IMPROVING MODEL-MODERATOR COLLABORATIVE PREDICTIONS WITH UNCERTAINTY ESTIMATION

Zi Lin[1]*, Ian D. Kivlichan[2]*, Jeremiah Zhe Liu[1], Lucy Vasserman[2]
[1]Google Research, [2]Jigsaw, Google
{lzi,kivlichan,jereliu,lucyvasserman}@google.com

ABSTRACT

As the scale of conversations online grows, content providers have increasingly incorporated machine learning models into moderation systems, operating in collaboration with human content moderators to detect toxic content given limited human time and attention. These machine learning models are typically evaluated using metrics like accuracy or AUROC. However, such metrics fail to capture the performance of the combined moderator-model system. Here, we propose an accuracy metric, Oracle-Model Collaborative Accuracy (OCA), that describes the overall system performance under constraints on human review capacity. Moreover, we present a challenging data benchmark, CoToMoD, for evaluating the performance of collaborative toxic comment moderation systems. Using this benchmark, we evaluate the performance of several models using OCA and other metrics: our results demonstrate the importance of metrics capturing the collaborative nature of the moderator-model system for this task, as well as the utility of uncertainty estimation in this domain.[1]

1 INTRODUCTION

Maintaining civil discussions online is a persistent challenge for online platforms. Due to the sheer scale of user-generated text every day, modern content moderation systems often employ machine learning algorithms to automatically classify user comments based on their toxicity, with the goal of flagging a sub-collection of likely policy-violating content for human experts to review (Etim, 2017). However, modern deep-learning NLP models have been shown to suffer from reliability and robustness issues, especially in the face of the rich and heterogeneous sociolinguistic phenomena in real-world online conversations. Examples include possibly generating confidently wrong predictions that latch on spurious lexical-level features (Wang & Culotta, 2020), or exhibiting undesired bias toward disadvantaged social subgroups (Dixon et al., 2018). This has raised questions about how the current toxicity detection models perform in realistic online environments, as well as the potential consequences for moderation systems (Rainie et al., 2017).

In this work, we study an approach to address these questions by incorporating model uncertainty into the collaborative model-moderator system's decision-making process. The intuition is that, by using uncertainty as a signal for the likelihood of model error, we can improve the efficiency and performance of the collaborative moderation system by prioritizing those examples that the model is most uncertain about for human review. Despite a plethora of uncertainty methods in the literature, there has been limited work studying their effectiveness in improving the performance of human-AI collaborative systems with respect to application-specific criteria of interest (Awaysheh et al., 2019; Dusenberry et al., 2020; Jesson et al., 2020). This is especially important for the content moderation task: real-world practice has a unique set of challenges and constraints, including label imbalance, distributional shift, and limited bandwidth of human experts; how these factors impact the collaborative system's effectiveness is not well understood.

To this end, we make foundational contributions to the study of the uncertainty-aware collaborative content moderation problem by (1) proposing a rigorous metric, *Oracle-Model Collaborative*

*Equal contribution. This work was done while the first author was an AI resident at Google Research.

[1] Code available at https://git.io/JOYSR.

Accuracy (OCA), to measure the overall system performance under capacity constraints on the human moderator. Then, (2) we introduce a challenging data benchmark, *Collaborative Toxicity Moderation in the Wild* (CoToMoD), for evaluating the effectiveness of a collaborative toxic comment moderation system. CoToMoD emulates the realistic train-deployment environment of a moderation system, in which the deployment environment contains a more diverse range of topics and richer linguistic phenomena than the training data, such that incorporating uncertainty signals is crucial for good system performance (Amodei et al., 2016). (3) Finally, we present a benchmark study evaluating the performance of four classic or state-of-the-art uncertainty approaches on CoToMoD (BERT, MC Dropout, DeepEnsemble, SNGP). We find that well-calibrated uncertainty models perform surprisingly well on the system-level OCA, and that the model's ability to handle class imbalance is crucial for performance. Supplementary Section B surveys related work.

2 CoToMoD: A Benchmark for effective content moderation using uncertainty

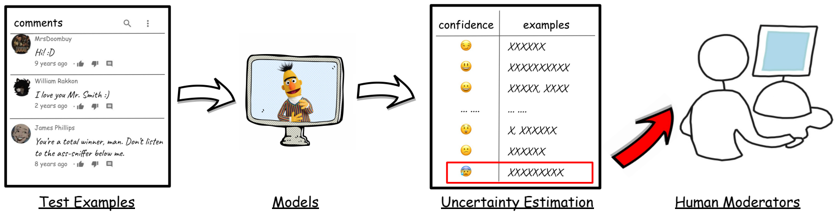

Figure 1: Schematic depicting the operation of the model-moderator system. We propose having the model send the comments about which it is most uncertain for human review.

Online content moderation is performed using review by a combination of humans and machine learning models. The moderator and model operate in collaboration, e.g. the model can flag a set of toxic posts to the moderator for further review. In a realistic industrial setting, toxicity detection models are often trained on a well-curated dataset with clean annotations, and then deployed to an environment that contains much richer sociolinguistic phenomenon and exhibits topical and distributional shifts when compared to the training data. To this end, we introduce a challenging data benchmark, the *Collaborative Toxicity Moderation in the Wild* (CoToMoD), to evaluate the practical effectiveness of uncertainty-aware models in improving the collaborative content moderation system in a realistic setting. Specifically, CoToMoD consists of a set of train, test, and deployment environments, where a toxicity model is trained and tested on a less diverse corpus (here, Wikipedia talk page comments (Wulczyn et al., 2017)), and then deployed to a much larger and more diverse environment (Borkan et al., 2019).

Oracle-Model Collaborative Accuracy Common metrics for evaluating the predictive or uncertainty perform of a toxicity detection model include accuracy, areas under the curve (AUC), or expected calibration error (ECE). However, these metrics do not account for the role of human experts in the moderation system, or specify how humans should utilize the model's predictions.

To capture the collaborative interaction between human moderators and machine learning models, we propose a metric *Oracle-Model Collaborative Accuracy* (OCA). Specifically, we consider a collaborative game between the neural model and an "oracle" human moderator with limited bandwidth in moderating online comments. Given a large number of examples to be moderated, the model first generates predictions for every example, and then sends a pre-specified number (we consider fixed fractions, but could also consider a constant limit) of these examples it is most uncertain about to the human experts, who then label them correctly. To this end, OCA measures the combined accuracy of this collaborative process, subject to a limited bandwidth α for the human oracle (i.e., they can only process α% of the total examples). Formally, given the dataset $D = \{(x_i, y_i)\}_{i=1}^n$, for a predictive model $f(x)$ equipped with an uncertainty estimator $u(x)$, the OCA per example is

$$\text{OCA}(x_i|\alpha) = \begin{cases} \mathbb{I}(f(x_i) = y_i) & \text{if } u(x_i) > q_{1-\alpha} \\ 1 & \text{o.w.} \end{cases},$$

so that $\text{OCA}(D) = \frac{1}{n}\sum_{i=1}^n \text{OCA}(x_i|\alpha)$ for the whole dataset. Here $q_{1-\alpha}$ is the $(1-\alpha)^{\text{th}}$ quantile of the model's uncertainty scores $\{u(x_i)\}_{i=1}^n$ over the entire dataset. The goal of this metric, then, is to understand how well the model and moderator can collaborate to achieve the highest overall accuracy, under capacity constraints on the human moderator. In this way, OCA quantifies the behavior of the full model-moderator system, and better captures its real-world usage. To improve the human-model collaborative accuracy using uncertainty, a toxicity model needs to generate well-calibrated uncertainty scores, such that the examples sent to the human moderator are those for which the model's predictions are most likely to be incorrect. Note that, as a population-level metric, OCA differs from the usual classification with rejection scenario; see Supplementary Section B.

Class Imbalance An important feature of the data distribution is the *rarity* of toxic content (Cheng et al., 2017; Wulczyn et al., 2017). This manifests in the datasets we use: 50,350 of the 1,999,514 examples in the CivilComments dataset are toxic ($\sim$$2.5\%$) (Borkan et al., 2019), and 21,384 out of 223,549 Wikipedia Talk Corpus examples are toxic ($\sim 9.6\%$) (Wulczyn et al., 2017). As we will show, failing to account for class imbalance can severely bias model predictions toward the majority (non-toxic) class, reducing the effectiveness of the overall collaborative system.

3 Benchmark Experiments

Uncertainty Methods We benchmark the performance of classic and the latest state-of-the-art methods from the probabilistic deep learning literature. We consider the BERT$_{\texttt{base}}$, a 12-layer deep transformer classifier model (Devlin et al., 2019), and select four methods based on their practical applicability for deep transformers. Specifically, we consider (1) **Deterministic** which computes the sigmoid probability $p(x) = \mathrm{sigmoid}(\mathrm{logit}(x))$ of a vanilla BERT model (Hendrycks & Gimpel, 2017), (2) *Monte Carlo Dropout* (**MC Dropout**) which estimates uncertainty using the Monte Carlo averaging of $p(x)$ from 10 dropout samples (Gal & Ghahramani, 2016), (3) **Deep Ensemble** which estimates uncertainty using the ensemble mean of $p(x)$ from 10 parallel-trained BERT models (Lakshminarayanan et al., 2017). Finally, we experiment with (4) *Spectral-normalized Neural Gaussian Process* (**SNGP**), a recent state-of-the-art approach that transforms a deep model into an approximate Gaussian process using simple regularization techniques (Liu et al., 2020). Given the predictive probability $p(x)$ from a model, we compute uncertainty as the predictive variance $u(x) = p(x)(1 - p(x))$, so that the uncertainty is maximized at $p(x) = 0.5$, and decreases toward 0 as $p(x)$ approaches 0 or 1.

To address class imbalance, we combine the uncertainty methods with **Focal Loss** (Lin et al., 2017). Focal loss reshapes the loss function to down-weight "easy" negatives (non-toxic examples), focusing training on more difficult examples, and thereby empirically improving predictive and uncertainty calibration performance on class-imbalanced datasets (Lin et al., 2017; Mukhoti et al., 2020).

Evaluation For each model, we evaluate their predictive performance, uncertainty quality, and collaborative effectiveness. For prediction, we compute test accuracy (Acc.), and also the areas under the receiver operating characteristic curve (AUC-ROC) and under the precision-recall curve (AUC-PR). For uncertainty, we compute the expected calibration error (ECE, see Supplementary Section A for definition) (Naeini et al., 2015) and the Brier score (i.e. mean-squared error). Finally, we measure collaborative effectiveness using the Oracle-Model Collaborative Accuracy (OCA). We measure OCA over a range of moderator review capacities, that is, the fraction of comments the model is permitted to pass to the moderator for further review. We compute the OCA for moderator review capacities of 1%, 5%, 10%, 15%, and 20%, sending comments to the moderator in order of increasing confidence. We discuss possible generalizations of this scheme in the Conclusion.

Results Table 1 shows the in-domain test performance training with the cross-entropy or focal loss, and Table 2 shows model performance in the deployment environment (the CivilComments dataset). We first observe that training with focal loss indeed helps mitigate the problem of class imbalance. However, this appears to come at the cost of uncertainty quality (as measured by ECE) for all methods except SNGP; this behavior persists both in- and out-of-domain. To study this phenomenon in more detail, we visualize the reliability diagram of the uncertainty models trained with and without focal loss in Supplementary Section C. We see that focal loss fundamentally changes the models' uncertainty behavior, systematically shifting the uncertainty curves from overconfidence (the lower right, below the diagonal) and toward the calibration line (the diagonal). However, the exact pattern of change is model dependent. We find that the deterministic model with focal loss is over-confident for predictions under 0.5, but under-confident for those above 0.5, while the SNGP models are still over-confident, although to a lesser degree compared to using cross-entropy loss.

	Model (Test)	AUC-ROC ↑	AUC-PR ↑	Acc. ↑	ECE ↓	Brier ↓	OCA (@0.01/0.05/0.10/0.15/0.20) ↑
XENT	Deterministic	0.972	0.782	0.922	0.0259	0.0565	0.927/0.944/0.962/0.975/0.985
	SNGP	0.971	0.778	0.924	0.0303	0.0550	0.929/0.947/0.965/0.979/0.986
	MC Dropout	0.971	0.787	0.926	0.0196	0.0505	0.931/0.950/0.969/0.982/0.991
	Ensemble	0.974	0.803	0.922	0.0261	0.0559	0.927/0.945/0.963/0.977/0.986
Focal	Deterministic	0.973	0.790	0.949	0.1489	0.0613	0.954/0.970/0.985/0.992/0.996
	SNGP	0.974	0.802	0.948	0.0128	0.0370	0.953/0.970/0.983/0.991/0.997
	MC Dropout	0.971	0.799	0.949	0.1479	0.0626	0.953/0.968/0.982/0.990/0.997
	Ensemble	0.973	0.806	0.948	0.1534	0.0638	0.952/0.968/0.982/0.991/0.996

Table 1: Metrics for models on the Wikipedia Talk Corpus (in-domain), given 10 models in the deterministic ensemble. We report OCA for several different moderator review capacities (fractions of examples that can be sent to moderators). XENT and Focal indicate models trained with the cross-entropy and focal losses, respectively. Arrows indicate which direction is better.

	Model (Deployment)	AUC-ROC ↑	AUC-PR ↑	Acc. ↑	ECE ↓	Brier ↓	OCA (@0.01/0.05/0.10/0.15/0.20) ↑
XENT	Deterministic	0.783	0.673	0.958	0.0129	0.0245	0.963/0.978/0.988/0.994/0.997
	SNGP	0.773	0.669	0.962	0.0084	0.0250	0.966/0.981/0.991/0.996/0.997
	MC Dropout	0.774	0.667	0.964	0.0125	0.0242	0.969/0.983/0.993/0.997/0.998
	Ensemble	0.786	0.676	0.961	0.0126	0.0244	0.966/0.980/0.990/0.994/0.997
Focal	Deterministic	0.804	0.679	0.982	0.1992	0.0372	0.985/0.992/0.996/0.997/0.998
	SNGP	0.802	0.684	0.981	0.0155	0.0278	0.984/0.993/0.996/0.998/0.998
	MC Dropout	0.799	0.679	0.980	0.2423	0.0379	0.985/0.992/0.996/0.997/0.998
	Ensemble	0.806	0.682	0.979	0.2020	0.0388	0.984/0.992/0.996/0.997/0.998

Table 2: Metrics for models on the CivilComments dataset (deployment), given 10 models in the deterministic ensemble. We report OCA for several different moderator review capacities (fractions of examples that can be sent to moderators). XENT and Focal indicate models trained with the cross-entropy and focal losses, respectively. Arrows indicate which direction is better.

A key question is whether classic uncertainty metrics like ECE reflect good model-moderator collaborative efficiency. To understand this, we compute the fraction of comments sent by the model to a moderator for which the model decision was overturned as a function of review bandwidth. We show this in Supplementary Section D for the models trained with cross-entropy in-domain, computed from Table 1. MC Dropout makes the most efficient use of moderator decisions across the range of bandwidths; it is also the best-calibrated model trained with cross-entropy (lowest ECE/Brier score) in Table 1, and moreover has the highest accuracy. However, low ECE does not directly translate to high OCA: the behavior at the extremes (predictions with maximum uncertainty) is most important for OCA, whereas e.g. ECE represents an average over the full dataset. Overall, the performance differences between models are small, and training with focal loss yields the greatest improvements.

4 Conclusion

In this work, we presented *CoToMoD*, a challenging benchmark for evaluating the practical effectiveness of collaborative (model-moderator) content moderation systems, along with the *Oracle-Model Collaborative Accuracy* (OCA), a rigorous metric for the effectiveness of a human-AI collaborative system that is distinct from classic measures of predictive performance or uncertainty calibration. Using CoToMoD, we evaluated a variety of different models on the content moderation task, and found that well-calibrated models perform better in terms of OCA than the model accuracy indicates, and discuss the link between calibration and OCA. Further theoretic development and qualitative study is needed to explore this connection in greater depth.

There exist three important future directions for this work. The first is to generalize OCA to "soft" (non-binary) labels, since in realistic moderation practice, it is common to use the fraction of annotators that rated an example toxic as the training label (Aroyo & Welty, 2015; Aroyo et al., 2019). The second is to evaluate the efficacy of uncertainty-based moderation policies under more realistic scenarios, such as sending a constant number rather than a fixed fraction of examples to the moderator. Finally, we could explore hybrid strategies for having the moderator review comments from the model, e.g. combining model uncertainty scores with other measures.

REFERENCES

Robert Adragna, Elliot Creager, David Madras, and Richard Zemel. Fairness and robustness in invariant learning: A case study in toxicity classification. *arXiv preprint arXiv:2011.06485*, 2020. URL `https://arxiv.org/abs/2011.06485`.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016. URL `https://arxiv.org/abs/1606.06565`.

Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, Mar. 2015. doi: 10.1609/aimag.v36i1.2564. URL `https://ojs.aaai.org/index.php/aimagazine/article/view/2564`.

Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. *Crowdsourcing Subjective Tasks: The Case Study of Understanding Toxicity in Online Discussions*, pp. 1100–1105. Association for Computing Machinery, New York, NY, USA, 2019. ISBN 9781450366755. URL `https://doi.org/10.1145/3308560.3317083`.

Abdullah Awaysheh, Jeffrey Wilcke, François Elvinger, Loren Rees, Weiguo Fan, and Kurt L. Zimmerman. Review of Medical Decision Support and Machine-Learning Methods. *Veterinary Pathology*, 56(4):512–525, July 2019. ISSN 1544-2217. doi: 10.1177/0300985819829524.

Peter L. Bartlett and Marten H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(59):1823–1840, 2008. URL `http://jmlr.org/papers/v9/bartlett08a.html`.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. *CoRR*, abs/1903.04561, 2019. URL `http://arxiv.org/abs/1903.04561`.

Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pp. 1217–1230, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450343350. doi: 10.1145/2998181.2998213. URL `https://doi.org/10.1145/2998181.2998213`.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In Hans Ulrich Simon, Sandra Zilles, and Ronald Ortner (eds.), *Algorithmic Learning Theory - 27th International Conference, ALT 2016, Proceedings*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 67–82. Springer Verlag, 2016. ISBN 9783319463780. doi: 10.1007/978-3-319-46379-7_5. 27th International Conference on Algorithmic Learning Theory, ALT 2016 ; Conference date: 19-10-2016 Through 21-10-2016.

Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, and Scott Yang. Online learning with abstention. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1059–1067, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL `http://proceedings.mlr.press/v80/cortes18a.html`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://www.aclweb.org/anthology/N19-1423`.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pp. 67–73, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278729. URL `https://doi.org/10.1145/3278721.3278729`.

Michael W. Dusenberry, Dustin Tran, Edward Choi, Jonas Kemp, Jeremy Nixon, Ghassen Jerfel, Katherine Heller, and Andrew M. Dai. Analyzing the role of model uncertainty for electronic health records. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, pp. 204–213, New York, NY, USA, April 2020. Association for Computing Machinery. ISBN 978-1-4503-7046-2. doi: 10.1145/3368555.3384457. URL `https://doi.org/10.1145/3368555.3384457`.

Bassey Etim. The times sharply increases articles open for comments, using google's technology. *The New York Times*, 13, 2017.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `http://proceedings.mlr.press/v48/gal16.html`.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=Hkg4TI9xl`.

Andrew Jesson, Sören Mindermann, Uri Shalit, and Yarin Gal. Identifying Causal-Effect Inference Failure with Uncertainty-Aware Models. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 11637–11649. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/860b37e28ec7ba614f00f9246949561d-Paper.pdf`.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, et al. Wilds: A benchmark of in-the-wild distribution shifts. *arXiv preprint arXiv:2012.07421*, 2020. URL `https://arxiv.org/abs/2012.07421`.

Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. Second opinion needed: communicating uncertainty in medical machine learning. *npj Digital Medicine*, 4(1):4, Jan 2021. ISSN 2398-6352. doi: 10.1038/s41746-020-00367-3. URL `https://doi.org/10.1038/s41746-020-00367-3`.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf`.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7498–7512. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/543e83748234f7cbab21aa0ade66565f-Paper.pdf`.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15288–15299. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/aeb7b30ef1d024a76f21a1d40e30c302-Paper.pdf`.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pp. 2901–2907. AAAI Press, 2015. ISBN 0262511290.

Lee Rainie, Janna Anderson, and Jonathan Albright. The Future of Free Speech, Trolls, Anonymity and Fake News Online, March 2017. URL https://www.pewresearch.org/internet/2017/03/29/the-future-of-free-speech-trolls-anonymity-and-fake-news-online/.

Zhao Wang and Aron Culotta. Identifying spurious correlations for robust text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3431–3440, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.308. URL https://www.aclweb.org/anthology/2020.findings-emnlp.308.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pp. 1391–1399, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. ISBN 9781450349130. doi: 10.1145/3038912.3052591. URL https://doi.org/10.1145/3038912.3052591.

A Expected calibration error

For completeness, we include a definition of the expected calibration error (ECE) (Naeini et al., 2015) here.

ECE can be computed by discretizes the probability range $[0, 1]$ into a set of B bins, and computes the weighted average of the difference between confidence (the mean probability within each bin) and the accuracy (the fraction of predictions within each bin that are correct),

$$\text{ECE} = \sum_{b=1}^{B} \frac{n_b}{N} |\text{conf}(b) - \text{acc}(b)|, \tag{1}$$

where $\text{acc}(b)$ and $\text{conf}(b)$ denote the accuracy and confidence for bin b, respectively, n_b is the number of examples in bin b, and $N = \sum_b n_b$ is the total number of examples.

B Related work

OCA draws on the idea of classification with a reject option, or learning with abstention (Bartlett & Wegkamp, 2008; Cortes et al., 2016; 2018; Kompa et al., 2021). In this classification scenario, the model has the option to reject an example instead of predicting its label. The challenge in connecting learning with abstention to OCA is to account for how many examples have already been rejected. The difficulty is that OCA is a dataset-level metric, i.e. the "reject" option is not at the level of individual examples. Moreover, this means OCA can be compared directly with traditional accuracy measures. This difference in focus enables us to consider human time as the limiting resource in the overall model-moderator system's performance.

Robustness to distribution shift has been applied to toxicity classification in other works (Adragna et al., 2020; Koh et al., 2020), emphasizing the connection between fairness and robustness. Our work focuses on how these methods connect to the human review process, and how uncertainty can lead to better decision-making for a model collaborating with a human. Dusenberry et al. (2020) analyzed how uncertainty affects optimal decisions in a medical context, though again at the individual example (rather than dataset) level.

C In-domain and Out-of-domain Reliability Diagrams for Deterministic and SNGP models

We plot the reliability diagrams for deterministic models and SNGP models with cross-entropy and focal cross-entropy. Figure 2 shows the reliability diagram in-domain and Figure 3 shows it out-of-domain.

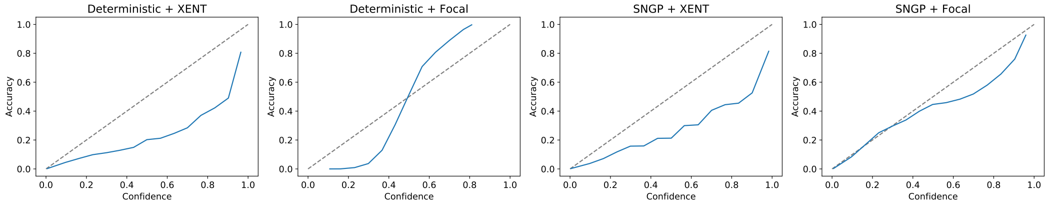

Figure 2: In-domain reliability diagrams for deterministic models and SNGP models with cross-entropy (XENT) and focal cross-entropy.

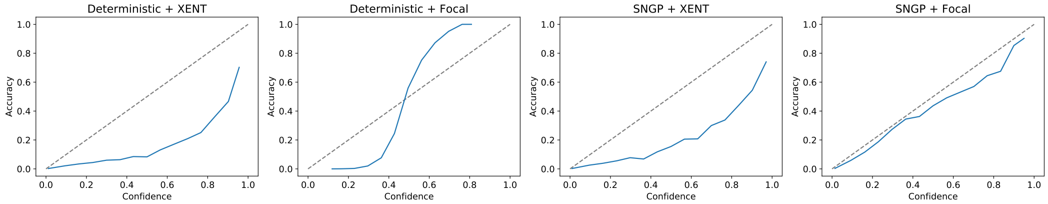

Figure 3: Reliability diagrams for deterministic models and SNGP models with cross-entropy (XENT) and focal cross-entropy on the CivilComments dataset.

D Moderation review efficiency

We plot the moderator review efficiency as a function of capacity for the models trained with cross-entropy in-domain in Figure 4, using the results shown in Table 1 of the main paper. In the limit of low moderator review capacity (the fraction of comments the moderator can review), all models have approximately 50% review efficiency; this decreases with increasing capacity.

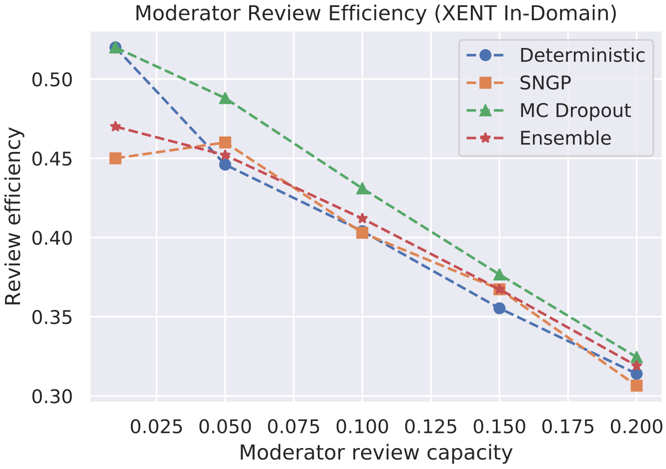

Figure 4: Moderator review efficiency (fraction of decisions the model sent to a moderator that were overturned) as a function of review capacity. All models were trained with cross-entropy and tested in-domain. MC Dropout makes most efficient use of moderator decisions for all tested moderator review capacities; from Table 1, it also has the best ECE. Dotted lines are to guide the eye.