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

Keito Inoshita

This document holds four analyses the main text refers to but does not display in full: the sampling-budget sweep behind the section on the sampling budget, the two further reporting degrees of freedom raised in the section on what a convention can move, and the axis-specific and leave-one-out results of the external replication. Nothing here is needed to follow the argument of the main text. Each section states the question it answers, how the numbers were produced, and what they do not establish. Parts of the main text are named rather than numbered, because the two documents are compiled independently; all other cross-references are internal to this document.

1 Tie rate against the sampling budget

A judge polled K times can return only a limited number of distinct values, so a sceptical reading of the main result is available: the ties that the conventions argue over are an artefact of polling cheaply, and they would disappear if the evaluation were run at a serious budget. This section tests that reading directly.

Each of the two sampled estimators was scored at $K \in \{1, 2, 4, 8, 16\}$ on the same items and the same constraints. The ballots at $K \leq 4$ are the ones used in the main text; $K = 8$ and $K = 16$ add further sampling seeds under an unchanged prompt and protocol. The item pool is fixed at the items the estimator covers at $K = 1$, which is $1{,}021$ items and 506 constraint pairs for GPT-5.6-luna and 882 items and 283 pairs for Haiku-4.5, so that the curve reports the budget rather than a coverage difference between budgets.

The sceptical reading does not survive. Quadrupling the budget from $K = 4$ to $K = 16$ moves the tie rate only from 0.389 to 0.338 and from 0.378 to 0.329. The tie rate is close to linear in the inverse budget, which is what the first panel of the budget figure in the main text shows, so it approaches a positive intercept rather than zero. The descriptive fit $T(K) = T_\infty + c/K$ gives $T_\infty = 0.334$ at $R^2 = 0.990$ for GPT-5.6-luna and $T_\infty = 0.333$ at $R^2 = 0.950$ for Haiku-4.5, within five thousandths of the floor $\Pr[p_i = p_j]$ that the binomial model of the main text derives with nothing fitted, namely 0.338 and 0.329. At the budget used in the main text the part of the tie rate that the budget could remove is 0.055 for GPT-5.6-luna and 0.045 for Haiku-4.5, against tie rates of 0.389 and 0.378. At most one part in seven of the tie rate is the evaluation, and the rest is the estimator on this corpus.

Two limits apply to this section. The straight line is a description and not a mechanism; it agrees with the binomial model because it is a summary of it, so the agreement between the fitted intercept and the derived floor is a consistency check and not a second piece of evidence. And each item's 16-ballot rate is what supplies p to that model, so the $K = 16$ point is where the estimate came from rather than an independent test of it.

2 Scores under four conventions at every budget

Table 1 gives the values behind the budget figure of the main text for five budgets and four conventions. Two readings matter beyond the tie rate itself.

The budget is not a neutral improvement. Raising K from 1 to 16 lifts the strict-increase score from 0.441 to 0.579 for GPT-5.6-luna and from 0.509 to 0.604 for Haiku-4.5, while over the same range the non-decreasing score falls from 0.951 to 0.917 and from 0.954 to 0.933. A ballot that resolves a tie can resolve it in either direction, and some resolve against the constraint: the inversion rate for GPT-5.6-luna rises from 0.049 at $K = 1$ to 0.083 at $K = 16$. Spending

Table 1: MON under four tie conventions as a function of the number of judgements drawn per item. Each estimator is scored on the item pool it covers at $K = 1$.

	K	Strict	Non-dec.	Tie rate	Distinct
GPT-5.6-luna	1	0.441	0.951	0.510	2
	2	0.512	0.933	0.421	3
	4	0.538	0.927	0.389	5
	8	0.561	0.917	0.356	9
	16	0.579	0.917	0.338	17
Haiku-4.5	1	0.509	0.954	0.445	2
	2	0.551	0.943	0.392	3
	4	0.562	0.940	0.378	7
	8	0.590	0.936	0.346	22
	16	0.604	0.933	0.329	45

more on the evaluation therefore moves the two extreme conventions in opposite directions, which is the argument of the main text restated at the level of the budget rather than of the scorer.

Resolution is a property of the protocol and the judge together. GPT-5.6-luna returns exactly $K + 1$ distinct values at every budget, which is what a fixed number of ballots allows. Haiku-4.5 returns 7, 22 and 45 at $K = 4$, 8 and 16, well above $K + 1$, because it abstains on part of the corpus: when a judgement is not a vote the denominator of the vote share varies from item to item, and the set of reachable values is correspondingly larger. A reader who infers an estimator's resolution from its nominal budget will therefore be wrong for any judge that is allowed to decline.

3 Population and constant ladders

The main text reports that tie handling is not the only undocumented choice in a score of this kind. Two others are available to anyone reimplementing the scorer. The first is the population: whether the score is computed over the interpolated rungs alone or over all rungs including the two endpoints on which the scenarios were selected. The second is what to do with a ladder on which a system returns the same value at every rung, which contributes only ties and can be kept or dropped. The choices are independent, because the first decides which pairs enter and the second decides which ladders do, so they form a 2×2 grid. Table 2 gives all four cells for the twenty estimators.

The grid is worth reporting in full because the two choices do not act on the same systems. For the seven full-resolution classifiers, dropping constant ladders changes nothing whatever, since a classifier that never returns the same value twice has no constant ladders to drop, while the population moves them by 0.051 to 0.097. For the twelve polled language models the exposure is reversed: dropping constant ladders moves the score by 0.078 to 0.143, and the population by at most 0.054. The lexicon is the extreme case of the second pattern, moving 0.295 under the ladder rule and 0.026 under the population. The range quoted in the main text is therefore not one effect but two, and which one a system is exposed to is decided by its output resolution before any comparison is made.

The released scorer fixes the first cell, interpolated rungs with constant ladders kept. That cell is the lowest of the four for nineteen of the twenty estimators, the exception being Haiku-4.5, whose all-rungs value is lower by 0.001. We fix it not because it is favourable but because it is the only cell that neither reports the selection rule back as a result nor removes the ladders on which a system fails to move at all. The consequence for a reader is the practical one: a published satisfaction rate that does not name its cell cannot be compared against these numbers, since the same predictions support four of them.

4 CHIARO by external axis, and leave-one-out

The external replication of the main text admits two external axes against which a satisfaction rate can be correlated: the benchmark's own multiple-choice accuracy, and accuracy on GoEmotions measured under our protocol. Table 3 gives the correlation under each convention against each axis, together with the coupling between the axis and the tie rate.

Table 2: Satisfaction rate under the 2×2 grid of population and constant-ladder handling. The released scorer reports the first column.

	Interpolated		All rungs		
Estimator	keep	drop	keep	drop	Range
RoBERTa-GoEmotions	0.660	0.660	0.757	0.757	0.097
Seed-1.6-flash	0.542	0.665	0.559	0.587	0.123
DistilRoBERTa-7	0.630	0.630	0.681	0.681	0.051
BERT-GoEmotions	0.652	0.652	0.721	0.721	0.069
Command-R7B	0.581	0.674	0.627	0.635	0.093
Nova-Micro	0.514	0.629	0.568	0.592	0.115
Nemotron-3-nano	0.573	0.671	0.601	0.617	0.098
Grok-4.3	0.536	0.655	0.578	0.587	0.119
Solar-Pro4	0.581	0.670	0.603	0.608	0.089
Kimi-K2.5	0.515	0.658	0.550	0.570	0.143
GPT-5.6-luna	0.538	0.652	0.581	0.593	0.115
Haiku-4.5	0.562	0.657	0.561	0.572	0.096
MiniMax-01	0.547	0.647	0.599	0.604	0.100
DistilRoBERTa-7c	0.652	0.652	0.728	0.728	0.076
GLM-5.3-flash	0.543	0.653	0.586	0.597	0.110
BERTweet-emotion	0.658	0.658	0.726	0.726	0.068
BERT-emotion	0.660	0.660	0.719	0.719	0.059
DistilBERT-emotion	0.623	0.623	0.699	0.699	0.077
Phi-4	0.617	0.695	0.643	0.652	0.078
VADER	0.314	0.609	0.341	0.678	0.364

Table 3: CHIARO, by external axis: Spearman's ρ between the satisfaction rate and each axis under the four conventions. Coupling is ρ between the axis and the tie rate. The sign changes on the axis whose coupling is negative and not on the other, which is the condition Proposition 3 of the main text states.

Axis	n	Coupling	Strict	Non-dec.	Tie = ½	Excl.	Flip
CHIARO MCQ	8	$+0.313$	$+0.78$	$+0.67$	$+0.69$	$+0.69$	no
GoEmotions	5	-0.975	$+0.10$	-0.60	$+0.10$	$+0.10$	yes

The two axes separate exactly as the external-axis proposition of the main text requires. On the benchmark's own multiple-choice axis, which is positively coupled with the tie rate at $+0.313$ over eight estimators, the convention moves the correlation by 0.114 and never changes its sign. On the GoEmotions axis, which is negatively coupled at -0.975, the correlation runs from $+0.10$ under strict increase to -0.60 under non-decreasing. The proposition states that a convention can reverse a verdict only against an axis coupled negatively with the tie rate, and these are the only two axes this benchmark offers; they fall on either side of that condition.

The GoEmotions axis nonetheless carries only five estimators, and Table 4 shows what that costs. Dropping any single estimator leaves every correlation on the multiple-choice axis positive, between $+0.500$ and $+0.955$ across all four conventions, so the absence of a sign change there is not the work of one system. On the GoEmotions axis the same operation moves three of the four conventions across zero, from -0.800 to $+0.400$. We therefore report the reversal on that axis as an instance of the narrow-panel caution of the main text, consistent with the proposition but not independently load-bearing, and we rest the replication on what does survive: the value dependence and the identity between spread and tie rate, which hold for all nine estimators of that benchmark.

Table 4: CHIARO, leave-one-estimator-out: the range each correlation takes when any single estimator is dropped.

Axis	Convention	Min	Max
CHIARO MCQ	Strict	$+0.673$	$+0.955$
CHIARO MCQ	Non-dec.	$+0.500$	$+0.864$
CHIARO MCQ	Tie $=$ ½	$+0.541$	$+0.901$
CHIARO MCQ	Excl.	$+0.541$	$+0.901$
GoEmotions	Strict	-0.800	$+0.400$
GoEmotions	Non-dec.	-0.800	-0.200
GoEmotions	Tie $=$ ½	-0.800	$+0.400$
GoEmotions	Excl.	-0.800	$+0.400$