How Much Consistency Is Your Accuracy Worth?

Jacob K. Johnson and **Ana Marasović**
Kahlert School of Computing
University of Utah
{jacob.k.johnson,ana.marasovic}@utah.edu

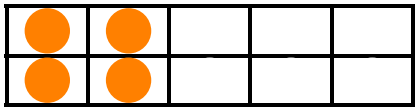
Abstract

Contrast set consistency is a robustness measurement that evaluates the rate at which a model correctly responds to all instances in a bundle of minimally different examples relying on the same knowledge. To draw additional insights, we propose to complement consistency with *relative consistency* — the probability that an equally accurate model would surpass the consistency of the proposed model, given a distribution over possible consistencies. Models with 100% relative consistency have reached a consistency peak for their accuracy. We reflect on prior work that reports consistency in contrast sets and observe that relative consistency can alter the assessment of a model's consistency compared to another. We anticipate that our proposed measurement and insights will influence future studies aiming to promote consistent behavior in models.

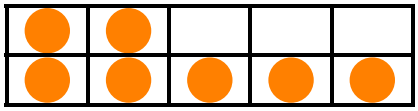
1 Introduction

Annotators introduce data shortcuts that allow models to solve tasks in unintended ways (Gururangan et al., 2018). In response, it has been proposed to measure whether a model correctly responds to a bundle (or a *contrast set*) of slightly modified instances that rely on the same knowledge (Gardner et al., 2020; Kaushik et al., 2020). The rate at which a model accomplishes this is termed *consistency*. We propose an additional measurement — *relative consistency* — that facilitates discussion about achievable consistency scores, enabling a more nuanced comparison.

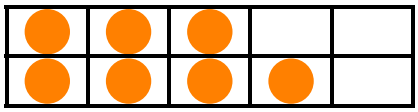
To demonstrate why this is desired, consider situations that are illustrated in Table 1. Both 1a–1b correctly solve two bundles, i.e., have the same consistency. 1b solves three additional instances but in a way that does not promote consistency; 1c shows that a higher consistency can be gained with the same accuracy. In contrast, although 1a is less accurate, everything it handled was done consistently, and higher consistency cannot be achieved with the same accuracy. This analysis sheds light on an upside of 1a and a limitation of 1b that might go unnoticed if we solely compare accuracy/consistency. Let us turn to examples 1d. Although it represents a model with an improved consistency relative to 1a, we could have achieved better consistency for the same accuracy (see 1e).[1]



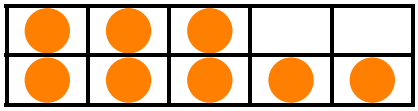
(a) Accuracy=4/10, Consistency=2/5, RelConsistency=100%



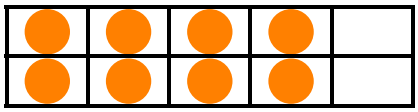
(b) Accuracy=7/10, Consistency=2/5, RelConsistency=66.7%



(c) Accuracy=7/10, Consistency=3/5, RelConsistency=100%



(d) Accuracy=8/10, Consistency=3/5, RelConsistency=88.9%



(e) Accuracy=8/10, Consistency=4/5, RelConsistency=100%

Table 1: Tables depict a dataset of 10 examples, where each column showcases a bundle of an original instance paired with its perturbed version. ● denotes that the instance is correctly predicted by a model. The relative consistency is the measurement we propose to complement the standard consistency.

Relative consistency (§2) measures whether the consistency of our model would likely be outper-

[1] Because this is a toy example, relative consistency is high, though not perfect, even in less-than-ideal cases 1b and 1d.

Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pages 250–260

formed by an equally accurate model, relative to the distribution of possible consistencies; see Eq. (5). Specifically, it is the probability that our model's consistency is (in most cases) higher or equal to the consistency scores that are achievable with the same accuracy. If relative consistency is 100% then our model is the most consistent it can be given its accuracy, as a more consistent, equally accurate model exists only with near-zero probability. In practice, the goal should be to increase the "standard consistency" while also achieving 100% relative consistency.

In light of this additional consistency metric, in §4 we revisit the findings of three publications that report consistency as a metric for their evaluations and point out some additional conclusions we might draw from these reported consistencies. Our code is available at https://github.com/jacobkj314/relative-consistency.

2 Relative Consistency

We first introduce background terminology (§2.1), then derive elements we need for defining relative consistency: (i) achievable consistency scores for a given accuracy (§2.2) and (ii) a distribution over achievable consistency scores (2.3).

2.1 Background

A *contrast set* or *bundle* is a set of minimally different instances that might admit different answers, thus testing a model across/near its decision boundary.[2] For example, these two HotpotQA instances (Yang et al., 2018) represent a contrast set:

- Q: Is the Marsilea or the Brabejum the genus of **more** individual species of plants? A: Marsilea
- Q: Is the Marsilea or the Brabejum the genus of **less** individual species of plants? A: Brabejum

The model is required to answer both of them correctly to be considered consistent in that bundle. Evaluation with contrast sets makes it harder for simple and inadequate models to perform highly (e.g, a model that has just learned a spurious correlation between the word "Marsilea" and "more"). Related studies construct bundles of paraphrases that have the same, not contrastive, labels (Elazar et al., 2021).

[2]Sometimes "contrast set" is used to refer to contrastive instances only (without the original ones).

The term *consistency* is overloaded in NLP and refers to different concepts (Li et al., 2019; Jang et al., 2022; Wang et al., 2023). In this work, we study *contrast set consistency* defined as the proportion of bundles where a model accurately labels every instance in a bundle:

$$\text{consistency} = \frac{|B \in \mathcal{B} : \forall x \in B, y_p(x) = y(x)|}{|\mathcal{B}|}, \tag{1}$$

where $\mathcal{B}$ is a set of all bundles of related instances in a given dataset, x is an example, $y_p(x)$ is the predicted label for x, and $y(x)$ is its gold label.

2.2 Achievable Consistency Scores

Consider a contrastive test set formed from n original instances, plus a contrastive instance derived from each original instance by varying along some pertinent dimension. There are $2n+1$ possible accuracies a that a model could achieve on this test set, namely $A = \{0, 1, \ldots, 2n-1, 2n\}$.[3] Similarly, there are $n+1$ possible consistencies c that a model could achieve, namely $C = \{0, 1, \ldots, n-1, n\}$.

Furthermore, for a given accuracy $a \in A$, only a subset $C_a \subseteq C$ of consistencies is achievable. Trivially, for $a = 0$, $C_a = \{0\}$ (because a model cannot consistently respond to a bundle without correctly responding to at least the instances within that bundle) and for $a = 2n$, $C_a = \{n\}$ (because a model that correctly responds to all instances has also consistently responded to all the bundles those instances comprise). C_a can then be defined in terms of n and a:

$$C_a = \{c \in C : c^{(a)}_{min} \leq c \leq c^{(a)}_{max}\} \tag{2}$$

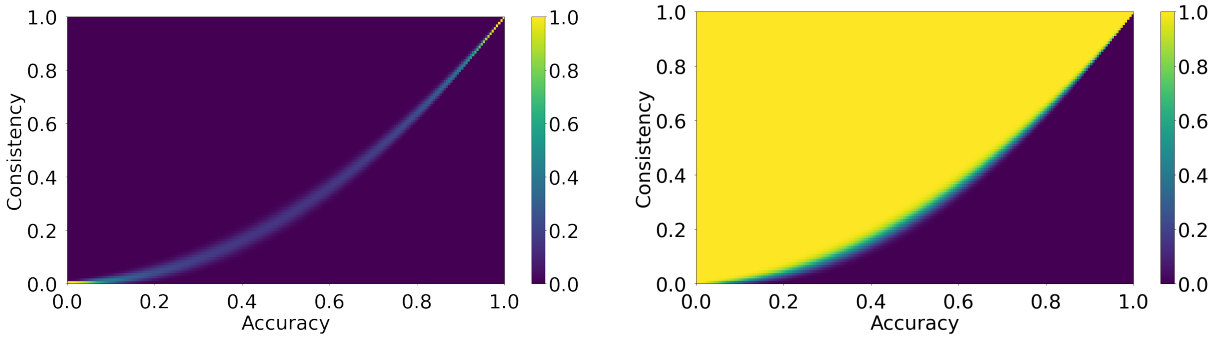
where $c^{(a)}_{min}$ and $c^{(a)}_{max}$ are defined as:

$$c^{(a)}_{min} = \begin{cases} 0 & \text{if } a \leq n \\ a - n & \text{if } a > n \end{cases} \tag{3}$$

$$c^{(a)}_{max} = \left\lfloor \frac{a}{2} \right\rfloor \tag{4}$$

Intuitively, if $a \leq n$ then it is possible that all bundles have one of their constituent instances incorrectly answered, in which case, $c^{(a)}_{min} = 0$. However, if $a > n$, then at least $a - n > 0$ of bundles must be fully correctly answered. Indeed, for a bundle to be inconsistent at least one item

[3]While accuracy is typically denoted as a proportion of correct instances, reporting absolute numbers simplifies our notation. It is easy to translate a quantity a to a corresponding proportion α via the identity $a = 2n\alpha$, while a consistency quantity c relates to the consistency proportion γ via $c = n\gamma$.



(a) Distributions of consistency scores.

(b) Relative consistency scores.

Figure 1: On the left is a heatmap of distributions of consistency at each accuracy for 100 bundles of 2 instances: each vertical slice corresponds to a separate distribution of different consistencies. Fig. 2 (Appendix) shows the $\log_{10}$ of this plot that better highlights the long tails of these distributions. On the right are relative consistency scores given a model's accuracy and consistency, i.e., the CDF of the figure on the left. Note that for a different number of bundles, these plots would look slightly different.

must be incorrectly answered, so for a given a, the number of incorrect items is $2n - a$. Thus, at most $2n - a$ bundles can be inconsistent, and $c^{(a)}_{min} = n - (2n - a) = n - 2n + a = a - n$.

The definition of $c^{(a)}_{max}$ follows from the observation that a maximally consistent model will consistently respond to the maximum number of bundles for which it is possible that both instances are correctly answered, and that equals $\lfloor \frac{a}{2} \rfloor$.

2.3 Distribution of Achievable Consistencies

Given an accuracy a, we construct a distribution of achievable consistencies $c \in C_a$ with:

$$\mathbb{P}(c|a) = \frac{m(c,a)}{M(a)} \tag{5}$$

where $M(a)$ is the number of ways a model can achieve accuracy a and is given by:

$$M(a) = \binom{2n}{a} \tag{6}$$

because there are $2n$ total instances, of which any a might be the ones to which a model correctly responds.[4] $m(c,a)$ represents the number of ways a model can achieve accuracy a and consistency c, and is given by:

$$m(c,a) = \binom{n}{c}\binom{n-c}{a-2c}2^{a-2c} \tag{7}$$

where:

- $\binom{n}{c}$ corresponds to the number of ways that c consistent bundles can be selected from n,
- $\binom{n-c}{a-2c}$ corresponds to the number of ways the remaining $a - 2c$ accurate instances can be distributed across the remaining $n - c$ bundles, giving each selected bundle only one correct instance (to avoid creating an additional consistent bundle),
- 2^{a-2c} represents the number of ways that these partially correct bundles could have either instance correct.

Using this, we can calculate $m(c,a)$ and $M(a)$ across all values of c and a for reasonable sizes of n. These distributions can be extended for bundle sizes above 2; see formulas in Appendix B. Figure 1a shows the distributions of consistency scores for a dataset with 100 bundles of 2 instances.

Note that this distribution is not uniform for different consistencies at a given accuracy. There will be some consistencies that have more ways to be achieved for a given accuracy. This is why the formula $m(c,a)$ is crucial to the computation of relative consistency that comes next.

This formulation assumes that all instances are equally difficult which is known to not be the case in practice (Swayamdipta et al., 2020). It also disregards any inductive biases of models/datasets that could skew the distribution.

Relative Consistency We measure the tendency to be consistent exhibited by a model that achieved accuracy a and consistency c on a contrastive set by computing the cumulative probability distribution over achievable consistencies in C_a up to c:

[4] It is possible to consider consistency to be the more underlying property of a model's behavior and compute a distribution over possible accuracies in the range $[2c, 2n - n + c]$. The corresponding accuracy by consistency distributions could then be computed given the above-defined consistency by accuracy distributions.

$$\mathrm{RC}(c, a) = \sum_{\substack{c_i \in C_a \\ c_i \leq c}} \mathbb{P}(c_i|a) \tag{8}$$

Intuitively, $\mathrm{RC}(c, a)$ indicates how likely the model's consistency is to outperform an equally accurate model relative to the distribution of achievable consistencies defined in (5). This allows us to quantify whether model consistency is below, at, or above chance, given its accuracy. In a good case, RC is high, meaning that it is unlikely that an equally accurate model will have higher consistency. Alternatively, if RC is low, then it is likely that an equally accurate model will have higher consistency (which is unwanted).

Although other measurements which contextualize consistency scores within a particular accuracy can be constructed — such as simply scaling the consistency between $c^{(a)}_{min}$ and $c^{(a)}_{max}$, or reporting the fraction of fully consistent of those that are at least partly correct — these approaches lack the probabilistic interpretation underlying RC. §3–4 highlight circumstances in which this probabilistic interpretation is useful, and Appendix C compares the score distributions obtained via these measurements to the score distributions obtained via RC.

3 Analysis with Simulated Contrastive Set

Suppose you evaluate a model on a contrastive test set containing 100 bundles of 2 instances. The distribution of consistencies for this dataset is shown in Figure 1a, with the CDF of that distribution (corresponding to the RC score) in Figure 1b.

Note that the highest-density region of the distribution moves upward as accuracy increases, and takes up only a very thin band. This means that, for a given accuracy, there is generally little room for improvement in consistency. This can be useful when discussing results: if a particular training approach yields a 5% improvement in consistency for an equally accurate model, that represents a substantial change in how the model tends to respond to inputs.

It can still happen that improving accuracy and consistency decreases relative consistency. As an example, consider comparing a model M_1, which achieves $a = 130, c = 45$ (65% accuracy, 45% consistency) against a model M_2 with $a = 150, c = 55$ (75% accuracy, 55% consistency). Clearly, model M_2 is more desirable for practical uses, if we are just comparing one model to another, but if we are comparing two different training approaches, and want to know which induces a stronger tendency for consistent responses, then we would be interested to know that M_1 has $\mathrm{RC} = 93.0\%$, while M_2 has $\mathrm{RC} = 37.1\%$. This insight, that one model is below chance consistency, while another is well above, is made possible by the probabilistic interpretation of RC.

Dataset	#Bundles	Acc	Cons	RC
UD Parsing	150	55.3	17.3	∼0.0
PERSPECTRUM	217	88.0	78.8	97.6
ROPES	974	40.1	17.6	97.8
MC-TACO	646	26.0	8.0	95.2

Table 2: Relative consistency scores computed for results reported in Gardner et al. (2020). In the 3rd column, we report the average of "Original Test" (original only) and "Contrast" (contrastive only) columns in their Table 2. That is the accuracy, a, we use in calculations in §2. Models with similar consistency (UD Parsing and ROPES) have different tendencies to respond consistently as revealed by their RC scores.

4 Meta-Analysis of Prior Work

In this section, we discuss results reported by prior works that conduct evaluation with contrast sets under the light of relative consistency.

4.1 Gardner et al. (2020)

They construct contrast sets for several common test sets by modifying a sample of the test set instances. They train a biaffine parser (Dozat and Manning, 2017) with ELMo embeddings (Peters et al., 2018) for UD parsing (Zeldes, 2017, Silveira et al., 2014, Basili et al., 2015, Ahrenberg, 2007), and RoBERTa (Liu et al., 2019) for reading comprehension tasks: ROPES (Lin et al., 2019), and MC-TACO (Zhou et al., 2019) and stance prediction: PERSPECTRUM (Chen et al., 2019). Table 2 shows the accuracy and consistency of these models for four of their contrast sets.[5] In the rightmost column, we report the relative consistency scores that we introduce.

Analysis We observe that the UD parsing and ROPES models have a similar consistency score

[5]We exclude contrast sets that do not have the bundle size of 2. They report the accuracy of the original instances and contrastive instances separately, so to obtain the accuracy in the contrast set (that we need to calculate RC) we average those. In doing so, we assume that the accuracy of the full original test set is similar to the accuracy of the sample of original test set instances.

Loss	Accuracy	Consistency	RC
MLE	65.7	52.1	100.0
↳ +UL	68.3	55.6	100.0
↳ +CE	76.6	64.7	100.0

Table 3: A comparison of relative consistency scores computed from results report in Dua et al. (2021) (in "Dev EM" and "Dev C" columns in their Table 3). The number of bundles is 844. The unlikelihood (UL) and contrastive estimation (CE) objectives improved the accuracy and consistency over MLE, *without decreasing relative consistency*. This is how consistency should be improved in this case.

(17.3 and 17.6). However, the UD parsing model's consistency has a near-zero chance to outperform an equally accurate model. On the other hand, the ROPES model is quite likely to do so.

Additionally, relative consistency shows that models with low consistency could nonetheless have a large tendency to respond to bundles consistently.[6] We see this with the results for MC-TACO, which, despite only achieving 8.0% consistency, is more consistent than an equally accurate model in 95.2% of cases. Intuitively, this means that the above chance model has at least generalized well within the few cases to which it correctly responds.

4.2 Dua et al. (2021)

They investigate whether training approaches that consider a full bundle of related instances together, instead of their constituent instances separately, improve consistency. Table 3 shows their report results obtained with T5 (Raffel et al., 2020) and the relative consistency scores we compute from their results, on the contrastive version of ROPES — a reading comprehension dataset for evaluating a model's ability to reason about "effects of the relationships in the background passage in the context of the situation".

Analysis We observe that the baseline model trained with the maximum likelihood estimation (MLE) is already at ceiling performance in terms of its tendency to produce consistent responses (i.e., its RC scores). Combining contrastive estimation (CE; Smith and Eisner, 2005), or unlikelihood training (UL; Welleck et al., 2020), with MLE not only improves the accuracy and consistency but also does so in a way that does not lower the relative consistency, which is desired. This emphasizes the effectiveness of these objectives.

[6]Note that high relative consistency does not guarantee that such a model will continue to respond to bundles consistently with improved accuracy.

4.3 Ravichander et al. (2022)

They introduce CondaQA, a contrastive dataset for studying reading comprehension models' effectiveness in reasoning about the implications of negation expressed in a given text. Each CondaQA instance comes with three minimally varied versions: one paraphrases the negation, another modifies what is negated (scope), and the last removes the negation. Ravichander et al. (2022) use UnifiedQA-v2 (Khashabi et al., 2022) as a backbone model. We explore the factors that might influence the consistency of the large and 3B versions of this model:

- The training objective: MLE, CE, or combined λ_1MLE+λ_2CE.
- The choice of hyperparemeters λ_1 and λ_2 (with UnifiedQA-large).

Table 4 shows accuracy, consistency, and relative consistency we obtain for bundles where the original instance is paired with its: (i) *scope*-edited version, and (ii) *affirmative* version (without negation). In Table 5 (Appendix), we also include the results with paraphrase-edits.

Analysis An increase in consistency does not necessarily indicate a heightened tendency to consistently respond to bundles (unless the accuracy stays the same). Compare CE with 1MLE+1CE (double underlined, in the upper part of the table). In this case, by training with MLE and CE, affirmative consistency has gone up slightly, however, the model's chance of outperforming an equally accurate model dropped down from 26% to 19%. This is an example of a suboptimal way of improving consistency, and MLE+CE is not necessarily superior to the standalone CE in this case. A similar, but less pronounced, situation occurs when comparing MLE against .33MLE+1CE for scope consistency in the bottom part of the table (italicized).

Conversely, even if standard consistency has not improved, a model's tendency to consistently respond to bundles may have. For example, compare MLE with 1MLE+1CE for scope consistency in the upper part of the table (wavy underlined). In this case, scope accuracy lowered slightly but absolute scope consistency remained the same, leading to a large improvement in Scope-RC. This may suggest that additional CE loss resulted in the model unlearning a few individual instances without unlearn-

Size	Loss	Scope-Acc	Aff-Acc	Scope-Cons	Aff-Cons	Scope-RC	Aff-RC
Large	MLE	66.84	67.09	42.86	42.35	17.10	10.06
	CE	64.80	66.84	40.31	43.37	20.10	26.64
	λ_1MLE + λ_2CE						
	↳ $\lambda_1, \lambda_2 = 1.0, 1.0$	66.33	68.11	42.86	44.39	30.43	19.37
	↳ $\lambda_1, \lambda_2 = 0.33, 1.0$	66.58	68.37	43.37	44.90	42.44	16.01
3B	MLE	74.23	76.79	*56.12*	60.71	*80.76*	88.68
	CE	74.23	77.55	56.12	61.73	80.76	92.03
	λ_1MLE + λ_2CE						
	↳ $\lambda_1, \lambda_2 = 0.33, 1.0$	74.23	77.04	56.12	60.71	80.76	88.68
	↳ $\lambda_1, \lambda_2 = 0.33, 1.0$	76.02	78.57	*58.67*	63.78	*79.32*	98.60

Table 4: Results of UnifiedQA-v2 (Khashabi et al., 2022) on the CondaQA contrastive dataset, with the expectation that including the Contrastive Estimation (CE) objective would improve consistency, as in Dua et al. (2021). RC scores are reported here only for some of the edit dimensions in CondaQA; see Table 5 for the rest.

ing any complete bundles it had learned. Similarly, 0.33MLE+1CE scope consistency in the upper part of the table (underlined once) increased slightly but the scope relative consistency has increased notably. If we compared only consistency we would conclude that the choice of hyperparameters λ_1, λ_2 is not vital, where actually they can affect model consistency behavior as shown by relative consistency.

5 Conclusion

We introduce relative consistency, which complements standard contrast consistency by allowing an accuracy and consistency score pair to be examined to determine whether a higher consistency was possible with that accuracy. This facilitates the comparison of consistencies achieved by models that achieved different levels of accuracy. We show that relative consistency enriches conclusions we make about whether a model is more consistent than another, and occasionally even leads us to different takeaways.

6 Limitations

This mathematical model is based on a simplified version of contrastive datasets. Contrastive datasets may have more than two edits for each original instance, which will result in a different distribution. Although we provide formulas for distributions of arbitrary bundle size in Appendix B, these distributions are less intuitive, more expensive to compute, and additionally have the drawback that, if a model achieves high pairwise RC on two of the elements of the bundle, it is likely to achieve high bundle RC, even if the other elements of the test set do not achieve high pairwise RC. In general, we recommend formulating questions of consistency in terms of bundles with one instance exhibiting a feature and the other instance lacking that feature. Moreover, contrastive datasets may include extra data that is not contrastive; e.g., CondaQA has a small number of bundles with a single instance because other instances in the bundle were filtered because they did not pass quality checks.

In §2.3, we state the drawbacks of the distribution (5). Namely, we do not consider that the distribution might be skewed due to the varying example difficulty and other inherent properties of datasets and models.

7 Acknowledgements

We thank anonymous reviewers for their thoughtful and constructive comments, members of the UtahNLP group for helpful feedback, and Petar Bakić for proofreading our formulas.

References

Lars Ahrenberg. 2007. LinES: An English-Swedish parallel treebank. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*, pages 270–273, Tartu, Estonia. University of Tartu, Estonia.

Roberto Basili, Cristina Bosco, Rodolfo Delmonte, Alessandro Moschitti, and Maria Simi. 2015. *Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project*, volume 589.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things

from a different angle:discovering diverse perspectives about claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 542–557, Minneapolis, Minnesota. Association for Computational Linguistics.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing.

Dheeru Dua, Pradeep Dasigi, Sameer Singh, and Matt Gardner. 2021. Learning with instance bundles for reading comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7347–7357, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. 2022. BECEL: Benchmark for consistency evaluation of language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3680–3696, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. 2020. Learning the difference that makes A difference with counterfactually-augmented data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. 2022. Unifiedqa-v2: Stronger generalization via broader cross-format training.

Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. 2019. A logic-driven framework for consistency of neural models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3924–3935, Hong Kong, China. Association for Computational Linguistics.

Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning over paragraph effects in situations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 58–62, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Abhilasha Ravichander, Matt Gardner, and Ana Marasovic. 2022. CONDAQA: A contrastive reading comprehension dataset for reasoning about negation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8729–8755, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Noah A. Smith and Jason Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 354–362, Ann Arbor, Michigan. Association for Computational Linguistics.

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. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Amir Zeldes. 2017. The gum corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51:581–612.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.

A Numerical Stability of Relative Consistency

To avoid numerical instability, especially when comparing RC scores for two models, (i.e. to determine whether a training approach improves a model's tendency to produce consistent responses, or determine which of two training approaches best improves a model's tendency towards consistent responses), we define:

$$\mu(c, a) = \sum_{\substack{c_i \in C_a \\ c_i \leq c}} m(c_i, a) \tag{9}$$

(i.e., the cumulative combinatoric mass) and then rephrase the definition of RC as:

$$\mathrm{RC}(c, a) = \frac{\mu(c, a)}{M(a)} \tag{10}$$

which relies on only one division, so is less prone to floating-point rounding errors.

This also allows us to compute:

$$\frac{\mu(c_1, a_1)}{M(a_1)} - \frac{\mu(c_2, a_2)}{M(a_2)} \tag{11}$$

(i.e., the improvement in $\mathrm{RC}(c_1, a_1)$ over $\mathrm{RC}(c_2, a_2)$ scores) as:

$$\frac{\mu(c_1, a_1)M(a_2) - \mu(c_2, a_2)M(a_1)}{M(a_1)M(a_2)} \tag{12}$$

which allows for comparisons between models that are very close in their RC scores, (i.e., in the long tail of consistency).

B Formulas for Bundle Sizes $b > 2$

Let us consider a contrastive test set containing n bundles of b instances each. There are $nb + 1$ possible accuracies a, but still $n + 1$ possible consistencies c.

C_a can then be defined in terms of n, b, and a as follows:

$$C_a = \{c \in C : c_{min}^{(a)} \leq c \leq c_{max}^{(a)}\} \tag{13}$$

where $c_{min}^{(a)}$ and $c_{max}^{(a)}$ are defined as:

$$c_{min}^{(a)} = \begin{cases} 0 & \text{if } a \leq n(b-1) \\ a - n(b-1) & \text{if } a > n(b-1 \end{cases} \tag{14}$$

$$c_{max}^{(a)} = \left\lfloor \frac{a}{b} \right\rfloor \tag{15}$$

Intuitively, if $a \leq n(b-1)$ then it is possible that all bundles have at least one of their constituent instances incorrectly answered, in which case, $c_{min}^{(a)} = 0$. However, if $a > n(b-1)$, then at least $a - n(b-1) > 0$ of bundles must be fully correctly answered. Indeed, for a bundle to be inconsistent at least one item must be incorrectly answered, so for a given a, the number of incorrect items is $nb - a$. Thus, at most $nb - a$ bundles can be inconsistent, and $c_{min}^{(a)} = n - (nb - a) = n - nb + a = a - n(b-1)$.

The definition of $c_{max}^{(a)}$ follows from the observation that a maximally consistent model will consistently respond to the maximum number of bundles for which it is possible that all b instances are correctly answered, and that equals $\left\lfloor \frac{a}{b} \right\rfloor$.

Now, $M(a)$ (the number of ways a model can achieve accuracy a) is given by:

$$M(a) = \binom{nb}{a} \tag{16}$$

and $m(c, a)$ (the number of ways a model can achieve accuracy a and consistency c) is given by:

$$m(c, a) = \binom{n}{c} \cdot G(n - c, b, a - cb) \tag{17}$$

where the first factor in the product still intuitively corresponds to the number of ways that c consistent bundles can be selected out of n, but the second refers to the number of ways the remaining correct instances could be distributed within responses to the test set such that no additional consistent bundles can be formed.

This second factor $G(m, b, k)$ is defined as:

$$G(m, b, k) = \sum_{r=0}^{R} (-1)^r \binom{m}{r} \binom{(m-r)b}{k - rb} \tag{18}$$

where $R = \min(m, \left\lfloor \frac{k}{b} \right\rfloor)$. This can be understood as the number of ways to select k elements of an $m \times b$ matrix such that no row contains a complete b elements selected. The first term (which simplifies to $\binom{mb}{k}$) is the total number of ways these k elements could be selected, ignoring the restriction on complete rows, and the remaining terms apply the principle of inclusion-exclusion to alternately subtract and add the number of ways that at least r rows could be filled (by multiplying the number of ways that r out of m rows could be selected by the number of ways the remaining $m - r$ rows and b columns could be filled by the remaining $k - rb$

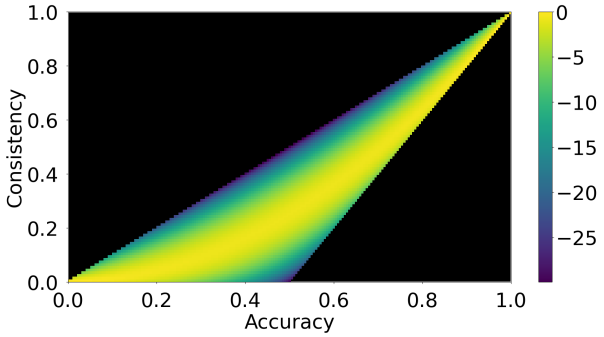


Figure 2: The $\log_{10}$ of the distributions of consistency scores in Figure 1a.

items to select), up to the maximal number of rows R that could be filled, whether that is determined by the total number of rows available m or the number of rows the items k could fill.

In general, we do not recommend using this measurement for bundle sizes above 2 except for evaluating consistency on three-valued features, as many consistency questions can be formulated as bundles with one instance exhibiting a feature and one instance lacking that feature.

C Distributions of Alternative Approaches

Figures 3 and 4 plot the distributions of consistency scores (for a 100-bundle dataset) obtained via simpler non-probabilistic alternatives and compare them to the distributions obtained via RC. Both of these characterizations lower the scores for consistencies that are above chance and raise the scores for consistencies that are below chance.

Size	Loss	B-A	P-A	S-A	A-A	B-C	P-C	S-C	A-C	B-RC	P-RC	S-RC	A-RC
Large	MLE	67.22	66.33	66.84	67.09	27.04	58.16	42.86	42.35	99.92	100.00	17.10	10.06
	CE	67.35	67.35	64.80	66.84	28.57	61.22	40.31	43.37	99.99	100.00	20.10	26.64
	λ_1MLE + λ_2CE												
	↳ $\lambda_1, \lambda_2 = 1.0, 1.0$	67.73	68.88	66.33	68.11	28.57	63.78	42.86	44.39	99.98	100.00	30.43	19.37
	↳ $\lambda_1, \lambda_2 = 0.33, 1.0$	68.24	68.37	66.58	68.37	30.10	63.27	43.37	44.90	100.00	100.00	42.44	16.01
3B	MLE	75.64	76.28	74.23	76.79	44.39	71.43	56.12	60.71	100.00	100.00	80.76	88.68
	CE	75.38	75.51	74.23	77.55	43.88	70.41	56.12	61.73	100.00	100.00	80.76	92.03
	λ_1MLE + λ_2CE												
	↳ $\lambda_1, \lambda_2 = 1.0, 1.0$	75.51	75.77	74.23	77.04	44.90	70.92	56.12	60.71	100.00	100.00	80.76	88.68
	↳ $\lambda_1, \lambda_2 = 0.33, 1.0$	76.53	77.55	76.02	78.57	45.92	73.47	58.67	63.78	100.00	100.00	79.32	98.60

Table 5: The full results of UnifiedQA-v2 (Khashabi et al., 2022) on the CondaQA contrastive dataset, with the expectation that including the Contrastive Estimation (CE) objective would improve consistency, as in Dua et al. (2021).

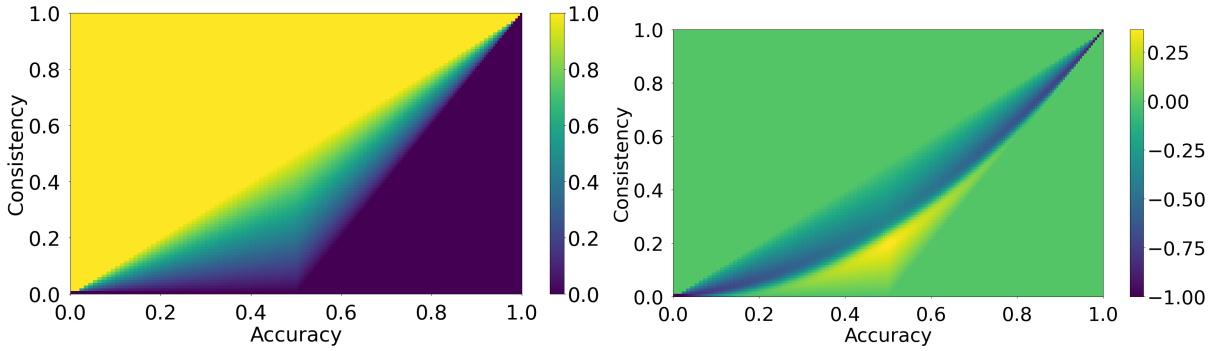


Figure 3: In this figure, the interval $\left[c_{min}^{(a)}, c_{max}^{(a)}\right]$ is simply scaled to cover $[0, 1]$ and the score is scaled accordingly. On the left is the score given a model's accuracy and consistency, on the right is shown the change in score when moving from RC to this formulation.

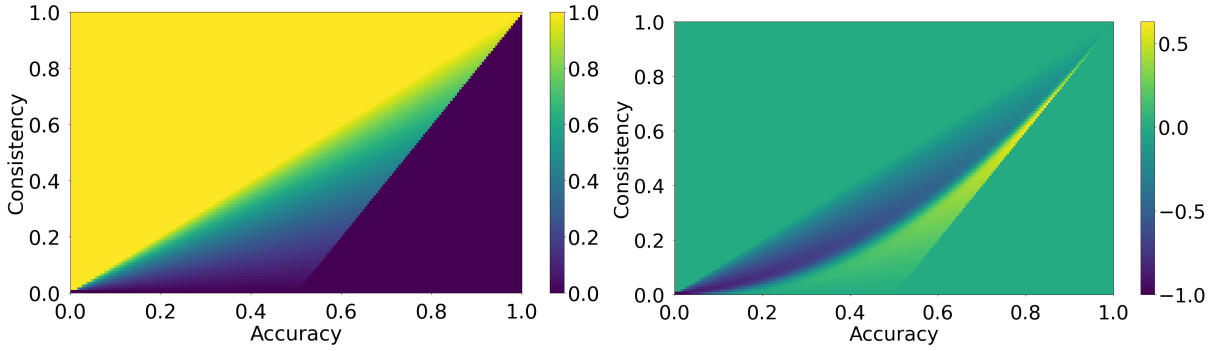


Figure 4: In this figure, of the bundles which are at least partially correct, the proportion of fully consistent bundles is reported. On the left is the score given a model's accuracy and consistency, on the right is shown the change in score when moving from RC to this formulation.