabian Gulf University. *Sultan Qaboos University Medical Journal*, 18(1), e68, e74. <https://pmc.ncbi.nlm.nih.gov/articles/PMC5892816/>
- Krippendorff, K. (2018). *Content Analysis: An Introduction to Its Methodology* (4th ed.). Sage Publications.
- Kuder, G. F., & Richardson, M. W. (1937). The Theory of the Estimation of Test Reliability. *Psychometrika*, 2(3), 151, 160. <https://doi.org/10.1007/BF02288391>
- Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., et al. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2), e0000198. <https://doi.org/10.1371/journal.pdig.0000198> (PMID: 36812645)
- Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159, 174. <https://doi.org/10.2307/2529310>
- Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., et al. (2023). Holistic Evaluation of Language Models (HELM). *Transactions on Machine Learning Research (TMLR)*. arXiv:2211.09110. <https://arxiv.org/abs/2211.09110>
- Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). *G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment*. arXiv:2303.16634. <https://arxiv.org/abs/2303.16634>
- National Board of Medical Examiners (NBME). (2020). *Constructing Written Test Questions for the Basic and Clinical Sciences* (NBME Item-Writing Guide, 6th ed.). NBME.
- Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). *Capabilities of GPT-4 on Medical Challenge Problems*. arXiv:2303.13375. <https://arxiv.org/abs/2303.13375>
- Pal, A., Umapathi, L. K., & Sankarasubbu, M. (2022). MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. *Conference on Health, Inference, and Learning (CHIL 2022)*, PMLR. arXiv:2203.14371. <https://arxiv.org/abs/2203.14371>
- Pal, A., Umapathi, L. K., & Sankarasubbu, M. (2023). Med-HALT: Medical Domain Hallucination Test for Large Language Models. *EMNLP 2023 (CoNLL)*. arXiv:2307.15343. <https://arxiv.org/abs/2307.15343>
- Panickssery, A., Bowman, S. R., & Feng, S. (2024). *LLM Evaluators Recognize and Favor Their Own Generations*. arXiv:2404.13076. <https://arxiv.org/abs/2404.13076>
- Scale AI. SEAL Leaderboards (Safety, Evaluations, and Alignment Lab). <https://scale.com/blog/leaderboard>
- Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., & Zettlemoyer, L. (2024). Detecting Pretraining Data from Large Language Models (Min-K% Prob). *International Conference on Learning Representations (ICLR)*. arXiv:2310.16789. <https://arxiv.org/abs/2310.16789>
- Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., et al. (2023a). Large Language Models Encode Clinical Knowledge (MultiMedQA, Med-PaLM). *Nature*, 620, 172, 180. arXiv:2212.13138. <https://arxiv.org/abs/2212.13138>

- Singhal, K., Tu, T., Gottweis, J., Sayres, R., et al. (2023b). *Towards Expert-Level Medical Question Answering with Large Language Models* (Med-PaLM 2). arXiv:2305.09617. <https://arxiv.org/abs/2305.09617>
- Šlapeta, J. (2023). Are ChatGPT and other pretrained language models good parasitologists? *Trends in Parasitology*, 39(5), 314, 316. <https://doi.org/10.1016/j.pt.2023.02.006> (PMID: 36872153)
- Srivastava, A., Rastogi, A., Rao, A., et al. (2022). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models (BIG-bench). *Transactions on Machine Learning Research* (TMLR). arXiv:2206.04615. <https://arxiv.org/abs/2206.04615>
- Sutawika, L., Schoelkopf, H., Gao, L., Abbasi, B., Biderman, S., Tow, J., et al. (2024). *EleutherAI/lm-evaluation-harness* (v0.4.3) [software]. Zenodo. <https://doi.org/10.5281/zenodo.12608602>
- UK AI Security Institute. *Inspect AI: An open-source framework for large language model evaluations* [software, MIT license]. <https://inspect.aisi.org.uk/> ; https://github.com/UKGovernmentBEIS/inspect_ai
- Verga, P., Hofstätter, S., Althammer, S., Su, Y., Piktus, A., Arkhangorodsky, A., Xu, M., White, N., & Lewis, P. (2024). *Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models* (PoLL). arXiv:2404.18796. <https://arxiv.org/abs/2404.18796>
- Xu, C., Sun, Q., Zheng, K., Geng, X., Zhao, P., Feng, J., Tao, C., & Jiang, D. (2023). *WizardLM: Empowering Large Pre-Trained Language Models to Follow Complex Instructions* (Evol-Instruct). arXiv:2304.12244. <https://arxiv.org/abs/2304.12244>
- Yang, S., Chiang, W.-L., Zheng, L., Gonzalez, J. E., & Stoica, I. (2023). *Rethinking Benchmark and Contamination for Language Models with Rephrased Samples*. arXiv:2311.04850. <https://arxiv.org/abs/2311.04850>
- Zhang, H., Da, J., Lee, D., Robinson, V., Wu, C., Song, W., Zhao, T., Raja, P., Zhuang, C., Slack, D., Lyu, Q., Hendryx, S., Kaplan, R., Lunati, M., & Yue, S. (2024). A Careful Examination of Large Language Model Performance on Grade School Arithmetic (GSM1k). *NeurIPS 2024 Datasets & Benchmarks*. arXiv:2405.00332. <https://arxiv.org/abs/2405.00332>
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *NeurIPS 2023 Datasets & Benchmarks*. arXiv:2306.05685. <https://arxiv.org/abs/2306.05685>
- Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W. X., Chen, X., Lin, Y., Wen, J.-R., & Han, J. (2023). *Don't Make Your LLM an Evaluation Benchmark Cheater*. arXiv:2311.01964. <https://arxiv.org/abs/2311.01964>
- Zhu, L., Wang, X., & Wang, X. (2025). JudgeLM: Fine-tuned Large Language Models are Scalable Judges. *International Conference on Learning Representations* (ICLR). arXiv:2310.17631. <https://arxiv.org/abs/2310.17631>

Benchmark and methodology paper. VetEval is operated by viggoVet; standings are published at veteval.ai.