HivemindEval: Construction and Provisional Validation of an Open-Weight Verifier for UK/EU Compliance Analysis, with a Disclosed Rating-Collapse Failure Mode

Giovanni De Lillo
Hypereum Ltd, London, United Kingdom
https://hypereum.tech · https://huggingface.co/Hypereum/HivemindEval

Technical Report v1.0 · 17 July 2026

Status notice. Every benchmark number in this report is provisional. Band-level gold labels are construction- and adjudication-based; numeric gold labels are AI-grader- and human-range-based at the low and mid score bands only. No result has yet been validated by an independent human expert panel. Section 11 specifies the planned validation; its outcome will be published regardless of what it shows. This report makes no claim that the released model is the best available, state of the art, or at expert parity.

Abstract

We describe the construction, evaluation, and release of HivemindEval v2.1, an open-weight verifier that scores compliance findings for six UK and EU regulatory frameworks on six quality dimensions, producing structured JSON output with per-dimension reasoning. The model is a supervised fine-tune of Qwen3-8B, released under Apache-2.0 as merged bfloat16 weights, and runs on a single 24 GB GPU at approximately 40 tokens per second. We evaluate it on a frozen, contamination-gated benchmark of 230 synthetic compliance findings (212 rubric-scorable), constructed from tiered quality classes and four injected defect classes, with the full item set committed by SHA-256 hash and a 68-item stratified subset published with per-item gold labels and the raw predictions of every compared model. On this benchmark the released model reaches a cross-tier rank concordance of 0.984 (95% CI 0.973–0.993), a statistical tie with a rubric-prompted Qwen2.5-72B baseline (0.985, CI 0.977–0.993), while leading on Kendall's τ_b (0.817 vs. 0.801) and running on hardware roughly one class smaller. A matched ablation isolates what the fine-tuning measurably buys: the same base model under the same native prompt produced zero parseable evaluations out of 212, whereas the fine-tuned model produced 211, from a prompt roughly ten times shorter than the scoring rubric a general model requires. The model is weakest exactly where we say it is: it ranks last among scored models on defect separation, and a rubric-prompted base 8B outperforms it on wrong-jurisdiction detection (1.00 vs. 0.90). We additionally disclose a complete failure mode of the withdrawn predecessor, v2.0, which emitted

an identical score vector for every input while producing syntactically flawless JSON, and we formalize the variance-based acceptance gates that now block such models from release. Finally, we present evidence for a distillation ceiling on defect discrimination: the AI reference grader itself separates the model's weakest defect class at only 0.532, which motivates the pre-registered human expert validation this report commits to. All evaluation artifacts required to reproduce the published leaderboard bit-for-bit are public.

1 Introduction

Automated compliance analysis is increasingly produced by large language models: systems draft findings, cite regulatory provisions, assign severities, and propose remediations. Whether such output can be trusted is a separate question from whether it can be produced, and in regulated sectors the two questions have different owners. A finding that cites the wrong jurisdiction's law, omits a required control, or assigns an unjustified severity is not merely low-quality text; it is an input to decisions that carry legal exposure.

Verification of this output is therefore its own task. The dominant approach, using a frontier model as a judge (Zheng et al., 2023; Liu et al., 2023), is effective but carries two costs that matter specifically in regulated deployments. First, the material being judged (the compliance posture of a bank, a hospital, a defence contractor) is often the most sensitive artifact the organisation produces, and sending it to an external API may itself be a compliance event. Second, judge behaviour behind a commercial API can change without notice, which is difficult to reconcile with audit requirements that assume a fixed, referenceable instrument.

Open-weight evaluator models (Kim et al., 2023, 2024; Li et al., 2024) address controllability and transparency for general evaluation tasks. They are not, however, specialised for regulatory verification, and the question of how to validate a domain evaluator honestly (in a setting where the natural source of training labels is itself a language model) has received less attention than the question of how to train one.

This report describes one attempt to do both, and to be precise about what was and was not achieved. We built HivemindEval, an 8-billion-parameter open-weight verifier for compliance findings under six UK/EU frameworks (PSD2 SCA-RTS; UK GDPR together with the NHS Data Security and Protection Toolkit; MOD JSP 440; Cyber Essentials Plus; DORA; and the EU AI Act). We constructed a frozen benchmark with explicit contamination controls, defined a gold-label policy that restricts each label source to the region where it was validated as sound, evaluated the model against larger open baselines and a frontier reference under identical conditions, and released the weights, the benchmark subset, the gold labels, every model's raw predictions, and the scoring code.

The results do not support a superiority claim, and we do not make one. They support a narrower and, we believe, more useful set of statements:

1. **Rank concordance at reduced scale.** On the frozen benchmark, the 8B fine-tune is statistically indistinguishable from a rubric-prompted 72B open model on cross-tier rank concordance, while leading on Kendall's τ_b, on hardware roughly one GPU class smaller (Section 7.1).
2. **Format reliability is the measurable purchase of fine-tuning, and it is not judgment quality.** A matched ablation (same base model, same native prompt) yields 0/212 parseable outputs; the fine-tune yields 211/212. We report format reliability and ranking quality separately throughout, because a model can produce perfect JSON and zero signal (Section 7.3), and our own v2.0 did exactly that (Section 8).

3. **Honest weakness accounting.** The model ranks last among scored models on defect separation; a rubric-prompted base 8B beats it on the wrong-jurisdiction class. If defect detection rather than quality ranking is the primary need, a rubric-prompted model is currently the better tool (Section 7.2).
4. **A distillation ceiling on defect discrimination.** Before attempting a targeted retrain of the weakest defect classes, we measured whether the AI reference grader that labels the training data could itself discriminate them. It could not (separation 0.532 on the target class). Distilling from a grader that shares the blind spot cannot repair it; only independent human numeric gold can. This finding converted a planned retrain into a pre-registered human validation (Sections 9 and 11).
5. **A disclosed failure mode with a formalized gate.** The withdrawn v2.0 emitted a constant score vector for every input while producing well-formed JSON. We describe the defect, its root cause in the label pipeline, and the variance-based acceptance gates that now constitute necessary release criteria (Section 8).

A note on scope. This report publishes the validation science in full: benchmark construction, contamination controls, gold policy, metric definitions, acceptance gates, results including losses, and the human-validation design. It withholds the training engineering: the data-generation pipeline, the identity and configuration of the teacher model, and the adapter hyperparameters. Section 12 states this boundary explicitly and gives the reasons. Reproducibility of the *evaluation* is complete and bit-exact from published artifacts; reproducibility of the *training* is not offered.

2 Related Work

LLM-as-a-judge. Zheng et al. (2023) established strong-model judging as a scalable proxy for human preference evaluation and documented its systematic biases. G-Eval (Liu et al., 2023) formalized rubric-guided scoring with chain-of-thought for natural-language generation. Both lines assume access to a frontier model at evaluation time, which conflicts with the data-residency constraint motivating this work.

Open-weight evaluators. Prometheus (Kim et al., 2023) and Prometheus 2 (Kim et al., 2024) train open evaluator models that approach proprietary-judge agreement on general tasks, supporting direct assessment and pairwise ranking under user-supplied criteria. Auto-J (Li et al., 2024) trains a generative judge for alignment evaluation. RewardBench (Lambert et al., 2024) benchmarks reward and judge models systematically. HivemindEval differs in three ways: it targets a single narrow domain with a fixed input and output schema rather than general evaluation; its validation protocol, rather than its training method, is the primary contribution; and its report treats the absence of human-anchored gold as a first-class limitation rather than an appendix caveat.

Evaluation integrity. Our contamination controls follow the general principle that test data must be provably disjoint from training data; we implement this with content-hash disjointness proofs over canonicalized items (Section 4.3). The commitment scheme (publishing the SHA-256 of a withheld full set alongside an open subset) borrows from standard cryptographic commitment practice to make later substitution detectable. The release documentation follows the intent of Model Cards (Mitchell et al., 2019) and Datasheets for Datasets (Gebru et al., 2021).

Calibration. Expected calibration error and its variants (Guo et al., 2017) motivate our planned score-calibration analysis; isotonic regression via the pool-adjacent-violators algorithm (Barlow et al., 1972) provides the rank-preserving recalibration method reserved for expert-anchored labels (Section 11.3).

Serving. Deployment measurements use vLLM (Kwon et al., 2023). Constrained decoding for schema enforcement follows the guided-generation line of work (Willard and Louf, 2023); we report observed validity separately from grammar-guaranteed validity (Section 6.5). The base model is Qwen3-8B (Qwen Team, 2025).

3 The Released Artifact

3.1 Task definition

The model performs single-item direct assessment. The input is one compliance finding serialized as JSON with fields `severity`, `category`, `title`, `description`, `evidence`, `regulatory_reference`, `remediation`, and optional `deadline` and `effort_days`. The output is one JSON object containing an integer `overall_score` in $[0, 100]$; a `dimensions` object with integer scores and free-text reasoning for six dimensions (`accuracy`, `completeness`, `regulatory_alignment`, `actionability`, `coherence`, `evidence_quality`); and an `improvement_suggestions` array. The full schema is reproduced in Appendix A.

The intended interpretation of the overall score is ordinal quality within the domain: higher scores indicate findings a competent reviewer would rank as better. Absolute calibration of the scale (whether a 72 "means" what a human expert's 72 means) is explicitly unvalidated at the high band (Section 5.3) and is the object of the planned expert study.

3.2 Model and interface

HivemindEval v2.1 is a supervised fine-tune of Qwen3-8B (Apache-2.0), released as full merged bfloat16 safetensors under Apache-2.0. The native interface is a fixed short system prompt (reproduced on the model card) followed by the raw finding JSON as the user message, greedy decoding at temperature 0, with the base model's thinking mode disabled. Outputs are verbose; a generation budget of at least 2,048 new tokens is required to avoid truncating the JSON before it closes. The published inference harness includes a validation-gated parser; for production use we recommend enforcing the schema at the serving layer with guided decoding, since the validity figures reported here are observed frequencies, not grammar guarantees.

3.3 What is released and what is withheld

Released: merged weights; the model card with all results in this report; a 68-item benchmark subset with per-item gold and the raw per-item predictions of all six compared models; the scoring implementation that reproduces every published number bit-for-bit offline; the inference harness; and the methodology documentation. Withheld: the full 230-item benchmark (committed by hash; retained as a private validation instrument), the training corpus and its generation pipeline, the teacher model's identity and configuration, and the adapter hyperparameters. Section 12 discusses this boundary; Section 4.5 gives the commitment hashes.

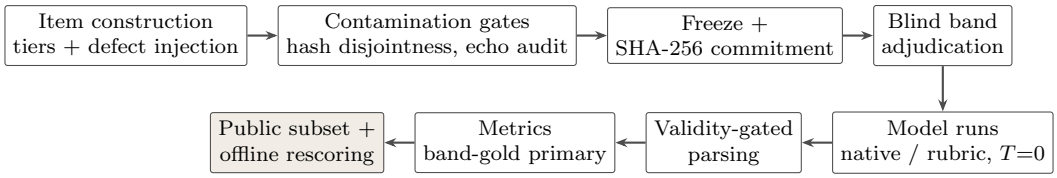

Figure 1: Validation pipeline. Items are constructed, contamination-gated, frozen under a hash commitment, and band-adjudicated before any model runs; parsing is validity-gated; primary metrics use band gold only; the public subset makes the leaderboard recomputable offline (shaded).

4 Benchmark Construction

4.1 Design goals

The benchmark was designed to answer ordinal questions (does the model rank better findings above worse ones, and does it rank defective findings below clean ones) rather than to certify absolute scores, because the available gold sources are only trustworthy for ordinal claims (Section 5). Four properties were treated as requirements: stratification over the quality range and over the frameworks; explicit defect classes with known ground truth by construction; provable disjointness from the model's training data; and a freeze protocol that makes post-hoc item substitution detectable.

4.2 Item construction

The frozen set contains 230 synthetic-but-realistic compliance findings, of which 212 are rubric-scorable and 18 are auxiliary items excluded from scoring (retained for provenance and instrument checks). Items were constructed along two axes.

Quality tiers. Clean items were written to one of four target tiers (excellent, good, mediocre, terrible) by controlled degradation of the same underlying scenario: an excellent item cites the correct provisions with dated evidence and a complete remediation; a terrible item is vague, uncited, and inactionable. The tier assignment is by construction and is later cross-checked by blind adjudication (Section 4.4).

Injected defect classes. Four defect classes were injected into otherwise plausible findings: `wrong_jurisdiction` (a citation from another jurisdiction's framework presented as operative: for example, a US health-privacy provision cited as governing an NHS integration); `hallucinated` (a fabricated control or citation); `missing_requirement` (omission of a control the framework requires for the described scenario); and `band_edge` (items constructed to sit just inside a tier boundary, where the correct tier is unambiguous but close). Defect items carry ground truth by construction: a correct evaluator should score a defective item below a clean good-tier item of the same scenario family.

Coverage spans all six frameworks; in the published subset each framework contributes 8–12 items (Appendix D). Item counts per stratum in the full set are recorded in the frozen manifest.

4.3 Contamination controls

Two controls were enforced before freezing.

Hash disjointness. Every scored item was canonicalized (whitespace- and key-order-normalized serialization) and hashed with SHA-256; the resulting content hashes were checked for intersection against the content hashes of every corpus used to train the released model, including all seed pools. The intersection is empty: 0 collisions across all 212 scored items. The disjointness check is part of the released tooling and was re-run as a release gate.

Echo-signal audit. An earlier internal instrument was found to leak the expected answer into the item text (an artifact we refer to as echo contamination: an item that names its own defect measures string matching, not judgment). Every benchmark item was audited for self-signal; items are required to present the defect without labelling it.

4.4 Freeze and adjudication pass

The item set was frozen before any compared model was run. A blind adjudication pass by an independent arbiter model (Google `gemini-2.5-flash-lite`; never a source of numeric labels or training signal) reviewed every item's construction-assigned band: 134 items were confirmed as constructed, 50 were retained with an adjacent-band annotation, 45 subtle-defect items were retained with their defect confirmed, and 1 disputed item was excluded from scoring. The arbiter's role is limited to band-level cross-checking; Section 5.1 explains why this preserves label-source independence for the primary metrics.

4.5 Commitments and the public subset

The full frozen set is committed by hash and withheld:

```
Full item set (230):     4f2dbff713ab6b56d2db0a8eb08a1296f86bca0229f55d8ca127a73d21e1a9c6
Published subset (68):   b51e721e37df6e9917f44550ccb2ea2376839d1c7744818a1fd432fc12a4eee1
```

The 68-item public subset was selected by stratified sampling with deterministic ordering by content hash (a mechanical rule chosen so that the subset cannot be cherry-picked) covering all four defect classes, all four quality tiers, and all six frameworks. The subset ships with per-item gold and the raw per-item predictions of all six compared models, which makes the published core-68 leaderboard exactly recomputable offline (Section 6.6). Withholding the remainder serves two purposes: it keeps a private validation instrument for future model versions, and it limits public disclosure of the full defect-coverage map. The hash commitment makes the withheld set verifiable if it is ever disclosed or audited.

5 Gold-Label Policy

The central methodological problem of this work is that the cheapest label source (a strong language model) is also an unreliable one in specific, measurable ways. Rather than average over that unreliability, the gold policy restricts each label source to the region where it was validated as sound.

5.1 Band gold: label-source independent

The primary metrics (Section 6.3) are rank- and band-based and are invariant under any strictly monotone transformation of a model's scores. Their gold is the *band*: the construction-assigned tier,

cross-checked by the blind adjudication pass of Section 4.4. No numeric label from any AI system enters the primary metrics. This is the design decision that allows the leaderboard to compare our own teacher-distilled model against frontier judges without structural favouritism: a capped or shifted numeric scale cannot change a model's concordance or τ.

5.2 Numeric gold: restricted to its validity domain

Numeric gold exists only where a source was validated: human score ranges on a set of graded probe items, and an LLM reference grader (identity withheld; Section 12) at the low and mid bands, where its rankings agree with the human-ranged probes. Mean absolute error against this low/mid numeric gold is reported as a *secondary* metric only, with an explicit conflict-of-interest disclosure: the released model's training labels share provenance with that gold, so its MAE is structurally favoured. We publish the number (Section 7.1) because hiding a favourable metric would be its own distortion; we instruct readers to treat it as a sanity check, never a ranking.

5.3 The high-band exclusion, and why it exists

During instrument validation we measured the reference grader against twelve human-range-graded probes and found a content-conditional compression at the top of the scale: on archetypal excellent items the grader scores in range, but on *realistic* excellent items (the kind the benchmark actually contains, with dated evidence and precise citations rather than textbook perfection) its probability of scoring at or above the human gold floor was zero on the probe set:

$$\widehat{P}\big(\hat{s}_{\text{ref}} \geq 78 \,\big|\, \text{realistic excellent}\big) = 0.00, \qquad \widehat{P}\big(\hat{s}_{\text{ref}} \geq 78 \,\big|\, \text{archetypal excellent}\big) = 1.00. \tag{1}$$

The bias is deterministic at temperature 0 (reproduced scores 69/72/76 on the three realistic excellent probes against human gold floors of 78–82) and its per-dimension reasoning approves what its numbers under-score, indicating score–reasoning decoupling rather than noise. Two consequences follow. First, the 75 high-band items carry no numeric gold: they are marked `PENDING_EXPERT` and are excluded from every provisional numeric metric. Publishing grader-capped high-band gold would flatter the grader-distilled model and penalize frontier judges: a leaderboard distorted in our own favour, which we consider worse than no leaderboard. Second, the released model inherits a version of the same compression (Section 7.4), which is disclosed as a limitation and is one of the two questions the expert study is designed to settle.

6 Evaluation Protocol

6.1 Compared models

Six systems were run on identical items: the released model under its native interface; three open-weight baselines, namely Qwen2.5-72B-Instruct-AWQ (4-bit, self-served), Llama-3.3-70B, and the base Qwen3-8B, each prompted with a ∼10.5k-token scoring rubric (the strongest general-model configuration we could produce for this task, hereafter *rubric-prompted*); the base Qwen3-8B under the released model's *native* short prompt, as the matched ablation; and one frontier reference (`claude-sonnet-5`, rubric-prompted) on the public subset, included as a not-open reference point rather than a competitor in the open-weight comparison.

6.2 Decoding, parsing, and validity accounting

All systems ran at temperature 0. Outputs were parsed by the released validation-gated parser; an item counts as *scored* for a model only if the output parses into the schema with all required fields. Validity is reported per model as scored/212 and is analysed as its own reliability property, separate from ranking quality. Reported ranking metrics for a model are computed over its scored items; the scored count is printed beside every row so that validity differences are never silently absorbed into ranking numbers.

6.3 Metrics

Let items be indexed by i with model score $s_i \in [0, 100]$ and construction band $t_i \in \{1, 2, 3, 4\}$ (terrible to excellent). Ties receive half credit throughout.

Cross-tier rank concordance ("tier concordance"): with $\mathcal{P} = \{(i, j) : t_i > t_j\}$,

$$C_{\text{tier}} \;=\; \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \Big[\mathbf{1}\{s_i > s_j\} \;+\; \tfrac{1}{2}\,\mathbf{1}\{s_i = s_j\}\Big]. \tag{2}$$

Excellent-band concordance. C_{exc} is C_{tier} restricted to $\mathcal{P}_E = \{(i, j) \in \mathcal{P} : \max(t_i, t_j) = 4\}$, isolating ordering behaviour at the top of the scale.

Defect separation. For defect class c with defect items Δ_c and clean good-tier items $\mathcal{G}$,

$$D_c \;=\; \frac{1}{|\mathcal{G}|\,|\Delta_c|} \sum_{g\in\mathcal{G}} \sum_{d\in\Delta_c} \Big[\mathbf{1}\{s_d < s_g\} \;+\; \tfrac{1}{2}\,\mathbf{1}\{s_d = s_g\}\Big], \tag{3}$$

the probability that a defective item is scored below a clean good item. Overall defect separation pools all defect items. $D_c = 1$ is perfect discrimination; $D_c = 0.5$ is chance.

Kendall's τ_b between scores and bands, with the standard tie correction:

$$\tau_b \;=\; \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}, \tag{4}$$

where n_c, n_d are concordant and discordant pairs, $n_0 = \binom{n}{2}$, and n_1, n_2 are the tie terms in scores and bands respectively.

Secondary numeric error. On the low/mid numeric-gold subset $\mathcal{N}_{\ell m}$ ($|\mathcal{N}_{\ell m}| = 54$) with gold g_i,

$$\text{MAE}_{\ell m} \;=\; \frac{1}{|\mathcal{N}_{\ell m}|} \sum_{i\in\mathcal{N}_{\ell m}} |\, s_i - g_i \,|. \tag{5}$$

These definitions are normative for this report; the released scorer is the reference implementation, and any discrepancy between prose and code resolves in favour of the code, which is public.

Table 1: Full board (212 items, open models). CI: 95% percentile bootstrap. MAE is the disclosed conflict-of-interest secondary metric of Section 5.2.

Model	Tier conc. (95% CI)	Excellent	Defect sep.	τ_b	MAE ℓm	Scored
Qwen2.5-72B-Instruct-AWQ (rubric)	**0.985** [0.977–0.993]	**0.990**	0.602	0.801	12.64	212/212
HivemindEval v2.1 (native)	0.984 [0.973–0.993]	0.982	0.591	**0.817**	4.00	211/212
Qwen3-8B (rubric)	0.973 [0.959–0.985]	0.973	**0.709**	0.751	10.00	208/212
Llama-3.3-70B (rubric)	0.965 [0.950–0.980]	0.968	0.616	0.764	17.52	173/212
Qwen3-8B (native prompt, no fine-tune)		no parseable output				0/212

6.4 Uncertainty

Confidence intervals on tier concordance are percentile bootstrap intervals over items: $B = 400$ resamples with replacement, seed 7, reporting the 2.5th and 97.5th percentiles. The released scorer reproduces the intervals exactly. τ_b is reported without an interval in the current scorer; this asymmetry is disclosed wherever τ_b is compared (Section 10).

6.5 Determinism bounds

Serving-stack outputs at temperature 0 are *verdict-stable* rather than byte-stable: identical requests can produce token-level divergence under continuous batching, because floating-point reduction order varies with batch composition and floating-point addition is not associative. We therefore define reproducibility at the verdict level (band assignments and gate verdicts invariant across repeats) and document byte-level reproducibility on a serving stack as a non-goal. In repeat probes, 13 of 14 items showed token-level jitter at temperature 0 while band verdicts remained stable away from tier boundaries; boundary-adjacent items are exactly where band flips concentrate, which motivates flagging boundary-adjacent scores in deployment rather than asserting hard bands there.

6.6 Offline reproducibility

The published core-68 leaderboard, including bootstrap intervals, is recomputable bit-for-bit from the released gold and prediction files with a single command and no GPU:

```
python3 eval/score_benchmark.py --gold data/gold_subset.json --predictions-dir
data/predictions
```

Regenerating predictions from the released weights on a 24 GB GPU reproduces the shipped predictions to verdict-stable agreement under the determinism bounds above; this was verified end-to-end against the published repository artifacts before release.

7 Results

7.1 Main boards

Table 1 reports the full frozen board (212 scored items, open models); Figure 2 shows the tier-concordance intervals. Table 2 reports the published 68-item subset, which adds the frontier reference and is exactly reproducible offline.

Reading the comparison honestly: the 72B baseline edges the released model on tier and excellent concordance, and the released model leads on τ_b. All tier-concordance gaps among the top models

Table 2: Core-68 public subset (adds the frontier reference; recomputable offline).

Model	Tier conc. (95% CI)	Excellent	Defect sep.	τ_b	MAE ℓm	Scored
Qwen3-8B (rubric)	0.991 [0.977–1.000]	0.988	0.639	0.788	9.32	67/68
Qwen2.5-72B-AWQ (rubric)	0.990 [0.980–1.000]	0.992	0.549	0.833	14.40	68/68
HivemindEval v2.1 (native)	0.979 [0.958–0.998]	0.976	0.615	**0.843**	4.30	67/68
claude-sonnet-5 (rubric; not open)	0.976 [0.951–0.996]	0.975	0.562	0.777	3.35	66/68
Llama-3.3-70B (rubric)	0.964 [0.932–0.988]	0.963	0.667	0.798	21.00	56/68

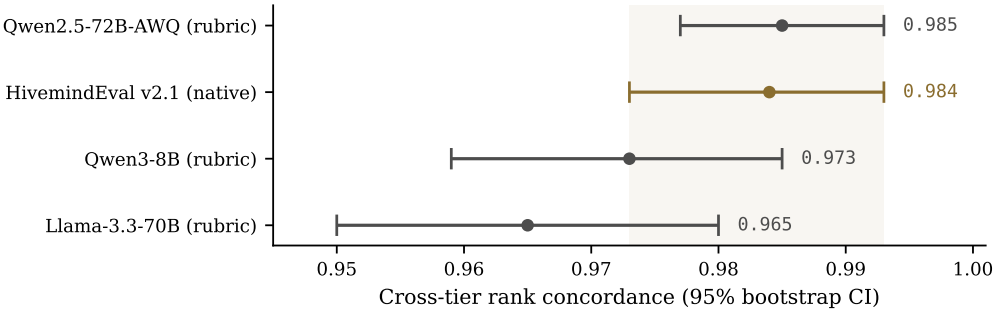

Figure 2: Cross-tier rank concordance with 95% bootstrap intervals, full board. The shaded band spans the released model's interval; the top two rows are a statistical tie. The released model (accent) is highlighted only where it leads or ties.

fall within overlapping 95% intervals. This is a statistical tie, not a win, and the near-ceiling clustering (0.96–0.99 across capable models) indicates that the benchmark's cross-tier ordering task is close to saturated for this model class: a property to correct in future benchmark versions with harder ordinal structure, not a property to advertise. The MAE column reflects shared label provenance for the released model (Section 5.2) and is printed for completeness, not for ranking.

7.2 Defect separation by class: the weakness, stated plainly

The released model is last among scored models on pooled defect separation (0.591), weakest on `missing_requirement` (0.42) and `band_edge` (0.36), and is beaten on `wrong_jurisdiction` (a core capability for this domain) by a rubric-prompted copy of its own base model (1.00 vs. 0.90). On the subset, the frontier reference scores 0.54 on `wrong_jurisdiction`, which indicates that frontier scale does not automatically confer this capability either; the rubric configuration appears to matter more than parameter count for this class. Practical guidance follows directly: for defect-screening workloads, use a rubric-prompted model; the released model's comparative value lies in ranking reliability, output structure, and deployment footprint, not in defect detection.

7.3 What the fine-tuning measurably buys: a matched ablation

The base Qwen3-8B, given the released model's exact native interface, produced prose rather than JSON on every one of 212 items: 0/212 parseable (Figure 4). The fine-tuned model produced valid schema output on 211/212 (99.5%), and on 12/12 acceptance-gate probes, from a system prompt roughly one paragraph long. The rubric that makes the base model competitive on ranking runs to

Table 3: Per-class defect separation, full board (D_c; 1.0 = perfect, 0.5 = chance).

Model	`wrong_jurisdiction`	`hallucinated`	`missing_requirement`	`band_edge`
Qwen3-8B (rubric)	**1.00**	0.59	**0.65**	**0.60**
Qwen2.5-72B-AWQ	0.89	0.59	0.46	0.46
Llama-3.3-70B	0.85	0.57	0.51	0.53
HivemindEval v2.1	0.90	**0.67**	0.42	0.36

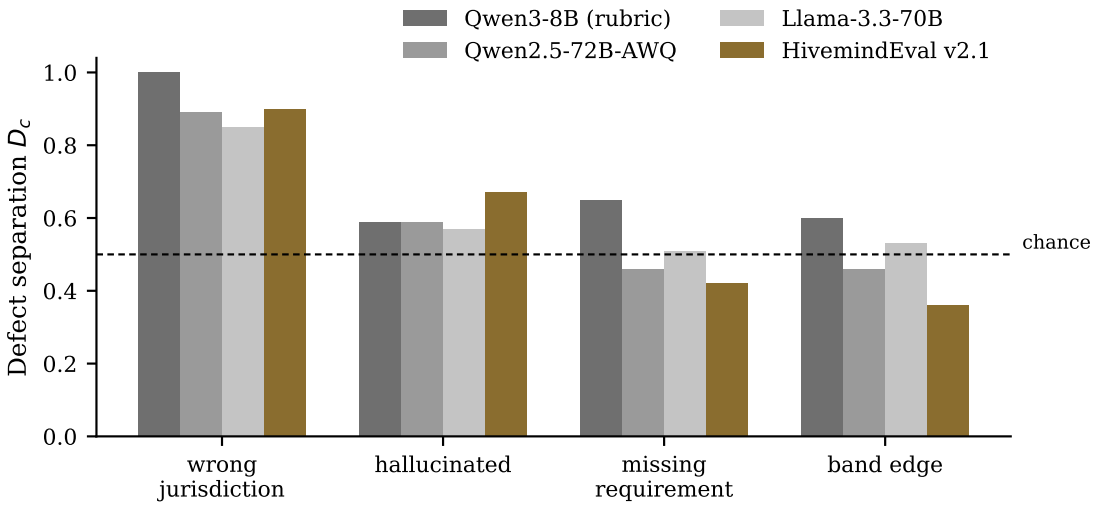

Figure 3: Defect separation by class (Table 3). The released model's two weakest classes, `missing_requirement` and `band_edge`, sit below every rubric-prompted baseline; a rubric-prompted copy of its own base model is perfect on `wrong_jurisdiction`.

~10.5k tokens per call (roughly ten times the prompt cost) and does not fit the 8k serving context used in the reference deployment. The fine-tuning therefore buys two things that can be measured: schema reliability at the native interface, and prompt economics; and it matches, within confidence intervals, the ranking behaviour of a model roughly nine times larger on hardware one class smaller. It does not buy ranking superiority, and Table 3 shows it costs defect-detection sharpness relative to the rubric-prompted base. We consider this trade profile a finding, not a defect of the analysis: it is precisely the decomposition (format versus judgment) that Section 8 shows can otherwise hide a completely broken model.

7.4 Acceptance gate and score-range behaviour

The release gate (formalized in Section 8.3) evaluates the final merged weights on twelve human-range-graded probes with a 2,048-token evaluation budget. v2.1 passes all criteria: overall score standard deviation 23.33 (threshold > 8), per-dimension standard deviations 21.1–31.9 (each > 5), best-minus-worst tier spread 51.33 (> 20), strict tier-mean monotonicity with zero violations, and jurisdiction-defect probe mean 22.0 (< 50), with 12/12 schema-valid outputs (Appendix B).

Per-tier means against human gold ranges show a conservative top of the range (Figure 5): excellent 72.0 against gold floors of 78–82; good 63.0; mediocre 35.0; terrible 20.7; jurisdiction-defect

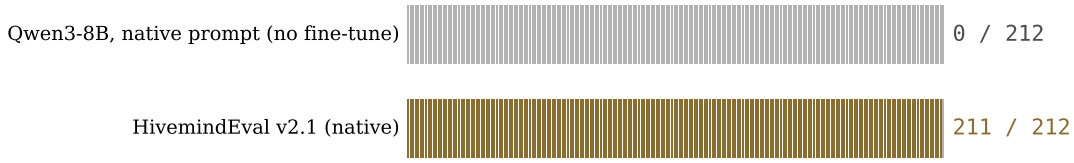

Figure 4: Format reliability under the identical native interface, 212 benchmark items. Each tick is one item; filled ticks parsed into the schema. The base model produced prose on every item.

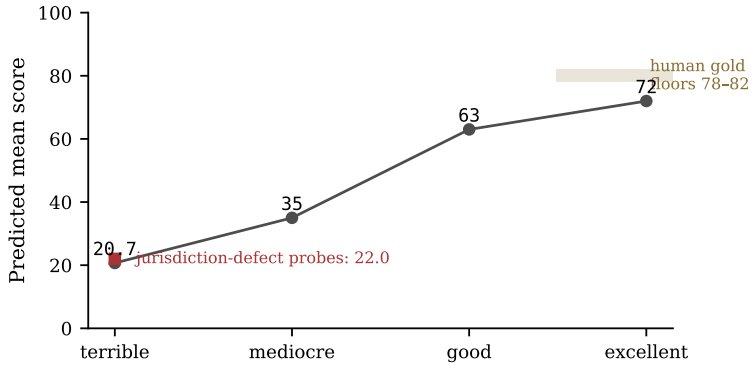

Figure 5: Per-tier predicted means (v2.1) against human gold floors. Ordering is strictly monotone with a 51.3-point spread; the excellent tier is compressed below the human floor band, a limitation inherited from the label source and awaiting expert-anchored recalibration (Section 11.3).

22.0. The ordering is correct and well separated; the absolute level at the top is compressed, consistent with the reference-grader compression the training labels inherit (Section 5.3). Deployment guidance on the model card reflects this: scores in the high 60s to low 70s should be treated as potentially excellent-tier pending expert-anchored recalibration.

7.5 Deployment profile

On a single 24 GB consumer GPU (RTX 4090 class) with vLLM at `--max-model-len 8192`, the model serves at approximately 40 tokens per second single-stream with 23.4 of 24.5 GB in use: a working configuration with minimal headroom, suitable for single-stream evaluation; concurrent serving requires a larger card and is unmeasured. Cold start from a fresh host measured 25 minutes 16 seconds, dominated by weight transfer (about 5 minutes when weights are staged in nearby object storage). These figures are single-configuration observations, not audited throughput benchmarks, and are reported as a hardware-tier comparison: the 72B baseline in 4-bit quantization requires at least a 46 GB-class device to serve at all.

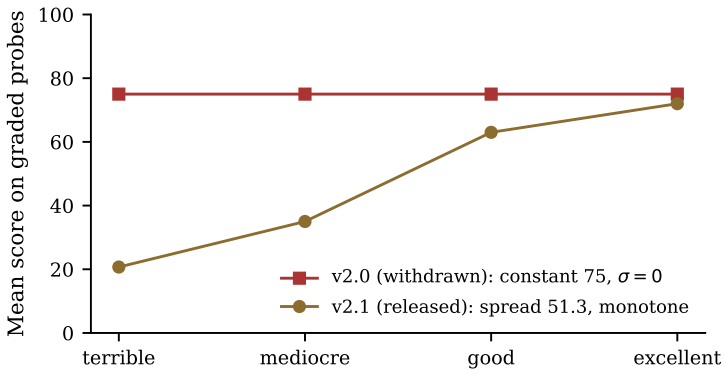

Figure 6: The v2.0 failure and the v2.1 repair on graded probes. v2.0 scored every tier identically at 75 with zero variance while emitting flawless JSON; v2.1 restores a monotone, well-separated response. Format validity is not judgment quality.

8 Case Study: The v2.0 Rating Collapse

8.1 The failure

The predecessor version, v2.0, was briefly visible on the public model hub beginning May 2026 before being withdrawn. In internal validation it exhibited complete rating collapse: it emitted the identical `overall_score` of 75, with an identical per-dimension score vector, for every input regardless of quality (standard deviation 0.0 across a graded probe set spanning excellent to terrible) while producing syntactically flawless, schema-conformant JSON on every call (Figure 6). The model was, in the operative phrase of the internal report, syntactically valid and semantically inert. Third-party copies made during the public window may persist; the model card instructs anyone holding v2.0 weights not to use them.

8.2 Root cause

The defect was traced to the training labels, not the architecture or the optimization: a data-generation fault propagated a constant placeholder score vector into the label set, so the corpus taught a constant function. No model trained on those labels could have learned to discriminate, and the failure is invisible to every check that inspects only output form: loss curves, JSON validity rates, and schema conformance all looked healthy. The general lesson, reflected throughout this report's separation of format from judgment, is that **format validity is not judgment quality**: a model can emit flawless structure and zero signal, and a validation battery that does not measure score variance against graded inputs will pass it.

8.3 The gates, formalized

The failure produced a set of variance- and ordering-based acceptance criteria that are now necessary conditions for release. Let $\{s_i\}$ be overall scores over the graded probe set, $s_i^{(d)}$ the score on

dimension d, and $\bar{s}_T$ the mean over probes of tier T:

$$\textbf{G1 (variance):} \quad \sigma(\{s_i\}) > 8. \tag{6}$$

$$\textbf{G2 (per-dimension variance):} \quad \sigma(\{s_i^{(d)}\}) > 5 \ \text{ for every dimension } d. \tag{7}$$

$$\textbf{G3 (spread):} \quad \bar{s}_{\text{exc}} - \bar{s}_{\text{terr}} > 20. \tag{8}$$

$$\textbf{G4 (ordering):} \quad \bar{s}_{\text{exc}} > \bar{s}_{\text{good}} > \bar{s}_{\text{med}} > \bar{s}_{\text{terr}}, \text{ zero pairwise violations.} \tag{9}$$

$$\textbf{G5 (defect response):} \quad \text{mean score on wrong-jurisdiction probes} < 50. \tag{10}$$

A collapsed model fails G1–G3 immediately. These gates are deliberately weak as quality measures (passing them proves a pulse, not competence) and deliberately strong as release blockers: they are cheap, deterministic, and they catch the entire class of degenerate models that format checks admit. v2.1's gate results appear in Section 7.4; the benchmark of Section 4 provides the competence measurement the gates do not.

8.4 Disclosure rationale

We disclose this failure in the model card and in this report for three reasons. Practically, third-party copies of the broken weights may still circulate, and users deserve an unambiguous do-not-use notice. Methodologically, the failure is the strongest available argument for the format/judgment separation this report maintains. Institutionally, an evaluation artifact asks to be trusted; a validation programme that has demonstrably caught its own worst failure is more credible than one that reports only successes.

9 Why There Is No v2.2: A Distillation Ceiling on Defect Discrimination

The natural response to Table 3 is a targeted retrain: supplement the training mix with dense `missing_requirement` and `band_edge` exemplars and fine-tune again. We prepared exactly that (92 supplement items, contamination-checked against the frozen benchmark with zero collisions) and then applied a viability pre-flight before spending any training compute: *can the label source discriminate the target class?*

It cannot. On the supplement's target class, the reference grader's own defect separation measured 0.532 on the overall score (0.604 on the completeness dimension): barely above the 0.42 it is meant to repair and close to chance. Its free-text reasoning identifies the omissions; its numeric scores barely penalize them, the same score–reasoning decoupling observed at the high band (Section 5.3). Distillation transfers the teacher's scoring function; a student distilled from a grader that does not numerically separate a defect class cannot learn to separate it, regardless of exemplar density. The planned retrain was therefore cancelled at zero training cost, and the conclusion generalizes: the released model's defect-detection weakness is bounded by its label source, not by its capacity. A larger student under the same labels inherits the same ceiling. The only route to genuine improvement on these classes is numeric gold from a source that does discriminate them, which, for this domain, means human experts, and which converts the engineering question into the validation study of Section 11.

A related scale analysis supports stopping rather than growing. On a benchmark where capable models cluster within overlapping confidence intervals at 0.96–0.99 concordance, moving to a 14B or 32B base can be expected to purchase a within-interval ranking change (statistically unresolvable

on the current instrument) while multiplying the per-deployment serving footprint from a 24 GB card into the 48–80 GB class. Under teacher-anchored gold, that is spending the product's principal deployment advantage to buy noise. We record this as the reasoned basis for freezing the model at 8B pending expert-anchored evidence of a real deficit.

10 Threats to Validity

Construct validity. Items are synthetic. They were written to be realistic and were adjudicated blind, but no claim is made that the tier structure of constructed findings matches the quality distribution of production compliance output. Tier membership is by construction; an item's "true" quality is defined by the construction protocol, not by external annotation.

Gold validity. Band gold is label-source independent, but the adjudication arbiter is itself a language model; its band-level cross-check could share systematic blind spots with other models. Numeric gold at the low/mid band descends from an AI grader validated only against a small human-ranged probe set. Inter-rater agreement with human experts (weighted κ, Krippendorff's α, ICC) is *unmeasured*, not merely unreported. Until the panel of Section 11 exists, every number here is provisional, and the report's own status notice governs.

Statistical validity. The primary ordering task is near saturation for capable models, compressing differences into overlapping intervals; conclusions about relative ranking quality among the top three systems are correspondingly weak. τ_b is reported without a confidence interval in the current scorer, an asymmetry that slightly over-privileges the one metric on which the released model leads; we flag rather than repair it in this version. Bootstrap intervals treat items as exchangeable and ignore the stratified construction; a stratified bootstrap is future work.

Internal validity. The MAE conflict of interest is disclosed in Sections 5.2 and 7.1. The rubric-prompted baselines were tuned in good faith to be strong, but rubric engineering is open-ended; a better rubric could move baseline numbers. All rubric materials ship with the harness so that this can be tested rather than trusted. Llama-3.3-70B's low scored count (173/212) reflects parsing failures under the rubric format; its ranking numbers are conditioned on its scored subset and are correspondingly less comparable.

External validity. English only; one input schema; one output schema; six frameworks. Nothing here transfers by assumption to free-text findings, other jurisdictions, or other languages. The model is not a general-purpose judge, and the card says so.

Reproducibility bounds. Offline rescoring is bit-exact. Regeneration from weights is verdict-stable, not byte-stable, for the reasons in Section 6.5; users comparing regenerated predictions to the shipped ones should expect token-level divergence with stable verdicts away from tier boundaries.

11 Pre-Registered Human Validation

This section specifies the study that converts the report's provisional numbers into validated ones, or refutes them. The design is fixed in advance and the pipeline that consumes its output is built

and mock-verified; the study's results will be published whichever direction they point.

11.1 Panel and protocol

At least three independent UK/EU compliance experts score benchmark items blind: experts see findings only, never any model's evaluation, and any expert who has previously seen an AI evaluation of an item abstains on it. Scoring uses the six-dimension instrument with per-dimension anchors, in sessions of at most 2–3 hours (approximately 6–7 hours total per expert), with abstention supported and an overlap design sufficient for inter-rater statistics. Label records conform to a fixed schema (expert identifier, cohort, per-dimension and overall scores, blindness attestation, session duration, and the item's content hash) and are ingested by a validating loader that rejects non-conformant or non-blind records.

11.2 Planned measurements

With expert labels x_{ie} (item i, expert e) and consensus gold $g_i = \text{median}_e\, x_{ie}$:

Inter-rater reliability. Quadratic-weighted κ between expert pairs (Cohen, 1968),

$$\kappa_w \;=\; 1 - \frac{\sum_{k,l} w_{kl}\, O_{kl}}{\sum_{k,l} w_{kl}\, E_{kl}}, \qquad w_{kl} = (k-l)^2; \tag{11}$$

Krippendorff's α with the interval metric, $\alpha = 1 - D_o/D_e$ (Krippendorff, 2004); and ICC(2,k) (Shrout and Fleiss, 1979). These establish the *human ceiling*: no model is asked to agree with experts more than experts agree with each other.

Expert-anchored model agreement. Every model in Tables 1–2 is scored against g_i with the same battery, producing the validated leaderboard. The high-band items excluded in Section 5.3 receive numeric gold for the first time.

Calibration. A binned calibration error against expert gold,

$$\text{CE} \;=\; \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \bar{g}_{B_m} - \bar{s}_{B_m} \right|, \tag{12}$$

adapted from expected calibration error (Guo et al., 2017) to the regression setting by binning on predicted score.

11.3 Recalibration, with a train/test firewall

If expert gold confirms the high-band compression, scores will be recalibrated by isotonic regression:

$$\hat{f} \;=\; \arg\min_{f\uparrow} \sum_{i \in \mathcal{A}} \big(f(s_i) - g_i\big)^2, \tag{13}$$

solved by the pool-adjacent-violators algorithm (Barlow et al., 1972). Because $\hat{f}$ is monotone, every rank-based metric in this report is provably unchanged by it; recalibration can repair the scale without retraining and without touching the leaderboard. Critically, the isotonic fit uses a **36-item**

anchor split $\mathcal{A}$, disjoint by content hash from the scoring gold: fitting the calibrator on the same expert labels used to score models would be evaluation on the training set of the calibrator and would void the validated claim. The same anchor-fit protocol will be offered identically to every model on the board, and primary metrics remain computed on raw scores.

11.4 Commitment

The panel's outputs (the human ceiling, the validated leaderboard, the calibration analysis, and the raw anonymized label statistics) will be published as an update to this report whatever they show, including the case in which the released model trails a rubric-prompted baseline on expert-anchored agreement. The recruiting call for the panel is public on the model repository.

12 Transparency and Responsible Release

The boundary, stated. This work publishes its validation science completely and withholds its training engineering. Public: benchmark construction and commitments, contamination proofs, the gold policy including its failures, metric code, all results including losses, the collapse disclosure, and the human-validation design. Withheld: the training corpus and generation pipeline, the teacher model's identity and configuration, and adapter hyperparameters. The reason is commercial and is stated rather than disguised: the training pipeline is the proprietary asset of a company whose other models are not released, and naming the teacher would disclose part of that pipeline. We accept the resulting asymmetry (the *evaluation* is reproducible by anyone; the *model* is reproducible by no one but us) and we consider a clearly drawn boundary more honest than a nominally open recipe with silent omissions. The arbiter model used for band adjudication *is* named (Section 4.4) because it belongs to the validation side of the boundary and never touches training.

Known distribution of the defective predecessor. v2.0 was publicly visible for a period beginning May 2026; at least one third-party inference catalogue captured a copy that survives the source repository's deletion. The model card carries a do-not-use notice for v2.0 weights, and a removal request for the stale third-party listing has been prepared. We flag the general hazard for the community: model-hub mirrors and derivative catalogues can outlive a withdrawal, so a public defect disclosure must assume the defective artifact remains obtainable.

Intended use and misuse. The model scores compliance findings for the six listed frameworks under the fixed schema. It is decision support for human reviewers; it is not legal or compliance advice, and its output must not be the sole basis for a compliance decision. It is unevaluated outside its schema, its frameworks, and English, and Section 7.2 identifies the workloads (defect screening) for which a different tool is currently better. Deployments requiring hard schema guarantees should enforce the schema at the serving layer rather than rely on observed validity.

Licensing. Model weights, benchmark subset, and code are released under Apache-2.0. The base model is Qwen3-8B (Apache-2.0, © Alibaba Cloud).

13 Conclusion

We set out to build a domain verifier that a regulated organisation could run on its own hardware, and to validate it without pretending our labels were better than they are. The resulting artifact is, by the measurements we can currently defend, an 8-billion-parameter model that matches a nine-times-larger open baseline on cross-tier rank concordance within overlapping confidence intervals, leads on rank correlation, produces schema-valid output at 99.5% where its base model produces none, and serves on a single 24 GB GPU. By the same measurements it is the weakest scored model at defect separation, it is out-detected on wrong-jurisdiction citations by a rubric-prompted copy of its own base model, and its score scale is compressed at the top in a way it plausibly inherited from its labels.

Three of this report's findings seem to us more durable than the leaderboard itself. A model can be syntactically perfect and semantically inert, so release gates must measure variance against graded inputs, not output form. A distilled evaluator is bounded by its label source in ways that more data and more parameters do not escape, so the honest response to a measured weakness is sometimes a human study rather than a retrain. And a benchmark whose gold partly descends from the system under test must restrict each label source to its validated domain and publish the conflict, or its leaderboard measures provenance rather than quality.

What this report cannot yet say is whether any of its numbers survive contact with independent human experts. That study is specified, its pipeline is built, and its results will be published either way. Until then, the correct reading of HivemindEval v2.1 is the one on its card: a provisionally validated, honestly bounded instrument: useful within its stated scope, and accompanied by everything needed to check it.

Acknowledgments

Model training and early experimentation were carried out under a UKRI AI Research Resource allocation on the Dawn supercomputer (University of Cambridge, Intel Data Center GPU Max 1550). This work was supported in part by an Emergent Ventures grant (Mercatus Center, George Mason University). We thank the Qwen team for the Apache-2.0 base model.

Reproducibility Statement

The 68-item public benchmark subset, per-item gold, the raw predictions of all six compared models, and the scoring code are released; `python3 eval/score_benchmark.py` recomputes every cell of Table 2, including bootstrap intervals (seed 7, B=400), bit-for-bit with no GPU. The full 230-item set and the training pipeline are withheld; the former is committed by SHA-256 (Section 4.5) so that any future disclosure is verifiable against this report. Prediction regeneration from the released weights is verdict-stable under the determinism bounds of Section 6.5. Model: https://huggingface.co/Hypereum/HivemindEval. Data: https://huggingface.co/datasets/Hypereum/hivemind-eval-benchmark. Code: https://github.com/hypereum-innovations/hivemind-eval.

References

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference under Order Restrictions*. Wiley.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), 213–220.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. arXiv:1803.09010.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of ICML 2017*. arXiv:1706.04599.

Kim, S., Shin, J., Cho, Y., Jang, J., Longpre, S., Lee, H., Yun, S., Shin, S., Kim, S., Thorne, J., and Seo, M. (2023). Prometheus: Inducing fine-grained evaluation capability in language models. arXiv:2310.08491.

Kim, S., Suk, J., Longpre, S., Lin, B. Y., Shin, J., Welleck, S., Neubig, G., Lee, M., Lee, K., and Seo, M. (2024). Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of EMNLP 2024*, 4334–4353. arXiv:2405.01535.

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology* (2nd ed.). Sage.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. (2023). Efficient memory management for large language model serving with PagedAttention. In *Proceedings of SOSP 2023*.

Lambert, N., Pyatkin, V., Morrison, J., Miranda, L. J., Lin, B. Y., Chandu, K., Dziri, N., Kumar, S., Zick, T., Choi, Y., Smith, N. A., and Hajishirzi, H. (2024). RewardBench: Evaluating reward models for language modeling. arXiv:2403.13787.

Li, J., Sun, S., Yuan, W., Fan, R.-Z., Zhao, H., and Liu, P. (2024). Generative judge for evaluating alignment. In *Proceedings of ICLR 2024*. arXiv:2310.05470.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023). G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of EMNLP 2023*.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. In *Proceedings of FAT* 2019*. arXiv:1810.03993.

Qwen Team (2025). Qwen3 technical report. arXiv:2505.09388.

Shrout, P. E., and Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428.

Willard, B. T., and Louf, R. (2023). Efficient guided generation for large language models. arXiv:2307.09702.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *NeurIPS 2023 Datasets and Benchmarks Track*. arXiv:2306.05685.

A Output Schema

```
{
  "overall_score": 0,
  "dimensions": {
    "accuracy":             {"score": 0, "reasoning": "..."},
    "completeness":         {"score": 0, "reasoning": "..."},
    "regulatory_alignment": {"score": 0, "reasoning": "..."},
    "actionability":        {"score": 0, "reasoning": "..."},
    "coherence":            {"score": 0, "reasoning": "..."},
    "evidence_quality":     {"score": 0, "reasoning": "..."}
  },
  "improvement_suggestions": ["..."]
}
```

All scores are integers in $[0, 100]$. Required fields are exactly as shown; the validation-gated parser rejects any output missing a field, containing a non-integer score, or wrapping the object in prose.

B Acceptance Gate, v2.1 Final Weights

Twelve human-range-graded probes; 2,048-token evaluation budget; thinking disabled; temperature 0.

Criterion	Threshold	Result	Verdict
Overall score variance (G1)	$\sigma > 8.0$	23.33	PASS
Per-dimension variance (G2)	each $\sigma > 5.0$	21.1–31.9 (all six)	PASS
Quality spread (G3)	> 20.0	51.33	PASS
Tier ordering (G4)	monotone, 0 violations	exc > good > med > terr	PASS
Jurisdiction-defect mean (G5)	< 50.0	22.0	PASS
Schema-valid outputs	n/a	12/12	informational

C Per-Tier Score Behaviour (v2.1)

Tier	Human gold range	Predicted mean
Excellent	floors 78–82	72.0
Good	n/a	63.0
Mediocre	n/a	35.0
Terrible	n/a	20.7
Jurisdiction-defect	n/a	22.0

Ordering is correct and well separated; the top of the range is compressed (Section 7.4). High-band absolute calibration awaits expert labels (Section 11.3).

D Published Subset Composition

68 items selected by stratified sampling with deterministic content-hash ordering: all four quality tiers; all four defect classes; each of the six frameworks contributing 8–12 items; zero overlap with any training corpus by content hash. Subset hash in Section 4.5; the exact per-stratum manifest ships with the dataset.

E Notation

Symbol	Meaning
s_i	model overall score for item i, integer in $[0, 100]$
t_i	construction band of item i, 1 = terrible … 4 = excellent
g_i	numeric gold for item i (where defined; Section 5.2)
$C_{\text{tier}}, C_{\text{exc}}$	cross-tier and excellent-band rank concordance
D_c	defect separation for defect class c
τ_b	Kendall's tau-b between scores and bands
$\text{MAE}_{\ell m}$	mean absolute error on the low/mid numeric-gold subset
$\mathcal{A}$	36-item expert anchor split for isotonic calibration (disjoint from scoring gold)
$\hat{f}$	monotone recalibration map (PAVA solution)