Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Output-Resolution-Driven Ranking Instability in Ordinal Emotion Benchmarks Based on a Tiered Invariance and Label-Order Testbed

Keito Inoshita

inosita.2865@gmail.com

Kansai University

Research Article

Keywords: Benchmark validity, Emotion recognition, Evaluation methodology, Ordinal constraints, Ranking instability, Tie conventions

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11448579/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Output-Resolution-Driven Ranking Instability in Ordinal Emotion Benchmarks Based on a Tiered Invariance and Label-Order Testbed

Keito Inoshita[1*]

[1*]Faculty of Business and Commerce, Kansai University, Osaka, 564-8680, Japan.

Corresponding author(s). E-mail(s): inosita.2865@gmail.com;

Abstract

Benchmark scores are read as properties of the systems they rank, and the rule turning a criterion into a number is treated as a detail of the scorer. For an ordinal criterion it is not: satisfying the constraint requires deciding what a tie is worth, a free parameter of the report. We construct the Tiered Invariance and Label-Order Testbed, an ordinal emotion benchmark whose relational structure is fixed by construction, not annotation: scenario cues are flipped one at a time along a five-rung ladder, so a system's leaning between two competing emotions must be non-decreasing along the ladder and invariant across register. Writing the tie weight as a single parameter makes the score linear in the fractions of ordered pairs a system separates and ties, yielding in closed form the range a convention can move a score, the point where two systems exchange places, and the condition under which the benchmark contradicts an external accuracy axis. On twenty estimators spanning supervised classifiers, polled language models and a lexicon, this one parameter overturns seven in ten pairwise verdicts within the defensible range. A measurement-only intervention isolates the cause: reading one classifier's own probability by polling rather than exactly, leaving model, corpus and constraints untouched, moves its score from 0.648 to 0.440 and drops it below every language model polled at that budget. Such a ranking therefore reports output resolution rather than independent accuracy, and comparative claims over ordinal benchmarks should state the convention and the tie rate that bounds its effect.

Keywords: Benchmark validity, Emotion recognition, Evaluation methodology, Ordinal constraints, Ranking instability, Tie conventions

1 Introduction

Comparative evaluation on shared benchmarks is how computer science decides which systems are worth adopting, and a position on such a benchmark is read as a property of the system and the test set. It is not only that. Every score is the output of a measurement chain, and the last link in that chain is a reporting rule that converts the satisfaction of a criterion into a number. For criteria that are *ordinal*, requiring an output to move in a prescribed direction as the input is varied, that rule contains a choice for which no canonical answer exists: what a tie between two outputs is worth. Rank correlations have carried competing tie conventions since Kendall [12, 16], and in machine translation meta-evaluation the choice has been shown to change which metric ranks best

and even to be gameable [8]. The choice is therefore known to matter. What is not known is what sets the quantity it acts on, and consequently how far a ranking can be moved by a decision that no benchmark is obliged to report. Put in those terms the question is one of information quality, and it concerns the report rather than the systems: a leaderboard is an information artefact on which adoption decisions are taken, and how much of its content is contributed by the recording rule is a property of that artefact which no amount of further modelling will settle.

Emotion recognition is where this question bites hardest, because the labels against which such systems are scored disagree with one another to a degree that has no parallel in most other classification tasks. In GoEmotions, three or more annotators agreed on at least one label in 17,763 of 58,009 examples, 31% [7]. Agreement varies sharply by category: the four emotions used in this article have interrater correlations between 0.207 and 0.346, against 0.645 for the most reliable category in that corpus. Disagreement is not an artefact to be cleaned; it is the normal condition of the measurement, and how a field handles that condition propagates into what the resulting systems are taken to establish about people [14]. A body of work therefore treats the disagreement as the object of study rather than as noise, predicting the label distribution instead of a majority vote, modelling annotators individually, and separating genuine ambiguity from error [20]. What these methods have in common is that the quantity they aim to recover is never observed. Verification falls back on proxies, agreement with a majority, held-out prediction, or recovery of artificially injected label noise, and in the disagreement regime even calibration becomes hard to define [1]. This motivates evaluation material in which the structure of the disagreement is fixed by construction rather than by annotation.

Constructing the ground truth solves that problem and introduces another. What a construction can fix is a *relational* constraint, namely how the output ought to behave as the input is varied, and turning the satisfaction of such a constraint into a number requires the tie convention above. Despite the established sensitivity of ordinal metrics to that convention, existing treatments leave three questions open: 1) the tie rate on which the convention acts is reported, if at all, as a property of the data, so the magnitude of the convention's effect on any particular benchmark cannot be bounded in advance; 2) the direction of causation is untested, because panels that mix full-resolution classifiers with language models polled a handful of times confound what a system knows with how finely its output was recorded; and 3) the companion criterion to monotonicity, invariance under a change of register, is scored with a tolerance that has the same free-parameter structure and the opposite bias, which no analysis has made explicit. Until these are settled, a comparative claim over an ordinal benchmark has no stated error bar with respect to its own scoring rule.

In this study, we construct the Tiered Invariance and Label-Order Testbed (TILT), an ordinal emotion benchmark whose relational structure is fixed by construction rather than by annotation, and use it to ask what a satisfaction rate over such a structure actually measures. Each item is written from a scenario with four cue slots that are flipped one at a time, the slots being bound to appraisal dimensions in the same-valence tier, producing a five-rung ladder along which a system's leaning toward one of two competing emotions must be non-decreasing, together with a matched pair in a second register along which it must be invariant. Scoring these criteria forces a reporting choice: a tie convention for ordering and a tolerance for invariance. We treat that choice as a parameter of the report rather than a detail of the scorer, which makes the reported score, the conditions under which two systems exchange places, and the conditions under which the benchmark contradicts an external accuracy axis analysable in closed form. Two interventions that alter only how finely a prediction is recorded, together with an external replication, then locate where the resulting variation comes from. The main contributions of this article are summarised as follows:

i) An *ordered* construction over cue slots is newly designed, with appraisal binding in the same-valence tier, extending the two-level contrastive items used in prior constructed benchmarks to a five-rung ladder, so that emotions sharing a valence, which lexical polarity cannot separate, can be placed in a graded order without human annotation.

ii) A reporting model is newly developed in which the conventions in use form a single axis, yielding in closed form the exact range a convention can move a score, the point at which two systems exchange places, and the condition under which the benchmark contradicts an external axis, and from which a convention-instability index is constructed that quantifies how much of a ranking is decided by the reporting rule rather than by the systems.

iii) Two measurement-only interventions are devised to establish the direction of causation, holding models, corpus, constraints and predictions fixed while changing only the resolution at which predictions are recorded, so that the confound between knowledge and recording is broken inside a single model, and are complemented by a parameter-free measurement model and a replication on an independent benchmark.

The rest of this article is organized as follows. Section 2 reviews related work, Section 3 presents the construction and the reporting model, Section 4 reports the experiments, Section 5 discusses what the results imply for benchmark reporting, and Section 6 concludes.

2 Related work

2.1 Disagreement-aware evaluation of emotion

Emotion annotation disagrees systematically, and the disagreement is not confined to hard cases. In GoEmotions, three or more annotators agreed on at least one label in 31% of examples [7], a level of divergence that a majority vote conceals rather than resolves. Crowd-enVENT collected self-reports from the writer of an event description alongside third-party estimates of the same text and found the two diverge systematically [26]: what an experiencer reports feeling and what a reader infers from the same words are different quantities, so the choice of whose judgement counts as the label is itself a design decision. The LeWiDi shared tasks take the consequence seriously and treat the divergence as a target to be learned rather than noise to be averaged away [17], which makes the distribution, not the mode, the object a system is asked to reproduce. Whether an automatic judge can stand in for that object is a separate question, and on emotion data the answer is currently negative: language models recover the majority label far better than the shape of the distribution around it [15]. That is one reason the present benchmark asserts relational properties rather than distributions, and one reason its reference panel is reported as evidence that the constraints are satisfiable rather than as a stand-in for annotators.

Proposals follow from that reframing: predicting the label distribution, modelling individual annotators, and separating inherent ambiguity from annotation error [20]. Each aims to recover a quantity that is never directly observed, the distribution a population of readers would produce, so verification necessarily relies on proxies, and the proxy is usually a second sample of annotators drawn from the same population. Where human labels disagree, even calibration, the most standard check available, is difficult to define [1], because the reference it would be calibrated against is itself an estimate. The question of what an evaluation number means therefore does not resolve by collecting more annotation; it requires fixing something about the data that annotation cannot fix.

This article works on the verification side rather than proposing an estimator. What a construction can fix is not a label but a relational property of the output, that one item should be scored above another, and the results below show that converting such a property into a number leaves a degree of freedom that decides the resulting ranking. The remainder of this section places that degree of freedom against work on scoring conventions and against the constructions the benchmark is built from.

2.2 Scoring conventions and the stability of rankings

How ties are charged in a rank correlation has been a live question since the statistic was formalised. Kendall devoted a paper to ties in ranking problems in 1945 [16], and the coefficient of Goodman and Kruskal drops tied pairs from the denominator altogether [12]; competing conventions have coexisted ever since, and the last of our four is exactly that coefficient restricted to one constraint. In machine translation meta-evaluation

the choice has been shown to change which metric ranks best, to be gameable, and to interact with whether a metric emits continuous or discrete scores, and that work responds by calibrating ties before comparing [8]. We take that phenomenon as given. What we add is an account of where the ties come from: the tie rate is set in part by the number of judgements the evaluator draws, a parameter of the evaluation procedure rather than a property of the system evaluated.

That a scoring convention can reverse a ranking is established, and the strongest demonstrations are outside emotion: holding collection, systems and cutoff fixed and changing only how per-user scores are aggregated turns a recommendation metric's rank correlation with its peers negative [27], replacing mean aggregation with social-choice rules changes the sign of the correlation against the mean ranking on NLP benchmarks [22], and aggregating the same scores three ways yields three different state-of-the-art systems on eight of nine datasets [6]. These studies establish that a benchmark's internal verdict is convention dependent.

Our question is different. The measurement here is not between two benchmark rankings but between the benchmark score and an accuracy measured outside it, on human-labelled data the construction never saw, so the convention decides not which internal ranking to believe but whether the benchmark agrees with the criterion it stands in for. Two results bound the expectation: reworking four commonsense benchmarks and testing them against eight downstream tasks leaves the original model rankings largely preserved [11], so a reversal must be demonstrated rather than assumed, while in estimating information theoretic ceilings on emotion predictability the choice of estimator governs the conclusion [13].

2.3 Constructed benchmarks, ordinal relations, and appraisal

Soft labels can be made known by construction by hiding input features and marginalising over a prior under an explicit ground-truth function [28]. The mechanism used here is the same one, transported from a tabular setting to language. In NLP the dominant way to build items whose relation is known is instead the contrast set: experts perturb a test instance so that its gold label changes, and the original and the perturbation are compared [10]. That gives a pair, not a series, which is the gap the ladder fills. The transport is not free: setting a tabular feature makes it faithful by definition, whereas a generator may simply not write an instructed fact into the text. Verifying that the manipulation landed is a manipulation check, formalised for language-model stimuli by Chen and Eger [4], of which our cue fidelity test is an instance. We differ in handling failures: that work regenerates until the expected label appears, whereas we quarantine, because regenerating selects for passing and makes the pass rate unreportable.

Testing a program by a relation that must hold between the outputs of several related inputs is metamorphic testing [3, 23], and our MON is a metamorphic relation of the monotonicity type used in NLP testing [5, 19, 21]. Existing formulations specify a strict order and leave tied outputs unresolved [19], and the one paper that notices ties records arbitrary tie-breaking as an unresolved source of false positives [29]. That is precisely the reporting degree of freedom studied here.

Which items can carry an order is a question about the emotions themselves. Describing emotion as a configuration of appraisals rather than as discrete categories was systematised by Smith and Ellsworth [24]; what matters here is that emotions sharing a valence separate on those dimensions, anger and fear dividing on where responsibility is attributed. Appraisal dimensions have been used for annotation and generation [18, 26], and cue slots bound to them have been used to build scenario-based items [9, 25], but in each case the manipulated levels form a two-condition contrast and carry no order. Constructed labels also have a measured distance from human judgement: on appraisal-generated scenarios, participants agreed with one another 69.38% of the time but with the pipeline's a priori labels only 41.49% [9]. That gap bounds what a constructed benchmark may claim, and is why the present claims are restricted to relational properties. Independently published appraisal-grounded contrastive benchmarks require one positive and one negative emotion per scene, and record that they therefore cannot speak to same-polarity-but-different-emotion cases [2]; the same-valence tier addresses that gap, and the ladder is what distinguishes the construction from a contrast.

Table 1 places these lines against one another on the properties the present construction needs. Comparing with these previous researches, no single line supplies all of them: the constructions that fix a relation stop at a pair, the testing literature that specifies an order leaves the tie unpriced, and the meta-evaluation literature that prices the tie works on metric rankings rather than on constructed items, so it has no occasion to ask what sets the tie rate. Where a line is plausibly capable of a property we mark it as having it, so the table understates rather than overstates what is new here.

3 Construction and reporting model

3.1 What is claimed as ground truth

Fig. 1 gives the construction. Before describing it, we state what it does and does not certify, because a constructed benchmark is only as good as the claim it makes for its labels.

We do not claim that the induced label distribution equals a human label distribution. A construction fixes what the designer put into an item; it cannot fix what a reader takes out of it. A reader's prior over the unstated parts of a situation need not match the designer's, and prior work has measured the resulting gap directly: on appraisal-generated scenarios, participants agreed with one another far more closely than they agreed with the labels the generation pipeline had assigned in advance [9]. A benchmark that asserted those labels as distributions would therefore be asserting something its own construction cannot support. Restricting the claim to relational properties is a response to that measurement rather than a hedge: it keeps only the part of the design that survives a mismatch between the designer's prior and the reader's.

What survives is comparative. If one item has more cue slots set toward a pole than another, a reader whose response to those cues is monotone should not rank the second above the first, and two items built from identical cues should not differ by whether the text is framed as a diary entry or a letter. Neither statement requires knowing the reader's prior, which is why they can be fixed by construction while a distribution cannot.

Two relational properties are asserted, and both are fixed by the construction rather than estimated:

- **MON**: within a fixed scenario, replicate and register, the leaning toward `pole_b` must be non-decreasing in the number of cue slots set toward that pole.
- **INV**: two items with identical cue content that differ only in register must receive the same leaning.

3.2 Scenarios and cue slots

A *scenario* is the unit of construction. It fixes a *frame*, one sentence of situation that stays constant across the whole ladder, together with an ordered pair of competing emotions and four *cue slots*. A cue slot is a single binary fact about the writer's circumstances, supplied as two mutually exclusive statements: one that pulls the reading toward `pole_a` and one that pulls it toward `pole_b`. Every item generated from the scenario is told exactly one statement from each slot, so an item is specified by a four-bit assignment and nothing else varies.

Table 2 shows one scenario in full. The frame places the writer in a running dispute with a neighbour, and the poles are anger and sadness. The four slots are the appraisal dimensions on which those emotions differ: who is responsible (*blame*), whether the situation can still be undone (*reversibility*), whether the writer is still acting on it (*control*), and where attention points (*orientation*). All four set toward `pole_a` give a writer who is blocked deliberately, can still file a complaint, and is planning next steps; all four toward `pole_b` give one whose driveway is blocked by a fallen tree, whose garden damage is irreversible, and who reminisces about better times. Neither is the worse situation, that is the constraint the slots are written under, but the first reads as anger and the second as sadness.

The design parameter is the number of slots set toward `pole_b`, $n_b = \text{round}(4\alpha)$ for $\alpha \in \{0, .25, .5, .75, 1\}$, which gives the five rungs of a ladder. *Which* slots are flipped at a given n_b is chosen by a seeded deterministic draw, so that the assignment is reproducible and so that no single dimension is privileged at intermediate rungs. The resulting claim is deliberately weak: we do not

Table 1 Properties of the nearest lines of prior work. A dash marks a property that does not apply to work on metric rankings rather than on constructed items, and the last column requires the recording resolution to be manipulated rather than observed.

Line of work	Relational ground truth	Graded order beyond a pair	No human annotation	Tie convention as a parameter	Resolution intervened on
Label-variation modelling [1, 20]					
Constructed soft labels [28]			✓		
Contrast sets [10]	✓				
Metamorphic monotonicity testing [19, 29]	✓	✓	✓		
Appraisal-grounded contrastive items [2, 9]	✓		✓		
Tie conventions in meta-evaluation [8, 16]	—	—	—	✓	
Convention-dependent aggregation [6, 22, 27]	—	—	—	✓	
TILT, this article	✓	✓	✓	✓	✓

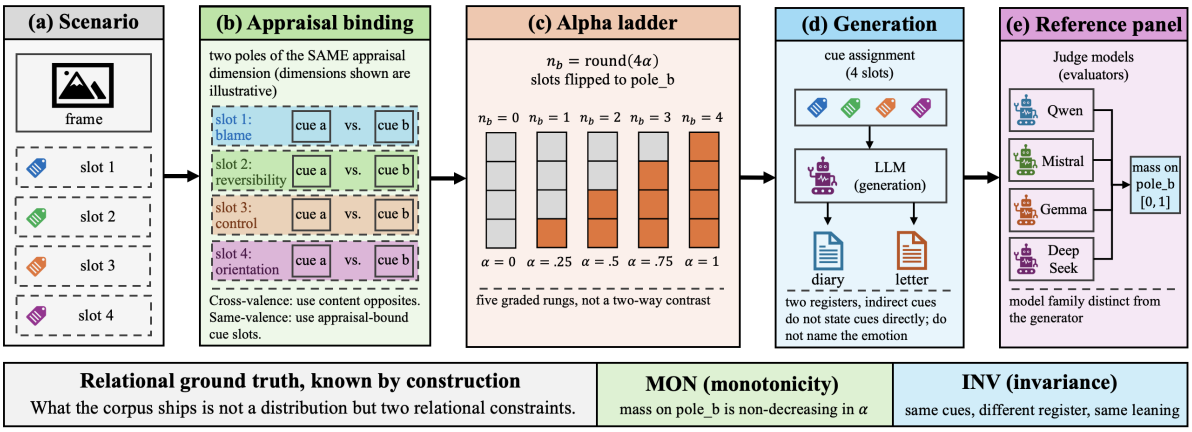

Fig. 1 The construction. The design parameter is the number of cue slots set toward `pole_b`, which defines a five-rung ladder; the dimensions shown are illustrative. What the corpus ships is not a distribution but two relational constraints.

Table 2 One same-valence scenario in full, as it ships in `tilt_scenarios.jsonl`. The frame is constant; each cue slot supplies two mutually exclusive statements that are opposite poles of one appraisal dimension. Both columns describe an equally unwelcome situation, so what separates them is the appraisal structure and not the severity of the event.

Frame. *You are writing about your ongoing conflict with a neighbour.* Poles. anger (`a`) vs. sadness (`b`)

Slot	Dimension	Statement toward `pole_a` (anger)	Statement toward `pole_b` (sadness)
1	blame	Your neighbour deliberately blocks your driveway every morning	A neighbour's tree has fallen and is still blocking your driveway
2	reversibility	You can still file a complaint to resolve the issue	The damage to your garden is irreversible and cannot be restored
3	control	You are seeking a court order to stop the harassment	You have accepted the loss and are moving on with your life
4	orientation	You discuss what actions to take next to protect your property	You reminisce about the good times when you got along with your neighbour

assert that an item at $\alpha = 0.5$ evokes any particular mixture, only that an item with more slots set toward `pole_b` should not read as *less* `pole_b` than one with fewer.

For the same-valence tier the two statements in each slot are opposite poles of *the same appraisal dimension*, chosen so that neither pole makes the situation obviously worse. For the cross-valence tier the statements are content opposites without an appraisal binding. This asymmetry is deliberate and is the result of a failure: an earlier version with ad hoc cues passed 18 of 23 cross-valence pairs but only 4 of 21 same-valence pairs, and 0 of 7 anger/fear pairs, with a pole-mass span of 0.012. A probe on real GoEmotions items showed the same judge panel separating same-valence pairs in natural text, which located the fault in the cue design rather than in the judges. Binding the four slots to appraisal dimensions raised same-valence acceptance to 15 of 39 and the span to 0.567. Valence-opposite pairs can be ordered by content alone; same-valence pairs cannot, and that is where the appraisal binding is required.

3.3 Generation and calibration

Item text is generated by an open-weight model at temperature 0.9, with the sampling seed derived from the item identifier by a cryptographic hash so that the corpus regenerates identically on any machine. The prompt supplies the frame and the four selected statements and asks for exactly two sentences that let every one of those facts shape the writing *without stating any of them directly and without naming an emotion*. The last two constraints matter: an item that quotes its own cues would make the constraint trivial, and an item that names the emotion would make it a lexical lookup. Each assignment is realised in two registers, a private diary entry and part of a letter to someone who already knows the situation, which are the items an invariance pair is built from, and the whole grid is repeated three times with different seeds.

Scenarios are not written by hand but proposed by a language model from a specification, then *screened*, because a scenario whose two poles are not actually anchored cannot carry a ladder. Screening generates only the two endpoints of each candidate and scores them with the reference panel, and keeps a candidate only if the mass on `pole_b` is at most 0.25 at $\alpha = 0$ and at least 0.75 at $\alpha = 1$. This is a criterion on the endpoints alone, which is why every headline figure in Section 4 is computed on the interpolated rungs: the endpoints restate the acceptance rule, and including them would report the selection back to the reader as a result. Of 44 cross-valence candidates 22 survived, and of 39 appraisal-bound same-valence candidates 15 did, giving the 37 scenarios that generate the corpus. The 46 rejected scenarios are not discarded; ten of them are given full ladders and released as a hard split, which is what makes the size of the selection effect measurable rather than assumed.

The released corpus is what remains after two exclusions applied before any system was scored. Items shorter than ten words are removed, because a generator that returns a fragment has not realised the assignment at all, and items whose cue fidelity falls at or below 0.5 are removed for the reason given in Section 3.4. That leaves 1,021 items over 37 scenarios, balanced across registers (511 diary, 510 letter), with the 89 excluded items shipped in a separate file that records which rule each one failed. Nothing is regenerated to replace a failure, for the reason given in Section 3.4.

3.4 Cue fidelity

In tabular construction, setting a feature makes it faithful by definition. In language it does not: a generator can decline to put an instructed fact into the text. Verifying that the manipulation landed is the role of a *manipulation check*, and its automated form for language-model stimuli has been formalised by Chen and Eger [4]; the procedure here is an instance of it. For every item and every slot, a reader model from a family disjoint from the generator is shown the text and the two competing statements for that slot and must say which holds. Fidelity is agreement with the statement actually used.

The check is what licenses the construction, so its outcome is reported rather than assumed. Fidelity over the corpus is 0.834 against a chance level of 0.5, which says that the generator does usually write the instructed fact into the text, but not always. It is highest at the two endpoints, 0.895 and 0.937, and falls to between 0.762 and 0.809 at the mixed rungs: an item told four facts that pull in the same direction is easier

to write faithfully than one told two of each. Fidelity also varies by dimension, from 0.943 for *orientation* down to 0.708 for *agency*, although Section 4 shows that dimension and slot position are confounded, so we do not interpret that breakdown.

Two consequences follow for the corpus. Items whose fidelity falls at or below 0.5 are *quarantined*: at that level the assignment cannot be said to have reached the text, so the constraint the item is supposed to carry is not supported by it. Where the manipulation-check literature regenerates a stimulus up to five times until the expected label appears, we do not regenerate at all. Regeneration selects for passing and makes the pass rate unreportable, and that rate is what a reader needs in order to judge how far the construction licenses its own constraints. The quarantined items ship alongside the corpus with the rule each one failed, so the exclusion can be audited rather than taken on trust.

3.5 Two reporting scales, and the tie convention

Construction validation and estimator evaluation ask different questions and take different statistics. Validation asks whether the intended order was induced, and uses continuous measures: MON_ρ, the mean Spearman correlation between α and the panel's leaning within a ladder, and $\mathrm{INV}_{\mathrm{gap}}$, the mean absolute difference across registers. Evaluation asks whether an estimator follows the order, and uses a satisfaction rate over ordered pairs.

The primary value of MON_ρ is 0.631 (95% CI $[0.541, 0.714]$, cluster bootstrap with scenario as the primary sampling unit), computed on interpolated levels only and on the released corpus, so that it reproduces from the released files alone. The screening that selected scenarios required a pole mass ≤ 0.25 at $\alpha = 0$ and ≥ 0.75 at $\alpha = 1$, so the endpoints are the acceptance criterion itself; including them raises the figure to 0.804 ($[0.768, 0.834]$), which we report as a measure of that selection effect and not as the headline.

3.5.1 A reporting model for ordinal constraints

Converting an ordinal constraint into a satisfaction rate requires deciding how to charge ties, and the shipped constraints assert that the leaning must be *non-decreasing*, under which a tie is satisfaction. A natural implementation nevertheless counts only strict increases, and the first version of our own scorer did exactly that. Four conventions are all defensible: charging ties as failures; as satisfaction, following the assertion; as half a point, following the usual treatment in τ_b; and excluding them from the denominator, which is Goodman and Kruskal's γ restricted to one constraint.

Rather than treat these as four unrelated choices, we place them on one axis. Write S_i, T_i and V_i for the fractions of ordered pairs on which estimator i strictly increases, ties, and inverts, so that $S_i + T_i + V_i = 1$. Let a convention charge a tie $\lambda \in [0, 1]$ points. The reported score is then

$$M_i(\lambda) \;=\; S_i + \lambda\, T_i, \tag{1}$$

which is S_i at $\lambda = 0$, $S_i + T_i$ at $\lambda = 1$ and $S_i + T_i/2$ at $\lambda = \frac{1}{2}$; the fourth convention, $S_i/(1 - T_i)$, is the same numerator on a reduced denominator. Three consequences follow. Each is elementary on its own and we claim no mathematical novelty for any of them; their value is joint, in separating the part of a verdict the system produces from the part the reporting procedure produces.

Proposition 1 (Convention range)**.** $\max_\lambda M_i(\lambda) - \min_\lambda M_i(\lambda) = T_i$.

Reporting the tie rate therefore bounds the arbitrariness exactly: a reader can reconstruct the score under any convention from one extra number.

Proposition 2 (Pairwise reversal)**.** *Estimators i and j exchange places at*

$$\lambda^*_{ij} \;=\; \frac{S_j - S_i}{T_i - T_j}, \qquad T_i \neq T_j, \tag{2}$$

and the exchange is reachable by a defensible convention if and only if $0 < \lambda^*_{ij} < 1$.

This replaces the observation that a ranking *can* move with a test of whether it moves for a given pair, computable from two numbers per system. Aggregating it over a panel of m systems gives a

single property of the benchmark,

$$R = \frac{\left|\{(i,j) : 0 < \lambda^*_{ij} < 1\}\right|}{\binom{m}{2}}, \qquad (3)$$

the share of the benchmark's pairwise verdicts that the choice of convention can overturn. We call R the convention-instability index.

Proposition 3 (Disagreement with an external axis)**.** *Let A_i be an accuracy measured outside the benchmark. At tie weight λ the benchmark contradicts A on the pair (i, j) if and only if*

$$(A_i - A_j)\big[(S_i - S_j) + \lambda\,(T_i - T_j)\big] < 0. \qquad (4)$$

A negative coupling between A and T across the panel makes Eq. (4) easy to satisfy and is what produces the aggregate sign change reported below, but it is pressure toward reversal rather than a sufficient condition; Eq. (4) is the exact statement, and Section 4 evaluates it pair by pair.

What the model leaves open is where T_i comes from. Section 4 answers that with two interventions on the measurement rather than on the system: quantising a continuous score to B levels, and polling a sampled score K times.

3.6 Tiers

Items are partitioned into a cross-valence tier, whose two poles differ in valence, and a same-valence tier, whose poles do not. The partition exists to separate a system that tracks valence from one that tracks emotion. A lexicon that scores only positivity can in principle order a joy/sadness ladder, because the cues that move the ladder also move the polarity; it cannot order an anger/fear ladder at all, because both poles are negative and the ordering is carried by who is responsible and whether the outcome is settled. The two tiers therefore ask different questions of the same system, and a score reported over their union conceals which of the two it answered. Appraisal-grounded benchmarks proposed independently cover the cross-valence case and record the same-valence case as out of scope [2]; the same-valence tier is where this construction is doing work that a polarity signal cannot.

The two tiers are also built differently, and the asymmetry is honest rather than incidental. Only the same-valence tier binds its cue slots to appraisal dimensions; the cross-valence tier uses content opposites, because a valence contrast does not need an appraisal account to be orderable. That leaves appraisal binding confounded with tier in the released corpus, 617 cross-valence against 404 same-valence items, so the binding cannot be credited from a tier comparison alone. Section 4.2 separates the two with a control arm that holds scenarios, frames, poles, ladder, generator and judges fixed and unbinds only the cue slots.

4 Experiments

4.1 Design

The released corpus contains 1,021 items over 37 scenarios, partitioned into 617 cross-valence and 404 same-valence items. It is the result of applying this article's own acceptance criteria to 1,110 generated items: 82 were removed for cue fidelity at or below 0.5 and 9 for falling under ten words, two items failing both tests. The 89 excluded items are released separately rather than deleted, so that the effect of the criteria can be checked. Constraints number 702.

A reference panel of four model families disjoint from the generator supplied 16,824 votes under a forced two-pole choice. The option order is counterbalanced because position bias, the order in which the alternatives are presented moving the verdict, is documented for language-model judges [30]. Twenty external estimators, disjoint from both the generator and the panel, are evaluated: seven full-resolution neural classifiers, whose scores are continuous and which never tie on this corpus, twelve zero-shot language models polled with a fixed ballot budget, and a valence-only lexicon. The panel is chosen to span output regimes rather than vendors, because the analysis below turns on how many distinct values a system can emit. Independent accuracy is measured on the GoEmotions *test* split under the same forced-choice protocol. Coverage differs across estimators, because the sampled models were queried with more repetitions per item than the deterministic ones; we report n with every accuracy, and the rank order of accuracy, which is what the correlations in Section 4.3 use, is unchanged when the

comparison is restricted to the items every system scored.

One estimator was dropped and one abstains. A third zero-shot model returned empty content whenever reasoning consumed its output budget, on about half of all calls, and three retry rounds did not reduce it; listing a system at that coverage would misattribute our own instrumentation to the model. The zero-shot Haiku model returned a non-pole answer on 27.0% of judgements, about 95% of them one-word answers naming a third emotion and the rest explicit refusals. Neither is converted into a vote: a parser that takes the first pole word it encounters would record the reasoning inside a refusal as a vote for the pole it mentions first.

4.2 Does the construction do what it claims?

Reporting construction validation before and after the fidelity filter tests whether the filter does any work. It does: MON_ρ rises from 0.787 to 0.804, the violation rate falls from 0.078 to 0.068 and INV_{gap} falls from 0.158 to 0.151. The null for that figure pairs each item with another from the same scenario and register but a different α, so the cue content genuinely differs; averaged over 500 random pairings it is 0.513 ($[0.479, 0.544]$). On the released corpus the panel's pole mass runs monotonically from 0.107 at $\alpha = 0$ to 0.972 at $\alpha = 1$ as shown in Fig. 2. Because quarantining is uneven across levels, 150 of the 222 shipped ladders retain all five rungs, and 145 of those also have a panel vote on every rung; each level's mean is taken over the items present at that level, and the order within a ladder is carried by the individual ladders.

Overall fidelity is 0.834 at the slot level over the 8,874 slots of the 1,110 generated items, the population the screen actually operated on; the corpus that survives it averages 0.863 per item. Per-dimension values range from 0.943 to 0.708, but this breakdown is confounded: the position of the true statement is fixed by slot index, so the highest and lowest dimensions differ in position as well as in content. We report the aggregate and withhold the per-dimension reading.

Regenerating eight scenarios with a second open-weight family, holding scenarios, cue assignment, ladder, register, seeds, prompts and judging

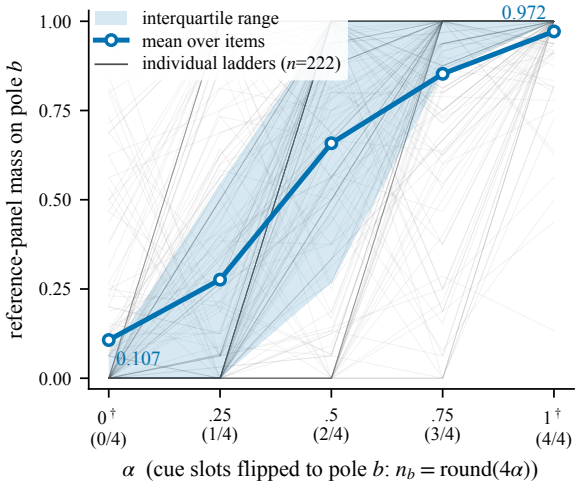

Fig. 2 The panel's leaning rises monotonically along the ladder. Thin lines are individual ladders, the heavy line their mean. The endpoints (†) were the screening criterion, so the interpolated levels carry the primary evidence.

protocol fixed, preserves the order while invariance degrades. Against a baseline computed on the same eight scenarios and the same three judges, MON_ρ moves from 0.810 to 0.675 and INV_{gap} from 0.136 to 0.223, the latter against the null of 0.513 above. Matching the scenarios and the judges removes two of the three confounds in this comparison; the replicate count still differs, so the magnitudes are not a clean generator effect and the direction is what we report.

Appraisal binding covers only the same-valence tier, so binding is confounded with tier. A control arm separates them, holding scenarios, frames, poles, ladder, generator and judges fixed and unbinding only the cue slots, so that the two statements in a slot become content opposites rather than poles of one appraisal dimension.

Binding the slots to appraisal dimensions raises MON_ρ from 0.311 to 0.865 on the same scenarios, and the difference is not the fidelity screen: applying the same screen to both arms leaves the control at 0.311, and applying it to neither leaves it at 0.323 as reported in Table 3. The per-level means show why. The bound arm sweeps the full range monotonically, from 0.168 at $\alpha = 0$ to 0.972 at $\alpha = 1$. The control arm starts at 0.296, never passes 0.832, and is *non-monotone* in the middle, falling from 0.564 at $\alpha = 0.25$ to 0.482 at $\alpha = 0.5$.

Two further observations are the mechanism rather than the outcome. Cue fidelity is lower without the binding (0.738 against 0.789), so more

Table 3 The appraisal binding against a control built from unbound cues, on the same fifteen scenarios. Columns are the reference panel's mean mass on `pole_b` at each rung.

	mass on `pole_b` at $\alpha =$						
Cue construction	0	.25	.5	.75	1	MON_ρ	span
Appraisal-bound	0.168	0.324	0.603	0.833	0.972	0.865	0.848
Control, same screen	0.296	0.564	0.482	0.688	0.832	0.311	0.542
Control, no screen	0.393	0.577	0.495	0.642	0.768	0.323	0.573

of the control arm fails the screen (26% against 15%): an unbound cue is harder for the generator to land in the text. And the generator does not keep the two sides equally severe when the dimension is not named, even though the prompt asks it to; instead, it reaches for contrasts of degree, writing the `pole_b` side as the milder situation. That is the failure the binding prevents: without a named dimension to vary, the generator separates same-valence poles by making one of them better, which is exactly the cue a valence signal reads and exactly what the same-valence tier exists to exclude.

4.3 How much can the tie convention move a score?

Table 4 scores the same corpus, the same constraints and the same twenty estimators under the four conventions of Section 3.5.1. Following Section 3, the population is the interpolated levels of the released corpus only, so that the numbers here are the ones the released scorer returns on the released files.

The same constraint moves each system by its own tie rate, and the two families move in opposite directions. The seven full-resolution classifiers never tie, so no convention touches them, but they invert 34–38% of ordered pairs. The 12 sampled models tie on 28–43% and invert only 6–11%. Charging ties as failures therefore measures the capacity not to tie, which is a property of output resolution; charging them as satisfaction, as the constraint asserts, promotes the systems that rarely invert. Which family a convention favours is fixed before any system is run.

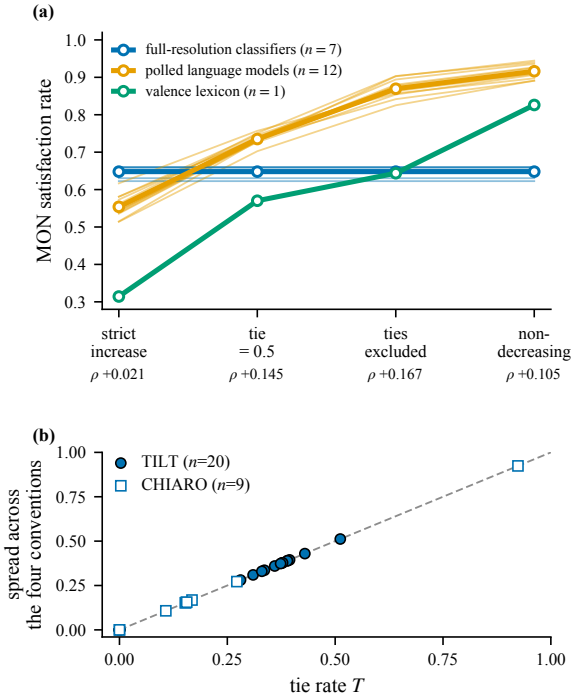

Fig. 3 (a) Twenty estimators under the four tie conventions, drawn as thin lines with the group means overlaid; ρ beneath each convention is its Spearman correlation with independent accuracy. (b) The spread available to each estimator against its tie rate, the dashed line being equality, which the estimators of both benchmarks meet to the precision of the released numbers.

Which verdicts a convention can overturn

Proposition 2 turns that bound into a test. Evaluating λ^*_{ij} for all $\binom{20}{2} = 190$ pairs of the panel, 134 of them cross inside $0 < \lambda^* < 1$, so the convention-instability index of Eq. (3) is $R = 0.705$: seven in ten of the benchmark's pairwise verdicts can be overturned by a convention that no one could call illegitimate. Because S_i and T_i are themselves estimated from finitely many ladders, λ^*_{ij} and R carry sampling uncertainty. Resampling ladders in scenario clusters ($B = 2000$, 37 scenarios) gives a 95% interval of $[0.605, 0.742]$, so the index stays above six in ten throughout. Classifying each pair by whether its interval for λ^* lies wholly inside $(0, 1)$, 79 pairs are reversible beyond resampling error and 89 more are consistent with reversal without establishing it. The unresolved pairs are those with nearly equal tie rates, where the denominator of Eq. (2) approaches zero and the crossing point is unstable by construction.

Table 4 The same ordinal constraint under four tie conventions. Accuracy is on the GoEmotions test split. The last row gives Spearman's ρ between accuracy and each column.

Estimator	Accuracy	Strict	Non-dec.	Tie $= \frac{1}{2}$	Excl.	Tie rate	Inversion	Distinct
Full-resolution classifiers								
RoBERTa-GoEmotions	0.941	0.660	0.660	0.660	0.660	0.000	0.340	1019
DistilRoBERTa-7	0.922	0.630	0.630	0.630	0.630	0.000	0.370	1019
BERT-GoEmotions	0.910	0.652	0.652	0.652	0.652	0.000	0.348	1019
DistilRoBERTa-7c	0.868	0.652	0.652	0.652	0.652	0.000	0.348	1019
BERTweet-emotion	0.832	0.658	0.658	0.658	0.658	0.000	0.342	1019
BERT-emotion	0.826	0.660	0.660	0.660	0.660	0.000	0.340	1019
DistilBERT-emotion	0.802	0.623	0.623	0.623	0.623	0.000	0.377	1019
Sampled language models								
Seed-1.6-flash	0.927	0.542	0.936	0.739	0.894	0.394	0.064	7
Command-R7B	0.907	0.581	0.891	0.736	0.842	0.310	0.109	7
Nova-Micro	0.888	0.514	0.891	0.703	0.825	0.377	0.109	7
Nemotron-3-nano	0.886	0.573	0.909	0.741	0.863	0.336	0.091	7
Grok-4.3	0.886	0.536	0.925	0.730	0.877	0.389	0.075	5
Solar-Pro4	0.883	0.581	0.911	0.746	0.867	0.330	0.089	7
Kimi-K2.5	0.883	0.515	0.945	0.730	0.903	0.430	0.055	7
GPT-5.6-luna	0.880	0.538	0.927	0.732	0.880	0.389	0.073	5
Haiku-4.5	0.876	0.562	0.940	0.751	0.903	0.378	0.060	7
MiniMax-01	0.873	0.547	0.907	0.727	0.854	0.360	0.093	7
GLM-5.3-flash	0.860	0.543	0.917	0.730	0.868	0.374	0.083	7
Phi-4	0.793	0.617	0.897	0.757	0.857	0.281	0.103	7
Score-valued lexicon								
VADER	0.691	0.314	0.826	0.570	0.644	0.512	0.174	171
Spearman vs. accuracy		+0.02	+0.11	+0.15	+0.17	—	—	—
Spearman vs. distinct values		+0.72	−0.84	—	—	−0.74	—	—

Applying Proposition 3 pair by pair, 159 of the 190 pairs admit a defensible λ at which the benchmark contradicts the external accuracy ordering. This is the result that strengthens as the panel widens: on five systems it was six pairs in ten, on twenty it is seven in ten.

The spread across conventions equals the tie rate

Proposition 1 says the range a convention can move the score is exactly the tie rate, which is what Fig. 3(b) plots, while Fig. 3(a) shows the two families moving in opposite directions. The identity held on an independent benchmark for all nine of its estimators to floating-point precision, and on our own twenty to 10^{-6}, which is the rounding of the released numbers. Reporting the tie rate therefore bounds the arbitrariness exactly.

A fifth degree of freedom

Tie handling is not the only reporting choice left undocumented. A ladder on which a system returns the same value at every rung contributes only ties, and whether such ladders are counted or dropped is a further convention; crossed with the choice of population it forms a 2×2 grid that moves the headline by 0.051 to 0.143 for the ordinary estimators and by 0.364 for the lexicon. The released scorer fixes one cell, interpolated rungs with constant ladders kept. The full grid is in the supplementary material.

4.4 Three winners from one set of predictions

The propositions above bound what a convention can do. The hard split shows what it does, on a table small enough to check by eye. Of the forty-six scenario candidates that failed the endpoint screen, ten were given full ladders and are released as a hard split of 300 items, so that the size of the

Table 5 The hard split under the four conventions. Three different estimators take first place. VADER returns one value on every item here, so all its pairs tie: it scores 1.000 under non-decreasing and has no ties-excluded value.

Estimator	Strict	Non-dec.	Tie = ½	Excl.	Tie rate
RoBERTa-GoEmotions	0.628	0.628	0.628	0.628	0.000
DistilRoBERTa-7	0.583	0.583	0.583	0.583	0.000
GPT-5.6-luna	0.244	0.911	0.578	0.733	0.667
VADER	0.000	1.000	0.500	—	1.000

selection effect is measurable rather than assumed; a constructed benchmark that reports only the scenarios it accepted gives the reader no way to separate a property of the systems from a property of the acceptance rule.

Table 5 scores that split with the released scorer. The predictions, the items and the constraints are the same in every column, and only the price of a tie changes. Three different estimators take first place: a supervised classifier under strict increase and under the half-credit rule, the lexicon under non-decreasing scoring, and a language model once ties are excluded. The lexicon's row is the clearest case, because it returns one value on every item of this split, so all of its pairs tie and it scores 0.000, 1.000, 0.500 and nothing at all depending only on what the report decides a tie is worth. The selection effect is visible in the same table and is not uniform across families: against the released corpus the strict-increase score falls by 0.032 for RoBERTa-GoEmotions and 0.047 for DistilRoBERTa-7 but by 0.293 for GPT-5.6-luna and 0.314 for the lexicon. That asymmetry follows from how the split was defined, since the endpoint screen uses a reference panel of language models, so scenarios enter the split precisely where judges of that kind failed to separate the poles while the supervised classifiers took no part in the selection. The split is accordingly a reference point for users of the split and not a difficulty ranking of the systems.

4.5 External replication on an independent benchmark

We repeated the analysis on CHIARO, an independently released appraisal-grounded contrastive emotion benchmark [2], after verifying provenance by recomputing its reported macro-F1 from its shipped predictions. Convention dependence, the

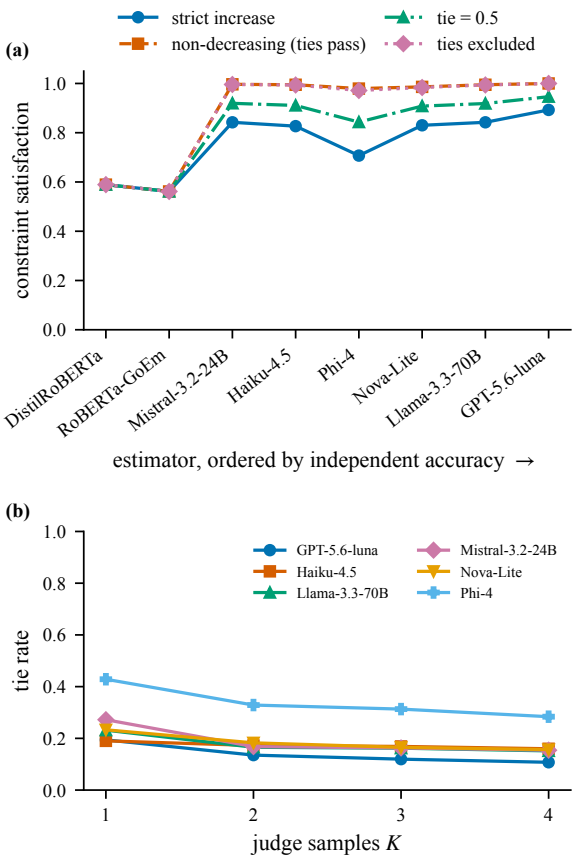

Fig. 4 The same analysis on an independently released benchmark. The value dependence and the identity between spread and tie rate replicate; whether the ranking reverses does not.

identity between tie rate and score spread, and the dependence on polling budget all replicate; the identity holds for all nine of its estimators to floating-point precision. Aggregate reversal does not replicate unconditionally, and Proposition 3 says why: it requires an external axis negatively coupled with the tie rate, which CHIARO's own accuracy axis is not. Its own protocol cannot exhibit the effect at all, because its options are restricted to the gold-polarity side and all eleven shipped systems score 1.000. Fig. 4 shows the conventions on that benchmark; axis-specific results and the leave-one-out analysis are in the supplementary material.

4.6 What the score tracks

Across the twenty estimators, no convention correlates with independent accuracy: Spearman is

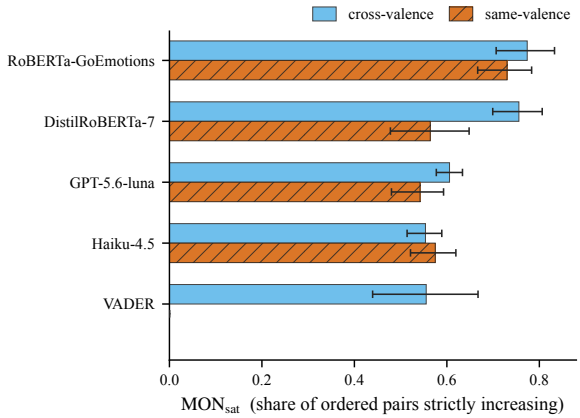

Fig. 5 Per-tier order following, with cluster-bootstrap intervals over scenarios. The lexicon's same-valence value is 0.000 by construction rather than by measurement.

+0.02 under strict increase and +0.11 under non-decreasing. What the score does track is the number of distinct values the estimator can emit: +0.72 under strict increase and −0.84 under non-decreasing, with the tie rate itself at −0.74 against the same axis. Accuracy and resolution are independent across this panel (+0.01).

Both statements carry intervals and they do not overlap: resampling scenarios leaves every accuracy interval containing zero and every resolution interval clear of it, a permutation null agrees ($p = 0.93$ against accuracy, $p = 0.0004$ against resolution), and dropping any one estimator leaves the resolution correlations in $[+0.68, +0.82]$ while the accuracy ones straddle zero throughout.

Under a fixed convention the reported score is dominated by output resolution rather than by independent task accuracy. That is not a new mechanism; it is the one Section 4.3 already identified, now visible because the panel is wide enough to separate resolution from ability. Section 4.7 is the same statement as an intervention: quantising a system's own predictions moves its score without touching what it knows.

This also retires a result we obtained on a narrower panel. On five estimators resolution and accuracy travelled together, and the conventions gave +0.90 under strict increase against −0.60 under non-decreasing. Five points cannot fix the sign of a rank correlation, and the wider panel removes it, so that pattern is a caution about narrow panels rather than a property of the benchmark.

The same conclusion survives a sharper test. Splitting the accuracy axis into cross-valence and same-valence halves matches TILT's own split and turns the correlation into a construct-validity test; matched tier for tier, Spearman is +0.03 and −0.01, so the benchmark tracks natural-text performance on neither. The split still shows the construction working on the one system whose failure mode is known in advance: the lexicon separates cross-valence pairs of natural text at 0.874 but is at chance at 0.506 on same-valence pairs, and on TILT the corresponding strict-order rates are 0.498 and 0.000. Fig. 5 gives the per-tier rates with cluster-bootstrap intervals.

It is noted that this absence of correlation fixes the scope of the instrument rather than exposing a defect in it. TILT measures compliance with a relational constraint that is true by construction, and it is not offered as a proxy for accuracy on naturally occurring text; a reader who needs that quantity should measure it directly, as we do here for the comparison. The absence is also what the article is about. Were the two to agree, the resolution result of Section 4.7 would still hold, because that result is established by intervening on a single system rather than by correlating across the panel.

Unlike λ, which is bounded at 1 because charging a tie more than a separation is not defensible, ϵ has no intrinsic upper bound: a large enough tolerance calls everything invariant. We therefore fix the domain in advance at $[0, 0.5]$, which on a $[0, 1]$ score already treats half the scale as invariant and so covers the range in which the criterion still asserts anything. R_{INV} below is defined over that pre-specified domain, and is comparable with R only in that sense.

4.7 Where the tie rate comes from

Section 3.5.1 leaves T_i unexplained. Two interventions fix it without touching any system: one changes how finely a score may be expressed, the other changes how many times it is sampled. The seven full-resolution classifiers and the score-valued lexicon are re-scored after rounding every prediction to B levels, $Q_B(p) = \mathrm{round}(Bp)/B$. The models, the corpus, the constraints and the predictions before rounding are identical; only the resolution at which those predictions are recorded changes, which separates output granularity from

anything the systems know. Table 6 reports the sweep.

The consequence for the ranking is the point. Rounding to eight levels moves a system 6.7 places out of twenty on average and the most accurate system down seventeen, while nothing any system predicts has changed. Under a fixed convention, resolution is not a nuisance parameter of the measurement; it is most of it.

The effect is not an artefact of uniform rounding. An equal-frequency quantiser, whose bin edges sit at each estimator's own empirical quantiles, gives the same direction throughout and a larger displacement at the coarsest setting. What moves the ranking is the number of values a system can express, not the shape of the map that reduces them.

4.8 The same model, measured two ways

Quantisation coarsens a score from outside. The complementary intervention coarsens it the way the panel's own protocol does: take each classifier's probability and, instead of reading it exactly, draw K ballots from it and score the resulting fraction. This is the treatment the sampled models receive, now applied to a system whose underlying prediction can also be read directly. The model, the corpus and the constraints are identical across K; only the measurement is. Table 7 reports the sweep and Fig. 6 shows it.

It is noted that the intervention is run in this direction, and not by reading the language models' own token probabilities, because the serving interface used for this panel returns no log-probabilities for any model in it. Reading the sampled models continuously would give the symmetric test, and that test remains undone; what the available direction establishes is that the resolution effect does not require two different model families to appear.

The effect is larger than any difference between systems. Polled at the panel's own budget of four ballots, the classifiers' strict score falls from 0.648 to 0.440 while the classifiers themselves are unchanged, as the group means in Fig. 6(a) trace. The displacement is not a property of the average. Fig. 6(b) separates the three regimes item by item: read exactly, the seven classifiers occupy 0.623 to 0.660 and sit above every polled language

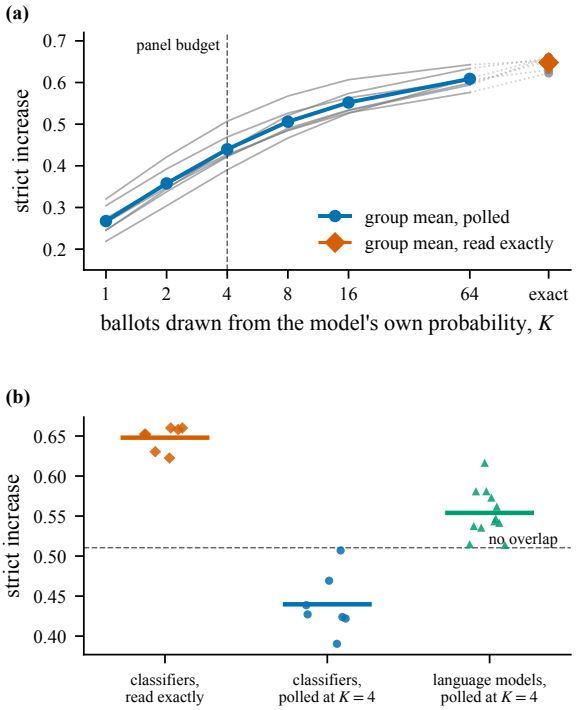

Fig. 6 The seven classifiers scored from their own probabilities and then polled from those same probabilities. (a) Strict increase against the ballot budget, grey curves per classifier; (b) the three regimes compared at the panel's own budget.

model, whereas polled at the same budget the best of them reaches only 0.507 against 0.514 for the weakest language model, so the two groups do not overlap and the entire panel inverts. A ranking that places sampled models above continuous ones is therefore reporting how each was measured.

4.9 The sampling budget

A judge polled K times can return only $K+1$ distinct values, so part of the tie rate belongs to the evaluation rather than to the system. Reducing K from four to one moves GPT-5.6-luna's satisfaction rate from 0.538 to 0.441 and its distinct value count from five to two while the model is unchanged.

Fig. 7(a) plots the tie rate against $1/K$, the axis on which that fit is a straight line, so the part of the tie rate no budget removes is read directly off the left edge. The intercept the curves reach there agrees to within five thousandths with $\Pr[p_i = p_j]$, which the measurement model below derives without fitting anything.

Table 6 Quantising the seven full-resolution classifiers and the score-valued lexicon to B levels. Predictions, models, corpus and constraints are held fixed. The last two rows give how far the twenty-system ranking moves under the strict-increase convention.

	levels B						
Tie rate T	2	4	8	16	32	64	∞
RoBERTa-GoEmotions	0.644	0.476	0.285	0.170	0.109	0.067	0.000
BERT-GoEmotions	0.526	0.379	0.241	0.158	0.099	0.065	0.000
BERTweet-emotion	0.542	0.344	0.188	0.136	0.079	0.045	0.000
VADER	0.794	0.674	0.573	0.545	0.524	0.516	0.512
Mean $\lvert\Delta\text{rank}\rvert$	8.4	8.2	6.7	3.1	1.4	0.8	0.0
Max $\lvert\Delta\text{rank}\rvert$	18	18	17	9	5	5	0

Table 7 The seven continuous classifiers, scored from their own probabilities and then polled K times from those same probabilities. Group means over 200 draws.

	continuous	polled at $K =$					
		1	2	4	8	16	64
Strict	0.648	0.267	0.358	0.440	0.506	0.552	0.609
Tie rate	0.000	0.591	0.448	0.321	0.219	0.146	0.058
Non-decreasing	0.648	0.858	0.806	0.761	0.725	0.697	0.667

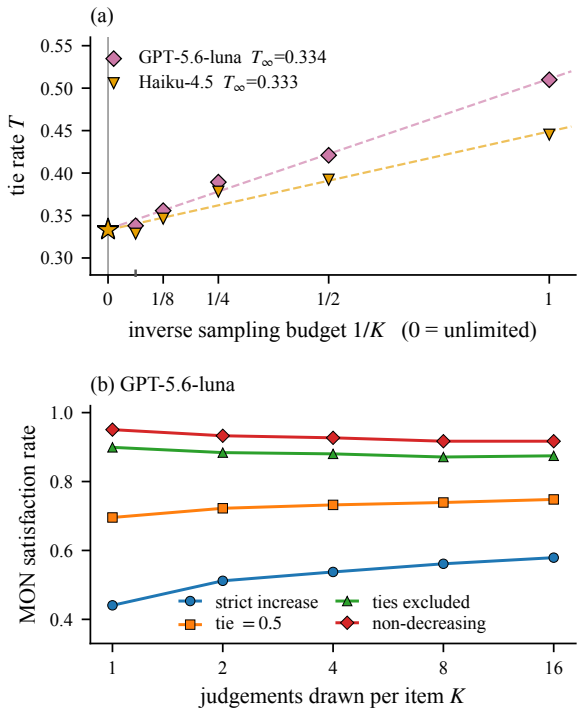

Fig. 7 (a) The tie rate against the inverse sampling budget, with the dashed descriptive fit $T(K) = T_\infty + c/K$ and stars at its intercept; (b) the four conventions separate by the tie rate at every budget.

The obvious reply, that the ties are an artefact of a cheap evaluation and would disappear under a larger budget, does not hold, and Fig. 7(b) shows why: the four conventions stay separated by the tie rate at every budget and converge only slowly. Scoring $K \in \{1, 2, 4, 8, 16\}$ on a fixed item pool, quadrupling the budget from $K = 4$ to $K = 16$ moves GPT-5.6-luna's tie rate only from 0.389 to 0.338 and Haiku's from 0.378 to 0.329, and over the whole sweep from $K = 1$ the gap between the extreme conventions narrows only from 0.510 to 0.338.

If an item's pole probability is p, polling K times returns Binomial$(K, p)/K$, so two items tie with probability

$$\begin{aligned} &\Pr\big[\hat{p}_K(x_i) = \hat{p}_K(x_j)\big] \\ &= \sum_{k=0}^{K} \Pr[X_i = k]\ \Pr[X_j = k], \end{aligned} \tag{5}$$

which has no free parameters once the pole probabilities are fixed.

Table 8 evaluates Eq. (5) with each estimator's own pole probabilities and no fitting, each item's 16-ballot rate serving as the estimate of its p, so the $K = 16$ column is not an independent test. It predicts the observed tie rate over $K \le 8$ to a mean absolute error of 0.015 for GPT-5.6-luna and 0.007 for Haiku-4.5, and its limit as $K \to \infty$ is $\Pr[p_i = p_j]$, the probability that two items carry the same pole probability. That floor is a property of the corpus and the estimator together, not of the budget, which is why the earlier $1/K$ fit is a summary of Eq. (5) rather than an independent finding.

The practical consequence is that the reporting problem cannot be spent away. A budget large enough to drive T to zero does not exist for these

Table 8 The binomial measurement model against the observed tie rate. The limit is a quantity the model derives rather than one we measured, so no observation is entered against it.

Tie rate T	K=1	2	4	8	16	limit
GPT-5.6-luna, observed	0.510	0.421	0.389	0.356	0.338	—
predicted	0.484	0.434	0.396	0.370	0.353	0.338
Haiku-4.5, observed	0.445	0.392	0.378	0.346	0.329	—
predicted	0.449	0.407	0.374	0.353	0.340	0.329

systems, so the convention that charges those ties keeps deciding the ranking no matter what the evaluation costs.

4.10 The same problem for the invariance constraint

MON asks whether one item outranks another and needs a tie weight. INV asks whether two items are scored the same, and exact equality is unreachable for a continuous score, so a benchmark must choose a tolerance ϵ and count $|f(x_{\text{diary}}) - f(x_{\text{letter}})| \leq \epsilon$ as satisfaction. That is the same kind of choice, and it behaves the same way: 138 of the 190 pairwise verdicts cross somewhere on $0 \leq \epsilon \leq 0.5$, a tolerance-instability of 0.726.

Table 9 gives the satisfaction rate as a function of ϵ. What the two constraints do not share is the direction of the bias. Under MON at $\lambda = 0$ a coarse estimator is punished because it cannot separate adjacent rungs; under INV at $\epsilon = 0$ the same coarseness is rewarded, because only a coarse estimator returns exactly equal values. The full-resolution classifiers satisfy exact invariance on 0.002 of pairs while the sampled estimators average 0.711, so the panel ordering is close to inverted between the two strictest conventions. A benchmark reporting MON under strict increase alongside INV under exact equality scores output granularity twice with opposite signs, which is why we report neither criterion alone and emit no composite of the two.

5 Discussion

5.1 What a satisfaction rate measures

A satisfaction rate is the natural way to compress conformity to an ordinal constraint into one number, and it is not solely a property of the estimator. Section 4 identifies two determinants on the reporting side. The first is the tie convention, which can overturn 134 of the 190 pairwise verdicts on this panel, an instability of $R = 0.705$ with a cluster-bootstrap interval of $[0.605, 0.742]$. The second is output resolution, which changes rankings even when models, predictions, corpus and constraints are all held fixed. Because quantising and polling are interventions rather than observations, they establish that the association between score and resolution is not merely correlational.

The direction of the bias depends on the convention. Charging ties as failures favours high-resolution outputs, which rarely tie; charging them as successes favours coarse outputs, which rarely invert. Neither convention is wrong in itself. What is wrong is reporting a ranking without the convention that produced it, and the remedy follows from Proposition 1: a benchmark of this kind should state the convention, the tie rate that bounds its effect, and R, and comparisons across systems of different output resolution should be made at a matched resolution.

Nothing in that remedy is specific to emotion. The mechanism needs only three ingredients: a criterion that is ordinal, a panel whose members differ in how many distinct values they can emit, and a scorer that must decide what a tie is worth. Those conditions are met wherever systems are compared by how often they order, prefer or rank a pair correctly, which now includes any leaderboard that places a model polled a handful of times beside one read from a continuous score. The replication of Section 4.5 is evidence that the account travels, and it is informative in both directions: on a benchmark built by other authors for a different purpose, convention dependence, the identity between tie rate and score spread, and the dependence on polling budget all reproduce, while the aggregate reversal does not, exactly where Proposition 3 says it cannot. A theory of a reporting artefact that predicts where the

Table 9 INV satisfaction against the tolerance ϵ, grouped by what the estimator's output can express. At $\epsilon = 0$ the ordering is the reverse of MON under strict increase.

	tolerance ϵ							
	0	.05	.10	.15	.20	.25	.30	.50
Continuous (n=7)	0.002	0.321	0.455	0.548	0.619	0.679	0.730	0.859
Sampled (n=12)	0.711	0.711	0.713	0.713	0.716	0.799	0.799	0.882
VADER	0.562	0.608	0.677	0.769	0.838	0.869	0.915	0.990
ρ vs. acc.	+0.14	+0.17	+0.09	+0.10	+0.10	−0.02	+0.10	−0.05

artefact is absent is more useful than one that only predicts where it appears. Emotion recognition is where we established the mechanism, because its extreme annotator disagreement is what forces constructed relational ground truth in the first place, but the reporting requirement that follows is a requirement on ordinal benchmarks in general.

Aggregate correlations are less stable than the pairwise condition. The reversal on our earlier five-system panel was real arithmetic on a panel where accuracy and resolution happened to travel together, and it disappeared once the panel was wide enough to separate them. Proposition 3 states the condition pair by pair and does not depend on that coupling, which is why the pairwise result strengthens with panel width while the aggregate one does not.

5.2 What the construction can and cannot certify

The tier split was introduced to separate tracking valence from tracking emotion. The most dramatic part of that separation, however, is not a measurement. The lexicon's same-valence value arises because the implementation assigns 0.5 to every same-valence item: a valence-only signal carries no information about which of two same-valence poles is meant, and the assignment expresses that design. It is not evidence that the split diagnoses anything. The evidence for that is the smaller, measured gap in the systems that consult more than valence, and it has none of the lexicon's starkness.

Only relational properties. We do not equate the induced distributions with human distributions, and prior work has measured that gap directly [9]. The control arm of Section 4.2 shows the construction is doing work rather than labelling it: removing the appraisal binding while holding everything else fixed drops MON_ρ from 0.865 to 0.311, and does so by making the generator separate the poles by severity instead of by appraisal. The generator swap shows that the guarantees are not equally strong even inside this restriction: order survives a change of generator family and invariance degrades.

The acceptance criteria deserve the same accounting. Excluding items in a constructed benchmark invites the suspicion that convenient ones were kept. The criteria here, cue fidelity and a minimum length, are defined independently of the reported outcomes, every construction statistic improves after they are applied, and the excluded items are shipped so that the reader can check both sides. Two caveats from review are retained: the fidelity readers overlap two of the four panel families, and restricting to the non-overlapping judges leaves the improvement intact but the strict-increase component of it is not significant against a stratified placebo.

5.3 Limitations

Five limitations remain. First, TILT certifies order and invariance, not human ambiguity: a separation criterion was part of the original design and did not hold. Second, appraisal binding is used only for the same-valence tier, so the reverse control, whether cross-valence items would also improve under a binding they do not have, remains untested. Third, an aggregate reversal and the value of R depend on panel composition and should not be treated as constants. Fourth, the reference panel consists of model judges and every stimulus is model-generated, so the constructed relations have not been validated as human percepts, and the same-valence tier rests on fifteen scenarios, which is the narrowest base in the construction. Fifth, the resolution intervention is one-directional. Polling a classifier that can also

be read exactly breaks the confound inside a single model, but the symmetric test, reading the sampled models continuously from their own token probabilities, could not be run because the serving interface exposes no log-probabilities for this panel. A panel served with them would settle whether the curve of Fig. 6(a) is traced in the other direction as well.

6 Conclusion

A benchmark score is usually read as a property of a system, but it is the output of a measurement chain in which the system's resolution determines how many ordered pairs tie and a reporting convention decides what those ties are worth. We made that chain explicit for ordinal constraints, using a benchmark whose order is fixed by construction rather than by annotation, and expressed the reporting choice as a single parameter. Writing the score as a linear function of that parameter turns three questions a benchmark cannot otherwise answer, namely how far a convention can move a score, when two systems exchange places, and when the benchmark contradicts an external axis, into closed-form conditions.

On a panel of twenty estimators the choice moves a score by exactly its tie rate and can overturn seven of every ten pairwise verdicts. The cause was then isolated inside a single model class rather than inferred across the panel: reading seven classifiers by polling rather than exactly, which changes no model, no prediction, no item and no constraint, drops their strict score from 0.648 to 0.440 and sends every one of them below every language model polled at the same budget, with the two groups no longer overlapping. Under a fixed convention the score therefore tracks output resolution far more closely than independent accuracy, the invariance criterion carries the same structure with the opposite bias, and binding the cue slots to appraisal dimensions is what makes the ladder gradeable at all. On an independently released benchmark the convention dependence and the tie-rate identity reproduce, while the aggregate reversal is absent exactly where the model says it must be.

The reporting requirement that follows is small and general: state the convention, state the tie rate that bounds its effect, and compare systems of different output resolution only at a matched resolution. In future studies, we will investigate whether the constructed relations hold as human percepts rather than as model judgements, run the symmetric resolution intervention on a panel whose serving interface exposes token probabilities, and examine how the instability index behaves on panels drawn from a different generation of systems. The corpus, the constraints and the scorer are released so that these questions can be asked of the same material.

Acknowledgements. This work was supported by JST SPRING (No. JPMJSP2150).

Statements and Declarations

Funding. This work was supported by JST SPRING (No. JPMJSP2150).

Author contributions. Keito Inoshita is the sole author. Stated in CRediT terms: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing — original draft, Writing — review and editing, Visualization, Project administration, Funding acquisition.

Competing interests. The author declares that he has no competing interests or financial conflicts to disclose.

Ethics approval. Not applicable. The study involves no human participants and no human annotation; the relational ground truth of the benchmark is fixed by construction.

Data and code availability. The corpus, the constraint files, the scorer and the analysis code are available at https://github.com/keito-git/tilt-benchmark.

References

[1] Baan J, Aziz W, Plank B, et al (2022) Stop measuring calibration when humans disagree. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp 1892–1915, https://doi.org/10.18653/v1/2022.emnlp-main.124

[2] Bommana D, Saim M, Jiang T (2026) Chiaroscuro for emotions: A contrastive emotion benchmark grounded in appraisal theory. arXiv

[3] Chen TY, Kuo FC, Liu H, et al (2018) Metamorphic testing: A review of challenges and opportunities. ACM Computing Surveys 51(1):4:1–4:27. https://doi.org/10.1145/3143561

[4] Chen Y, Eger S (2025) Do emotions really affect argument convincingness? a dynamic approach with LLM-based manipulation checks. In: Findings of the Association for Computational Linguistics: ACL 2025, pp 24357–24381, https://doi.org/10.18653/v1/2025.findings-acl.1250

[5] Cho S, Ruberto S, Terragni V (2025) Metamorphic testing of large language models for natural language processing. In: 2025 IEEE International Conference on Software Maintenance and Evolution, pp 174–186, https://doi.org/10.1109/ICSME64153.2025.00025

[6] Colombo P, Noiry N, Irurozki E, et al (2022) What are the best systems? new perspectives on NLP benchmarking. In: Proceedings of the 36th International Conference on Neural Information Processing Systems, pp 26915–26932

[7] Demszky D, Movshovitz-Attias D, Ko J, et al (2020) GoEmotions: A dataset of fine-grained emotions. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp 4040–4054, https://doi.org/10.18653/v1/2020.acl-main.372

[8] Deutsch D, Foster G, Freitag M (2023) Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp 12914–12929, https://doi.org/10.18653/v1/2023.emnlp-main.798

[9] Gandhi K, Lynch Z, Fränken JP, et al (2024) Human-like affective cognition in foundation models. arXiv:2409.11733, arXiv:2409.11733

[10] Gardner M, Artzi Y, Basmov V, et al (2020) Evaluating models' local decision boundaries via contrast sets. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp 1307–1323, https://doi.org/10.18653/v1/2020.findings-emnlp.117

[11] Gevers I, Daelemans W (2026) Benchmarking the benchmarks: Testing the predictive validity of commonsense benchmarks. arXiv https://doi.org/10.48550/arXiv.2608.03340

[12] Goodman LA, Kruskal WH (1954) Measures of association for cross classifications. Journal of the American Statistical Association 49(268):732–764. https://doi.org/10.2307/2281536

[13] Inoshita K (2026) Bias-corrected ceilings of emotion predictability from human label variation based on instance-level Fano bounds. arXiv:2608.15619, arXiv:2608.15619

[14] Inoshita K (2026) Bridging the silos in affective AI: A critical perspective from data to society. AI & Society https://doi.org/10.1007/s00146-026-03324-y

[15] Inoshita K, Zhou X, Kawai A, et al (2026) LLMs capture emotion labels, not emotion uncertainty: Distributional analysis and calibration of human–LLM judgment gaps. arXiv https://doi.org/10.48550/arXiv.2604.27345

[16] Kendall MG (1945) The treatment of ties in ranking problems. Biometrika 33(3):239–251. https://doi.org/10.1093/biomet/33.3.239

[17] Leonardelli E, Abercrombie G, Almanea D, et al (2023) SemEval-2023 task 11: Learning with disagreements (LeWiDi). In: Proceedings of the 17th International Workshop on Semantic Evaluation, pp 2304–2318, https://doi.org/10.18653/v1/2023.semeval-1.314

[18] Liu JM, Cao H, Sun R, et al (2024) CAPE: A Chinese dataset for appraisal-based emotional generation using large language models. arXiv:2410.14145, arXiv:2410.14145

[19] Manino E, Rozanova J, Carvalho D, et al (2022) Systematicity, compositionality and transitivity of deep NLP models: a metamorphic testing perspective. In: Findings of the Association for Computational Linguistics: ACL 2022, pp 2355–2366, https://doi.org/10.18653/v1/2022.findings-acl.185

[20] Plank B (2022) The "problem" of human label variation: On ground truth in data, modeling and evaluation. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp 10671–10682, https://doi.org/10.18653/v1/2022.emnlp-main.731

[21] Ribeiro MT, Wu T, Guestrin C, et al (2020) Beyond accuracy: Behavioral testing of NLP models with CheckList. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp 4902–4912, https://doi.org/10.18653/v1/2020.acl-main.442

[22] Rofin M, Mikhailov V, Florinsky M, et al (2023) Vote'n'rank: Revision of benchmarking with social choice theory. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pp 670–686, https://doi.org/10.18653/v1/2023.eacl-main.48

[23] Segura S, Fraser G, Sanchez AB, et al (2016) A survey on metamorphic testing. IEEE Transactions on Software Engineering 42(9):805–824. https://doi.org/10.1109/TSE.2016.2532875

[24] Smith CA, Ellsworth PC (1985) Patterns of cognitive appraisal in emotion. Journal of Personality and Social Psychology 48(4):813–838. https://doi.org/10.1037/0022-3514.48.4.813

[25] Sun Z, Xu H, Uusberg A, et al (2026) CAREBench: Evaluating LLMs' emotion understanding by assessing cognitive appraisal reasoning. arXiv https://doi.org/10.48550/arXiv.2605.17176

[26] Troiano E, Oberländer L, Klinger R (2023) Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. Computational Linguistics 49(1):1–72. https://doi.org/10.1162/coli_a_00461

[27] Valcarce D, Bellogín A, Parapar J, et al (2020) Assessing ranking metrics in top-N recommendation. Information Retrieval Journal 23(4):411–448. https://doi.org/10.1007/s10791-020-09377-x

[28] de Vries S, Thierens D (2025) Generating the ground truth: Synthetic data for soft label and label noise research. International Journal of Data Science and Analytics 20(6):5603–5615. https://doi.org/10.1007/s41060-025-00786-z

[29] Xie X, Ho JWK, Murphy C, et al (2011) Testing and validating machine learning classifiers by metamorphic testing. Journal of Systems and Software 84(4):544–558. https://doi.org/10.1016/j.jss.2010.11.920

[30] Zheng L, Chiang WL, Sheng Y, et al (2023) Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In: Advances in Neural Information Processing Systems 36: Datasets and Benchmarks Track

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TILTJDSAsupplementary.pdf