Your Model is Overconfident, and Other Lies We Tell Ourselves

Timothee Mickus[1,*] **Aman Sinha**[2,3,*] **Raúl Vázquez**[1]
[1]University of Helsinki [2]Université de Lorraine [3]ICANS Strasbourg
firstname.lastname@{[1]helsinki.fi,[2]univ-lorraine.fr}

Abstract

The difficulty intrinsic to a given example, rooted in its inherent ambiguity, is a key yet often overlooked factor in evaluating neural NLP models. We investigate the interplay and divergence among various metrics for assessing intrinsic difficulty, including annotator dissensus, training dynamics, and model confidence. Through a comprehensive analysis using 29 models on three datasets, we reveal that while correlations exist among these metrics, their relationships are neither linear nor monotonic. By disentangling these dimensions of uncertainty, we aim to refine our understanding of data complexity and its implications for evaluating and improving NLP models.

1 Introduction

A central, but often overlooked, concept in natural language processing is the consensus of annotators when labeling a specific datapoint. Annotators often express different perspectives, and prior literature provides strong evidence that this dissensus is a legitimate characteristic of language data, rather than a consequence of noisy annotation processes (Plank et al., 2014; Plank, 2022; Uma et al., 2021).

Annotator dissensus is often linked to *data complexity*: if humans do not agree as to whether a pair of sentences contradict each other, then we expect this pair to be hard to label with a neural network. Data complexity can be quantified through training dynamics (e.g., how early a training example is learned), model confidence, or performance variability across models (Guo et al., 2017; Swayamdipta et al., 2020; Hendrycks and Dietterich, 2019). This relationship shapes the design of evaluation benchmarks, which increasingly incorporate multiple annotations per datapoint (e.g., Bowman et al., 2015; Nie et al., 2020). As such, matching human uncertainty with model uncertainty is an explicit desideratum laid out in numerous NLP applications.

Common drivers of data uncertainty include annotation noise, semantic ambiguity, and overlapping class boundaries (Hu et al., 2023). These factors not only lower inter-annotator agreement but also highlight the presence of linguistic phenomena such as semantic indeterminacy — which lead to label ambiguity, as extensively documented across NLP tasks (Bowman et al., 2015). Hence, prior work has expected annotator dissensus to be a reasonable proxy for data complexity (e.g., Hachey et al., 2005; Lalor et al., 2018).

Our findings reveal non-linear and conflicting relationships between annotator dissensus and model-derived complexity metrics. Through experiments using 29 models on the ChaosNLI and DynaSent datasets (Nie et al., 2020; Potts et al., 2021), we observe that various metric indicators of data complexity often correlate with one another — and yet that the relationship they hold with human-based assessments of linguistic dissensus is far from linear or monotonic. Moreover, indicators of data complexity derived from model behavior tend to conflict depending on whether they account for the correctness of the models' predictions — for instance, assessments derived from conformal prediction methods do not align with the training dynamics approach of Swayamdipta et al. (2020), and both fail to adequately capture the true human label variation collected by Nie et al. (2020).

2 Background

Our study falls at the intersection of data complexity and uncertainty in NLP, particularly in relation to annotator disagreement, label variation, and the metrics used to evaluate them (Jiang and de Marneffe, 2022; Uma et al., 2021; Lorena et al., 2019;

[*] These authors contributed equally to this work.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5401–5417

Baan et al., 2023). Though closely related, these concepts involve distinct challenges.

Data uncertainty or aleatoric uncertainty describes the *randomness or noise inherent to the data* (Hu et al., 2023). This type of uncertainty is *irreducible* and cannot be eliminated through model improvements or tuning (Kiureghian and Ditlevsen, 2009; Kendall and Gal, 2017; Hüllermeier and Waegeman, 2021). Sources of data uncertainty include noisy observations, overlapping classes, ground truth errors or inherent randomness.

Annotator disagreement is highlighted as *a fundamental characteristic of linguistic data*, stemming from both annotation noise and the inherent ambiguity of language (Plank et al., 2014; Aroyo and Welty, 2015; Pavlick and Kwiatkowski, 2019; Fornaciari et al., 2021). It often correlates with *label uncertainty*, where no single correct label exists. High-disagreement examples contain valuable signals for classifiers (Basile et al., 2021; Palomaki et al., 2018). However, disagreement does not necessarily indicate annotation noise—it may instead reflect genuine linguistic or contextual ambiguity.

Data complexity or data difficulty refers to the *characteristics of a data sample that make classification inherently difficult*. It is related to the structural properties of the data, not to randomness or noise, as is the case with data uncertainty. Several factors contribute to data complexity: proximity to decision boundaries and class overlap, semantic indeterminacy, and task-specific challenges, such as requiring world knowledge (Plank, 2022; Jiang and de Marneffe, 2022).

Metrics for evaluating data uncertainty and complexity include model confidence (probability of the predicted class), entropy of predicted probabilities (measuring classification uncertainty), and confidence calibration (aligning confidence with performance). These help assess label uncertainty caused by overlapping class boundaries (e.g., Geng et al., 2024; Zhou et al., 2022; Xiao and Wang, 2019). Training dynamics, such as how quickly a model learns to classify an example or the shape of the loss curve, further reveal the relative difficulty of datapoints (Swayamdipta et al., 2020; Toneva et al., 2019; Baldock et al., 2021). These metrics offer tools to analyze the challenges of data complexity, yet they do not provide a nuanced perspective on the interplay between annotator disagreement and model uncertainty.

Complexity, disagreement & uncertainty. Data complexity, annotator disagreement, and data uncertainty are intertwined phenomena with similar root causes. For instance, while authors differ in their terminology, 'overlapping classes' (Ho and Basu, 2002; Peterson et al., 2019; Lorena et al., 2019), 'absence of a single ground truth' (Aroyo and Welty, 2015; Baan et al., 2022), or 'linguistic ambiguity' due to semantic and social factors (Plank, 2022) all refer to the fact that some datapoints can have different labels — which drives up complexity, disagreement and uncertainty.

Hence, these three phenomena have been conflated in the literature. For example, uncertainty is often measured through label distribution entropy, used as a proxy for annotator disagreement (Zhang et al., 2022; Baumler et al., 2023). Similarly, Lalor et al. (2018) has found alignment between human difficulty and model-assigned probability mass, suggesting that both perceive difficulty similarly. This viewpoint is also prevalent in active learning (e.g., Hachey et al., 2005) and attested less directly in quality estimation (e.g., Jamison and Gurevych, 2015). Additionally, anecdotal evidence has been used to support this connection, especially in studies exploring the underlying causes of data complexity (Swayamdipta et al., 2020; Baldock et al., 2021; Rajpurkar et al., 2018).

This perspective has also been employed as a working hypothesis. For example, Weinshall et al. (2018) assume that knowledge distillation from a model can play the same role as humans in active learning scenarios, while Beigman and Beigman Klebanov (2009) propose replacing annotator-based disagreement assessment with classifier proxies. Authors adopting this approach often treat it as a simplifying initial assumption, and frequently include modeling work to better capture human variation (e.g. Reidsma and Carletta, 2008) or discussions of the limitations of this working hypothesis (as in Beigman and Beigman Klebanov).

Treating these three phenomena as interchangeable oversimplifies their relationships. Disagreement often signals semantic complexity but can also stem from bias, expertise variance, or cultural differences (e.g., Jiang and de Marneffe, 2022). Similarly, uncertainty overlaps with complexity but also arises from noise that is not tied to structural data complexity (Kendall and Gal, 2017), while example difficulty has been linked to factors that are *a priori* not linguistic, such as class imbalance or distributional shifts (Ho and Basu, 2002; Gawlikowski

et al., 2023). While some studies find overlap between human disagreement and model uncertainty (Swayamdipta et al., 2020; Baldock et al., 2021), others challenge this view (Reidsma and Carletta, 2008). Our findings highlight the need to distinguish these concepts, as we show that complexity metrics do not map linearly to human assessments.

3 Indicators of data complexity

We formalize several means of assessing how difficult it is to assign one of k possible labels $\{y_1, \ldots, y_k\} = Y$ to a specific instance x.

3.1 Human-based indicators

Empirical population dissensus. The simplest way to quantify disagreement on a specific datapoint is to ask multiple annotators $a_1, \ldots, a_n$, and compute how unpopular the majority opinion is. As such, if annotator a_j would assign the label y_{a_j} to the observation x, we can define the probability $\Pr_{\mathbb{H}}(y_i|x)$ on label y_i as the proportion of annotators agreeing on the label y_i, or formally

$$\Pr{}_{\mathbb{H}}(y_i|x) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\left\{y_{a_j} = y_i\right\}$$

and denote the dissensus among annotators as:

$$\mathbb{H}_{\text{dis}} = 1 - \max_{y_i \in Y} \Pr{}_{\mathbb{H}}(y_i|x), \tag{1}$$

$\mathbb{H}_{\text{dis}}$ is hence inversely related to the popularity of the most common label. If all annotators agree, there is a strong consensus, implying that $\mathbb{H}_{\text{dis}} = 0$ because $\max_i \Pr_{\mathbb{H}}(y_i|x) = 1$.

Empirical population entropy. The empirical dissensus $\mathbb{H}_{\text{dis}}$ has the drawback of not factoring in minority opinions: there is a distinction to be made between having the opinions split among a handful of well-supported alternatives versus a total lack of consensus and annotators maximally split across all possible alternatives. To account for such differences, we consider the empirical entropy of opinions (Nie et al., 2020), or

$$\mathbb{H}_{\text{ent}} = -\sum_{y_i \in Y} \Pr{}_{\mathbb{H}}(y_i|x) \log \Pr{}_{\mathbb{H}}(y_i|x) \tag{2}$$

Entropy measures uncertainty or diversity in the label distribution, better accounting for both dominant and minority labels. As before, $\mathbb{H}_{\text{ent}} = 0$ when all annotators agree on one label. Contrastingly, $\mathbb{H}_{\text{dis}}$ is maximal when $\Pr_{\mathbb{H}} \sim \text{Unif}$, i.e., when annotators are evenly split across all labels.

3.2 Reference-free model-based indicators

Model pool dissensus and model pool entropy. Given a set of models parametrized by $\theta_1, \ldots, \theta_m$, we can easily extend the concepts of dissensus ($\mathbb{H}_{\text{dis}}$) and entropy ($\mathbb{H}_{\text{ent}}$) to models' predictions, instead of relying on human annotators. To do this, we evaluate the predictions of a model θ_j by selecting the label $\text{argmax}_{y_i \in Y}\ p(y_i|x, \theta_j)$. Next, we define the probability $\Pr_{\mathbb{M}}(y_i|x)$ of this datapoint being labeled as y_i, by tallying the number of models that predict y_i as the most likely label:

$$\Pr{}_{\mathbb{M}}(y_i|x) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\left\{y_i = \operatorname*{argmax}_{y_k \in Y}\ p(y_k|x, \theta_j)\right\}$$

Using this distribution, we can analogously define both metrics for the models' predictions:

$$\mathbb{M}_{\text{dis}} = 1 - \max_{y_i \in Y} \Pr{}_{\mathbb{M}}(y_i|x) \tag{3}$$

$$\mathbb{M}_{\text{ent}} = -\sum_{y_i \in Y} \Pr{}_{\mathbb{M}}(y_i|x) \log \Pr{}_{\mathbb{M}}(y_i|x) \tag{4}$$

Averaged model entropy. Entropy has also been used to assess the confidence of a model in its own prediction (e.g., Malinin and Gales, 2021; Schröder et al., 2022; Baumler et al., 2023). A reasonable line of thought is that lower confidence scores reflect data complexity. To evaluate the difficulty of labeling x, we can average the label distribution entropy across multiple models:

$$\mathbb{M}_{\text{avg ent}} = -\frac{1}{m}\sum_{j=1}^{m}\sum_{y_i \in Y} p(y_i|x, \theta_j) \times \log p(y_i|x, \theta_j) \tag{5}$$

Conformal prediction set size. A more elaborate statistical estimator than entropy consists in quantifying the ambiguity necessary for a probabilistic classifier to meet a certain statistical guarantee; an approach known as conformal prediction (CP, Vovk et al., 2005; Angelopoulos and Bates, 2022). In practice, we can also use a probabilistic classifier parametrized with θ to derive a set of possible labels $\mathcal{C}_\theta(x) \subseteq Y$ for every input x such that the true label y^* is likely to be in $\mathcal{C}_\theta(x)$, with a budget tolerance for failure $1 - \alpha$. Formally, we want to construct a set-valued function $\mathcal{C}_\theta$ such that

$$\forall x \qquad \Pr\left(y^* \in \mathcal{C}_\theta(x)\right) \geq 1 - \alpha$$

We can then capture the ambiguity inherent to a prediction by considering the size of the prediction

set, $|\mathcal{C}_\theta(x)|$: a larger CP set size ought to reflect a greater uncertainty as to what the true label is. To convert a probabilistic classifier $p(Y|X,\theta)$ to such a set-valued classifier, we rely on a least-ambiguous set-valued classifier method (Sadinle et al., 2019). This consists in identifying the value t_θ such that, for all calibration datapoints x' with their label y' in a held-out calibration dataset $\mathcal{D}_{\text{cal}}$:

$$\hat{q} = \frac{|\mathcal{D}_{\text{cal}}| + 1}{|\mathcal{D}_{\text{cal}}|}(1-\alpha)$$
$$t_\theta = \sup\left\{t \middle| \Pr\left(p\left(y'|x',\theta\right) \geq t\right) \geq \hat{q}\right\}$$

Using t_θ, we can construct the set

$$\mathcal{C}_\theta(x) = \{y \mid p\left(y|x,\theta\right) \geq t_\theta\}$$

which provides the expected statistical guarantee. Here, we convert CP sets into uncertainty indicators by considering their average size across models:

$$\mathbb{M}_{\text{CP}} = \frac{1}{m}\sum_{i=1}^{m} |\mathcal{C}_{\theta_i}(x)| \tag{6}$$

Here, we experiment with three variants, based on different risk tolerances with $\alpha = 0.05$, $\alpha = 0.1$ or $\alpha = 0.2$. While conformal prediction algorithms require labeled calibration sets $\mathcal{D}_{\text{cal}}$, their predictions are made without label information. Hence we consider CP set size indicators to be reference-free, as they can estimate uncertainty for unlabeled datapoints. We use as $\mathcal{D}_{\text{cal}}$ all other datapoints in the test set (i.e., a leave-one-out process).

3.3 Reference-dependent model-based indicators

Model pool failure rate. Since it is in principle possible for models to broadly agree on a label that human annotators would not have picked, one value worth considering is the proportion of models that fail to produce the reference label y^* we would expect given our annotations. Defining

$$\mathbb{M}^{\text{ref}}_{\text{fail}} = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\left\{\underset{y_j \in Y}{\text{argmax}}\, p(y_j|x,\theta_i) \neq y^*\right\} \tag{7}$$

highlights the disconnect between model predictions and human annotations. A low value for $\mathbb{M}^{\text{ref}}_{\text{fail}}$ implies a strong alignment between the model pool and the human-provided reference label; a high value suggests that many models fail to predict y^*.

Early computation termination. Baldock et al. (2021) propose to estimate the difficulty of an example through the computational cost of a correct prediction. They first compute the hidden representations $\mathbf{h}_i^1, \ldots \mathbf{h}_i^l$ for a specific input $\mathbf{x}_i$ and then assess which of these representations lie in label-specific subspaces using kNN classifiers, since datapoints that are easier to process ought to be mapped onto unambiguous subspaces earlier.

This approach assumes there is a meaningful distance metric between the different representations — an assumption that is not easy to meet with sequence-level classification tasks, where inputs can have different matrix shapes. We can however leverage the fact that Transformer layers can be viewed as functions mapping from and unto the same space (Elhage et al., 2021): Earlier work has suggested to interpret hidden representations for a specific layer by directly projecting them onto the label-space, skipping over all subsequent layers (nostalgebraist, 2020; Geva et al., 2022). We therefore replace Baldock et al.'s kNN classifiers with the learned classifier head. More formally, if a model parametrized with θ_i is made of l_i layers of the form $f_{\theta_i,j}(\mathbf{X}) = \phi(\mathbf{X}, \theta_{i,j})$ and a projection head $f_{\theta_i,\text{proj}}(\mathbf{X}) = \text{argmax}\,\psi(\mathbf{X}, \theta_{i,l_i+1})$, let us denote all early predictions from layer j onward as

$$\hat{Y}_{ij} = \left\{ f_{\theta_i,\text{proj}} \circ f_{\theta_i,k} \circ \cdots \circ f_{\theta_i,1}\left(\mathbf{X}\right) \;\middle|\; j \leq k \leq l_i \right\}$$

which allows us to retrieve the first layer k such that all predicts from layer k to layer l are correct, according to a reference label y^*:

$$\text{s}^{\text{ref}}_{1^{\text{st}}\ \text{layer}}(\theta_i) = \begin{cases} 1 & \text{if } p(y|x,\theta_i) \neq y^* \\ \dfrac{\min_j\{j \mid \hat{Y}_{ij} = \{y^*\}\}}{l+1} & \text{otherwise} \end{cases}$$

$$\mathbb{M}^{\text{ref}}_{1^{\text{st}}\ \text{layer}} = \frac{1}{m}\sum_{i=1}^{m} \text{s}^{\text{ref}}_{1^{\text{st}}\ \text{layer}}(\theta_i) \tag{8}$$

We average across our pool of models so that the indicator is not too sensitive to one specific model's idiosyncratic behavior. This also leads us to normalizing according to the number of layers so that we maintain consistent ranges across models with different layer counts. We also make the practical choice of setting examples that models do not label correctly to the higher end of the scale.

Early training termination. One can also consider that easier items require less training (Swayamdipta et al., 2020). If for a given model θ_i we have access to different checkpoints across

training $\theta_i^1, \ldots, \theta_i^p$, we can simply assess when the model starts making reliable predictions. Consider the set of predictions from all future checkpoints:

$$F_{ij} = \left\{ \underset{y}{\text{argmax}}\, p(y|x, \theta_i^j), \ \ldots, \ \underset{y}{\text{argmax}}\, p(y|x, \theta_i^p) \right\}$$

which we use to define:

$$\text{s}^{\text{ref}}_{1^{\text{st}}\ \text{ckpt}}(\theta_i^p) = \begin{cases} 1 & \text{if } p(y|x, \theta_i^p) \neq y^* \\ \frac{\min_j \{j \mid F_{ij} = \{y^*\}\}}{p+1} & \text{otherwise} \end{cases}$$

$$\mathbb{M}^{\text{ref}}_{1^{\text{st}}\ \text{ckpt}} = \frac{1}{m} \sum_{i=1}^{m} \text{s}^{\text{ref}}_{1^{\text{st}}\ \text{ckpt}}(\theta_i^p) \quad (9)$$

Here again, we normalize according to the number of checkpoints, average across all models, and manually penalize models that do not ultimately learn to produce the target reference.

Failure rate through training. We can also assume that easier items are likely to be attributed the expected reference at any stage of training, whereas more complex observations will only be labeled properly during the later stages. We can therefore quantify the proportion of checkpoints where the model failed to produce the expected label y^*:

$$\mathbb{M}^{\text{ref}}_{\text{avg}\ \text{ckpt}} = \frac{1}{mp} \sum_{i=1}^{m} \sum_{j=1}^{p} \mathbb{1} \left\{ p(y|x, \theta_i^j) \neq y^* \right\} \quad (10)$$

Again, we average across a pool of models to mitigate idiosyncrasies.

Probability mass through training. One problem with the approach in eq. (10) is that it does not distinguish between cases where the classifier correctly predicts y^* and assigns no weight to any other options from cases where the probability assigned to y^* is only within a small margin from that of an incorrect class. Swayamdipta et al. (2020) propose to consider the probability mass assigned by the classifier across training,[1] or formally:

$$\mathbb{M}^{\text{ref}}_{\text{avg}\ \text{ckpt}\ p} = 1 - \frac{1}{mp} \sum_{i=1}^{m} \sum_{j=1}^{p} p(y^*|x, \theta_i^j) \quad (11)$$

Eq. (11) is minimized when the gold label y^* is assigned a probability of 1 throughout training.

4 ChaosNLI

4.1 Experimental setup

We first study classifiers trained on SNLI (Bowman et al., 2015) or MNLI (Williams et al., 2018) and evaluated on ChaosNLI (Nie et al., 2020). NLI as a task is a ternary classification problem, which involves classifying pairs of sentences depending on whether the second sentence contradicts the first; whether the first entails the second; or whether they are neutral with respect to one another, i.e., the second sentence neither derives from nor contradicts the first. We list some statistics as to the number of instances in these datasets in Table 1.

Dataset	Variant	Train	Val	Test
SNLI	All labeled (<1B)	549 367	9842	(*unused*)
	5-splits (1B)	109 873	9842	(*unused*)
MNLI	All labeled (<1B)	392 702	9815	(*unused*)
	5-splits (1B)	78 540	9815	(*unused*)
ChaosNLI	SNLI split	—	—	1000
	MNLI split	—	—	1000

Table 1: Dataset statistics for NLI datasets.

We might expect the family of models we consider to define our indicators to weigh on results. In particular, the homogeneity of the pool of models considered — in terms of pretraining and finetuning data, algorithmic and architectural designs, or parameter counts — is a factor of interest.

Heterogeneous training, similar parameter counts (1B group). One may expect that data complexity indicators should be established by considering a large swath of models trained in conditions as varied as possible — i.e., using different training data and algorithms. To this end, we consider 5 different LLMs in the 1B parameter range; OLMo (Groeneveld et al., 2024), Pythia (Biderman et al., 2023), Llama 3.2 (Grattafiori et al., 2024), Falcon (Almazrouei et al., 2023), and BLOOM (Scao et al., 2023). So as to further maximize the difference across the different models we consider, we partition the NLI training set (either SNLI or MNLI) into five equally sized subsets $s_1, \ldots, s_5$ and train one model for each pair of LLM and NLI subset, or 25 classifiers on SNLI and MNLI each.

Homogeneous training data, different parameter counts (<1B group). Conversely, we might expect that the model pool should be established with a fixed training data — on the one hand, this corresponds to an assumption frequently made when measuring aleatoric uncertainty in the Bayesian literature; on the other hand, we might expect that difficulty should be intimately linked to the data a model has been exposed to. To that tend, we consider a family of smaller BERT-type models

[1] This indicator corresponds to what Swayamdipta et al. call "confidence."

	<1B pool		1B pool	
	$\mathbb{H}_{\text{ent}}$	$\mathbb{H}_{\text{dis}}$	$\mathbb{H}_{\text{ent}}$	$\mathbb{H}_{\text{dis}}$
$\mathbb{M}_{\text{dis}}$	0.2440	0.2179	0.1947	0.1772
$\mathbb{M}_{\text{ent}}$	0.2784	0.2433	0.2183	0.1970
$\mathbb{M}_{\text{avg ent}}$	0.3901	0.3490	0.2811	0.2398
$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.3737	0.3186	0.2767	0.2315
$\mathbb{M}_{\text{CP } \alpha=0.1}$	0.3763	0.3379	0.2819	0.2393
$\mathbb{M}_{\text{CP } \alpha=0.2}$	0.3248	0.3064	0.2482	0.2157
$\mathbb{M}^{\text{ref}}_{\text{fail}}$	0.3990	0.3959	0.3497	0.3330
$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	0.3719	0.3796	0.3624	0.3387
$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	0.4357	0.4244	0.3682	0.3443
$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	0.3969	0.3904	0.3477	0.3274
$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$	0.4386	0.4241	0.3670	0.3428

Table 2: Spearman correlation between human-based and model-based indicators on SNLI.

	<1B pool		1B pool	
	$\mathbb{H}_{\text{ent}}$	$\mathbb{H}_{\text{dis}}$	$\mathbb{H}_{\text{ent}}$	$\mathbb{H}_{\text{dis}}$
$\mathbb{M}_{\text{dis}}$	−0.0022	−0.0045	0.1419	0.1074
$\mathbb{M}_{\text{ent}}$	0.0023	−0.0011	0.1587	0.1201
$\mathbb{M}_{\text{avg ent}}$	−0.0077	−0.0158	0.1329	0.1095
$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.0101	−0.0085	0.0788	0.0425
$\mathbb{M}_{\text{CP } \alpha=0.1}$	−0.0073	−0.0173	0.1798	0.1164
$\mathbb{M}_{\text{CP } \alpha=0.2}$	−0.0184	−0.0231	0.1581	0.0936
$\mathbb{M}^{\text{ref}}_{\text{fail}}$	0.1174	0.1508	0.1726	0.2246
$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	0.0682	0.0966	0.2040	0.2514
$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	0.1132	0.1479	0.1829	0.2307
$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	0.1168	0.1498	0.1764	0.2261
$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$	0.1094	0.1434	0.1813	0.2307

Table 3: Spearman correlation between human-based and model-based indicators on MNLI.

(Turc et al., 2019) so as to verify how the indicators in behave with respect to a family of different models trained homogeneously on the same data and under the same conditions, varying in terms of architecture designs and parameter counts. We fine-tune all of Turc et al.'s BERT models on each of the NLI training sets in their entirety.

4.2 Results

Human-based and model-based indicators do not agree with each other. A straightforward first approach consists in computing how the different indicators correlate with one another — in particular, we start by focusing on comparing human-based indicators to model-based indicators.

The corresponding Spearman correlation values are shown in Tables 2 and 3. Reference-free indicators defined without factoring in the majority label among human annotators (eqs. (3) to (6)) almost systematically yield lower correlations than reference-dependent indicators (eqs. (7) to (11)).

Yet, while we observe positive and significant trends throughout, the correlation itself is somewhat low. For a sense of scale, if we are to focus on SNLI for which we observe the highest correlations, two human-based indicators or two reference-dependent indicators, tend to yield correlation scores of $\rho \geq 0.9$. When comparing two reference-free indicators, we can observe two sub-groups: namely, $\mathbb{M}_{\text{avg ent}}$ and the CP set size indicators yield correlations of $\rho \geq 0.9$),[2] whereas $\mathbb{M}_{\text{dis}}$ and $\mathbb{M}_{\text{ent}}$ yield a correlation of $\rho \approx 0.95$, and comparisons across these two sub-groups are in the range $0.64 < \rho < 0.88$. The observation also holds on MNLI: We observe a correlation of $\rho \approx 0.90$ for $\mathbb{H}_{\text{dis}}$ and $\mathbb{H}_{\text{ent}}$, correlations systematically greater than $\rho \geq 0.9$ between any two reference-dependent indicators, and correlations between $0.46 \leq \rho \leq 0.96$ for reference-free indicators (with again $\mathbb{M}_{\text{dis}}$ and $\mathbb{M}_{\text{ent}}$ forming a subgroup distinct from $\mathbb{M}_{\text{CP}}$ and $\mathbb{M}_{\text{avg ent}}$). In sum, all three groups of indicators portray different pictures, echoing findings from prior works (esp. Pavlick and Kwiatkowski, 2019): The difficulty associated to the samples is *not* the same for the humans and models, regardless of the pool considered.[3]

The behavior of model-based indicators also appears contingent on the exact setup. For instance, observations derived from our 1B model pool on MNLI would suggest $\mathbb{M}_{\text{CP } \alpha=0.1}$ to be quite in line with human label variation assessments — whereas the corresponding coefficient in the <1B pool on MNLI is about 0. In the same vein, the choice of α for CP has different effects on SNLI and MNLI insofar the 1B pool is concerned: Whereas $\mathbb{M}_{\text{CP } \alpha=0.2}$ yields higher results than $\mathbb{M}_{\text{CP } \alpha=0.05}$ on MNLI, the opposite is true for SNLI classifiers.

Reference-free indicators conflate model success and model failure. We can also remark that reference-free and reference-dependent indicators do not agree either. This is evident, for instance, by looking at Figure 1, which exemplifies one such comparison. We can see that the joint distribution of the indicators forms an inverted U-shape distribution, i.e., the reference-free indicator rates as equally good items that the reference-dependent does discriminates. Generally speaking, reference-free indicators tend to assign similar scores to datapoints rated as either maximal or minimal by reference-dependent indicators: In fact, if we partition datapoints according to whether a majority of the models fail to predict the annotator

[2] Except $\mathbb{M}_{\text{CP } \alpha=0.05}$ and $\mathbb{M}_{\text{CP } \alpha=0.2}$, where $\rho \approx 0.80$.

[3] We can stress this relationship is non-linear, see §B.1.

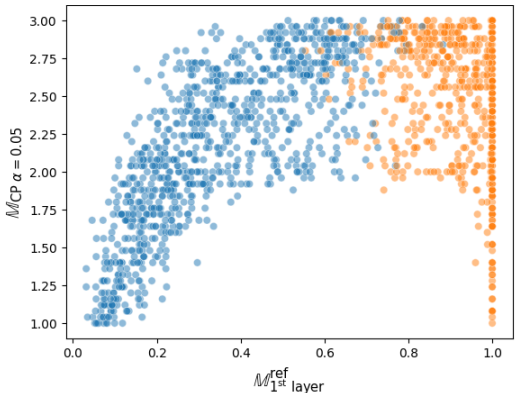

Figure 1: Example of joint distribution between a reference-free and a reference-dependent indicator (SNLI 1B pool, $\mathbb{M}_{\text{CP, }\alpha=0.05}$ vs. $\mathbb{M}^{\text{ref}}_{1^{\text{st}}\text{ layer}}$). Datapoints in orange are misclassified by 50% of the pool, blue datapoints aren't. See also tables 9 and 10 (§B.2).

majority label (corresponding to the orange and blue hues in Figure 1), we can observe systematic *anti*-correlations when the models do tend to fail.[4] One major factor at play here is that models fail more often on samples with a high human dissensus. This can be shown with Mann-Whitney U tests. On SNLI, for the 1B models, we observe a p-value of $p < 10^{-27}$ and a common language effect size $f = 66.7\%$; as for the <1B models, we have $p < 10^{-42}$, $f = 72.2\%$. On MNLI, the 1B model pool yields $p < 10^{-14}$ and $f = 61.3\%$, whereas the <1B pool yields $p < 10^{-7}$ and $f = 58.1\%$.[5]

5 DynaSent

5.1 Experimental setup

An overlooked aspect of our discussion so far is whether the discrepancy between indicators highlighted in §4 can be mitigated. Since we trained our classifiers to match the consensus among annotators, any probability mass not assigned to the majority label is penalized. As a result, classifiers default to assigning most of their probability mass to a single label. That is, they may be poorly calibrated in the sense of Guo et al. (2017), and the probabilities that they assign might not reflect the probability of the model being correct. As Baan et al. (2022) discuss, measuring calibration for ambiguous labels is not possible in practice as metrics such as ECE explicitly assume the existence of a single valid label. Baan et al. instead advocate to measure how well the model's distribution aligns with human label variation.[6] We thus expect that directly optimizing the classifiers to match the empirical human label variation will remedy this issue. We refer to this approach as training with 'soft' continuous labels, instead of 'hard' categorical labels.

For these experiments we use DynaSent (Potts et al., 2021), which also allows us to assess whether out results generalize beyond NLI. DynaSent is a sentiment analysis dataset built incrementally through several rounds of adversarial data collection, featuring four sentiment classes: positive, negative, neutral or mixed. Each datapoint is annotated by five crowd-workers, providing a first-order approximation of human label variation. We focus on the round 1 data to train classifiers using either hard or soft labels as targets. A potential challenge is label imbalance, which has been linked to example difficulty (Ho and Basu, 2002). Therefore, we convert the dataset to a ternary label scheme by removing all datapoints that at least one annotator marked as 'mixed,' then subsample the dataset to guarantee an equal distribution of positive, neutral and negative labels. We defer a replication of this experiment on the original dataset to Appendix B.4 and focus on this re-balanced version below.

Dataset	Variant	Train	Val	Test
DynaSent	Re-balanced	32 001	3066	3027
	All labeled (§B.4)	84 388	3600	3600

Table 4: Dataset statistics for Dynasent.

This re-balancing procedure severely limits the size of our dataset to roughly 32K training instances, see Table 4. We report results on this re-balanced dataset, using the same PLM pools as in our previous experiments. Given the limited data, we train 1B pool models on the full dataset, and report results across 3 runs, whereas we only train 1 seed for each of the 24 models from Turc et al. (2019) in our homogeneous pool.

5.2 Results

Table 5 summarizes the correlation between model-based and human-based indicators on Dynasent. Due to the dataset's structure—specifically the limited number of annotations per datapoint and their majority label selection—$\mathbb{H}_{\text{ent}}$ and $\mathbb{H}_{\text{dis}}$ are perfectly correlated. We replicate observations made for NLI in §4: When training the classifiers using hard labels, we observe a clear divide between reference-free and reference-dependent indicators.

[4] See in Tables 9 and 10 in §B.2 for detailed results.

[5] Conversely, we can still identify a small subset of datapoints with low human dissensus but high model failure rates.

[6] This is similar to what we present in Tables 2 and 3.

	<1B pool		1B pool	
	soft	hard	soft	hard
$\mathbb{M}_{\text{dis}}$	0.0679	0.0654	0.1553	0.1435
$\mathbb{M}_{\text{ent}}$	0.0708	0.0631	0.1606	0.1463
$\mathbb{M}_{\text{avg ent}}$	0.1395	0.1266	0.1996	0.1206
$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.1223	0.1075	0.1904	0.1386
$\mathbb{M}_{\text{CP } \alpha=0.1}$	0.1177	0.0998	0.1836	0.1349
$\mathbb{M}_{\text{CP } \alpha=0.2}$	0.1100	0.0813	−0.0871	−0.1262
$\mathbb{M}^{\text{ref}}_{\text{fail}}$	0.1286	0.1156	0.1858	0.1764
$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	0.1319	0.1235	0.2016	0.1928
$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	0.1313	0.1137	0.2016	0.1907
$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	0.1268	0.1133	0.1865	0.1780
$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$	0.1504	0.1360	0.2257	0.1893

Table 5: Spearman's ρ between model-based indicators vs. human-based (entropy) indicator on DynaSent.

Using soft labels yields obvious improvements for some of the reference-free indicators: In particular $\mathbb{M}_{\text{CP}}$ and $\mathbb{M}_{\text{avg ent}}$ are in some case competitive with reference-dependent metrics. It is however crucial to highlight that results remain volatile, as attested by the anti-correlations yielded by $\mathbb{M}_{\text{CP } \alpha=0.2}$;[7] remark also that the optimal risk tolerance α we observe on DynaSent is 0.05, instead of 0.1 as we observed for NLI models. Nor is the gap between reference-free and reference-dependent metrics fully mitigated: In all cases, the reference-dependent indicator $\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$ outperforms all other indicators. Taking stock of which indicators strongly benefit from soft-label trainings (viz. $\mathbb{M}_{\text{CP}}$, $\mathbb{M}_{\text{avg ent}}$ and $\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$), we remark that they are derived from the probability distribution, rather than its argmax. Soft labels foster distributions that are more in line with human label variation, but this might not suffice to fully bridge the gap between reference-free and reference-dependent indicators.[8]

6 Discussion

Our study of how different indicators of data complexity correlate to one another has shown a somewhat perplexing picture worth diving into. As we have established, model-based indicators align poorly with human-based indicators — while we often observe positive correlations, their magnitudes are low. Defining model-based indicators with respect to human majority labels partially narrows the gap between the two, primarily because reference-free indicators often converge on a single label, regardless of its alignment with human preferences or the strength of consensus within the annotator pool. Within the reference-free indicators, we can also tentatively distinguish two subgroups: assessments that rely only on the pool of models considered (eqs. (3) and (4)) appear to have a distinct profile from those which rely on more complex statistics, such as CP set sizes or entropy (eqs. (5) and (6)). For CP specifically, it is worth stressing that desiderata in terms of coverage can also entail significant variability.

Training classifiers to directly predict human label variation does not fully bridge the gap between reference-free and reference-dependent indicators, and only improves correlations with human assessment for indicators that do not summarize an model's distribution to its argmax. Models often overwhelmingly agree on labels that lack humans annotator consensus, and factoring in human preferences in indicators is necessary though not sufficient for bridging the gap between human-based and model-based assessments difficulty. This underscores a critical limitation of the current research landscape: Reference-free approaches such as CP or entropy are at odds with reference-dependent approaches (e.g., Swayamdipta et al., 2020; Baldock et al., 2021), in that the former conflate failures and successes.[9]

Practical engineering recommendations also emerge from our observations. Authors interested in developing automated assessments of data complexity in line with human assessments should favor (i) training models on soft labels, (ii) factoring in the actual probability distribution of the model, and (iii) leveraging the human label distribution, e.g., through the majority label.

In all, the present observations highlight a disconnect in the current literature. If data uncertainty is to be accounted for by factors such as noise, ambiguity or label overlap during data collection — factors that we also expect to weigh in on measurements of linguistic disagreement — then there is a need to reconcile this line of thought with the limited predictability of model-based assessments of data complexity from annotators' preferences.

[7] There are several possibility as to what causes this unexpected pattern for $\mathbb{M}_{\text{CP } \alpha=0.2}$; we consider in particular the small size of the dataset as the behavior is not reproduced when training on the non-subsampled dataset, cf. §B.4.

[8] This is also in line with the fact that reference-free indicators derived from classifiers trained on soft labels still conflate success and failure, as shown in §B.2, Table 11.

[9] A related train of thought that can shed more light on our observations consists in considering which factors shape model decisions. See §B.3 for a discussion.

7 Conclusions

We present a study with 11 indicators and 29 models, which show that human-based assessments of difficulty need not align with model-based assessments (Pavlick and Kwiatkowski, 2019) and that model-based assessments exhibit stark differences according to whether they factor in human preferences. Data complexity and annotator disagreements, as assessed by model-based indicators or annotator label distribution, have clearly distinct behaviors, despite the overlap the literature posits (Lalor et al., 2018). This calls for replication of our study in other settings, other tasks, other languages, etc.: Establishing the prevalence of the confound we identify remains a topic for future work.

Lastly, our findings also question practices adopted by the field. If we are to posit a sharp distinction between data complexity as exemplified by Swayamdipta et al. (2020) or Baldock et al. (2021), vs. uncertainty as captured by e.g. conformal prediction methods, then we need to explain why said data complexity is more in line with annotator disagreement than CP-based estimates of uncertainty. Likewise, model-based estimates used in active learning (e.g. Schröder et al., 2022; Baumler et al., 2023) do not align with all definitions of uncertainty, especially label uncertainty as assessed through inter-annotator agreements. Such an exercise in terminology is a necessary step forward if we are to address challenges such as disentangling sources of uncertainty (Mucsányi et al., 2024) or leveraging uncertainty as a richer training signal (Basile et al., 2021; Palomaki et al., 2018).

Acknowledgements

This work was supported by the ICT 2023 project "Uncertainty-aware neural language models" funded by the Academy of Finland (grant agreement № 345999).

Limitations

We identify two core limitations on our findings.

First, the present study relies on two datasets, namely the ChaosNLI re-annotation by Nie et al. (2020) and the DynaSent dataset of Potts et al. (2021). While this limits the usefulness of our findings, and entails that our results might not carry on to other setups, we believe this choice is practically necessary (in that very few datasets are available with a training split large enough to easily train classifiers). It is in fact debatable whether DynaSent squarely meets all desiderata, since its validation split might not contain a rich enough set of annotations to accurately capture human label variation: DynaSent only collects five judgments from crowd workers, which Nie et al. (2020) shows to be unreliable. On a practical level, this also means that there are many human-based indicators that we have ignored; e.g., the 'complicated' label of Jiang and de Marneffe (2022) or other explicit self-reports of uncertainty from the annotators could yield valuable insight that would contrast with the label distribution–based indicators we consider in eqs. (1) and (2); conversely, we have not considered metrics defined with respect to the dataset its in entirety (e.g. Ethayarajh et al., 2022). Of course, all studies need to define their scope: In our case, more can always be done to integrate other data uncertainty/difficulty indicators from a wider range of studies, beyond the key ones we study here (viz. Nie et al., 2020; Vovk et al., 2005; Baldock et al., 2021; Swayamdipta et al., 2020).

Second, we rely on pool of models that have not been individually optimized for the task they are tested on. This point bears further discussion: As we identify in Table 8, a major driver of the difference between reference-free and reference-dependent indicators is whether or not the classifier correctly identifies the gold label; and it therefore stands to reason that better trained classifiers may exhibit different patterns. There are however three key facts to stress here. First, hyperparameter tuning over a large pool of models (24 BERT variants from (Turc et al., 2019), plus 25 1B PLMs–based classifiers, on three different datasets) is computationally prohibitive and would actively hinder the reproducibility of our experiments, which justifies the practice of limiting hyperparameter searches. Second, our discussion pertains to the general usefulness of the indicators, rather than the fitness of the models — or in other words, it is reasonable to expect of indicators of data complexity that they be robust enough to be deployed with less-than-top-of-the-leaderboard models. Third, going by Table 8, the main driver for the limited correlation between the different groups of model-based indicators is their failure, i.e., the main insight is that we would observe higher correlations if the models never failed, which is not a very realistic standard to expect from NLP systems. While we believe this justifies our approach, it is quite plausible that the exact results as reported here would shift towards higher correlations with human assessments should

the models reach higher accuracy scores.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *Preprint*, arXiv:2311.16867.

Anastasios N. Angelopoulos and Stephen Bates. 2022. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *Preprint*, arXiv:2107.07511.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. pages 1892–1915, Abu Dhabi, United Arab Emirates.

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications. *Preprint*, arXiv:2307.15703.

Robert John Nicholas Baldock, Hartmut Maennel, and Behnam Neyshabur. 2021. Deep learning through the lens of example difficulty. In *Advances in Neural Information Processing Systems*.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. pages 15–21, Online.

Connor Baumler, Anna Sotnikova, and Hal Daumé III. 2023. Which examples should be multiply annotated? active learning when annotators may disagree. pages 10352–10371, Toronto, Canada.

Eyal Beigman and Beata Beigman Klebanov. 2009. Learning with annotation noise. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 280–287, Suntec, Singapore. Association for Computational Linguistics.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: a suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. pages 632–642, Lisbon, Portugal.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. pages 2591–2597, Online.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(1):1513–1589.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. pages 6577–6595, Mexico City, Mexico.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. pages 30–45, Abu Dhabi, United Arab Emirates.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, and 24 others. 2024. OLMo: Accelerating the science of language models. pages 15789–15809, Bangkok, Thailand.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Ben Hachey, Beatrice Alex, and Markus Becker. 2005. Investigating the effects of selective sampling on the annotation task. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 144–151, Ann Arbor, Michigan. Association for Computational Linguistics.

Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*.

Tin Kam Ho and M. Basu. 2002. Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):289–300.

Mengting Hu, Zhen Zhang, Shiwan Zhao, Minlie Huang, and Bingzhe Wu. 2023. Uncertainty in natural language processing: Sources, quantification, and applications. *arXiv preprint arXiv:2306.04459*.

Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506.

Emily Jamison and Iryna Gurevych. 2015. Noise or additional information? leveraging crowdsource annotation item agreement for natural language tasks. pages 291–297, Lisbon, Portugal.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5580–5590, Red Hook, NY, USA. Curran Associates Inc.

Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112. Risk Acceptance and Risk Communication.

John P Lalor, Hao Wu, Tsendsuren Munkhdalai, and Hong Yu. 2018. Understanding deep learning performance through an examination of test set difficulty: A psychometric case study. *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2018:4711–4716.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, and 13 others. 2021. Datasets: A community library for natural language processing. pages 175–184, Online and Punta Cana, Dominican Republic.

Ana C. Lorena, Luís P. F. Garcia, Jens Lehmann, Marcilio C. P. Souto, and Tin Kam Ho. 2019. How complex is your classification problem? a survey on measuring classification complexity. *ACM Comput. Surv.*, 52(5).

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh. 2024. Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. *Preprint*, arXiv:2402.19460.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? pages 9131–9143, Online.

nostalgebraist. 2020. interpreting GPT: the logit lens.

Jennimaria Palomaki, Olivia Rhinehart, and Michael Tseng. 2018. A case for a range of acceptable annotations. In *Proceedings of the 1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper Proceedings of the 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management (SAD 2018 and CrowdBias 2018) co-located the 6th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2018), Zürich, Switzerland, July 5, 2018*, volume 2276 of *CEUR Workshop Proceedings*, pages 19–31. CEUR-WS.org.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. pages 10671–10682, Abu Dhabi, United Arab Emirates.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? pages 507–511, Baltimore, Maryland.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2021. DynaSent: A dynamic benchmark for sentiment analysis. pages 2388–2404, Online.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. pages 784–789, Melbourne, Australia.

Dennis Reidsma and Jean Carletta. 2008. Squibs: Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. 2019. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, and 373 others. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Christopher Schröder, Andreas Niekler, and Martin Potthast. 2022. Revisiting uncertainty-based query strategies for active learning with transformers. pages 2194–2203, Dublin, Ireland.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. pages 9275–9293, Online.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. 2019. An empirical study of example forgetting during deep neural network learning. In *ICLR*.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*.

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

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic Learning in a Random World*. Springer-Verlag.

Daphna Weinshall, Gad Cohen, and Dan Amir. 2018. Curriculum learning by transfer learning: Theory and experiments with deep networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5238–5246. PMLR.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. pages 1112–1122, New Orleans, Louisiana.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. pages 38–45, Online.

Yijun Xiao and William Yang Wang. 2019. Quantifying uncertainties in natural language processing tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7322–7329.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022. A survey of active learning for natural language processing. pages 6166–6190, Abu Dhabi, United Arab Emirates.

Xinlei Zhou, Han Liu, Farhad Pourpanah, Tieyong Zeng, and Xizhao Wang. 2022. A survey on epistemic (model) uncertainty in supervised learning: Recent advances and applications. *Neurocomputing*, 489:449–465.

A Implementation details

Our use of all preexisting research artifacts is consistent with their corresponding licenses. We trust creators of said artifacts to have handled any personally identifying information that the artifacts may contain.

We also provide our code for replication purposes at this link.

A.1 Data

As noted above, we use SNLI (Bowman et al., 2015: retrieved from HuggingFace), MNLI (Williams et al., 2018, retrieved from HuggingFace), ChaosNLI (Nie et al., 2020; retrieved from GitHub) and DynaSent (Potts et al., 2021, retrieved from GitHub, round 1 data). We remove items without public labels from SNLI and MNLI, as well as datapoints with no majority label from DynaSent.

Data shuffling was seeded (with fixed random seeds per runs) for DynaSent experiments so as to guarantee strictly comparable training conditions between soft and hard label experiments.

A.2 Models

All models are implemented with HuggingFace (HF; Wolf et al., 2020; Lhoest et al., 2021). As per default HF implementations, for the 1B pool of models, classifiers rely on the last token in the input; for the <1B model pool, we use the first token. All experiments are supervised full fine-tuning processes using learned linear projections as

classification heads. Models are trained on a V100 NVIDIA GPU, for an individual runtime of $\leq$15 hours for any individual model.

All classifier heads for DynaSent were initialized (with fixed random seeds per run) so as to guarantee strictly comparable training conditions between soft and hard label experiments.

Number of epochs	2
Batch size	16

(a) Hyperparameters, <1B SNLI models

Number of epochs	10
Batch size	16

(b) Hyperparameters, 1B SNLI models

Number of epochs	5
Batch size	1
Gradient accumulation	16
Warmup ratio	0.1

(c) Hyperparameters, all DynaSent models

Number of epochs	5
Batch size	1
Gradient accumulation	16
Warmup ratio	0.1
Learning rate	1e-6

(d) Hyperparameters, all MNLI models

Table 6: Hyperparameters for all models considered

Hyperparameters are listed in Table 6. Any hyperparameter not listed in Table 6 was left to its default value as listed in the HF documentation.

B Supplementary results

B.1 Non-linear relationship of human-based indicators and model-based indicators

To get a better grasp on the magnitude of the difference highlighted in Tables 2 and 3, we can turn to residual analyses. We fit a linear regression, attempting to predict one indicator from another, and measure the proportion of variance that this linear model can explain using a coefficient of determination R^2. Corresponding values are shown in Table 7, with Table 7a focusing on the <1B group and Table 7b the 1B group. In short, R^2 are never above 20%, and often below 10% for reference-free metrics, suggesting that at least 80% of the behavior of our model-based indicators cannot be accounted for with human-based indicators alone. In this case as well, we can observe a difference between

	$\mathbb{H}_{\text{dis}}$	$\mathbb{H}_{\text{ent}}$
$\mathbb{M}_{\text{dis}}$	0.0475	0.0595
$\mathbb{M}_{\text{ent}}$	0.0592	0.0775
$\mathbb{M}_{\text{avg ent}}$	0.1218	0.1521
$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.1015	0.1396
$\mathbb{M}_{\text{CP } \alpha=0.1}$	0.1142	0.1416
$\mathbb{M}_{\text{CP } \alpha=0.2}$	0.0939	0.1055
$\mathbb{M}^{\text{ref}}_{\text{fail}}$	0.1568	0.1592
$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	0.1441	0.1383
$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	0.1802	0.1898
$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	0.1524	0.1576
$\mathbb{M}^{+\text{ref}}_{\text{avg ckpt } p}$	0.1799	0.1924

(a) <1B models

	$\mathbb{H}_{\text{dis}}$	$\mathbb{H}_{\text{ent}}$
$\mathbb{M}_{\text{dis}}$	0.0314	0.0379
$\mathbb{M}_{\text{avg ent}}$	0.0575	0.0790
$\mathbb{M}_{\text{ent}}$	0.0388	0.0477
$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.0536	0.0766
$\mathbb{M}_{\text{CP } \alpha=0.1}$	0.0573	0.0795
$\mathbb{M}_{\text{CP } \alpha=0.2}$	0.0465	0.0616
$\mathbb{M}^{\text{ref}}_{\text{fail}}$	0.1109	0.1223
$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	0.1147	0.1313
$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	0.1186	0.1356
$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	0.1072	0.1209
$\mathbb{M}^{+\text{ref}}_{\text{avg ckpt } p}$	0.1175	0.1347

(b) 1B models

Table 7: Proportion of explained variance (R^2) of linear regressions predicting a model-based indicator from a human-based indicator.

reference-free and reference-dependent indicators: As one would expect, reference-dependent indicators yield quantitatively higher R^2 scores, suggesting they are (marginally) more in line with human indicators. It is worth highlighting that out of all the reference-free indicators, conformal prediction set sizes and average model entropy scores tend to be the most in line with human judgments. This echoes our earlier remarks on the reference-free indicators being partitioned in two sub-groups, and suggests that more elaborate statistical estimators may mitigate some of the discrepancy we observe between human-based and model-based indicators.

B.2 Interaction of model-based indicators and model success

A more formal statement of the argument shown in Figure 1 is that we observe higher correlations when comparing two model-based indicators than when comparing human-based to model-based indicators, although correlations remain smaller than what we observe when comparing indicators within the same group. This can be seen in Table 8 for SNLI, where the values are clearly below what we observe within any subgroup of indicators, but higher than what we summarized in Table 2.

To show that all of our reference-free indicators conflate model-success and failure, we can break down observations depending on whether over 50% of the model pool produces the majority label of the human annotator pools. Recomputing correlations for each subgroup yields systematic negative correlations when the models tend to fail, and systematic positive correlations when the models tend to succeed according to the majority labels, as summarized in Tables 9 to 11 — informally, corre-

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	0.6190	0.5098	0.6053	0.6091	0.5986
$\mathbb{M}_{\text{ent}}$	0.6212	0.5133	0.6083	0.6117	0.6035
$\mathbb{M}_{\text{avg ent}}$	0.5904	0.4973	0.6221	0.5876	0.6428
$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.4602	0.3762	0.4845	0.4585	0.5206
$\mathbb{M}_{\text{CP } \alpha=0.1}$	0.4601	0.3748	0.4881	0.4572	0.5044
$\mathbb{M}_{\text{CP } \alpha=0.2}$	0.3546	0.3170	0.3747	0.3504	0.3623

(a) <1B models

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	0.5154	0.4966	0.5029	0.5035	0.5048
$\mathbb{M}_{\text{ent}}$	0.5168	0.4984	0.5047	0.5073	0.5127
$\mathbb{M}_{\text{avg ent}}$	0.5292	0.5419	0.5468	0.5292	0.5560
$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.4860	0.4958	0.4972	0.4916	0.5286
$\mathbb{M}_{\text{CP } \alpha=0.1}$	0.5216	0.5290	0.5353	0.5241	0.5527
$\mathbb{M}_{\text{CP } \alpha=0.2}$	0.5232	0.5338	0.5437	0.5188	0.5338

(b) 1B models

Table 8: Spearman correlation between reference-dependent and reference-free indicators.

lations form the 'left leg' of the inverted U-shape distributions, and anti-correlation the 'right leg.'

We have already mentioned the surprising anti-correlation of $\mathbb{M}_{\text{CP } \alpha=0.2}$ on DynaSent within the 1B pool in the main body of the text, here we see that this unexpected behavior also impacts its joint distributions. Another case worth highlighting concerns $\mathbb{M}_{\text{CP } \alpha=0.05}$ on MNLI within the 1B pool: correlations and anti-correlations are of remarkably lower magnitudes, which suggests that this specific indicator is not in line with any of the reference-dependent indicators we consider. This is in line with our remarks that the exact setup considered always plays a key role. More broadly, these observations largely confirm our claims in the main body of this article: We find that in most cases, the discrepancy between reference-free and reference-dependent indicators is due to the former conflating model success and model failure.

B.3 Factors shaping model dissensus

We can also leverage the different pools of models to assess how their factors of variation might impact data complexity metrics. In particular, our heterogeneous group of 1B models was defined with respect to different PLMs and training subsets, and therefore we can measure whether pretraining conditions are more impactful than supervised fine-tuning data. In practice, we can measure how likely it is that two predictions for a specific datapoint will match, given that they were made by classifiers trained from the same model or by classifiers trained on the same training subset of SNLI. This can be measured using common language effect sizes derived from a Mann-Whitney U test. Doing so suggests a statistically significant effect from both splits and models ($p < 10^{-44}$) with a very small effect size ($f \approx 51.2\%$) on SNLI. On MNLI, we find a somewhat stronger effect ($f = 53.10\%, p < \epsilon$) when considering classifiers derived from the same PLM; as for the training data, it appears to yield the opposite effect, though with a much higher p-value ($f = 49.77\%, p < 10^{-3}$); i.e., any two different PLMs trained on the same split tend to disagree more than other pair of models. For DynaSent, recall we have no sub-splits to experiment with; however we do find a positive effect when considering classifiers derived from the same PLMs ($f = 53.63\%, p < \epsilon$). In short, there is some evidence that classifiers derived from the same PLM tend to make similar predictions.

Our homogeneous <1B pool also allows us to look into whether responses are more likely to differ for two models with a larger difference in number of parameters. To test this, we can measure the likelihood of the parameter count difference being larger when the predictions differ using U tests. Doing so, we can observe a common language effect size of $f = 45.96\%$ on SNLI, $f = 45.63\%$ on MNLI, and $f = 45.69\%$ on DynaSent. We can likewise observe a similar effect when focusing on our heterogeneous pool: we find a common language effect size of $f = 48.90\%$ for SNLI, $f = 45.63\%$ for MNLI, and $f = 43.72\%$ on DynaSent. In other words, predictions that match tend to come from models with more similar parameter counts.

B.4 Replication of DynaSent experiments without re-balancing

For the sake of exhaustiveness, we include experimental results derived all usable datapoints in DynaSent, i.e., without applying the label re-balancing step we detail in §5. This entails several key differences. In particular, we now consider a classification with imbalanced labels among four possible classes (instead of three evenly split classes as per §5), but have access to more data to fit our classifiers. Also note that by construction Potts et al. (2021) exclude datapoints classified by a majority of annotators as 'mixed' from the test sets, meaning there is a clear distributional shift between train and test.

Comparisons between model-based and human-based indicators are shown in Table 12. The main point to be stressed is that we replicate the core findings stressed in the main body of the text, with two key differences. First, we do not observe an

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	−0.8508	−0.8518	−0.8461	−0.8085	−0.7631
$\mathbb{M}_{\text{ent}}$	−0.7933	−0.7930	−0.7863	−0.7551	−0.7007
$\mathbb{M}_{\text{avg ent}}$	−0.5969	−0.5617	−0.5390	−0.5828	−0.5941
$\mathbb{M}_{\text{CP } \alpha=0.05}$	−0.3958	−0.3876	−0.3874	−0.3736	−0.3552
$\mathbb{M}_{\text{CP } \alpha=0.1}$	−0.4965	−0.4640	−0.4670	−0.4833	−0.4824
$\mathbb{M}_{\text{CP } \alpha=0.2}$	−0.4392	−0.4014	−0.3974	−0.4297	−0.4403

(a) <1B models, datapoints where most models fail

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	0.9507	0.6920	0.9187	0.9213	0.8859
$\mathbb{M}_{\text{ent}}$	0.9489	0.6912	0.9177	0.9201	0.8873
$\mathbb{M}_{\text{avg ent}}$	0.8202	0.5998	0.8760	0.8123	0.9371
$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.5833	0.4026	0.6342	0.5763	0.7053
$\mathbb{M}_{\text{CP } \alpha=0.1}$	0.6237	0.4339	0.6775	0.6156	0.7166
$\mathbb{M}_{\text{CP } \alpha=0.2}$	0.4985	0.4016	0.5317	0.4866	0.5165

(b) <1B models, datapoints where most models succeed

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	−0.7761	−0.7593	−0.7737	−0.7136	−0.6838
$\mathbb{M}_{\text{ent}}$	−0.7131	−0.7075	−0.7174	−0.6539	−0.6140
$\mathbb{M}_{\text{avg ent}}$	−0.5615	−0.5264	−0.5283	−0.5303	−0.5111
$\mathbb{M}_{\text{CP } \alpha=0.05}$	−0.3670	−0.3633	−0.3515	−0.3389	−0.3037
$\mathbb{M}_{\text{CP } \alpha=0.1}$	−0.4761	−0.4565	−0.4575	−0.4453	−0.4182
$\mathbb{M}_{\text{CP } \alpha=0.2}$	−0.6427	−0.5836	−0.5967	−0.6156	−0.6116

(c) 1B models, datapoints where most models fail

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	0.9536	0.8928	0.9159	0.9003	0.8934
$\mathbb{M}_{\text{ent}}$	0.9464	0.8891	0.9107	0.8980	0.8962
$\mathbb{M}_{\text{avg ent}}$	0.8803	0.8955	0.9116	0.8694	0.9315
$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.7748	0.7939	0.7971	0.7759	0.8546
$\mathbb{M}_{\text{CP } \alpha=0.1}$	0.8546	0.8601	0.8816	0.8491	0.9103
$\mathbb{M}_{\text{CP } \alpha=0.2}$	0.8996	0.8982	0.9320	0.8799	0.9166

(d) 1B models, datapoints where most models succeed

Table 9: Spearman correlation on SNLI data between reference-dependent and reference-free indicators of data complexity, broken down by average model success or failure.

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	−0.7676	−0.7625	−0.7812	−0.6911	−0.7380
$\mathbb{M}_{\text{ent}}$	−0.7322	−0.7301	−0.7469	−0.6487	−0.7034
$\mathbb{M}_{\text{avg ent}}$	−0.5241	−0.5130	−0.5148	−0.5189	−0.5044
$\mathbb{M}_{\text{CP } \alpha=0.05}$	−0.3463	−0.3307	−0.3484	−0.3289	−0.3320
$\mathbb{M}_{\text{CP } \alpha=0.1}$	−0.4079	−0.3930	−0.4083	−0.3964	−0.3930
$\mathbb{M}_{\text{CP } \alpha=0.2}$	−0.5070	−0.4949	−0.5013	−0.5041	−0.4915

(a) <1B models, datapoints where most models fail

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	0.9747	0.8189	0.9593	0.9276	0.9345
$\mathbb{M}_{\text{ent}}$	0.9623	0.8053	0.9484	0.9294	0.9281
$\mathbb{M}_{\text{avg ent}}$	0.7828	0.7236	0.8000	0.9118	0.7530
$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.5770	0.5819	0.5857	0.7120	0.5664
$\mathbb{M}_{\text{CP } \alpha=0.1}$	0.6581	0.6282	0.6680	0.7971	0.6466
$\mathbb{M}_{\text{CP } \alpha=0.2}$	0.7559	0.6753	0.7693	0.8888	0.7386

(b) <1B models, datapoints where most models succeed

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	−0.8044	−0.7910	−0.8041	−0.7773	−0.7784
$\mathbb{M}_{\text{ent}}$	−0.7594	−0.7457	−0.7640	−0.7346	−0.7320
$\mathbb{M}_{\text{avg ent}}$	−0.5422	−0.4933	−0.4884	−0.5259	−0.5365
$\mathbb{M}_{\text{CP } \alpha=0.05}$	−0.1415	−0.1186	−0.1614	−0.1246	−0.1106
$\mathbb{M}_{\text{CP } \alpha=0.1}$	−0.3913	−0.3610	−0.4014	−0.3724	−0.3706
$\mathbb{M}_{\text{CP } \alpha=0.2}$	−0.5293	−0.4843	−0.5179	−0.5155	−0.5182

(c) 1B models, datapoints where most models fail

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	0.9767	0.8951	0.9413	0.9447	0.9459
$\mathbb{M}_{\text{ent}}$	0.9635	0.8875	0.9309	0.9337	0.9358
$\mathbb{M}_{\text{avg ent}}$	0.7414	0.7345	0.7863	0.7339	0.7532
$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.0547	0.0540	0.0366	0.0432	0.0330
$\mathbb{M}_{\text{CP } \alpha=0.1}$	0.5058	0.5204	0.5220	0.4995	0.5159
$\mathbb{M}_{\text{CP } \alpha=0.2}$	0.6734	0.6806	0.7091	0.6661	0.6936

(d) 1B models, datapoints where most models succeed

Table 10: Spearman correlation on MNLI data between reference-dependent and reference-free indicators of data complexity, broken down by average model success or failure.

anti-correlation for $\mathbb{M}_{\text{CP } \alpha=0.2}$ within the 1B pool pof models. Second, the use of soft labels does not appear to be beneficial to any indicator derived from the <1B pool of models. Soft labels do remain useful to 1B models, and we still observe that soft labels are not sufficient for reference-free indicators to systematically bridge the gap separating them from reference-dependent indicators.

We further include evidence for the inverted U-shapes of the joint distributions of reference-free and reference-dependent indicators in Table 13. Remarks similar to what we already discussed in §B.2 can be made.

Overall, this supplementary experiment suggests an interesting perspective for future work: Label-imbalance does impact our indicators, as Ho and Basu (2002) suggest, but its exact effects appear to be contingent on the exact pool of models under consideration.

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	−0.7329	−0.6892	−0.7323	−0.6989	−0.6435
$\mathbb{M}_{\text{ent}}$	−0.6539	−0.6259	−0.6519	−0.6287	−0.5509
$\mathbb{M}_{\text{avg ent}}$	−0.2106	−0.2094	−0.1926	−0.1969	−0.2094
$\mathbb{M}_{\text{CP } \alpha=0.05}$	−0.2056	−0.2213	−0.1903	−0.2089	−0.1463
$\mathbb{M}_{\text{CP } \alpha=0.1}$	−0.2449	−0.2595	−0.2248	−0.2454	−0.2067
$\mathbb{M}_{\text{CP } \alpha=0.2}$	−0.3730	−0.3801	−0.3465	−0.3658	−0.3743

(a) <1B models, soft labels, datapoints where most models fail

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	0.9777	0.9115	0.9666	0.9452	0.7375
$\mathbb{M}_{\text{ent}}$	0.9502	0.9039	0.9406	0.9231	0.7442
$\mathbb{M}_{\text{avg ent}}$	0.4917	0.5037	0.5107	0.4592	0.9306
$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.5433	0.5904	0.5592	0.5280	0.8927
$\mathbb{M}_{\text{CP } \alpha=0.1}$	0.6237	0.6439	0.6423	0.6005	0.9367
$\mathbb{M}_{\text{CP } \alpha=0.2}$	0.7351	0.7151	0.7542	0.7032	0.8975

(b) <1B models, soft labels, datapoints where most models succeed

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	−0.7420	−0.7208	−0.7437	−0.7178	−0.6310
$\mathbb{M}_{\text{ent}}$	−0.6792	−0.6705	−0.6804	−0.6559	−0.5609
$\mathbb{M}_{\text{avg ent}}$	−0.3213	−0.3350	−0.3053	−0.2978	−0.2991
$\mathbb{M}_{\text{CP } \alpha=0.05}$	−0.3062	−0.3126	−0.2906	−0.2945	−0.2432
$\mathbb{M}_{\text{CP } \alpha=0.1}$	−0.3848	−0.3805	−0.3664	−0.3747	−0.3465
$\mathbb{M}_{\text{CP } \alpha=0.2}$	−0.4875	−0.4821	−0.4647	−0.4793	−0.4771

(c) <1B models, hard labels, datapoints where most models fail

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	0.9832	0.8659	0.9580	0.9454	0.6924
$\mathbb{M}_{\text{ent}}$	0.9692	0.8547	0.9448	0.9337	0.7022
$\mathbb{M}_{\text{avg ent}}$	0.4628	0.4438	0.4835	0.4602	0.9265
$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.4725	0.4379	0.4943	0.4837	0.8639
$\mathbb{M}_{\text{CP } \alpha=0.1}$	0.6116	0.5831	0.6318	0.6129	0.9209
$\mathbb{M}_{\text{CP } \alpha=0.2}$	0.7380	0.7528	0.7509	0.7194	0.8124

(d) <1B models, hard labels, datapoints where most models succeed

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	−0.8089	−0.7620	−0.7934	−0.7572	−0.7499
$\mathbb{M}_{\text{ent}}$	−0.7167	−0.6897	−0.7127	−0.6785	−0.6616
$\mathbb{M}_{\text{avg ent}}$	−0.4201	−0.3365	−0.3031	−0.4281	−0.4268
$\mathbb{M}_{\text{CP } \alpha=0.05}$	−0.4268	−0.3579	−0.3555	−0.4238	−0.4122
$\mathbb{M}_{\text{CP } \alpha=0.1}$	−0.4366	−0.3594	−0.3407	−0.4470	−0.4430
$\mathbb{M}_{\text{CP } \alpha=0.2}$	0.3924	0.3417	0.2990	0.4006	0.3996

(e) 1B models, soft labels, datapoints where most models fail

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	0.8681	0.7185	0.8442	0.8292	0.7854
$\mathbb{M}_{\text{ent}}$	0.8666	0.7170	0.8425	0.8279	0.7849
$\mathbb{M}_{\text{avg ent}}$	0.7937	0.7843	0.8146	0.7943	0.9399
$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.7935	0.7715	0.8166	0.7923	0.9239
$\mathbb{M}_{\text{CP } \alpha=0.1}$	0.8056	0.7311	0.8328	0.7980	0.8688
$\mathbb{M}_{\text{CP } \alpha=0.2}$	−0.3794	−0.3613	−0.3880	−0.3715	−0.4033

(f) 1B models, soft labels, datapoints where most models succeed

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	−0.8121	−0.7822	−0.8183	−0.7466	−0.5481
$\mathbb{M}_{\text{ent}}$	−0.7233	−0.7044	−0.7384	−0.6693	−0.4481
$\mathbb{M}_{\text{avg ent}}$	−0.4584	−0.4193	−0.4118	−0.4364	−0.2267
$\mathbb{M}_{\text{CP } \alpha=0.05}$	−0.4693	−0.4369	−0.4026	−0.4593	−0.3371
$\mathbb{M}_{\text{CP } \alpha=0.1}$	−0.4678	−0.3731	−0.3376	−0.4844	−0.5469
$\mathbb{M}_{\text{CP } \alpha=0.2}$	0.2795	0.2400	0.2432	0.2634	0.0669

(g) 1B models, hard labels, datapoints where most models fail

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	0.8522	0.7383	0.8303	0.8231	0.8084
$\mathbb{M}_{\text{ent}}$	0.8510	0.7372	0.8289	0.8220	0.8073
$\mathbb{M}_{\text{avg ent}}$	0.7860	0.6693	0.8113	0.7770	0.8519
$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.7091	0.5812	0.7301	0.7074	0.7697
$\mathbb{M}_{\text{CP } \alpha=0.1}$	0.7794	0.6588	0.8057	0.7729	0.8225
$\mathbb{M}_{\text{CP } \alpha=0.2}$	−0.5255	−0.4794	−0.5538	−0.5172	−0.5332

(h) 1B models, hard labels, datapoints where most models succeed

Table 11: Spearman correlation on DynaSent re-balanced data between reference-dependent and reference-free indicators of data complexity, broken down by average model success or failure.

	<1B pool		1B pool	
	soft	hard	soft	hard
$\mathbb{M}_{\text{dis}}$	0.0695	0.0699	0.1380	0.1332
$\mathbb{M}_{\text{ent}}$	0.0712	0.0801	0.1449	0.1368
$\mathbb{M}_{\text{avg ent}}$	0.1415	0.1301	0.2081	0.1320
$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.1278	0.1265	0.1881	0.1533
$\mathbb{M}_{\text{CP } \alpha=0.1}$	0.1292	0.1251	0.1952	0.1557
$\mathbb{M}_{\text{CP } \alpha=0.2}$	0.1247	0.1087	0.0294	0.0709
$\mathbb{M}^{\text{ref}}_{\text{fail}}$	0.1399	0.1426	0.1709	0.1725
$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	0.1239	0.1205	0.2013	0.1843
$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	0.1411	0.1438	0.1990	0.2006
$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	0.1344	0.1370	0.1740	0.1749
$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$	0.1627	0.1605	0.2214	0.1885

Table 12: Spearman's ρ between model-based indicators vs. human-based (entropy) indicator on DynaSent.

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	−0.7786	−0.7574	−0.7875	−0.7709	−0.7041
$\mathbb{M}_{\text{ent}}$	−0.7339	−0.7202	−0.7402	−0.7327	−0.6448
$\mathbb{M}_{\text{avg ent}}$	−0.5347	−0.5086	−0.5059	−0.5460	−0.5116
$\mathbb{M}_{\text{CP } \alpha=0.05}$	−0.4914	−0.4610	−0.4673	−0.5049	−0.4600
$\mathbb{M}_{\text{CP } \alpha=0.1}$	−0.4963	−0.4701	−0.4672	−0.5090	−0.4741
$\mathbb{M}_{\text{CP } \alpha=0.2}$	−0.5433	−0.5288	−0.5146	−0.5535	−0.5470

(a) <1B models, soft labels, datapoints where most models fail

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	0.9449	0.9100	0.9416	0.9312	0.8865
$\mathbb{M}_{\text{ent}}$	0.9366	0.9106	0.9342	0.9246	0.8953
$\mathbb{M}_{\text{avg ent}}$	0.7411	0.7966	0.7566	0.7283	0.9544
$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.7115	0.7870	0.7242	0.7040	0.9270
$\mathbb{M}_{\text{CP } \alpha=0.1}$	0.7390	0.7973	0.7539	0.7286	0.9453
$\mathbb{M}_{\text{CP } \alpha=0.2}$	0.7873	0.8189	0.8038	0.7734	0.9493

(b) <1B models, soft labels, datapoints where most models succeed

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	−0.7997	−0.7526	−0.7957	−0.7885	−0.7217
$\mathbb{M}_{\text{ent}}$	−0.7568	−0.7154	−0.7514	−0.7528	−0.6680
$\mathbb{M}_{\text{avg ent}}$	−0.5807	−0.5388	−0.5483	−0.5938	−0.5507
$\mathbb{M}_{\text{CP } \alpha=0.05}$	−0.5090	−0.4579	−0.4805	−0.5251	−0.4749
$\mathbb{M}_{\text{CP } \alpha=0.1}$	−0.5107	−0.4721	−0.4826	−0.5213	−0.4944
$\mathbb{M}_{\text{CP } \alpha=0.2}$	−0.5565	−0.5299	−0.5244	−0.5622	−0.5710

(c) <1B models, hard labels, datapoints where most models fail

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	0.9432	0.8761	0.9376	0.9227	0.8878
$\mathbb{M}_{\text{ent}}$	0.9361	0.8817	0.9321	0.9181	0.8963
$\mathbb{M}_{\text{avg ent}}$	0.7925	0.8090	0.8104	0.7838	0.9691
$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.7427	0.7817	0.7596	0.7381	0.9363
$\mathbb{M}_{\text{CP } \alpha=0.1}$	0.7608	0.7840	0.7788	0.7509	0.9467
$\mathbb{M}_{\text{CP } \alpha=0.2}$	0.8146	0.8096	0.8327	0.7989	0.9506

(d) <1B models, hard labels, datapoints where most models succeed

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	−0.7581	−0.6938	−0.7591	−0.6973	−0.4618
$\mathbb{M}_{\text{ent}}$	−0.6731	−0.6259	−0.6856	−0.6202	−0.3616
$\mathbb{M}_{\text{avg ent}}$	−0.4073	−0.3155	−0.3829	−0.4030	−0.1180
$\mathbb{M}_{\text{CP } \alpha=0.05}$	−0.3711	−0.2924	−0.3530	−0.3667	−0.1134
$\mathbb{M}_{\text{CP } \alpha=0.1}$	−0.5794	−0.4849	−0.4691	−0.5806	−0.5778
$\mathbb{M}_{\text{CP } \alpha=0.2}$	−0.0654	−0.0862	0.0103	−0.0663	−0.3327

(e) 1B models, soft labels, datapoints where most models fail

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	0.8672	0.6965	0.8362	0.8318	0.7920
$\mathbb{M}_{\text{ent}}$	0.8654	0.6956	0.8342	0.8304	0.7923
$\mathbb{M}_{\text{avg ent}}$	0.7803	0.7605	0.8099	0.7799	0.9346
$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.7380	0.7007	0.7687	0.7377	0.8975
$\mathbb{M}_{\text{CP } \alpha=0.1}$	0.7746	0.7023	0.8199	0.7700	0.8823
$\mathbb{M}_{\text{CP } \alpha=0.2}$	0.1686	0.0856	0.1970	0.1647	0.1396

(f) 1B models, soft labels, datapoints where most models succeed

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	−0.7523	−0.7232	−0.7450	−0.7109	−0.6883
$\mathbb{M}_{\text{ent}}$	−0.6754	−0.6490	−0.6772	−0.6430	−0.6104
$\mathbb{M}_{\text{avg ent}}$	−0.4985	−0.4198	−0.3943	−0.4953	−0.4814
$\mathbb{M}_{\text{CP } \alpha=0.05}$	−0.4228	−0.3493	−0.3663	−0.4268	−0.4001
$\mathbb{M}_{\text{CP } \alpha=0.1}$	−0.4657	−0.3925	−0.3999	−0.4716	−0.4440
$\mathbb{M}_{\text{CP } \alpha=0.2}$	−0.3062	−0.2508	−0.2484	−0.3039	−0.3090

(g) 1B models, hard labels, datapoints where most models fail

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	0.8574	0.7466	0.8239	0.8221	0.8170
$\mathbb{M}_{\text{ent}}$	0.8550	0.7449	0.8212	0.8196	0.8148
$\mathbb{M}_{\text{avg ent}}$	0.7639	0.6358	0.7988	0.7506	0.8091
$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.7023	0.6103	0.7234	0.7026	0.7357
$\mathbb{M}_{\text{CP } \alpha=0.1}$	0.7482	0.6241	0.7769	0.7438	0.7884
$\mathbb{M}_{\text{CP } \alpha=0.2}$	0.4830	0.4154	0.4977	0.4782	0.4898

(h) 1B models, hard labels, datapoints where most models succeed

Table 13: Spearman correlation on DynaSent data (without label re-balancing) between reference-dependent and reference-free indicators of data complexity, broken down by average model success or failure.