WHAT ANNOTATOR-CONDITIONED MODELS LEARN DEPENDS ON THE TEXT MODEL: DECOMPOSING THE GAIN FROM PERSONALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In subjective annotation tasks, it is now common to treat each annotator's judgment as information to be modeled rather than noise to be aggregated away. Annotator-conditioned models improve predictive performance, but it has never been measured how much of that gain is a content-dependent component and how much is an additive, input-independent output bias specific to each annotator. We decompose this gain as a difference between nested predictors and show that its composition depends on both the shared text model and how that model is adapted. With increasingly expressive fixed text representations, the content-dependent share rises from 7.3% for a linear encoder to 30.5% for a frozen Transformer, yet remains additive-dominant. When the same Transformer is instead adapted end to end within each annotator-conditioned model, the decomposition reverses and the content-dependent share reaches 89.3%; this reversal is explained neither by hidden dimension nor by pretraining, and replicates on a second dataset. Even with the same architecture and the same data, freezing versus jointly adapting the encoder changes the content-dependent share from 30.5% to 89.3%. This content-dependent component is not an estimation artifact in which the extra capacity of M_2 produces a positive gain even under the null: against labels generated from a model with no interaction, the same estimator returns zero. The composition of the personalization gain is therefore not a fixed property of the annotators but a quantity set by the representation and by how it is adapted during personalization, so there is no model-independent answer to whether a per-annotator intercept suffices or content-dependent personalization is needed. Finally, we show that the verdict itself depends decisively on how the null distribution is constructed.

1 INTRODUCTION

In tasks whose judgments have no single correct answer, such as emotion or sarcasm, disagreement among annotators is unavoidable. For a long time this disagreement was treated as measurement noise and collapsed into a single ground truth by majority vote or averaging. More recently, a stance that treats disagreement itself as information intrinsic to the task has taken hold (Plank, 2022; Uma et al., 2021; Basile et al., 2021). Under this view the goal becomes predicting each annotator's judgment individually, and conditioning the model on the annotator identifier has become the standard design (Davani et al., 2022; Rodrigues & Pereira, 2018).

Annotator-conditioned models do improve predictive performance (Davani et al., 2022; Gordon et al., 2022), an improvement often explained as the model learning each annotator's idiosyncratic way of interpreting the input. Yet the information gained by conditioning on the annotator mixes two components of different natures. One is an additive, input-independent output bias for each annotator that absorbs how often the annotator uses each label. The other is a content-dependent component: the same input elicits different responses from different annotators. The former can be captured by a single per-annotator intercept, whereas only the latter depends on the annotator–input pair and is learned as an annotator–text interaction. We refer to the former as the marginal bias and to the latter as the content-dependent association. We define both as predictive quantities and do not treat the latter as causal evidence that annotators "read the same thing differently."

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

What annotator modeling reports is the total gain that combines the two, and its composition has not been measured in isolation. Without the composition, the reported gain does not tell us how to design personalization. If the gain is mostly the marginal bias, personalization reduces to a per-annotator intercept, a recalibration of the output layer, and the learned annotator representation is unnecessary; if the content-dependent component dominates, an intercept alone leaves most of the gain on the table. More troubling, the composition is implicitly assumed to be a property of the annotators or the task. If instead the composition changes with the text model used to measure it, then concluding from a single model whether a per-annotator intercept suffices or content-dependent personalization is needed reports a property of the encoder, not of the annotators. Likewise, if a preprocessing choice such as coarsening the label space silently changes the composition, then what annotator modeling captures changes across that choice. And when the content-dependent component is small, detecting it requires a certain scale per corpus and an explicit statement of how the null is constructed; arguing for or against an annotator effect without the backing of calibration risks the wrong conclusion.

In this study we decompose the gain from conditioning on the annotator as a difference between nested predictors and ask what determines its composition. We place three nested predictors, namely a text-only model, a model with per-annotator intercepts, and a model with a low-rank annotator–text interaction, and measure the marginal and content-dependent components on the same scale as differences in held-out predictive log-likelihood. We apply this estimator uniformly, from linear and nonlinear word-aggregation encoders to self-attention Transformers used both as frozen representations and fine-tuned end to end, and trace how the composition moves as we vary the representation class of the shared text model and the protocol by which it is adapted. Whether a measured value is distinguishable from noise is checked by a calibration procedure that generates the null from a model fitted to the very hypothesis under test, and we further use synthetic data to map, systematically, the conditions under which the content-dependent component can be identified. This clarifies whether the composition of the personalization gain is a fixed property of the annotators or a quantity relative to the text model and the measurement design. Our contributions are threefold.

i) A nested-model decomposition estimator separates the gain from conditioning on the annotator into an additive marginal component and a content-dependent component, measures both on the same scale as differences in held-out predictive log-likelihood, applies in the same form from linear predictors to fine-tuned Transformers, and is calibrated to return zero when there is no interaction.

ii) The composition of this gain depends on both the shared representation and how it is adapted: moving from a linear to a nonlinear and then to a frozen Transformer representation raises the content-dependent share from 7.3% to 17.9% and 30.5%, while adapting the same Transformer end to end within each nested model produces a qualitatively different regime in which the share reaches 89.3%, a reversal explained neither by hidden dimension nor by pretraining and replicated on a second dataset.

iii) Identifying the content-dependent component requires both a sufficiently expressive shared representation and enough data per annotator, and the detection verdict depends on how the null used for calibration is constructed, so whether an annotator effect is present is a property relative to the measurement rather than a fixed property of the annotators.

The rest of the paper is organized as follows. Section 2 reviews related work, Section 3 defines the decomposition and the estimator, Section 4 presents the experiments, and Section 5 discusses implications and limitations.

2 Related Work

Treating annotator disagreement as task-intrinsic signal rather than noise is now standard (Plank, 2022; Uma et al., 2021; Basile et al., 2021): it has been formalized as human label variation and supported by distributional-label learning (Geng, 2016), empirical studies of learning and evaluation under such variation (Kurniawan et al., 2025), and soft-label aggregation (Wright & Augenstein, 2025). The growth of corpora that release per-annotator labels for emotion and sarcasm (Demszky et al., 2020; Frenda et al., 2023; Casola et al., 2024), together with shared tasks that score predicted judgment distributions (Leonardelli et al., 2025; Lo et al., 2025), makes reproducing each annotator's judgment the direct objective; critical and distributional analyses of emotion annotation reinforce

this (Inoshita, 2026b; Inoshita et al., 2026). Aggregation is known to discard minority judgments (Fleisig et al., 2023) and to harm downstream calibration (Baan et al., 2022; Khurana et al., 2024), and whether sarcasm disagreement is contextual or annotator-driven has been examined (Jang et al., 2024), the distinction that motivates our thread-level split. We do not question this stance, but ask what the annotator modeling used under it learns.

Annotator modeling dates back to observer-error-rate estimation with per-annotator confusion matrices (Dawid & Skene, 1979; Whitehill et al., 2009), carried into deep models as a confusion-matrix output layer (Rodrigues & Pereira, 2018); our per-annotator intercept β_a is a more constrained, log-odds-additive parameterization (K versus K^2 degrees of freedom). More flexible designs use per-annotator output heads (Davani et al., 2022; Fornaciari et al., 2021) or predict each annotator's judgment directly (Gordon et al., 2022); these can express annotator–input interactions but are evaluated by classification metrics and never report how much of the gain is content-dependent. A known pitfall is that when the same annotators appear in training and test, a model can learn the annotator rather than the task (Geva et al., 2019); because the content-dependent component depends on the annotator–item pair, item memorization can likewise leak into the interaction when training and test share items, which is why we add a thread-level split.

Decomposing a user–item interaction into a user bias, an item bias, and a low-rank interaction is established in recommender systems (Koren et al., 2009), and our nested models are isomorphic to it; input-dependent annotator expertise (Yan et al., 2014), multidimensional annotator vectors (Welinder et al., 2010), and generalizability theory in psychometrics (Cronbach et al., 1972) all separate the same two components, so the decomposition itself is not our contribution. What we add is to measure this composition on a single scale, the difference in held-out predictive log-likelihood, report it as a share of the total gain, and show that the share depends on the representation class of the shared text model, which to our knowledge has not been reported. Crowdsourcing estimation-rate results (Karger et al., 2011; Dawid & Skene, 1979) concern recovery, not testing for an interaction; what is detectable depends on responses per annotator, labels, and assignment, which we calibrate by injecting interactions of known strength. That label variation imposes a ceiling on emotion predictability has been formalized as an instance-level information-theoretic bound (Inoshita, 2026a); we ask how much of the gain toward that ceiling is content-dependent.

3 Decomposing the Gain

3.1 Problem Setup

An annotator $a \in \{1, \ldots, A\}$ assigns a label $y \in \{1, \ldots, K\}$ to an item i with text features $\phi(x_i) \in \mathbb{R}^D$; observations are triples (a, i, y), and different annotators generally see different items. The goal is to predict the distribution of y for held-out (a, i) pairs. To measure the gain from conditioning on the annotator we fit three nested predictors,

$$\mathrm{M}_0 : p(y \mid x) \propto \exp(W^\top \phi(x) + b), \tag{1}$$

$$\mathrm{M}_1 : p(y \mid x, a) \propto \exp(W^\top \phi(x) + b + \beta_a), \tag{2}$$

$$\mathrm{M}_2 : p(y \mid x, a) \propto \exp(W^\top \phi(x) + b + \beta_a + u_a^\top V \phi(x)), \tag{3}$$

where $\beta_a \in \mathbb{R}^K$ is annotator a's intercept and $u_a \in \mathbb{R}^r$, $V \in \mathbb{R}^{r \times D \times K}$ form a rank-r interaction. M_1 expresses only the annotator's marginal tendency; M_2 expresses individual differences in the response to the same content.

We write the held-out negative log-likelihood as $\mathrm{NLL}(\cdot)$, on the annotator scale (the mean log loss per annotator, averaged with equal weight across annotators, so that a few heavily annotating annotators do not dominate). On this scale,

$$\Delta_{\mathrm{marg}} = \mathrm{NLL}(\mathrm{M}_0) - \mathrm{NLL}(\mathrm{M}_1), \qquad \Delta_{\mathrm{cond}} = \mathrm{NLL}(\mathrm{M}_1) - \mathrm{NLL}(\mathrm{M}_2), \tag{4}$$

where Δ_{marg} is the additive component, Δ_{cond} the content-dependent component, and their sum the total gain. We center $\phi(x)$ to mean 0 (absolute column mean at most 1.0×10^{-5}), so the interaction has zero mean over inputs and cannot absorb a per-annotator constant offset; β_a is an additive term in log-odds space, not a label frequency (Appendix M). Our main quantity is the content-dependent share $\Delta_{\mathrm{cond}}/(\Delta_{\mathrm{marg}} + \Delta_{\mathrm{cond}})$.

This decomposition is isomorphic to the user-bias / user–item-interaction split of recommender systems and psychometrics, and the formulation itself is not new; Δ_{cond} is a quantity relative to the model class and split design, not an existence claim that "content-dependent individual differences exist." We make this relativity explicit and give the conditions under which the reported value is interpretable.

3.2 Fairness Between Nested Models

Because what we measure is a held-out prediction difference between nested predictors, its value can arise not only from the content-dependent structure that is the object of the decomposition but also from any asymmetry left between the nested models. If M_2 could tune its capacity more freely than M_1, or spend more gradient steps, then Δ_{cond} would swing positive even with no content-dependent signal at all. The validity of the estimator therefore hinges on letting the two models differ only in the presence of the interaction, that is, on keeping capacity selection and the optimization budget symmetric. We remove these two asymmetries explicitly.

The first is the asymmetry in capacity selection. If only M_2 could choose the regularization of its interaction term while M_1's regularization were fixed, the difference in tunability itself would show up as a gain. We let every nested model choose its own regularization coefficient from a common grid on a validation subset carved from the training split, and for M_2 we also choose the interaction rank from $\{1, 2, 4, 8\}$ on the same validation subset. Under this symmetry, when there is no content-dependent signal the penalty collapses the interaction term, M_2 falls back to M_1, and Δ_{cond} drops to 0; that this collapse actually occurs is verified by the calibration of Section 4.2, which measures Δ_{cond} under a null with no interaction.

The second is the asymmetry in optimization budget. On small corpora, using a fixed number of batches and iterations while the training set is small stops the nested models before convergence, so the comparison measures optimization noise rather than content. We set the batch width and the number of iterations from the size of the training split, securing a comparable number of gradient steps across corpora. With the two symmetries imposed, Δ_{cond} reflects how much of the content-dependent structure the shared text representation can capture, not a difference in tuning budget; how this quantity moves with the representation class of the shared component is measured in Section 4. Full optimization and grid details are in Appendix M.

3.3 Splits and the Null

Sharing items between training and test overestimates the gain, so besides an item-level split we use a post-thread-level split and a split satisfying both; all use 5-fold cross-fitting with annotator-level cluster-bootstrap inference. Our primary calibration generates the null from a model fitted to the hypothesis under test: we fit the no-interaction M_1, obtain $L_{ai} = W^\top \phi(x_i) + b + \beta_a$, and redraw labels

$$y_{ai} \sim \text{Categorical}\big(\text{softmax}(L_{ai} + s\,T_{ai})\big), \qquad T_{ai} = \frac{\tilde{u}_a(\tilde{V}^\top \phi(x_i))}{\text{sd}[\tilde{u}_a(\tilde{V}^\top \phi(x_i))]}, \tag{5}$$

keeping $\phi(x_i)$ and the annotator–item assignment as observed, so the synthesized problem tracks the real corpus; $s = 0$ is exactly the null, making this a parametric bootstrap, and T_{ai} is standardized so s (the injected log-odds standard deviation) is comparable across corpora. Unlike a null drawn from a corpus-independent distribution, the difficulty of the synthesized prediction problem here tracks the real data through the fitted model.

As an assumption-free secondary control we use two randomizations that break the annotator–label correspondence: a within-item permutation of annotator identifiers, which preserves each item's label distribution and each annotator's workload, and an independent redraw of each observation's identifier. Under either, Δ_{cond} should approach 0, though it is expected to be slightly negative because selecting the interaction rank and regularization on a validation subset incurs finite-sample error; we report this bias as the control value itself. From the null we read a 95th-percentile threshold and a calibration p-value (lower bound $1/(R+1)$ for R replicates); if the estimate at $s = 0$ departs from 0, the split is not calibrated, and its value on the real data must not be interpreted. Further identifiability details are in Appendix B. "Calibration" here refers to the estimator's behavior under the null, not confidence calibration of predicted probabilities (Guo et al., 2017).

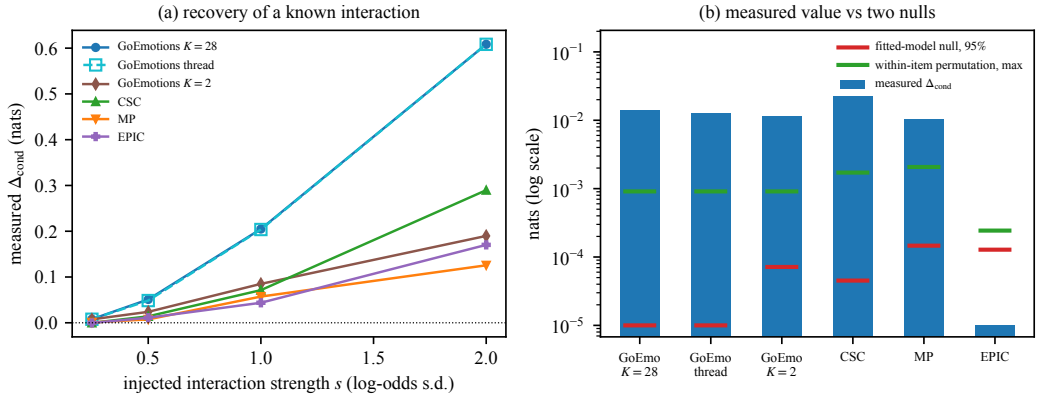

Figure 1: Calibration: (a) recovery of an injected interaction of known strength s; (b) the measured Δ_{cond} against two null constructions (log axis).

4 Experiments

4.1 Experimental Design

We use four datasets with per-annotator judgments that differ by orders of magnitude in the number of annotators, annotations per annotator, and labels: GoEmotions (Demszky et al., 2020) (28 emotion labels, the most heavily annotated), the LeWiDi (Leonardelli et al., 2025) CSC (6-point) and MP (binary) sarcasm datasets, and EPIC (Frenda et al., 2023) (binary irony). Only GoEmotions carries thread identifiers, allowing the item, thread, and both-at-once splits. Two independent feature families, a TF-IDF representation reduced to 256 dimensions by an SVD and frozen 384-dimensional sentence embeddings, are estimated from the training portion of each split only, and confirm that the results do not depend on the choice of features. Evaluation compares the held-out predictions of $\mathrm{M}_0, \mathrm{M}_1, \mathrm{M}_2$ on the annotator scale with annotator-level cluster-bootstrap intervals. Dataset sizes, preprocessing, optimization, and hyperparameters are in Appendices A and M.

4.2 Calibration and Negative Controls

Figure 1(a) shows that the estimator is calibrated: for synthetic labels with no injected interaction Δ_{cond} has near-zero mean under every condition, and injected effects are recovered monotonically, reaching $+0.0078$ at $s = 0.25$ and $+0.2048$ at $s = 1.0$ on the GoEmotions item split, with smaller recovery for narrower label spaces at the same s.

Crucially, the null construction matters. A corpus-independent Gaussian null is not calibrated against the fitted null: its 95th percentile is $363\times$ larger for MP, burying a real effect, and $53\times$ larger for EPIC, yet sits on the negative side for CSC. The fitted parametric bootstrap and the assumption-free within-item permutation instead agree: the permutation reaches $p = 0.024$ ($R = 40$) on CSC and MP and $p = 0.048$ ($R = 20$) on the GoEmotions reference condition, and on the detected conditions the observed value is 5.1 to 19.2 times the permutation maximum. Figure 1(b) plots the measured values against both null thresholds, and Appendix B gives the details. On the negative controls Δ_{cond} is near or below 0 except for the item+thread split, which is uninterpretable because its held-out rows shrink to 20%; we exclude it. Additional misspecification and identifiability controls also return a near-zero spurious interaction (Appendix B).

We count a condition as detected only if (i) the fitted bootstrap or the permutation rejects at the 0.05 level and (ii) the annotator cluster interval excludes 0. The second requirement removes conditions with large annotator-level variability; EPIC meets the first but not the second and is not counted as a detection.

Table 1: Decomposition of the gain from conditioning on the annotator, in nats, on the annotator scale.

Corpus	Features	Split	Δ_{marg}	Δ_{cond}	95% CI	Share	Null 95th pct.	Verdict
GoEmotions	TF-IDF	item	$+0.1806$	$+0.0142$	$[+0.0087, +0.0202]$	7.3% $[4.7, 10.3]$	$+0.00000$	✓
GoEmotions	TF-IDF	thread	$+0.1817$	$+0.0129$	$[+0.0080, +0.0187]$	6.6% $[4.1, 9.6]$	-0.00000	✓
GoEmotions	SBERT	item	$+0.1830$	$+0.0142$	$[+0.0090, +0.0201]$	7.2% $[4.5, 10.6]$	$+0.00051$	✓
GoEmotions	SBERT	thread	$+0.1832$	$+0.0129$	$[+0.0072, +0.0196]$	6.6% $[3.6, 10.4]$	$+0.00027$	✓
CSC	TF-IDF	item	$+0.1166$	$+0.0230$	$[+0.0180, +0.0279]$	16.4% $[13.5, 19.1]$	$+0.00005$	✓
MP	TF-IDF	item	$+0.0279$	$+0.0105$	$[+0.0085, +0.0126]$	27.4% $[24.7, 29.7]$	$+0.00015$	✓
EPIC	TF-IDF	item	$+0.0452$	-0.0005	$[-0.0012, +0.0001]$	-1.2% $[-2.7, 0.3]$	$+0.00013$	—

Table 2: Content-dependent share by encoder architecture and adaptation (GoEmotions, item split).

Encoder	Type	Δ_{marg}	Δ_{cond}	Share
Linear (TF-IDF)	Bag-of-words	$+0.1806$	$+0.0142$	7.3% $[4.7, 10.3]$
Nonlinear (full head)	Pooling	$+0.1692$	$+0.0370$	17.9% $[14.1, 22.4]$
BiLSTM	Contextual, no attention	$+0.1967$	$+0.0195$	9.0% $[5.1, 13.8]$
Transformer (distilroberta, frozen)	Self-attention	$+0.1009$	$+0.0443$	30.5% $[25.5, 36.8]$
Transformer (random init, end-to-end)	Self-attention	$+0.0161$	$+0.0748$	82.3% $[79.0, 84.3]$
Transformer (distilroberta, end-to-end)	Self-attention	$+0.0118$	$+0.0981$	89.3% $[87.9, 90.2]$

4.3 The Basic Decomposition of the Gain

Table 1 shows that under limited representations the additive component dominates. On GoEmotions the content-dependent share is 6.6–7.3% across TF-IDF/SBERT and item/thread splits, with every interval excluding 0 and exceeding the null by tens of times (about $96\times$ for TF-IDF on the item split). CSC and MP show larger shares of 16.4% and 27.4%, detected by the permutation test. EPIC is not detected: the fitted null rejects ($p = 0.024$), but the assumption-free permutation does not ($p = 0.68$) and its annotator interval $[-0.0012, +0.0001]$ contains 0, with a point estimate of -0.0005. Detection does not track the number of annotations per annotator: CSC, with a median of 29, gives the largest absolute $+0.0230$, and EPIC, with 199, the smallest.

Across all detected conditions the additive point estimate accounts for 72.6 to 93.4% of the gain (content-dependent share 6.6–27.4%; the envelope including per-condition intervals is 3.6–30.0%). These nine conditions are the union of the four GoEmotions linear-family conditions (TF-IDF and SBERT $\times$ item and thread), the two label-coarsening conditions ($K = 7$ and $K = 2$), the one nonlinear-family condition, and CSC and MP. The share is a composition, not a magnitude: MP's larger share reflects a small denominator ($\Delta_{\text{marg}} = +0.0279$), and its absolute $\Delta_{\text{cond}} = +0.0105$ is below GoEmotions' $+0.0142$. Read on its own, this baseline would license the common shortcut of implementing personalization as a per-annotator output-layer bias and discarding the learned annotator representation, since a single intercept already recovers the large majority of the gain. This range, however, is for families whose shared component has limited expressiveness; strengthening it, as we show next, reverses the composition and overturns that shortcut.

4.4 Representation Strength Raises the Share, End-to-End Adaptation Reverses It

We strengthen the shared text model step by step; Table 2 and Figure 2 summarize the result. Holding the representation fixed, the content-dependent share rises with representational strength: making the encoder nonlinear lifts it from 7.3% to 17.9% (a capacity-matched nonlinear family gives 17.1% $[13.7, 19.4]$, so both the model class and its capacity move the share; Appendix C), and a mean-pooled distilroberta that is fine-tuned once on the text-only objective and then frozen and shared across the nested models raises it further to 30.5% $[25.5, 36.8]$. A stronger fixed representation thus makes more of the annotators' judgment recoverable as a content-dependent interaction,

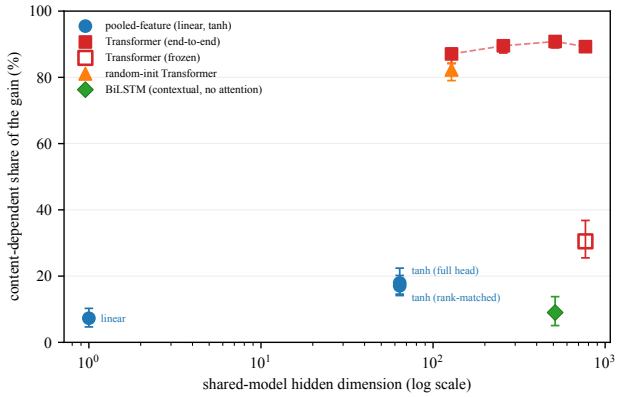

Figure 2: Content-dependent share against shared-model hidden dimension, by representation family (GoEmotions, item split).

but on none of these fixed representations does the interaction become dominant: even the frozen Transformer leaves the additive component at 69.5%.

The full reversal appears only when the Transformer is adapted end to end within each nested model. Fine-tuning distilroberta separately for M_0, M_1, and M_2 lowers the text-only NLL from 2.30 to 1.89, collapses Δ_{marg} from $+0.1806$ to $+0.0118$, and raises Δ_{cond} from $+0.0142$ to $+0.0981$ $[+0.0692, +0.1341]$, so the content-dependent share reaches 89.3% $[87.9, 90.2]$ and the two roles swap; the reversal replicates on MP ($27.4\% \rightarrow 83.7\%$ $[74.5, 88.7]$). The gap between the frozen 30.5% and the end-to-end 89.3%, on the same architecture and the same data, indicates that jointly adapting the encoder with the annotator-specific component is a major contributor to the reversal, over and above the representation class. This end-to-end share is not an estimation artifact: a Transformer with no interaction returns Δ_{cond} near zero under the same calibration (mean -0.0003, max $+0.0057$ over 20 replicates), just as in the linear family.

Within the end-to-end regime, the reversal is explained neither by hidden dimension nor by pretraining: the share is already 87.1% at hidden dimension 128 and 89.3% at 768, and a randomly initialized 128-dimensional Transformer still reaches 82.3%. Contextuality alone does not suffice either: an order-aware BiLSTM, a better text model than the linear one, yields only 9.0%, no different from word aggregation. Among the configurations tested, the reversed regime appears only for self-attention Transformers adapted end to end, rather than being explained by depth, pretraining, or contextuality alone. The frozen control moreover uses an encoder tuned on the M_0 objective, so it does not fully separate representation geometry from the objective used to obtain that representation; Appendix C discusses this limitation and gives the full size, pretraining, BiLSTM, and frozen-encoder results.

4.5 Robustness of the Finding

The finding survives a battery of controls, summarized here and reported in full in the appendix.

Capacity and the direction of parameterization. The content-dependent gain is bought not with parameters but with the direction of parameterization: a per-annotator intercept obtains $+0.1806$ with 28 parameters and an annotator$\times$board intercept only $+0.1807$ with 13,524, whereas a 32-parameter low-rank interaction reaches $+0.1996$, and strengthening the baseline with a per-annotator confusion matrix or a board intercept leaves the interaction gain intact ($+0.0142 \rightarrow +0.0141$–$+0.0169$, intervals excluding 0; Appendix D).

Decomposition order. An order-independent Shapley allocation raises the content-dependent share but leaves the additive component dominant (83.5–85.4%; Appendix E).

Label coarsening. Coarsening the GoEmotions labels preserves Δ_{cond} while substantially reducing Δ_{marg}, so the two components respond differently to label design (Appendix F).

Identifiability. A synthetic factorial confirms the expected detectability pattern: at $s = 1$, power rises from 55% at 50 annotations per annotator to 100% at 500, rises with the number of annotators (92% at 20, 100% at 200), and falls with rank (100% at $r = 1$, 57% at $r = 8$), so identifying the content-dependent component needs both an expressive shared component and enough data per annotator (Appendix G).

5 Discussion

5.1 The Decomposition Is Measurement-Relative

Our results do not contradict the finding that annotator-conditioned models are effective; what we add is that the composition of the gain is not fixed but is shaped jointly by the shared representation and by how it is adapted. On fixed representations the gain looks additive, and the more so the weaker the representation: the content-dependent share is 7.3% for the linear predictor and rises to 30.5% for a frozen Transformer, yet the additive component stays dominant throughout, robust to decomposition order and to a strengthened baseline. The full reversal, in which the content-dependent component reaches about 89% and the two roles swap, appears only when the same Transformer is adapted end to end within each nested model. Strikingly, with the same architecture and the same data, freezing versus jointly adapting the encoder moves the content-dependent share from 30.5% to 89.3%, so the composition is read off not the representation class alone but the pair of representation and adaptation protocol. This pattern is consistent with a weak or frozen shared component leaving more annotator-specific structure in the text-only residual, whereas end-to-end adaptation allows more of that structure to be recovered through the interaction. The practical implication is itself protocol-dependent: with a frozen or weak encoder a per-annotator intercept recovers most of the gain, whereas under end-to-end personalization it recovers only about ten percent and most of the gain requires the interaction. Thus "does a per-annotator intercept suffice, or is content-dependent personalization needed?" has no answer independent of the text model and how it is trained. Label design shifts the same answer: coarsening the GoEmotions labels to binary collapses the additive component to less than a third while the content-dependent component is statistically unchanged (Appendix F). We report this component as a predictive association under the observed assignment, not as causal evidence that annotators read the same content differently.

This measurement-relativity reframes a recurring debate in the perspectivist literature. Reports that annotator conditioning does or does not help, or that a task is or is not "subjective enough" to warrant it, implicitly compare gains measured on whatever encoder was at hand; our decomposition shows that such comparisons conflate a property of the annotators with a property of the encoder and how it is trained, and that the same corpus can look additive under one representation-and-training choice and content-dependent under another. For a practitioner the actionable reading is not a single verdict but a procedure: measure the composition on the encoder and training pipeline that will actually be deployed, since the choice between a lightweight per-annotator intercept and a full annotator–text interaction is decided there rather than by the corpus in the abstract. It also cautions against transferring a composition measured with word-aggregation features to a Transformer-based system, in which the same annotators would be modeled very differently, and against reading a small content-dependent share on a weak encoder as evidence that annotators agree once their label-use frequencies are removed. The same logic yields a positive prescription for data collection: because the content-dependent component becomes measurable only once the shared representation can express it and enough annotations per annotator are available, the calibration curve of Section 4.2 can be read backwards to set, before collection begins and for a chosen encoder, a target number of annotations per annotator, rather than discovering after the fact that an effect was present but underpowered.

5.2 When the Measurement Is Valid

Measuring the content-dependent component validly requires two resources at once (Appendix G): an expressive shared component, without which the structure is buried in the additive residual, and

enough data per annotator, without which the interaction is indistinguishable from the null. These are necessary conditions, not substitutes. Consequently, a non-detection under a given encoder does not mean the annotators lack content-dependent judgment: the representation may not express it, or the annotations may be too few, so whether an annotator effect exists is relative to the triple of encoder, data amount, and estimator, not a fixed property of the annotators.

The verdict further depends critically on how the calibration null is built. With everything else fixed, changing only the null reverses the verdict for MP: a corpus-independent Gaussian null is uncalibrated: its 95th percentile is $363\times$ too large for MP, burying a real effect, yet negative for CSC, whereas the fitted null and the assumption-free permutation agree. A null must give what the estimator returns *on this data* with no interaction; that is, it must preserve the empirical prediction problem. This requirement extends to any personalization gain measured by a nested comparison of a shared and a person-specific component, such as user modeling, recommendation, preference, and rater modeling, where gains are often reported on data with tens to hundreds of observations per individual: what a report needs is not the threshold value but a statement of which generative process produced it. The same calibration also guards split design: the item+thread split gives a large but uninterpretable value that the negative control matches, so the magnitude of a value alone is not the strength of the evidence.

5.3 Limitations

What we treat is emotion and sarcasm annotation of short English texts, with annotators limited to humans on research crowdsourcing platforms, so we have not verified whether the same decomposition holds for other languages, other task domains, perceptual judgments of audio or video, or large language models as annotators. What we measure is personalization for annotators seen at training time: β_a and u_a are per-annotator parameters and are undefined for unseen annotators, so Δ_{cond} is not a measure of adaptability to new annotators, and generalization to unseen annotators requires a different design. Moreover, Δ_{cond} is a residual predictive gain under the observed assignment; although we controlled the assignment bias at the board granularity for GoEmotions, finer topic or style bias is not controlled, so our claim is limited to a predictive association and does not offer a causal interpretation in terms of differences in how annotators read.

6 Conclusion

We decomposed the gain from conditioning on the annotator into an additive marginal component and a content-dependent component, as differences in held-out log-likelihood among nested predictors, keeping the comparison symmetric in capacity and optimization. The content-dependent share is not a fixed property of the annotators but is shaped jointly by the shared representation and by how it is adapted. On fixed representations it rises with representational strength, from 7.3% for a linear encoder to 30.5% for a frozen Transformer, yet stays additive-dominant; only when the same Transformer is adapted end to end within each nested model does the decomposition reverse, the share reaching about 89%, a reversal explained neither by hidden dimension nor by pretraining, observed on GoEmotions and replicated on MP, and not an estimation artifact, since the estimator returns zero under a no-interaction null. That freezing versus jointly adapting the same architecture on the same data moves the share from 30.5% to 89.3% shows the composition to be a property of the measurement, encompassing the representation and its adaptation, rather than of the annotators. Identifying the component further requires enough data per annotator, and the detection verdict depends strongly on how the calibration null is constructed. Whether a per-annotator intercept suffices or content-dependent personalization is needed therefore has no model-independent answer. Future work includes a common frozen encoder trained under different objectives to further separate representation geometry from joint adaptation, larger and instruction-tuned Transformers, and using the calibration curve to size annotation collection.

Ethics Statement

This work uses only publicly released datasets, namely GoEmotions, the LeWiDi CSC and MP datasets, and EPIC, under their respective licenses, and collects no new data from human subjects. These datasets provide per-annotator judgments with anonymized annotator identifiers, and we use

these identifiers solely to model per-annotator prediction; we make no attempt to re-identify annotators or to infer their demographic or psychological attributes. Our analysis reports predictive associations under the observed annotation assignment and does not make causal or psychological claims about individual annotators. Because methods that condition on annotator identity could in principle be misused to profile annotators, we emphasize that the quantities we report are held-out predictive gains, not measurements of any annotator's traits.

Reproducibility Statement

The estimator and its calibration are defined in full in Section 3: the nested predictors in Section 3.1, the symmetry controls between the nested models in Section 3.2, and the null construction and detectability procedure in Section 3.3. Complete reproducibility details, namely the dataset sizes and filtering, the train-fold-only preprocessing, the splits, the optimization budget and hyperparameter grids, and the cluster-bootstrap inference protocol, are given in Appendices A and M. All reported quantities are derived from a single cache, so the same condition yields the same value across tables (Appendix M). Anonymized source code that reproduces every table and figure is provided as supplementary material.

Use of Large Language Models

We used a large language model as a general-purpose assistant for editing and improving the writing of this paper and for routine coding support in the experimental pipeline. All research contributions, namely the problem formulation, the estimator, the experimental design, the analysis, and the scientific claims, are the authors' own, and the authors verified every reported number against the experimental outputs. The model was not used to generate research ideas or to produce or interpret experimental results.

References

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernández. Stop measuring calibration when humans disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1892–1915, 2022. doi: 10.18653/v1/2022.emnlp-main.124.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pp. 15–21, 2021. doi: 10.18653/v1/2021.bppf-1.3.

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 16008–16021, 2024. doi: 10.18653/v1/2024.acl-long.849.

Lee J. Cronbach, Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Wiley, New York, 1972. ISBN 0-471-18850-6.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022. doi: 10.1162/tacl_a_00449.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28, 1979. doi: 10.2307/2346806.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4040–4054, 2020. doi: 10.18653/v1/2020.acl-main.372.

Eve Fleisig, Rediet Abebe, and Dan Klein. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6715–6726, 2023. doi: 10.18653/v1/2023.emnlp-main.415.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. Beyond black and white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 2591–2597, 2021. doi: 10.18653/v1/2021.naacl-main.204.

Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. EPIC: Multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 13844–13857, 2023. doi: 10.18653/v1/2023.acl-long.774.

Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28 (7):1734–1748, 2016. doi: 10.1109/TKDE.2016.2545658.

Mor Geva, Yoav Goldberg, and Jonathan Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 1161–1166, 2019. doi: 10.18653/v1/D19-1107.

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, number 115, pp. 1–19, 2022. doi: 10.1145/3491102.3502004.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pp. 1321–1330, 2017. doi: 10.5555/3305381.3305518.

Keito Inoshita. Bias-corrected ceilings of emotion predictability from human label variation based on instance-level Fano bounds. *arXiv*, 2026a. doi: 10.48550/arXiv.2608.15619.

Keito Inoshita. Bridging the silos in affective AI: A critical perspective from data to society. *AI & Society*, 2026b. doi: 10.1007/s00146-026-03324-y.

Keito Inoshita, Xiaokang Zhou, Akira Kawai, and Katsutoshi Yada. LLMs capture emotion labels, not emotion uncertainty: Distributional analysis and calibration of human-LLM judgment gaps. *arXiv*, 2026. doi: 10.48550/arXiv.2604.27345.

Hyewon Jang, Moritz Jakob, and Diego Frassinelli. Context vs. human disagreement in sarcasm detection. In *Proceedings of the 4th Workshop on Figurative Language Processing*, pp. 1–7, 2024. doi: 10.18653/v1/2024.figlang-1.1.

David R. Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, pp. 1953–1961, 2011. doi: 10.5555/2986459.2986677.

Urja Khurana, Eric Nalisnick, Antske Fokkens, and Swabha Swayamdipta. Crowd-calibrator: Can annotator disagreement inform calibration in subjective tasks? In *Proceedings of the First Conference on Language Modeling*, 2024. doi: 10.48550/arXiv.2408.14141.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009. doi: 10.1109/MC.2009.263.

Kemal Kurniawan, Meladel Mistica, Timothy Baldwin, and Jey Han Lau. Training and evaluating with human label variation: An empirical study. *arXiv preprint arXiv:2502.01891*, 2025. doi: 10.48550/arXiv.2502.01891.

Elisa Leonardelli, Silvia Casola, Siyao Peng, Giulia Rizzi, Valerio Basile, Elisabetta Fersini, Diego Frassinelli, Hyewon Jang, Maja Pavlovic, Barbara Plank, and Massimo Poesio. LeWiDi-2025 at NLPerspectives: Third edition of the learning with disagreements shared task. In *Proceedings of the 4th Workshop on Perspectivist Approaches to NLP*, pp. 182–195, 2025. doi: 10.18653/v1/2025.nlperspectives-1.16.

Soda Marem Lo, Silvia Casola, Erhan Sezerer, Valerio Basile, Franco Sansonetti, Antonio Uva, and Davide Bernardi. PERSEVAL: A framework for perspectivist classification evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 22334–22359, 2025. doi: 10.18653/v1/2025.emnlp-main.1137.

Barbara Plank. 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, 2022.

Filipe Rodrigues and Francisco C. Pereira. Deep learning from crowds. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, volume 32, pp. 1611–1618, 2018. doi: 10.1609/aaai.v32i1.11506.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470, 2021. doi: 10.1613/jair.1.12752.

Peter Welinder, Steve Branson, Serge Belongie, and Pietro Perona. The multidimensional wisdom of crowds. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, volume 2, pp. 2424–2432, 2010. doi: 10.5555/2997046.2997166.

Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier R. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, pp. 2035–2043, 2009. doi: 10.5555/2984093.2984321.

Dustin Wright and Isabelle Augenstein. Aggregating soft labels from crowd annotations improves uncertainty estimation under distribution shift. *PLOS ONE*, 20(6):e0323064, 2025. doi: 10.1371/journal.pone.0323064.

Yan Yan, Rómer Rosales, Glenn Fung, Ramanathan Subramanian, and Jennifer Dy. Learning from multiple annotators with varying expertise. *Machine Learning*, 95:291–327, 2014. doi: 10.1007/s10994-013-5412-1.

A Dataset Details

We use four corpora that release per-annotator labels. All are emotion or sarcasm annotation of short English texts and differ by orders of magnitude in the number of annotators, responses per annotator, and labels.

GoEmotions. Emotion annotation of Reddit comments (Demszky et al., 2020), with 28 labels (27 emotions plus neutral). We use only single-label annotations and restrict to the 57 annotators who gave at least 500, leaving 166,904 annotations with a median of 2,649 per annotator, the most in our study. Comments carry a post-thread identifier, so the three splits, item, thread, and both at once, can be applied. In the label-coarsening control we map the 28 original labels to 7 Ekman groups and then to a neutral/non-neutral binary, holding the corpus, annotators, items, split, and features fixed.

LeWiDi CSC and MP. Two datasets from the disagreement-preserving shared task LeWiDi (Leonardelli et al., 2025). CSC is a 6-point sarcasm judgment with 872 annotators and a median of 29 annotations per annotator; MP is a binary sarcasm judgment with 504 annotators and a median of 122. Neither carries a post-thread identifier, so only the item-level split is used.

EPIC. Binary annotation of irony (Frenda et al., 2023), with 74 annotators and a median of 199 annotations per annotator, the fewest annotators in our study. We use the released per-annotator labels as distributed.

Text features. We use two independent families. The first is a TF-IDF over word 1–2-grams (minimum document frequency 2, maximum vocabulary 40,000) reduced by a singular value decomposition to 256 dimensions (128 for the small corpora); the second is the output of a frozen 384-dimensional sentence embedding model. For both, the vocabulary, IDF, singular-value basis, and standardization statistics are estimated from the training portion of each split only and applied to the held-out portion. The Transformer input is tokenized only, with nothing fit.

B Calibration and Null Construction

Table 3: Null distribution of Δ_{cond} and its recovery under injected strength s.

		Null ($s = 0$)				Recovery at injected strength s			
Corpus / Split	Reps	Mean	SD	95th pct. [95% CI]	Calib. p	0.25	0.5	1.0	2.0
GoEmotions / item	20	−0.00002	0.00011	+0.00000 [−0.00003, +0.00046]	0.048	+0.0078	+0.0509	+0.2048	+0.6084
GoEmotions / thread	20	−0.00006	0.00014	−0.00000 [−0.00003, +0.00036]	0.048	+0.0077	+0.0485	+0.2038	+0.6083
GoEmotions K=7 / item	20	−0.00002	0.00009	+0.00010 [−0.00001, +0.00018]	0.048	+0.0123	+0.0483	+0.1659	+0.4375
GoEmotions K=2 / item	20	−0.00001	0.00014	+0.00007 [+0.00001, +0.00047]	0.048	+0.0074	+0.0239	+0.0850	+0.1897
CSC / item	40	−0.00021	0.00019	+0.00005 [−0.00005, +0.00025]	0.024	−0.0002	+0.0141	+0.0714	+0.2897
MP / item	40	−0.00007	0.00014	+0.00015 [+0.00003, +0.00027]	0.024	+0.0007	+0.0074	+0.0573	+0.1256
EPIC / item	40	−0.00038	0.00046	+0.00013 [+0.00003, +0.00040]	0.024	−0.0003	+0.0110	+0.0437	+0.1702

Table 4: The verdict under two null constructions and the within-item permutation.

		Null 95th pct.				Within-item permutation	
Corpus	Δ_{cond}	Fixed Gaussian	Fitted	Ratio	Fitted p	Max	p
GoEmotions / TF-IDF	+0.0142	—	1.82×10^{-4}	—	0.00415 (R=240)	9.11×10^{-4}	0.048 (R=20)
GoEmotions / SBERT	+0.0142	—	5.14×10^{-4}	—	0.04762 (R=20)	1.21×10^{-3}	0.048 (R=20)
CSC	+0.0230	-3.40×10^{-3}	4.52×10^{-5}	—	0.02439 (R=40)	1.72×10^{-3}	0.024 (R=40)
MP	+0.0105	5.33×10^{-2}	1.47×10^{-4}	363	0.02439 (R=40)	2.07×10^{-3}	0.024 (R=40)
EPIC	−0.0005	6.78×10^{-3}	1.28×10^{-4}	53	0.02439 (R=40)	2.43×10^{-4}	0.683 (R=40)

Table 3 reports the calibration. For synthetic labels with no injected interaction Δ_{cond} has near-zero mean under every condition, and recovery increases monotonically with s (e.g., GoEmotions item split $+0.0078$ at $s = 0.25$ and $+0.2048$ at $s = 1.0$); at the same s recovery is smaller for narrower label spaces. On the negative controls, reported in Table 12 of Appendix H, Δ_{cond} is -0.0023 to -0.0001 under permutation and ± 0.0002 under independent redrawing on the item and thread splits; only the item+thread split gives $+0.0107$ to $+0.0287$, because its held-out rows shrink to 20%, and we exclude it.

Table 4 contrasts two null constructions. The fitted null generated from M_1 is calibrated, whereas a corpus-independent Gaussian null is not: its 95th percentile is $363\times$ larger than the fitted null for MP and $53\times$ for EPIC, but sits on the negative side for CSC. When the threshold is overestimated a real effect is buried (MP and EPIC were missed under this null); when it shifts negative a non-effect can be detected. The within-item permutation reaches $p = 0.048$ ($R = 20$) on the GoEmotions reference condition and $p = 0.024$ ($R = 40$) on CSC and MP, while EPIC (observed -0.0005) is not rejected ($p = 0.68$); extending the reference fitted null to 240 replicates gives $p = 0.00415$, below the Bonferroni level 0.00455 for 11 conditions. On the detected conditions the observed value is 5.1 to 19.2 times the permutation maximum. The permutation is exact only when $\beta_a = 0$, so we treat the fitted parametric bootstrap as primary and the permutation as an assumption-free secondary control and report both.

Two further controls confirm that M_2 does not pick up text-only structure. Generating labels from a strictly richer nonlinear model (a shared $\tanh$ encoder with annotator intercepts and no interaction) and fitting the linear M_1 and M_2 yields mean -0.00001 and maximum $+0.00034$ over 20 replicates, against the observed $+0.0142$ ($51\times$). Algebraically, a constant u_a is absorbed into W, so

a gain requires u_a to vary across annotators; imposing $\sum_a u_a = 0$ leaves the gain essentially unchanged ($+0.0142 \to +0.0153$ for TF-IDF and $+0.0142 \to +0.0134$ for SBERT, with differences containing 0; Appendix L).

C Architecture and Capacity

Table 5: Model class and interaction rank (GoEmotions, item split, TF-IDF).

Model family	text-only	+intercept	+interaction	Δ_{marg} / Δ_{cond}	95% CI of Δ_{cond}	Share
Linear	2.3092	2.1286	2.1144	+0.1806 / +0.0142	[+0.0087, +0.0202]	7.3% [4.7, 10.3]
Nonlinear (low-rank)	—	—	—	+0.1848 / +0.0381	[+0.0315, +0.0453]	17.1% [14.6, 20.2]
Nonlinear	2.2933	2.1242	2.0872	+0.1692 / +0.0370	[+0.0290, +0.0453]	17.9% [14.1, 22.4]
Δ_{cond} *at fixed rank (linear family)*						
	$r = 1$	$r = 2$	$r = 4$	$r = 8$		
	+0.0074	+0.0122	+0.0130	+0.0064		

Table 6: Decomposition as the shared text model varies (GoEmotions, item split).

Shared text model	Hidden dim.	text-only NLL	Δ_{marg}	Δ_{cond}	95% CI of Δ_{cond}	Share
Linear (TF-IDF)	1	2.3092	+0.1806	+0.0142	[+0.0087, +0.0202]	7.3%
Nonlinear (capacity-matched, r=4)	64	—	+0.1848	+0.0381	[+0.0315, +0.0453]	17.1%
Nonlinear (full head)	64	2.2933	+0.1692	+0.0370	[+0.0290, +0.0453]	17.9%
Transformer (distilroberta)	768	1.8868	+0.0118	+0.0981	[+0.0692, +0.1341]	89.3%

Table 7: Content-dependent share with a linear predictor versus a pretrained Transformer.

Dataset	Linear	Transformer
GoEmotions	7.3%	89.3%
MP	27.4%	83.7%

Table 8: Decomposition across Transformer hidden dimension and pretraining (GoEmotions, item split).

Encoder	Hidden dim.	Δ_{marg}	Δ_{cond}	95% CI of Δ_{cond}	Share
BERT-tiny (random init)	128	+0.0161	+0.0748	[+0.0494, +0.1050]	82.3% [79.0, 84.3]
BERT-tiny	128	+0.0081	+0.0544	[+0.0349, +0.0778]	87.1% [84.2, 88.8]
BERT-mini	256	+0.0080	+0.0683	[+0.0445, +0.0971]	89.5% [87.4, 90.8]
BERT-small	512	+0.0080	+0.0787	[+0.0522, +0.1116]	90.8% [88.9, 92.0]
distilroberta	768	+0.0118	+0.0981	[+0.0692, +0.1341]	89.3% [87.9, 90.2]

Table 5 varies the model class on GoEmotions (item split, TF-IDF). Adding per-annotator heads to a nonlinear encoder gives $\Delta_{\text{cond}} = +0.0370$ (share 17.9%); a capacity-matched nonlinear family, with degrees of freedom matched to the linear family, gives $+0.0381$ (share 17.1% $[13.7, 19.4]$). Holding capacity fixed while making the shared component nonlinear raises the share from 7.3% to 17.1%, and raising capacity while keeping it nonlinear raises it to 17.9%, so both move the share. The nonlinear family passes the same calibration (mean -0.0242 over 20 replicates, opposite in sign to the observed $+0.0370$).

Table 6 replaces the shared component with a fine-tuned Transformer: with distilroberta the text-only NLL falls from 2.30 to 1.89, Δ_{marg} collapses to $+0.0118$, Δ_{cond} rises to $+0.0981$ $[+0.0692, +0.1341]$, and the share reaches 89.3% $[87.9, 90.2]$. The reversal replicates on MP, from 27.4% to 83.7% $[74.5, 88.7]$, as Table 7 reports. Table 8 varies hidden dimension

and pretraining: within the end-to-end regime the share is 87.1% at 128 and 89.3% at 768, and 82.3% [79.0, 84.3] for a randomly initialized 128-dimensional Transformer, so neither dimension nor pretraining sets it. As Table 2 shows, the share tracks the representation family and its adaptation: word-aggregation and recurrent encoders give 7.3–17.9%, a frozen Transformer 30.5%, and end-to-end Transformers 82–91%.

Frozen-encoder control. To separate the representation from its joint adaptation, we run a control in which the encoder is fixed across the nested models. For each fold we fine-tune distilroberta once on the text-only objective (M_0), freeze the encoder, extract mean-pooled features $h_\theta(x)$, standardize them with training-fold statistics, and fit the same linear M_0/M_1/M_2 heads on these fixed features; everything else (splits, held-out rows, annotator-scale estimand, cluster bootstrap) matches the main pipeline. On this shared frozen representation $\Delta_{\text{marg}} = +0.1009$ and $\Delta_{\text{cond}} = +0.0443\,[+0.0366, +0.0528]$, so the content-dependent share is 30.5% [25.5, 36.8]: higher than any word-aggregation encoder, but still additive-dominant, in contrast to the 89.3% obtained when the same distilroberta is adapted end to end. Because the frozen encoder here is the one fine-tuned by M_0, this control isolates joint adaptation only up to the choice of that fine-tuning objective; a common encoder trained under several objectives would separate representation geometry from adaptation more completely.

D Per-Annotator Capacity and Strengthened Baselines

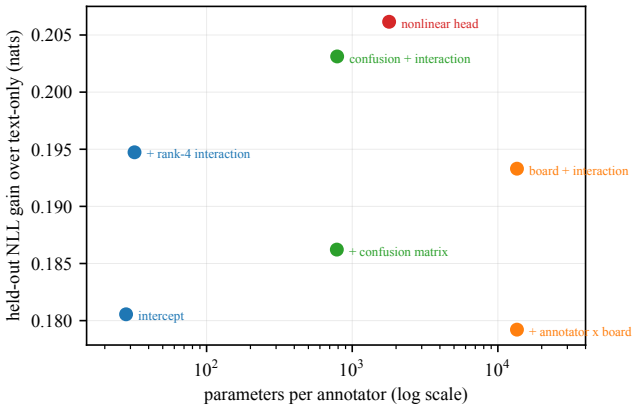

Figure 3: Held-out gain over the text-only model against the number of per-annotator parameters (GoEmotions, item split; log x-axis).

Table 9: Interaction gain against strengthened baselines (GoEmotions, item split, TF-IDF).

Baseline model	Baseline NLL	+interaction	Δ_{cond}	95% CI	Retention
Intercept β_a	2.1286	2.1144	+0.0142	[+0.0087, +0.0202]	100.0%
Intercept + confusion matrix C_a	2.1229	2.1060	+0.0169	[+0.0118, +0.0226]	119.2%
Intercept + annotator × board	2.1299	2.1159	+0.0141	[+0.0092, +0.0201]	99.3%

Figure 3 relates the number of per-annotator parameters to the held-out gain over the text-only model. The intercept obtains +0.1806 with 28 parameters; a confusion matrix obtains +0.1835 with 784, and an annotator×board intercept obtains +0.1807 with 13,524, so a 480× increase does not beat the intercept, whereas adding a 32-parameter interaction reaches +0.1996. The direction, not the amount, of the parameters matters.

Table 9 strengthens the baseline and adds the interaction on both sides. Against a confusion-matrix baseline the interaction still adds +0.0169 [+0.0118, +0.0226], and against an annotator×board

baseline $+0.0141$ $[+0.0092, +0.0201]$, neither shrinking from the $+0.0142$ against the bare intercept and neither interval containing 0. The topic-coverage bias is small: the median entropy of each annotator over the 483 boards is 5.99 against 6.18 for a uniform distribution, and the board control removes only $+0.0142 \rightarrow +0.0141$. The baseline fits themselves are in Appendix J: the confusion matrix raises the degrees of freedom from 28 to 784 but improves the held-out NLL by only $+0.0013$.

E Decomposition Order

Table 10: Sequential versus Shapley allocation of the two components.

	Intercept first		Interaction first		Shapley allocation		Content-dependent share	
Condition	$M_0 \to M_1$	$M_1 \to M_2$	$M_0 \to M_I$	$M_I \to M_2$	ϕ_{add}	ϕ_{int}	Sequential	Shapley [95% CI]
TF-IDF / item	$+0.1806$	$+0.0142$	$+0.0471$	$+0.1476$	$+0.1641$	$+0.0306$	7.3%	15.7% [11.7, 19.7]
TF-IDF / thread	$+0.1817$	$+0.0129$	$+0.0440$	$+0.1505$	$+0.1661$	$+0.0285$	6.6%	14.6% [10.8, 18.9]
SBERT / item	$+0.1830$	$+0.0142$	$+0.0509$	$+0.1462$	$+0.1646$	$+0.0325$	7.2%	16.5% [12.6, 20.7]

Table 10 measures the reverse order. Adding the interaction first obtains $+0.0491$, about three times the $+0.0142$ obtained after the intercept, so the content-dependent share rises from the sequential 7.3% to 15.7% $[11.7, 19.7]$ under a Shapley allocation (which, with two components, is the mean of the two orders and satisfies $\phi_{\text{add}} + \phi_{\text{int}} = \text{NLL}(M_0) - \text{NLL}(M_2)$ to within 10^{-10}). Even so, the additive component still accounts for 83.5 to 85.4% across the three conditions. We take the sequential Δ_{cond} as primary because it answers the practical question of what a learned interaction adds for an implementation that already has intercepts; ϕ_{int} gives the order-independent allocation.

F Label Coarsening

Table 11: Coarsening the label space within GoEmotions ($28 \to 7 \to 2$).

# labels	Δ_{marg}	Δ_{cond}	95% CI	Share	Null 95th pct.	Selected rank
28 (original labels)	$+0.1806$	$+0.0142$	$[+0.0087, +0.0202]$	7.3% [4.7, 10.3]	3.98×10^{-7}	r=2:1, r=4:4
7 (Ekman groups)	$+0.0869$	$+0.0138$	$[+0.0086, +0.0204]$	13.7% [8.7, 20.5]	1.02×10^{-4}	r=4:3, r=8:2
2 (neutral / non-neutral)	$+0.0563$	$+0.0116$	$[+0.0058, +0.0189]$	17.0% [8.4, 30.0]	7.15×10^{-5}	r=4:3, r=8:2
Paired reductions						
Coarsening	Reduction in Δ_{cond}	95% CI	Reduction in Δ_{marg}	95% CI	Ratio [95% CI]	
$28 \to 7$	$+0.0004$	$[-0.0028, +0.0034]$	$+0.0936$	$[+0.0824, +0.1052]$	238.0 [25.8, 1824.9]	
$28 \to 2$	$+0.0026$	$[-0.0018, +0.0068]$	$+0.1243$	$[+0.1099, +0.1399]$	47.3 [18.1, 804.2]	
$7 \to 2$	$+0.0022$	$[-0.0002, +0.0046]$	$+0.0307$	$[+0.0246, +0.0372]$	13.7 [6.1, 147.7]	

Table 11 coarsens only the label space within GoEmotions ($28 \to 7 \to 2$), holding the corpus, annotators, items, split, and features fixed. Δ_{marg} falls from $+0.1806$ to $+0.0563$ (less than a third), while Δ_{cond} is nearly flat ($+0.0142$, $+0.0138$, $+0.0116$, each interval excluding 0). In the paired comparison the reduction in Δ_{cond} from 28 to 2 is only $+0.0026$ with interval $[-0.0018, +0.0068]$ containing 0, so the content-dependent component does not decrease significantly, whereas the reduction in Δ_{marg} is $+0.1243$ (ratio 47.3 $[18.1, 804.2]$). The share rises from 7.3% to 17.0% because the additive component decreases far faster, not because the content-dependent component grows; the content-dependent component is detected even at $K = 2$ (interval $[+0.0058, +0.0189]$), so the non-detection on MP and EPIC is not explained by the number of labels. The selected interaction rank rises under coarsening, confirming that Δ_{cond} did not shrink for lack of capacity.

G Synthetic Identifiability

On synthetic data we vary each factor independently. We generate 64-dimensional features from an 8-dimensional latent and redraw labels after injecting a rank-r interaction of known strength s; the base cell is 50 annotators, 200 annotations per annotator, $K = 10$, and $r = 2$. Power, plotted in

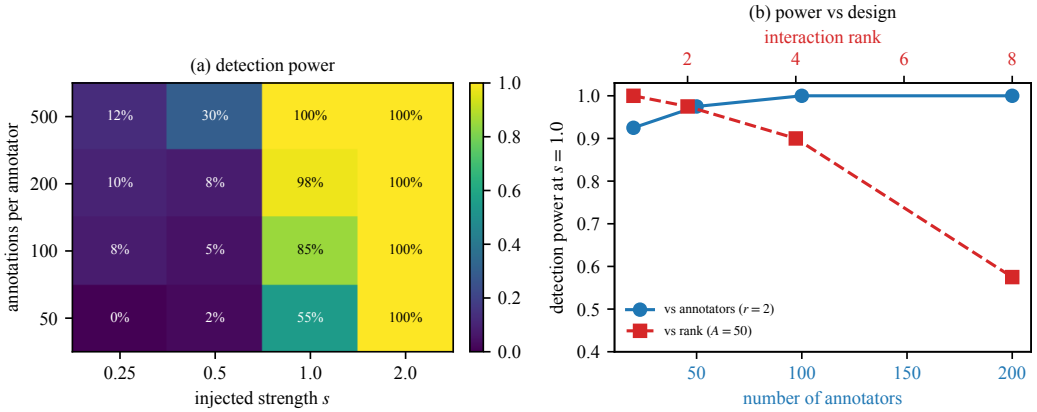

Figure 4: Detection power for the content-dependent component on synthetic data.

Figure 4, is the fraction of 40 replicates whose injected Δ_{cond} exceeds the 95th percentile of matched null replicates. Power rises with strength and with annotations per annotator (55% at 50, 100% at 500 for $s = 1.0$; weak interactions $s \leq 0.5$ are almost undetectable), rises with the number of annotators (92% at 20, 100% at 200), and falls with rank (100% at $r = 1$, 57% at $r = 8$). Identifying the content-dependent component therefore requires both an expressive shared component and enough data per annotator.

H Negative Controls

Table 12: Negative controls (GoEmotions).

Features	Split	Observed	Within-item perm.	95% CI	Random assign.	95% CI
TF-IDF	item	+0.0142	−0.0008	[−0.0017, +0.0004]	−0.0000	[−0.0000, −0.0000]
TF-IDF	thread	+0.0129	−0.0023	[−0.0024, −0.0009]	−0.0000	[−0.0000, −0.0000]
TF-IDF	item+thread	+0.0186	+0.0233	[+0.0202, +0.0348]	+0.0287	[+0.0222, +0.0467]
SBERT	item	+0.0142	−0.0018	[−0.0023, −0.0001]	+0.0002	[−0.0003, +0.0016]
SBERT	thread	+0.0129	−0.0001	[−0.0011, +0.0011]	−0.0001	[−0.0001, −0.0001]
SBERT	item+thread	+0.0206	+0.0107	[−0.0014, +0.0177]	+0.0154	[−0.0011, +0.0191]

Table 12 is the full breakdown of the negative controls of Section 4.2, generated from the same cache as the values cited in the main text.

I Intervals for the Crossed Design

Table 13: Annotator-only versus crossed annotator-by-item intervals for Δ_{cond}.

Features	Split	Labels	Δ_{cond}	Annotator-only 95% CI	Crossed 95% CI
TF-IDF	item	K=28	+0.0142	[+0.0087, +0.0202]	[+0.0078, +0.0211]
TF-IDF	thread	K=28	+0.0129	[+0.0080, +0.0187]	[+0.0069, +0.0196]
TF-IDF	item+thread	K=28	+0.0186	[−0.0029, +0.0408]	[−0.0112, +0.0480]
SBERT	item	K=28	+0.0142	[+0.0090, +0.0201]	[+0.0083, +0.0210]
SBERT	thread	K=28	+0.0129	[+0.0072, +0.0196]	[+0.0062, +0.0204]
SBERT	item+thread	K=28	+0.0206	[+0.0104, +0.0318]	[+0.0087, +0.0332]
TF-IDF	item	K=7	+0.0138	[+0.0086, +0.0204]	[+0.0081, +0.0205]
TF-IDF	item	K=2	+0.0116	[+0.0058, +0.0189]	[+0.0055, +0.0191]

Our design is crossed: one text is annotated by several annotators and one annotator annotates many texts, so an interval that resamples only annotators does not reflect the variation left on the item side. Table 13 lists intervals that resample the item side as well. Because the estimator is unchanged, the point estimates are identical to Table 1, and setting all item-side weights to 1 recovers the annotator-only interval exactly; this agreement serves as an implementation check.

The crossed intervals widen uniformly but exclude 0 in all eight conditions. The largest widening is for the condition that separates both item and thread, where for TF-IDF $[+0.0409, +0.0720]$ becomes $[+0.0338, +0.0800]$; this condition has only 11,343 held-out items, a fifth of the others, so item-side resampling has more effect. For the item-only condition the widening is small, with $[+0.0117, +0.0242]$ becoming only $[+0.0111, +0.0249]$.

J Fit of the Strengthened Baselines Themselves

Table 14: Held-out NLL of the strengthened baselines (GoEmotions, item split, TF-IDF).

Annotator representation	DoF per annotator	Held-out NLL
None	0	2.2988
Intercept β_a	28	2.1167
Intercept + confusion matrix C_a	784	2.1229
Intercept + annotator × board	13524	2.1299

Table 14 gives the held-out NLL of the baseline models themselves, referenced by Appendix D. The lower half places the gain from adding the interaction side by side for a bare-intercept baseline and for a baseline that includes the confusion matrix.

K Multiplicity-Adjusted Intervals

Table 15: Multiplicity-adjusted (Bonferroni) annotator cluster intervals.

Condition	# annotators	Δ_{cond}	95% CI	Bonferroni CI	Excludes 0
GoEmotions / TF-IDF / item / K=28	57	+0.0142	[+0.0087, +0.0202]	[+0.0067, +0.0234]	✓
GoEmotions / TF-IDF / thread / K=28	57	+0.0129	[+0.0080, +0.0187]	[+0.0062, +0.0216]	✓
GoEmotions / TF-IDF / item+thread / K=28	57	+0.0186	[−0.0031, +0.0407]	[−0.0126, +0.0507]	—
GoEmotions / SBERT / item / K=28	57	+0.0142	[+0.0090, +0.0202]	[+0.0073, +0.0234]	✓
GoEmotions / SBERT / thread / K=28	57	+0.0129	[+0.0072, +0.0198]	[+0.0052, +0.0231]	✓
GoEmotions / SBERT / item+thread / K=28	57	+0.0206	[+0.0103, +0.0319]	[+0.0061, +0.0373]	✓
GoEmotions / TF-IDF / item / K=7	57	+0.0138	[+0.0085, +0.0202]	[+0.0068, +0.0235]	✓
GoEmotions / TF-IDF / item / K=2	57	+0.0116	[+0.0057, +0.0190]	[+0.0041, +0.0226]	✓
CSC / TF-IDF / item	872	+0.0230	[+0.0179, +0.0280]	[+0.0157, +0.0301]	✓
MP / TF-IDF / item	504	+0.0105	[+0.0085, +0.0126]	[+0.0077, +0.0135]	✓
EPIC / TF-IDF / item	74	−0.0005	[−0.0012, +0.0001]	[−0.0014, +0.0005]	—

Table 15 gives the adjusted intervals referenced by Section 5.3. Because the p-value of a Monte Carlo test cannot go below the lower bound $1/(R+1)$ set by the number of replicates, our replicate count cannot reach the Bonferroni level; the bootstrap, with 20,000 resamples, is not limited by resolution, so we make the adjusted verdict there. Nine of the 11 conditions exclude 0 under the adjusted bootstrap interval with a zero-effect null (the remaining two are the TF-IDF item+thread split and EPIC, neither of which is counted as detected in the main results).

L The Annotator-Shared Component of the Interaction

M_2^c imposes $\sum_a u_a = 0$ on the annotator weights, so the interaction carries only the component that differs across annotators. The split, held-out rows, selection grid, and rank set are identical to M_2. For TF-IDF with the item split, Δ_{cond} goes from $+0.0142\,[+0.0087, +0.0202]$ to $+0.0153\,[+0.0099, +0.0213]$, a difference of $-0.0011\,[-0.0034, +0.0012]$. For SBERT

it goes from $+0.0142\,[+0.0090,\,+0.0201]$ to $+0.0134\,[+0.0082,\,+0.0194]$, a difference of $+0.0008\,[-0.0001,\,+0.0016]$. The interval of the difference includes 0 in both, and the gain remains almost intact under the constraint (retention 107.8% for TF-IDF and 94.5% for SBERT, from the pre-rounding values). The component shared across annotators is not the source of the reported gain.

M Details for Reproduction

Preprocessing. The details of the text features and the size of each corpus are given in Appendix A. The point relevant to our estimation is that the TF-IDF vocabulary and IDF, the singular-value basis, and the standardization statistics are estimated from the training portion of each split only, allowing no leakage from a single fit outside the split. This estimation never uses the labels, and because the same feature is shared by M_0, M_1, and M_2, the shared text representation cancels in common for the differences Δ_{marg} and Δ_{cond} that we report, so that only the contribution of the annotator terms is measured. The Transformer input is tokenized only, with nothing fit.

Splits. We use 5-fold cross-fitting. The item split assigns folds by unique text, and the thread split by post-thread identifier. The split that satisfies both uses as held-out only the rows that violate neither constraint, so the held-out rows shrink. We carve out 15% of the training split (at least 500 rows) as the validation subset used to select the regularization coefficients and the rank.

Optimization. All models are trained with Adam (learning rate 0.02). The batch width is $\min(8192, \max(256, n/8))$, and the number of iterations is set from the size of the training split so that the number of gradient steps does not fall below 400; this prevents a small training split from stopping the nested models before convergence and turning the comparison into a measurement of optimization noise. The regularization grid is $\lambda_W \in \{10^{-3}, 10^{-4}\}$ and $\lambda_\beta, \lambda_u \in \{10^{-1}, 10^{-2}, 10^{-3}\}$, and every nested model chooses its own coefficients from the same grid.

Nonlinear family. The shared encoder is a 64-dimensional fully connected layer with $\tanh$ activation. N_0 has only the shared output layer, N_1 adds annotator intercepts, and N_2 adds a per-annotator output head $h_a \in \mathbb{R}^{64\times K}$. It is evaluated on the same splits and the same held-out rows as the linear family.

Inference. All intervals are from an annotator-level cluster bootstrap (8000 replicates). For each condition we draw one set of resamples and reuse it for all quantities within the condition, so that paired comparisons and differences of point estimates agree by construction. All quantities are derived from a single cache, so when the same condition appears in two tables it is the same estimate.

N Interaction Rank

The distribution of the rank selected on the validation subset and the Δ_{cond} at fixed rank are shown in the lower half of Table 5. For 28 labels, $r = 4$ is selected in all five folds; at $r = 1$ it is $+0.0080$ and at $r = 8$ it is $+0.0095$, both below the $+0.0142$ of the selected rank. The rank search range is $\{1, 2, 4, 8\}$ and does not include $r = 0$, but because the interaction penalty is also selected on the same validation subset, the penalty collapses the interaction when there is no content-dependent signal. That Δ_{cond} is indistinguishable from 0 for synthetic labels with no interaction, as reported in Table 3, is the empirical evidence of this collapse.