Interpreting Predictive Probabilities: Model Confidence or Human Label Variation?

Joris Baan, **Raquel Fernández**, **Barbara Plank**, **Wilker Aziz**
University of Amsterdam, IT University of Copenhagen, MCML Munich, LMU Munich
{j.s.baan,raquel.fernandez,w.aziz,}@uva.nl, b.plank@lmu.de

Abstract

With the rise of increasingly powerful and user-facing NLP systems, there is growing interest in assessing whether they have a good *representation of uncertainty* by evaluating the quality of their predictive distribution over outcomes. We identify two main perspectives that drive starkly different evaluation protocols. The first treats predictive probability as an indication of model confidence; the second as an indication of human label variation. We discuss their merits and limitations, and take the position that both are crucial for trustworthy and fair NLP systems, but that exploiting a single predictive distribution is limiting. We recommend tools and highlight exciting directions towards models with disentangled representations of uncertainty about predictions and uncertainty about human labels.

1 Introduction

In common language, uncertainty refers to "a state of not being definitely known or perfectly clear; a state of doubt".[1] In statistics and machine learning, uncertainty is taken as a state to be represented (Lindley, 2013; Halpern, 2017)—the state of the world as a function of inherently stochastic experiments or the state of knowledge of an agent observing or interacting with the world—and its mathematical representation requires prescribing a probability measure (Kolmogorov, 1960).

In modern NLP, neural networks are the de-facto standard to predict complex probability measures from available context (Goldberg and Hirst, 2017): given an input (or prompt), a neural network prescribes a representation of uncertainty over the space of responses (*e.g.*, strings or classes), typically, by mapping the input to the parameter of a probability mass function (*e.g.*, in text classification, inputs are mapped to the probability masses of each outcome in the label space).

[1] Oxford English Dictionary, accessed October 13th 2023.

Recently, transformer-based large language models (LLMs) are becoming increasingly powerful and display remarkable abilities on complex classification tasks, leading to an increased deployment in user-facing applications. This motivates the need for models that can signal when they are likely to be wrong (**P1**; an aspect of trustworthiness), and models that can capture different linguistic and human interpretations (**P2**; an aspect of language including fairness).

In this position paper, we identify that the exact same representation of uncertainty—the predictive distribution over outcomes—is sometimes interpreted as an indication of confidence in model predictions (**P1;** Desai and Durrett, 2020; Dan and Roth, 2021; Jiang et al., 2021a) and other times as an indication of variation in human perspectives (**P2;** Plank, 2022).

We hope to provide clarity and accelerate progress by:

(i) Identifying these two perspectives on the predictive distribution and examining how each evaluates the quality of predictive distribution in Section 2.

(ii) Discussing their merits and limitations, and relating them to popular notions of *aleatoric* and *epistemic* uncertainty in Section 3.

(iii) Taking the position that both perspectives contribute to trustworthy and fair NLP systems, but that exploiting a single predictive distribution is limiting—*e.g.*, does a uniform predictive distribution represent uncertainty about human perspectives, or rather about the correctness of that prediction itself?—and highlighting exciting directions towards models that can predict distributions over human or linguistic interpretations, and simultaneously abstain from answering when lacking such knowledge or skills in Section 4.

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 268–277

2 Two Perspectives on Uncertainty

Consider a user-facing question answering (QA) system. Ideally, this model is able to abstain on questions that it is likely to get wrong (a.k.a. selective answering or prediction; Kamath et al., 2020; Yoshikawa and Okazaki, 2023), for which its probabilities should reflect confidence in predictions (*i.e.*, predictive probabilities help us determine whether the model is right or wrong). Now consider that various NLP tasks, including QA, are being acknowledged as supporting human label variation (Plank, 2022), and that some questions can be underspecified, ambiguous or subjective (there are many such datasets, for QA see for example Min et al. (2020) and Amouyal et al. (2023), and for other tasks see Section 3.2). Different annotators might therefore provide a different reference answer. From this perspective, probabilities should reflect the relative frequency of each answer assigned to that particular question by the pool of annotators (*i.e.*, predictive probabilities help us determine what answers represent the views of a certain population). These two perspectives on the role of predictive probabilities in fact aim at different sources of uncertainty: uncertainty about model error (*e.g.*, due to imperfect design and estimation) and uncertainty about human labels (*e.g.*, due to label variation in a population). So, if a model predicts a uniform distribution, does this mean that all answers are plausible or that this prediction should not be trusted?

2.1 Background

Most text classifiers chain two building blocks: i) a parametric model which, given input text x, prescribes the probability mass function (pmf) $f(y; x)$ of the conditional random variable $Y|X = x$ taking on values in a set $\{1, \ldots, K\}$ of K class labels; and ii) a decision rule $\delta_f(x)$ to map from $f(\cdot; x)$ to a single label. For most modern models, the map $x \mapsto f(\cdot; x)$ is realised by a neural network and the most common decision rule $\delta_f(x) = \arg\max_{k \in [K]} f(k; x)$ returns the mode of the pmf. Next, we identify two main perspectives on predictive probability $f(y; x)$, with starkly different evaluation frameworks.[2]

[2]We use capital letters for random variables (*e.g.*, X, Y) and lowercase letters for outcomes (*e.g.*, x, y). As standard, $X = x$ denotes random variable (rv) assignment. For logical predicates we use the Iverson bracket $[A = B]$ to denote a new rv whose outcome is 1, when A and B are assigned the same outcome, and 0 otherwise. A deterministic function of an rv defines a new rv; for example, the rv $\delta_f(X) = \arg\max_{k \in [K]} f(k; X)$ captures the mode of the conditional distribution as a function of the random input X. We use $\Pr$ to denote an implicit probability measure capturing the data generation process; we do not possess an explicit representation for this measure, but we can estimate its assessment via Monte Carlo–that is, the relative frequency of the relevant events in a dataset of labelled inputs.

2.2 P1: Uncertainty about Model Error

The first and arguably more common perspective interprets predictive probabilities as predictive of *classification performance* and is often explained as evaluating the extent to which "a model knows when it does not know" (*e.g.*, in NLP: Desai and Durrett, 2020; Dan and Roth, 2021; Jiang et al., 2021a). An increasingly popular evaluation framework taking this perspective is calibration.

The core desideratum behind *confidence calibration* (Naeini et al., 2015; Guo et al., 2017) is that, **in expectation over inputs**, a classifier's predictive mode probability $\pi_f(X) = \max_{k \in [K]} f(k; X)$ matches the relative frequency of predictions $\delta_f(X) = \arg\max_{k \in [K]} f(k; X)$ being judged as correct $[Y = \delta_f(X)] = 1$. So, $\forall q \in [0, 1]$,

$$\Pr\left([Y = \delta_f(X)] = 1 \mid \pi_f(X) = q\right) \stackrel{?}{=} q \,. \quad (1)$$

For example, if 100 predictions are made with probability 0.9, then 90 should be judged as correct.[3] In practice Equation (1) is hard to MC estimate (for it requires observing multiple predictions with identical probability), so the probability space is partitioned into M bins. For each bin B_m, the calibration error is the difference between accuracy and average probability of the predictions in it. The expected calibration error (ECE) is the weighted average over bins:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} (\mathrm{acc}(B_m) - \mathrm{conf}(B_m)) \,. \quad (2)$$

2.3 P2: Uncertainty about Human Labels

Crucially, the above interpretation is different from evaluating, **for each individual input** x, whether the predictive probability $f(k; x)$ matches the relative frequency with which (a population of) humans would pick that same label k: $\forall k \in [K]$,

$$\Pr(Y = k | X = x) \stackrel{?}{=} f(k; x) \,. \quad (3)$$

Although there is no standard evaluation protocol yet (Lovchinsky et al., 2020; Basile et al., 2021;

[3]Other notions assess calibration for fixed classes (*class-wise*; Nixon et al., 2019) or probability vectors (*multi-class*; Vaicenavicius et al., 2019; Kull et al., 2019).

Plank, 2022), researchers use datasets with multiple annotations per input to estimate a *human distribution*, and compare that to the predictive distribution through statistical divergence (*e.g.*, Kullback-Leibner or Jensen-Shannon Divergence; Total Variation Distance), or summary statistics like entropy (Pavlick and Kwiatkowski, 2019; Nie et al., 2020; Baan et al., 2022; Glockner et al., 2023).

2.4 Ambiguity in Explaining Calibration

The language that is often used to explain calibration allows (quite ironically) for both perspectives **P1** and **P2**.

Desai and Durrett (2020): "*If a model assigns 70% probability to an event, the event should occur 70% of the time if the model is calibrated*". The word "event" can refer to observing a class given an input (**P2**) or a model prediction matching the observed class (**P1**).

Jiang et al. (2021b): "*the property of a probabilistic model's predictive probabilities actually being well correlated with the probabilities of correctness*". The word "correctness" can refer to the probability of observing that class in the data (**P2**) or to the probability of a predicted class matching the data (**P1**).

Gupta et al. (2021): "*a classifier is said to be calibrated if the probability values it associates with the class labels match the true probabilities of correct class assignments*" and "*It would be desirable if the numbers z_k output by a network represented true probabilities*". Human annotators could assign the class (**P2**), or a model could (**P1**). The phrase "true probabilities" could refer to observed class (**P2**) or model error (**P1**) frequencies.

The examples above illustrate well that one may regard predictive probabilities one way or another, each interpretation tracking a different type of event (*i.e.*, correctness, assessed marginally for a collection of inputs, or label frequency, assessed conditionally against a population of annotators). Crucially, however, most models are trained to approximately recover the maximum likelihood solution—a single realisation of the map $x \mapsto f(\cdot; x)$, with no room for quantification of uncertainty about its correctness. Therefore, without special incentives (*e.g.*, regularisation, change of loss or supervision; some of which we discuss in Section 4.1), our predictive distributions are not meant to inherently support **P1**, and they *may* support **P2**, as we discuss in the next section.

3 Merits and Limitations

The predictive distribution for an input x is sometimes taken as a representation of uncertainty about **a model's future classification performance** ("knowing when it knows"); other times as a representation of uncertainty about **label frequency in a population of human annotators** (human label variation). We now discuss merits and limitations for each perspective.

3.1 P1: Uncertainty about Model Error

From a statistical perspective, most NLP systems are trained on single annotations using regularised maximum likelihood estimation (MLE), without mechanism or incentive to represent uncertainty about their own correctness (MLE recovers a single realisation of the map $x \mapsto f(\cdot; x)$). This is unlike, for instance, Bayesian estimation (where the map $x \mapsto f(\cdot; x)$ is given random treatment; more in Section 4).

In addition, regardless of whether *models* represent uncertainty about their own correctness, calibration *metrics*, and ECE in particular, are known to have limitations, e.g., problems with binning (Nixon et al., 2019; Vaicenavicius et al., 2019; Gupta et al., 2021), evaluating only the mode probability rather than the entire distribution (Kumar et al., 2019; Vaicenavicius et al., 2019; Widmann et al., 2019; Kull et al., 2019), and being minimised by global label frequencies (Nixon et al., 2019). Moreover, Baan et al. (2022) recently demonstrate that ECE disregards plausible instance-level label variation and pose that such calibration metrics are ill-suited for tasks with human label variation.

Finally, the sense of trustworthiness from verifying that Equation (1) holds (for a given confidence level q) in a given dataset, might not transfer to any one future prediction in isolation. Though some studies examine the effect of communicating predictive probability to human decision makers (Zhang et al., 2020; Wang and Yin, 2021; Vodrahalli et al., 2022; Vasconcelos et al., 2023; Dhuliawala et al., 2023), to the best of our knowledge, none verified the user-impact of models with various calibration scores, raising the question: can calibration metrics like ECE discriminate systems perceived as more trustworthy?

3.2 P2: Uncertainty about Human Labels

The idea that gold labels are too simplistic has been around for some time (Poesio and Artstein, 2005;

Aroyo and Welty, 2015) and is gaining traction with increasing evidence that annotators can plausibly pick different class labels for an input (Plank, 2022). Examples include subjective tasks such as hate speech detection (Kennedy et al., 2022) and textual emotion recognition (Demszky et al., 2020); and ambiguous or difficult tasks like object naming (Silberer et al., 2020), textual entailment (Pavlick and Kwiatkowski, 2019; Nie et al., 2020), part-of-speech tagging (Manning, 2011; Plank et al., 2014) and discourse relation classification (Scholman et al., 2022). However, the connection to uncertainty is relatively new (Pavlick and Kwiatkowski, 2019; Nie et al., 2020; Baan et al., 2022).

From a statistical perspective, text classifiers predict a distribution for $Y|X = x$, and are *precisely* mechanisms to represent uncertainty about a given input's label. However, given that they are parametric models trained with regularised MLE, they can at best learn to predict *observed* label variability (which is often not present in NLP datasets since most record only single annotations), or label variability as a *byproduct* of parametric bottlenecks, regularisation and other inductive biases that reserve (conditional) probability for unseen labels.

Evaluating whether probability mass is indeed allocated coherently with plausible variability is limited by: 1) datasets lacking multiple high quality annotations per input, 2) unclarity about how many annotations are sufficient to reliably estimate the human distribution (Zhang et al., 2021), 3) how to separate plausible variation from noise—for example due to spammers (Raykar and Yu, 2011; Beigman Klebanov and Beigman, 2014; Aroyo et al., 2019), and 4) the assumption of one unique human distribution being a simplification: subpopulations can cause the marginal distribution not to be representative of its individual components (Baan et al., 2022; Jiang et al., 2023).

3.3 Sources of Uncertainty

These two perspectives on the predictive distribution in NLP can be put in a broader context of statistics and machine learning by considering that there can be many sources that lead to uncertainty (Der Kiureghian and Ditlevsen, 2009; Hüllermeier and Waegeman, 2021; Gruber et al., 2023; Jiang et al., 2023; Baan et al., 2023). For example, underspecified input, ambiguity, noise or lack of training data can all be considered sources that may lead to uncertainty.

Such sources are often categorised as *aleatoric* (irreducible; inherent to data) or *epistemic* (reducible, inherent to modelling). In that sense, **P1** regards the predictive distribution as epistemic uncertainty, whereas **P2** as aleatoric uncertainty. Armed with this knowledge, one can pick the right modeling tools for each, and tap into this broader literature. In the next section, we make several recommendations.

4 Best of Both Worlds

We argue that the desiderata behind both perspectives are equally important for trustworthy and fair NLP systems, but that expecting the predictive distribution to represent both is limiting. Rather than calibrating the predictive distribution to better indicate model error, we outline alternative directions to capture uncertainty about predictions (towards more trustworthy NLP) *and* uncertainty about human perspectives (towards fairer NLP)—where the latter can, and in our view *should* be represented by the predictive distribution.

4.1 Towards More Trustworthy NLP Systems

Inspired by machine translation quality estimation (*e.g.* Blatz et al., 2004; Specia et al., 2009; Fomicheva et al., 2020) and the observation that models fail in predictable ways, one could train a (separate) module to predict errors. Ideally, this module is uncertainty-aware (Glushkova et al., 2021), and predicts fine-grained errors (Dou et al., 2022). Predictive probabilities (or summaries like entropy) are features that can be combined with, for example, model explainability features (Li et al., 2022; Ye and Durrett, 2022; Park and Caragea, 2022) or input properties (Dong et al., 2018; Kamath et al., 2020).

Alternatively, the event space can be expanded beyond only the target variable to include parameters too, thus allowing for uncertainty about them. Since this leads to intractability, some (approximate) Bayesian solutions in NLP include Langevin dynamics (Gan et al., 2017; Shareghi et al., 2019), Monte Carlo dropout (Shelmanov et al., 2021; Vazhentsev et al., 2022), ensembling (Ulmer et al., 2022), variational inference (Ponti et al., 2021), and stochastic attention (Pei et al., 2022). Other directions rely on the distance of a new input to the training data, like conformal prediction (Maltoudoglou et al., 2020; Giovannotti and Gammerman, 2021; Zerva and Martins, 2023) or feature space density

(Van Amersfoort et al., 2020; Vazhentsev et al., 2022; Mukhoti et al., 2023).

Evaluating model error uncertainty is challenging, in part because ground truth is difficult to find. Proxy tasks like selective answering (Dong et al., 2018; Kamath et al., 2020; Yoshikawa and Okazaki, 2023) are useful due to their flexibility in defining quality (other than accuracy), and error indicators (other than predictive probability), and we encourage more principled evaluation methods.

Rottger et al. (2022) propose two annotation paradigms: encouraging the *description* of multiple beliefs or *prescription* of one consistent belief. Prescriptive datasets, by definition, have no data uncertainty, and although that does not change merits of the model-error perspective, one could now safely supervise models to be more coherent with this interpretation (the goal of calibration), *e.g.* by minimising ECE directly, or through other regularisation objectives (Kong et al., 2020).

4.2 Towards Fairer NLP Systems

To represent uncertainty about plausible human interpretations, data is crucial. For example: how are annotators recruited, what are their backgrounds, how diverse is the population, what guidelines do they follow, what is their incentive, how focused are they, what is their prior experience or expertise, how many annotations per input are collected?

In NLP, these factors are commonly not controlled for. However, recently, researchers use annotator information to model sub-populations (Al Kuwatly et al., 2020; Akhtar et al., 2020) or even individual annotators (Geva et al., 2019; Mostafazadeh Davani et al., 2022; Gordon et al., 2022). Without access to such information, others collect and train on multiple annotations per instance (Peterson et al., 2019; Uma et al., 2020; Fornaciari et al., 2021; Uma et al., 2021; Zhang et al., 2021; Meissner et al., 2021), or individual annotator confidence scores (Chen et al., 2020; Collins et al., 2022).

Besides data, an appealing but non-trivial alternative (for some tasks, like textual entailment) is to encourage models to generalise to the linguistic phenomena that give rise to label variation, despite supervising with single annotations Pavlick and Kwiatkowski (2019). Yet another direction is to isolate and understand specific sources of label variation, for example, linguistic ambiguity, and design targeted methods to model them (Beck et al., 2014; Jiang and Marneffe, 2022; Liu et al., 2023).

Not all variability is desirable. However, detecting or even defining annotation errors when variation is plausible is difficult. Annotation error detection methods exist, however currently focus on gold labels (Wei et al., 2022; Klie et al., 2022; Weber and Plank, 2023). We encourage studying noise in label variation settings (Paun et al., 2018; Gordon et al., 2021).

5 Conclusion

In this position paper, we identified two important perspectives on the predictive distribution in NLP. We believe that the desiderata behind both are crucial for fair and trustworthy NLP systems, but that exploiting the same predictive distribution is limiting. We recommend exiting tools and directions to represent uncertainty about predictions (model confidence) and about label variation (human perspectives). We hope to facilitate a better understanding of uncertainty in NLP, and encourage future work to acknowledge, represent and evaluate multiple sources of uncertainty with principled design decisions.

Limitations

Evaluation along a specific axis can be useful regardless of whether a model has been explicitly designed to meet this goal. One could argue this is true for both calibration as well as human label variation. It is certainly also true in other sub-fields, like interpretability. For example, probing hidden representations or specific linguistic information, without having explicitly trained models to store them. Furthermore, although we focus on classification systems in the language domain, the topics we highlight and discuss are equally important in other domains, such as computer vision (*e.g.*, affective computing), or language generation (*e.g.*, story telling).

Acknowledgements

JB is supported by the ELLIS Amsterdam Unit. RF and BP are supported by the European Research Council (ERC), grant agreements No. 819455 (DREAM) and No. 101043235 (DIALECT), respectively. WA is supported by the EU's Horizon Europe research and innovation programme (grant agreement No. 101070631, UTTER).

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 151–154.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

Samuel Joseph Amouyal, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig, and Jonathan Berant. 2023. Qampari: An open-domain question answering benchmark for questions with many answers from multiple paragraphs.

Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing subjective tasks: The case study of understanding toxicity in online discussions. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 1100–1105, New York, NY, USA. Association for Computing Machinery.

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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2307.15703*.

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. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Daniel Beck, Trevor Cohn, and Lucia Specia. 2014. Joint emotion analysis via multi-task Gaussian processes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1798–1803, Doha, Qatar. Association for Computational Linguistics.

Beata Beigman Klebanov and Eyal Beigman. 2014. Difficult cases: From data to learning, and back. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 390–396, Baltimore, Maryland. Association for Computational Linguistics.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321, Geneva, Switzerland. COLING.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online. Association for Computational Linguistics.

Katherine M Collins, Umang Bhatt, and Adrian Weller. 2022. Eliciting and learning with soft labels from every annotator. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 40–52.

Soham Dan and Dan Roth. 2021. On the effects of transformer size on in- and out-of-domain calibration. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2096–2101, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

Shehzaad Dhuliawala, Vilém Zouhar, Mennatallah El-Assady, and Mrinmaya Sachan. 2023. A diachronic perspective on user trust in AI under uncertainty. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5567–5580, Singapore. Association for Computational Linguistics.

Li Dong, Chris Quirk, and Mirella Lapata. 2018. Confidence modeling for neural semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 743–753, Melbourne, Australia. Association for Computational Linguistics.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. Is GPT-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7250–7274, Dublin, Ireland. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

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. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

Zhe Gan, Chunyuan Li, Changyou Chen, Yunchen Pu, Qinliang Su, and Lawrence Carin. 2017. Scalable Bayesian learning of recurrent neural networks for language modeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 321–331, Vancouver, Canada. Association for Computational Linguistics.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Patrizio Giovannotti and Alex Gammerman. 2021. Transformer-based conformal predictors for paraphrase detection. In *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 152 of *Proceedings of Machine Learning Research*, pages 243–265. PMLR.

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2023. Ambifc: Fact-checking ambiguous claims with evidence. *arXiv preprint arXiv:2104.00640*.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André F. T. Martins. 2021. Uncertainty-aware machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3920–3938, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yoav Goldberg and Graeme Hirst. 2017. *Neural Network Methods in Natural Language Processing*. Morgan & Claypool Publishers.

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

Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Association for Computing Machinery*, CHI '21, New York, NY, USA.

Cornelia Gruber, Patrick Oliver Schenk, Malte Schierholz, Frauke Kreuter, and Göran Kauermann. 2023. Sources of uncertainty in machine learning–a statisticians' view. *arXiv preprint arXiv:2305.16703*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Kartik Gupta, Amir Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, Cristian Sminchisescu, and Richard Hartley. 2021. Calibration of neural networks using splines. In *International Conference on Learning Representations*.

Joseph Y Halpern. 2017. *Reasoning about uncertainty*. MIT press.

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

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.

Nan-Jiang Jiang, Chenhao Tan, and Marie-Catherine de Marneffe. 2023. Understanding and predicting human label variation in natural language inference through explanation. *arXiv preprint arXiv:2304.12443*.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021a. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021b. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2022. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, pages 1–30.

Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. 2022. Annotation error detection: Analyzing the past and present for a more coherent future. *Computational Linguistics*, pages 1–42.

Andrey N. Kolmogorov. 1960. *Foundations of the Theory of Probability*, 2 edition. Chelsea Pub Co.

Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. 2020. Calibrated language model fine-tuning for in- and out-of-distribution data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1326–1340, Online. Association for Computational Linguistics.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32.

Ananya Kumar, Percy S Liang, and Tengyu Ma. 2019. Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32.

Dongfang Li, Baotian Hu, and Qingcai Chen. 2022. Calibration meets explanation: A simple and effective approach for model confidence estimates. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2784, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dennis V Lindley. 2013. *Understanding uncertainty*. John Wiley & Sons.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity. *arXiv preprint arXiv:2304.14399*.

Igor Lovchinsky, Alon Daks, Israel Malkin, Pouya Samangouei, Ardavan Saeedi, Yang Liu, Swami Sankaranarayanan, Tomer Gafner, Ben Sternlieb, Patrick Maher, et al. 2020. Discrepancy ratio: Evaluating model performance when even experts disagree on the truth. In *International Conference on Learning Representations*.

Lysimachos Maltoudoglou, Andreas Paisios, and Harris Papadopoulos. 2020. Bert-based conformal predictor for sentiment analysis. In *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 128 of *Proceedings of Machine Learning Research*, pages 269–284. PMLR.

Christopher D Manning. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *Computational Linguistics and Intelligent Text Processing: 12th International Conference, CICLing 2011, Tokyo, Japan, February 20-26, 2011. Proceedings, Part I 12*, pages 171–189. Springer.

Johannes Mario Meissner, Napat Thumwanit, Saku Sugawara, and Akiko Aizawa. 2021. Embracing ambiguity: Shifting the training target of NLI models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 862–869, Online. Association for Computational Linguistics.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

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

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. 2023. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24384–24394.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring calibration in deep learning. In *CVPR workshops*, 7.

Seo Yeon Park and Cornelia Caragea. 2022. On the calibration of pre-trained language models using mixup guided by area under the margin and saliency. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5364–5374, Dublin, Ireland. Association for Computational Linguistics.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing Bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

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

Jiahuan Pei, Cheng Wang, and György Szarvas. 2022. Transformer uncertainty estimation with hierarchical stochastic attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11147–11155.

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*, pages 9617–9626.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 76–83, Ann Arbor, Michigan. Association for Computational Linguistics.

Edoardo M. Ponti, Ivan Vulić, Ryan Cotterell, Marinela Parovic, Roi Reichart, and Anna Korhonen. 2021. Parameter space factorization for zero-shot learning across tasks and languages. *Transactions of the Association for Computational Linguistics*, 9:410–428.

Vikas C Raykar and Shipeng Yu. 2011. Ranking annotators for crowdsourced labeling tasks. *Advances in neural information processing systems*, 24.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Merel Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2022. DiscoGeM: A crowdsourced corpus of genre-mixed implicit discourse relations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3281–3290, Marseille, France. European Language Resources Association.

Ehsan Shareghi, Yingzhen Li, Yi Zhu, Roi Reichart, and Anna Korhonen. 2019. Bayesian learning for neural dependency parsing. 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 3509–3519, Minneapolis, Minnesota. Association for Computational Linguistics.

Artem Shelmanov, Evgenii Tsymbalov, Dmitri Puzyrev, Kirill Fedyanin, Alexander Panchenko, and Maxim Panov. 2021. How certain is your Transformer? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1833–1840, Online. Association for Computational Linguistics.

Carina Silberer, Sina Zarrieß, Matthijs Westera, and Gemma Boleda. 2020. Humans meet models on object naming: A new dataset and analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1893–1905, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Lucia Specia, Marco Turchi, Nicola Cancedda, Nello Cristianini, and Marc Dymetman. 2009. Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, Barcelona, Spain. European Association for Machine Translation.

Dennis Ulmer, Jes Frellsen, and Christian Hardmeier. 2022. Exploring predictive uncertainty and calibration in NLP: A study on the impact of method & data scarcity. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2707–2735, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft loss functions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 173–177.

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.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. 2019. Evaluating model calibration in classification. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3459–3467. PMLR.

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. 2020. Uncertainty estimation using

a single deep deterministic neural network. In *International conference on machine learning*, pages 9690–9700. PMLR.

Helena Vasconcelos, Gagan Bansal, Adam Fourney, Q Vera Liao, and Jennifer Wortman Vaughan. 2023. Generation probabilities are not enough: Exploring the effectiveness of uncertainty highlighting in ai-powered code completions. *arXiv preprint arXiv:2302.07248*.

Artem Vazhentsev, Gleb Kuzmin, Artem Shelmanov, Akim Tsvigun, Evgenii Tsymbalov, Kirill Fedyanin, Maxim Panov, Alexander Panchenko, Gleb Gusev, Mikhail Burtsev, Manvel Avetisian, and Leonid Zhukov. 2022. Uncertainty estimation of transformer predictions for misclassification detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8237–8252, Dublin, Ireland. Association for Computational Linguistics.

Kailas Vodrahalli, Tobias Gerstenberg, and James Y Zou. 2022. Uncalibrated models can improve human-ai collaboration. *Advances in Neural Information Processing Systems*, 35:4004–4016.

Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, IUI '21, page 318–328, New York, NY, USA. Association for Computing Machinery.

Leon Weber and Barbara Plank. 2023. ActiveAED: A human in the loop improves annotation error detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8834–8845, Toronto, Canada. Association for Computational Linguistics.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. 2022. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*.

David Widmann, Fredrik Lindsten, and Dave Zachariah. 2019. Calibration tests in multi-class classification: A unifying framework. *Advances in neural information processing systems*, 32.

Xi Ye and Greg Durrett. 2022. Can explanations be useful for calibrating black box models? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6199–6212, Dublin, Ireland. Association for Computational Linguistics.

Hiyori Yoshikawa and Naoaki Okazaki. 2023. Selective-lama: Selective prediction for confidence-aware evaluation of language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1972–1983.

Chrysoula Zerva and André FT Martins. 2023. Conformalizing machine translation evaluation. *arXiv preprint arXiv:2306.06221*.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Learning with different amounts of annotation: From zero to many labels. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7620–7632, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 295–305, New York, NY, USA. Association for Computing Machinery.