Investigating the Multilingual Calibration Effects of Language Model Instruction-Tuning

Jerry Huang[♣♠♡*] **Peng Lu**[♠] **Qiuhao Zeng**[♢]
Yusuke Iwasawa[♡] **Yutaka Matsuo**[♡] **Sarath Chandar**[♣✠■]
Edison Marrese-Taylor[♡★†] **Irene Li**[♡††]
[♣]Mila - Quebec AI Institute [♠]Université de Montréal
[♡]The University of Tokyo [♢]Western University
[✠]Polytechnique Montréal [■]CIFAR AI Chair [★]AIST
[†]**Corresponding Author:** irene.li@weblab.t.u-tokyo.ac.jp

Abstract

Ensuring that deep learning models are well-calibrated in terms of their predictive uncertainty is essential in maintaining their trustworthiness and reliability, yet despite increasing advances in foundation model research, the relationship between such large language models (LLMs) and their calibration remains an open area of research. In this work, we look at a critical gap in the calibration of LLMs within multilingual settings, in an attempt to better understand how the data scarcity can potentially lead to different calibration effects and how commonly used techniques can apply in these settings. Our analysis on two multilingual benchmarks, over 29 and 42 languages respectively, reveals that even in low-resource languages, model confidence can increase significantly after instruction-tuning on high-resource language SFT datasets. However, improvements in accuracy are marginal or non-existent, resulting in mis-calibration, highlighting a critical shortcoming of standard SFT for multilingual languages. Furthermore, we observe that the use of label smoothing to be a reasonable method alleviate this concern, again without any need for low-resource SFT data, maintaining better calibration across all languages. Overall, this highlights the importance of multilingual considerations for both training and tuning LLMs in order to improve their reliability and fairness in downstream use.

1 Introduction

Tremendous progress has been made in building models that follow natural language instructions (Ouyang et al., 2022; Sanh et al., 2022; Chung et al., 2024) through the use of LLMs pre-trained on large amounts of data as well as high-quality datasets that enable them to learn to interact in a human-like manner (Bach et al., 2022; Wang et al., 2022, 2023a). However, such models have demonstrated a propensity for over-confidence in their predictions (Zhao et al., 2021; Jiang et al., 2021; Xiong et al., 2024), eliciting concerns over their use in more high-stakes decision-making scenarios. Such observations are not new with respect to neural networks, which have consistently been shown to suffer from over-confident predictions and over-estimate the likelihood of their correctness (Guo et al., 2017; Szegedy et al., 2016; Müller et al., 2019; Naeini et al., 2015; Minderer et al., 2021). To improve this, methods such as temperature scaling (Guo et al., 2017) and label smoothing (Müller et al., 2019) have been proposed as solutions with varying effectiveness, spurring additional work in ensuring that predictions and confidence remain matching (Lin et al., 2017; Mukhoti et al., 2020; Pereyra et al., 2017; Liu et al., 2022).

Investigations into the calibration of LLMs remains limited (Zhao et al., 2021), but studies have shown them to behave similarly to prototypical networks in terms of calibration (Huang et al., 2025a). However, LLMs remain unique in their application, one in particular is their ability to be used on different languages that can vary resource scarcity. This raises an interesting question:

Do solutions to calibrating LLMs act on multiple languages simultaneously?

In this work, we provide a preliminary empirical answer to this question. Our work and contributions are summarized as follows:

a) We evaluate a series of evaluations on multiple multilingual classification tasks, evaluating whether or not the downstream calibration of LLMs remains unchanged between languages.

b) Interestingly, all languages exhibit an interesting characteristic: instruction-tuned models are less calibrated than their base counterparts. Yet an additional observation comes on those not used for instruction-tuning, where confidence

[*]Work done as a visiting student at the University of Tokyo.
[†]Equal supervision.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 1–59
March 24-29, 2026

increases without improvements in accuracy, leading to greater overconfidence.

c) Using label smoothing only on the instruction-tuning data, we observe both better calibration and minimal drops in accuracy compared to models not using smoothing.

Overall, our work demonstrates a simple yet effective solution for ensuring better calibration for multilingual language models, potentially opening a path towards better methods for ensuring their robustness and reliability for downstream use.

2 Related Work

Uncertainty Calibration. Uncertainty calibration (Brier, 1950; Murphy, 1972; DeGroot and Fienberg, 1983) attempts to match the prediction probabilities yielded for different inputs to the expected accuracy on these inputs. Widely used metrics for measuring such properties include the expected calibration error (ECE) (Naeini et al., 2015). Additional metrics that have been proposed include the Root Mean Square (Hendrycks et al., 2019) and Static/Adaptive Calibration Errors (SCE/ACE) (Nixon et al., 2019), offering complementary perspectives that enable a more comprehensive assessment of uncertainty alignment.

Nevertheless, calibration of LLMs remains underexplored, particularly from a statistical perspective. Zhao et al. (2021) first show a general lack of calibration in models, with a simple solution being to prompt models using additional content-free in-context samples (i.e. examples with input "N/A") and calibration parameters. Huang et al. (2025a) meanwhile show that instruction-tuning itself can lead to a significant loss in calibration from a base model, with a proposed solution being the use of label smoothing during training.

Multilingual LLMs. Multilingual NLP have greatly evolved beyond the initial English-centric paradigm to address the linguistic diversity of our world. Recent LLMs (Grattafiori et al., 2024; Yang et al., 2024; OpenAI, 2023; Gemma Team, 2024) have demonstrated remarkable multilingual capabilities by leveraging massive pre-training datasets spanning dozens to hundreds of languages. However, research indicates persistent challenges in these systems, particularly cases where models internally process non-English inputs through English-like representations, and consistent performance gaps between high-resource and low-resource languages (Zhong et al., 2024).

3 Methodology

Models and Setup. To investigate our hypothesis, we train and test the `Mistral-7B` (Jiang et al., 2023), `Llama3.1-8B` (Grattafiori et al., 2024) and `Gemma2-2B` (Gemma Team, 2024) models. We tune on the `Alpaca` (Dubois et al., 2023), `Tulu3Mixture` (Wang et al., 2023a) and `OpenHermes` (Teknium, 2023) datasets, using recommendations from Wang et al. (2023a). We employ the AdamW optimizer using a grid search over learning rates $\{$`5e-6`, `1e-5`, `5e-5`, `1e-4`$\}$, linear warm-up over the first 2% of training, a batch size of 128 and dropout 0.1.

Tasks. To evaluate calibration fairly across multiple languages, we use the **`MMLU-ProX`** dataset (Xuan et al., 2025) and **`GlobalMMLU`** (Singh et al., 2024), two comprehensive benchmark covering diverse sets of languages (29 and 42 respectively), built upon the English-only MMLU dataset (Hendrycks et al., 2021). Each language version consists of 12k and 14k identical questions, enabling more direct cross-linguistic comparisons. For each question, an input is given and the model must select between K different candidate answers (up to 10 for **`MMLU-ProX`** and 4 for **`GlobalMMLU`**). More specifically, the perplexities over candidate generations is computed and normalized to form a probability distribution over the different options, which is then used for classification.

4 Results

4.1 Calibration Across Languages

The first question we attempt to answer is whether or not LLM calibration is only relevant for the languages that appear in LLM training data. For this purpose, we conduct an initial validation by using a number of open models and their instruction-tuned variants and evaluating on a number of languages. This is illustrated in Figure 1. Interestingly, we observe a consistent phenomenon across all languages: instruction-tuning leads to worsening calibration across all languages, even those which are unlikely to appear in the tuning datasets.

4.2 Accounting For Over-Confidence

With over-confidence becoming an issue after tuning, this enables us to pose a further question: *How can we reduce the risk of over-confidence without*

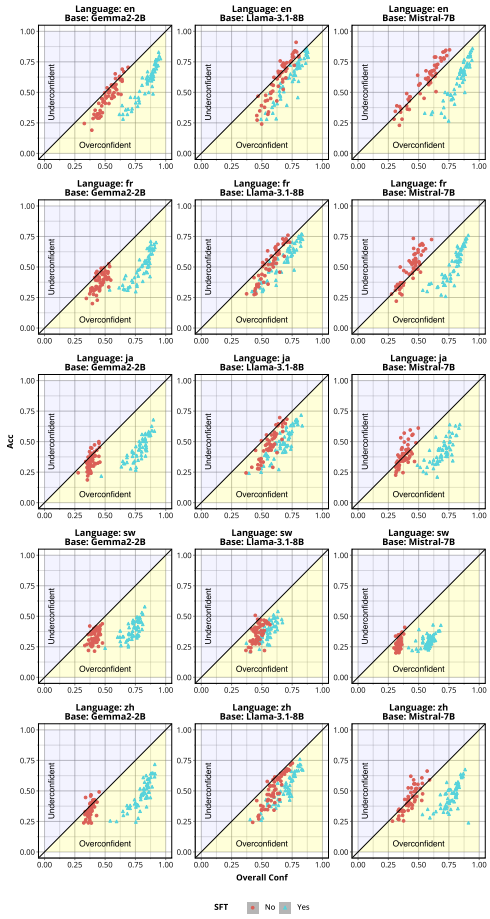

Figure 1: Comparison of **Base** models (red) and instruction-tuned (blue) models on various languages on **GlobalMMLU**. Deviation from the straight line indicates under or over-confidence; models show increasing overconfidence across all languages.

performance degradations? To this end, we can look towards label smoothing as a solution.

Label smoothing (LS) has previously been demonstrated to be a promising paradigm in settings to prevent models from becoming overconfident (Szegedy et al., 2016; Müller et al., 2019) or when noise exists in the provided labels (Lukasik et al., 2020; Wei et al., 2022b; Lu et al., 2023). Consider a model parameterized by $\boldsymbol{\theta}$ to output a conditional distribution $P(\cdot|\boldsymbol{x};\boldsymbol{\theta})$ over a label set. Models are usually trained by minimizing a cross-entropy (CE) loss on a dataset $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^N$ sampled from an unknown distribution $p(\boldsymbol{x}, y)$,

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}^{\text{CE}}(\boldsymbol{\theta}) &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\delta_{y_n}^{\gamma_k}\log P(\gamma_k|\boldsymbol{x};\boldsymbol{\theta}) \\ &\approx -\mathbb{E}_{p(\boldsymbol{x},y)}\left[\sum_{k=1}^{K} p(\gamma_k|\boldsymbol{x})\log P(\gamma_k|\boldsymbol{x};\boldsymbol{\theta})\right] \\ &= -\mathbb{E}_{p(\boldsymbol{x},y)}[\text{KL}\left[\boldsymbol{\sigma}(\boldsymbol{x})\|\widehat{\boldsymbol{\sigma}}(\boldsymbol{x};\boldsymbol{\theta})\right]] + c \\ &= \mathcal{L}_{p(\boldsymbol{x},y)}^{\text{CE}}(\boldsymbol{\theta}),\end{aligned}$$

where δ_i^j is the Kronecker delta and $\widehat{\boldsymbol{\sigma}}(\boldsymbol{x};\boldsymbol{\theta}) \in [0,1]^K$ is the output distribution. Label smoothing mixes the original distribution with a discrete uniform distribution $\mathcal{U} = [1/K]^K \in \mathbb{R}^K$ using a smoothing rate $\beta \in [0,1]$. Thus, label smoothing can be understood to regularize towards a uniform distribution over the output labels, preventing overfitting by encouraging a higher entropy and can therefore become particularly effective for countering overconfidence. This claim is illustrated in Section B.1. Furthermore, this can be viewed as a constrained optimization problem that encourages equality among the logits for each class to ensure that over-confidence is penalized (Section B.2).

Can LS help with Multilingual Calibration? A natural question is whether or not LS is sufficient to help with multilingual calibration as a whole.Table 1 shows that models using smoothing become on average much better calibrated across both tasks, while any decrease in accuracy remains minimal (≤ 0.005 accuracy points). This is further consistent across individual languages (Section C), highlighting the versatility it has for maintaining more robust and reliable language models. However, high amounts of smoothing can also lead to some decreases in accuracy, highlighting a need to properly account for such a hyperparameter, either in a fixed or adaptive manner.

Figure 2 offers a specific illustration on the Yoruba language, which is not present within tuning data; while accuracy does not change (no vertical shift), confidence increases (horizontal shift towards the right), leading to greater miscalibration. Label smoothing visibly mitigates this over-confidence without any change in accuracy.

5 Discussion

Why does SFT Lead to Mis-Calibration? Oh et al. (2024) offer a perspective from the lens of out-of-distribution (OOD) generalization. To better

Table 1: Result of instruction-tuning various LLMs with and without label smoothing. Using smoothing can reduce calibration error (ECE and RMS) for all model/dataset combinations, with minimal decrease in accuracy. The better performing setting is **bolded** and further ***italicized*** when the gap is statistically significant (using a student's t-test).

Model	SFT Dataset	Smoothing	MMLU-ProX Accuracy	MMLU-ProX Entropy	MMLU-ProX ECE	MMLU-ProX RMS	GlobalMMLU Accuracy	GlobalMMLU Entropy	GlobalMMLU ECE	GlobalMMLU RMS
Gemma2-2B	Base Model		0.137	1.361	0.040	0.064	0.326	1.787	0.051	0.057
	OpenHermes	0.0	**0.253**	0.945	0.171	0.126	**0.372**	0.949	0.317	0.071
		0.1	**0.253**	1.324	***0.087***	***0.099***	0.370	1.673	***0.044***	***0.054***
	Tulu3Mixture	0.0	**0.223**	1.185	0.100	0.097	0.359	1.247	0.185	0.064
		0.1	0.222	1.246	***0.059***	***0.087***	**0.360**	1.717	***0.043***	***0.049***
Mistral-7B	Base Model		0.238	1.498	0.150	0.118	0.354	1.855	0.041	0.041
	OpenHermes	0.0	**0.275**	1.275	0.132	0.116	0.372	1.561	0.053	0.036
		0.1	0.273	1.360	***0.113***	**0.113**	**0.374**	1.699	***0.043***	***0.030***
	Tulu3Mixture	0.0	0.312	1.160	0.129	0.123	0.391	1.467	0.066	0.041
		0.1	***0.319***	1.297	***0.107***	***0.117***	**0.394**	1.568	***0.046***	***0.035***
LLama-3.1-8B	Base Model		0.192	0.391	0.143	0.119	0.431	1.593	0.023	0.031
	OpenHermes	0.0	0.259	0.435	0.159	0.131	**0.439**	1.345	0.071	0.041
		0.1	***0.350***	0.831	***0.153***	***0.123***	0.438	1.521	***0.032***	***0.030***
	Tulu3Mixture	0.0	0.296	0.583	0.163	0.127	0.440	1.363	0.073	0.041
		0.1	***0.315***	0.772	***0.148***	**0.126**	**0.441**	1.489	***0.038***	***0.034***

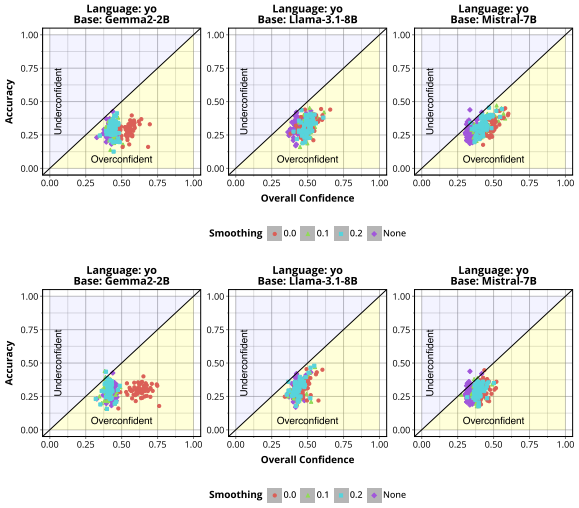

Figure 2: Reliability diagrams for the yo split of **GlobalMMLU** after instruction-tuning on **Tulu3Mixture** (top) and **OpenHermes** (bottom).

understand this, the SFT data can be interpreted as consisting of an in-distribution (ID) dataset whereas the downstream dataset on which generalization and calibration are tested constitutes an OOD dataset. Under such a setting, simultaneously maintaining accuracy and calibration of the final classifier (in the case of an auto-regressive LLM, this is the language modeling head) has a direct relationship to the diversity of the feature embeddings. In particular, they show that there exists a dependence of the bound on the minimal singular value of the covariance matrix, indicating that as the set of learnt feature embeddings (the embedding of the prompt in this context) becomes less mutually dependent, both calibration error and classification error can be minimized. Prior works have shown that fine-tuning can significantly reduce the diversity of such features (Mukhoti et al., 2024; Kumar et al., 2022; Huh et al., 2024), hinting why SFT can significantly degrade calibration.

On Difference Confidence Measures for LLMs. In additional to statistical properties of models, LLMs have the additional axis of being able to be directly prompted for their internal confidence measures, called *verbalized confidence* (Geng et al., 2024). However, such approaches often rely on the LLM producing an internalized measure of confidence (Tao et al., 2024), for example a raw value $c \in [0, 1]$, for which it remains poorly understood whether or not such responses are true measures of the model confidence or whether or not such values can be trusted. overall, this motivates further needs for better understanding confidence/uncertainty calibration in LLMs, both from statistical as well as user-facing perspectives.

6 Conclusion

Through controlled tuning experiments with various LLMs, we reveal a surprising finding where LLMs become significantly uncalibrated on low-resource language predictions even without having observed them during this tuning. In particular,

models demonstrate an innate tendency to become overconfident, even without improvements in performance, a particular setting which enables the use of label smoothing as a generalized solution. In sum, our research reveals that language models can observe adverse downstream effects in subtle manners unobservable through traditional evaluation, and despite the possibility of solutions for mitigating these concerns, it highlights a need to better consider how datasets are used for training and how to adequately evaluate models.

Limitations

Lack of Explicit Solutions

As an empirical study, our work only investigates the existence of different calibration phenomena depending on the input language. As such, we do not provide a whole-sale solution to calibrating models on any arbitrary language.

Ethical Considerations

This paper proposes a method to improve calibration in large-vocabulary language models. We anticipate minimal societal impact or ethical concerns that may stem from the findings of this work.

Acknowledgments

This work was supported by JST ACT-X (Grant JPMJAX24CU) and JSPS KAKENHI (Grant 24K20832). Jerry Huang is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Canada Graduate Scholarships (reference number 589326) as well as the Japan Society for the Promotion of Science (JSPS) Postdoctoral Fellowship for Research in Japan program (reference number SP25408). Sarath Chandar is supported by a Canada CIFAR AI Chair, the Canada Research Chair in Lifelong Machine Learning and a NSERC Discovery Grant. The authors thank Yufei Cui for discussions that prompted the initial direction of this work. This work was made possible in part thanks to computational resources from Calcul Québec[1], the Digital Research Alliance of Canada (DRAC)[2] and the AI Bridging Cloud Infrastructure (ABCI)[3].

[1] https://www.calculquebec.ca/
[2] https://alliancecan.ca/en
[3] https://abci.ai/ja/

References

Stephen H. Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M. Saiful Bari, Thibault Févry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, and 8 others. 2022. Promptsource: An integrated development environment and repository for natural language prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022 - System Demonstrations, Dublin, Ireland, May 22-27, 2022*, pages 93–104.

Nematollah Kayhan Batmanghelich, Gerald T. Quon, Alex Kulesza, Manolis Kellis, Polina Golland, and Luke Bornn. 2014. Diversifying sparsity using variational determinantal point processes.

Dimitri P. Bertsekas. 1999. *Nonlinear Programming*.

Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Ta-Chung Chi, Ting-Han Fan, and Alexander Rudnicky. 2024. Attention alignment and flexible positional embeddings improve transformer length extrapolation. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 132–148.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. Scaling instruction-finetuned language models. *Journal on Machine Learning Research*, 25:70:1–70:53.

Tri Dao. 2024. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Morris H. DeGroot and Stephen E. Fienberg. 1983. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(1):12–22.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacafarm: A simulation framework for methods that

learn from human feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6577–6595. Association for Computational Linguistics.

Asish Ghoshal, Xilun Chen, Sonal Gupta, Luke Zettlemoyer, and Yashar Mehdad. 2021. Learning better structured representations using low-rank adaptive label smoothing. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Aaron Grattafiori and 1 others. 2024. The llama 3 herd of models.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. 2019. Deep anomaly detection with outlier exposure. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Pin-Lun Hsu, Yun Dai, Vignesh Kothapalli, Qingquan Song, Shao Tang, Siyu Zhu, Steven Shimizu, Shivam Sahni, Haowen Ning, and Yanning Chen. 2024. Liger kernel: Efficient triton kernels for llm training.

Jerry Huang. 2025. How well can a long sequence model model long sequences? comparing architectural inductive biases on long-context abilities. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 29–39, Abu Dhabi, UAE. Association for Computational Linguistics.

Jerry Huang, Peng Lu, and Qiuhao Zeng. 2025a. Calibrated language models and how to find them with label smoothing. In *Proceedings of the 42nd International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, 13-19 July 2025*, Proceedings of Machine Learning Research.

Jerry Huang, Prasanna Parthasarathi, Mehdi Rezagholizadeh, and Sarath Chandar. 2024a. Context-aware assistant selection for improved inference acceleration with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5817–5830, Miami, Florida, USA. Association for Computational Linguistics.

Jerry Huang, Prasanna Parthasarathi, Mehdi Rezagholizadeh, and Sarath Chandar. 2024b. Towards practical tool usage for continually learning llms. *Preprint*, arXiv:2404.09339.

Jerry Huang, Prasanna Parthasarathi, Mehdi Rezagholizadeh, Boxing Chen, and Sarath Chandar. 2025b. Do robot snakes dream like electric sheep? investigating the effects of architectural inductive biases on hallucination. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1075–1096, Vienna, Austria. Association for Computational Linguistics.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*.

Like Hui, Mikhail Belkin, and Stephen Wright. 2023. Cut your losses with squentropy. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 14114–14131.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know *When* language models know? on the calibration of language models for question answering. *Trans. Assoc. Comput. Linguistics*, 9:962–977.

Andrew Kerr, Duane Merrill, Julien Demouth, and John Tran. 2017. Cutlass: Fast linear algebra in cuda c++.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Will LeVine, Benjamin Pikus, Pranav Raja, and Fernando Amat Gil. 2023. Enabling calibration in the zero-shot inference of large vision-language models.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007.

Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, and Jose Dolz. 2022. The devil is in the margin: Margin-based label smoothing for network calibration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 80–88.

Peng Lu, Ahmad Rashid, Ivan Kobyzev, Mehdi Rezagholizadeh, and Philippe Langlais. 2023. LABO: towards learning optimal label regularization via bi-level optimization. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5759–5774.

Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. 2020. Does label smoothing mitigate label noise? In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 6448–6458.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 15682–15694.

Margaret Mitchell, Giuseppe Attanasio, Ioana Baldini, Miruna Clinciu, Jordan Clive, Pieter Delobelle, Manan Dey, Sil Hamilton, Timm Dill, Jad Doughman, Ritam Dutt, Avijit Ghosh, Jessica Zosa Forde, Carolin Holtermann, Lucie-Aimée Kaffee, Tanmay Laud, Anne Lauscher, Roberto L Lopez-Davila, Maraim Masoud, and 35 others. 2025. SHADES: Towards a multilingual assessment of stereotypes in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11995–12041, Albuquerque, New Mexico. Association for Computational Linguistics.

Jishnu Mukhoti, Yarin Gal, Philip Torr, and Puneet K. Dokania. 2024. Fine-tuning can cripple your foundation model; preserving features may be the solution. *Transactions on Machine Learning Research*, 2024.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip H. S. Torr, and Puneet K. Dokania. 2020. Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 4696–4705.

Allan H. Murphy. 1972. Scalar and vector partitions of the probability score: Part i. two-state situation. *Journal of Applied Meteorology and Climatology*, 11(2):273–282.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2901–2907.

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring calibration in deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 38–41.

Changdae Oh, Hyesu Lim, Mijoo Kim, Dongyoon Han, Sangdoo Yun, Jaegul Choo, Alexander G Hauptmann, Zhi-Qi Cheng, and Kyungwoo Song. 2024. Towards calibrated robust fine-tuning of vision-language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, December 10-15, 2024, Vancouver, BC, Canada*.

OpenAI. 2023. GPT-4 Technical Report. *CoRR*, abs/2303.08774. ArXiv: 2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*.

Gabriele Prato, Jerry Huang, Prasanna Parthasarathi, Shagun Sodhani, and Sarath Chandar. 2023. EpiK-eval: Evaluation for language models as epistemic models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*,

pages 9523–9557, Singapore. Association for Computational Linguistics.

Gabriele Prato, Jerry Huang, Prasanna Parthasarathi, Shagun Sodhani, and Sarath Chandar. 2024. Do large language models know how much they know? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6054–6070, Miami, Florida, USA. Association for Computational Linguistics.

PyTorch. 2024. torchtune: Pytorch's finetuning library.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, and 21 others. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826.

Shuchang Tao, Liuyi Yao, Hanxing Ding, Yuexiang Xie, Qi Cao, Fei Sun, Jinyang Gao, Huawei Shen, and Bolin Ding. 2024. When to trust llms: Aligning confidence with response quality. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 5984–5996. Association for Computational Linguistics.

Teknium. 2023. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A. Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. 2019. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13888–13899.

Philippe Tillet, Hsiang-Tsung Kung, and David D. Cox. 2019. Triton: an intermediate language and compiler for tiled neural network computations. In *Proceedings of the 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages, MAPL@PLDI 2019, Phoenix, AZ, USA, June 22, 2019*, pages 10–19.

Xinyu Wang, Linrui Ma, Jerry Huang, Peng Lu, Prasanna Parthasarathi, Xiao-Wen Chang, Boxing Chen, and Yufei Cui. 2025. Resona: Improving context copying in linear recurrence models with retrieval. In *Second Conference on Language Modeling*.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023a. How far can camels go? exploring the state of instruction tuning on open resources. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13484–13508.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, and 16 others. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5085–5109.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Jiaheng Wei, Hangyu Liu, Tongliang Liu, Gang Niu, Masashi Sugiyama, and Yang Liu. 2022b. To smooth or not? when label smoothing meets noisy labels. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23589–23614.

Erik Wijmans, Brody Huval, Alexander Hertzberg, Vladlen Koltun, and Philipp Krähenbühl. 2025. Cut

your losses in large-vocabulary language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*.

Ronald J. Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.*, 1(2):270–280.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. abs/1910.03771.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Yun Xing, Junjue Wang, Huitao Li, Xin Li, Kunyu Yu, Nan Liu, Qingyu Chen, Douglas Teodoro, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. 2025. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *CoRR*, arXiv:2503.10497.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 technical report.

QIUHAO Zeng, Jerry Huang, Peng Lu, Gezheng Xu, Boxing Chen, Charles Ling, and Boyu Wang. 2025. ZETA: Leveraging z-order curves for efficient top-k attention. In *The Thirteenth International Conference on Learning Representations*.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M. Susskind. 2023. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 40770–40803.

Chang-Bin Zhang, Peng-Tao Jiang, Qibin Hou, Yunchao Wei, Qi Han, Zhen Li, and Ming-Ming Cheng. 2021. Delving deep into label smoothing. *IEEE Trans. Image Process.*, 30:5984–5996.

Zhilu Zhang and Mert R. Sabuncu. 2020. Self-distillation as instance-specific label smoothing. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2024. Beyond english-centric llms: What language do multilingual language models think in?

Łukasz Rajkowski. 2019. Analysis of the Maximal a Posteriori Partition in the Gaussian Dirichlet Process Mixture Model. *Bayesian Analysis*, 14(2):477 – 494.

A Extended Related Works

A.1 Uncertainty Calibration

In a K-way classification setting, let $\mathcal{X} \in \mathbb{R}^D$ and $\mathcal{Y} \in \{\gamma_k\}_{k=1}^K$ indicate the input and label space, respectively. Let f be a classifier and $f(\widehat{y}|\boldsymbol{x}) = \widehat{c}$ be its confidence, i.e., the maximum of probabilities among K dimensions corresponding to its prediction $\widehat{y}$. A model is *perfectly-calibrated* when

$$P\left(\widehat{y} = y|\widehat{c} = c\right) = c \;\; \forall c \in [0, 1]. \tag{1}$$

Model calibration can be expressed as $\mathbb{E}\left[|P\left(\widehat{y} = y|\widehat{c} = c\right) - c|\right]$.

Expected Calibration Error. Widely used calibration metrics include the expected calibration error (ECE) (Naeini et al., 2015), which divides the confidence scores of N samples into M uniform confidence bins $\{B_m\}_{m=1}^M$ and takes a weighted sum over the bin-wise errors.

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\text{acc}\left(B_m\right) - \text{conf}\left(B_m\right)\right|. \tag{2}$$

Multiclass & Static Calibration Error. Static Calibration Error (SCE), which is a simple extension of Expected Calibration Error to every probability in the multiclass setting. SCE bins predictions separately for each class probability, computes the calibration error within the bin, and averages across bins:

$$\text{SCE} = \frac{1}{K} \sum_{k=1}^{K} \sum_{m=1}^{M} \frac{|B_{mk}|}{N} \left|\text{acc}\left(m, k\right) - \text{conf}\left(m, k\right)\right|. \tag{3}$$

Here, $\text{acc}(m, k)$ and $\text{conf}(m, k)$ are the accuracy and confidence of bin m for class label k, respectively; $|B_{mk}|$ is the number of predictions in bin b for class label k.

Adaptivity & Adaptive Calibration Error. Adaptive calibration ranges are motivated by the bias-variance tradeoff in the choice of ranges, suggesting that in order to get the best estimate of the overall calibration error the metric should focus on the regions where the predictions are made (and focus less on regions with few predictions). This leads to the Adaptive Calibration Error (ACE), which uses an adaptive scheme which spaces the bin intervals so that each contains an equal number of predictions.

In detail, ACE takes as input the predictions P (usually out of a softmax), correct labels, and a number of ranges R.

$$\text{ACE} = \frac{1}{K \cdot R} \sum_{k=1}^{K} \sum_{r=1}^{R} \left|\text{acc}\left(r, k\right) - \text{conf}\left(r, k\right)\right|. \tag{4}$$

Here, $\text{acc}(r, k)$ and $\text{conf}(r, k)$ are the accuracy and confidence of adaptive calibration range r for class label k. Calibration range r defined by the $\lfloor N/R \rfloor$-th index of the sorted and thresholded predictions.

RMS and MAD Calibration Error. The Root Mean Square Calibration Error measures the square root of the expected squared difference between confidence and accuracy at a confidence level. A similar formulation which less severely penalizes large confidence-accuracy deviations is the Mean Absolute Value (MAD), which is a lower bound of the RMS Calibration Error.

To empirically estimate these mis-calibration measures, the N samples of are partitioned again into M bins. Here, bins are not equally spaced since the distribution of confidence values is not uniform but dynamic. Concretely, the RMS Calibration Error is estimated with the numerically stable formula

$$\text{RMSCE} = \sqrt{\sum_{m=1}^{M} \frac{|B_m|}{N} \left(\frac{1}{|B_m|} \sum_{k \in B_i} \mathbb{1}\left(y_k = \widehat{y}_k\right) - \frac{1}{|B_m|} \sum_{k \in B_m} c_k\right)^2}. \tag{5}$$

Along similar lines, the MAD Calibration Error — which is an improper scoring rule due to its use of absolute differences rather than squared differences — is estimated with

$$\text{MAD} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left| \frac{1}{|B_m|} \sum_{k \in B_m} \mathbb{1}\left(y_k = \widehat{y}_k\right) - \frac{1}{|B_m|} \sum_{k \in B_m} c_k \right|. \tag{6}$$

B Proofs

B.1 Proof of Label Smoothing Claim

Remark B.1. Label smoothing can be understood to regularize towards a uniform distribution over the output labels, preventing over-fitting by encouraging a higher entropy and can therefore become particularly effective for countering overconfidence.

Proof. Re-using our notation from Section 4.2, consider a smoothing rate $\beta \in [0, 1]$ and the cross-entropy loss. If label smoothing is used, the loss then becomes

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}}^{\text{LS}}(\boldsymbol{\theta}) &= -\frac{1}{N} \sum_{i=1}^{N} \left[\sum_{k=1}^{K} \left[(1-\beta)\delta_{y_n}^{\gamma_k} + \frac{\beta}{K} \right] \log P(\gamma_k | \boldsymbol{x}; \boldsymbol{\theta}) \right] \\
&= (1-\beta)\mathcal{L}_{\mathcal{D}}^{\text{CE}}(\boldsymbol{\theta}) + \frac{\beta}{K} \sum_{i=1}^{N} \text{KL}[\boldsymbol{u} \| \widehat{\boldsymbol{\sigma}}(\boldsymbol{x}_n; \boldsymbol{\theta})] + c \\
&\approx -\mathbb{E}_{p(\boldsymbol{x},y)} \left[\text{KL}\left[(1-\beta)\boldsymbol{\sigma}(\boldsymbol{x}) + \beta \boldsymbol{u} \| \widehat{\boldsymbol{\sigma}}(\boldsymbol{x}; \boldsymbol{\theta}) \right] \right] + c \\
&= \mathcal{L}_{p(\boldsymbol{x},y)}^{\text{LS}}(\boldsymbol{\theta}).
\end{aligned} \tag{7}$$

□

B.2 Proof of Constraint Claim

To better understand the underlying effects of LS, one can repose its effects from a constraint optimization perspective, where the constraints are imposed from regularization penalties (Bertsekas, 1999). First, define

Definition B.2. The **logit distance** vector for $\boldsymbol{x}$, $\boldsymbol{d}(\boldsymbol{x})$, is

$$\boldsymbol{d}(\boldsymbol{x}) = \left[\max_{1 \le i \le K} \boldsymbol{\ell}(\boldsymbol{x})_i - \boldsymbol{\ell}(\boldsymbol{x})_k \right]_{k=1}^{K} \in \mathbb{R}^K. \tag{8}$$

One way of ensuring that a model does not over-estimate a specific class is to enforce this as a hard constraint, which results in equal logits among all classes and a softmax output of $\boldsymbol{o} = f(\boldsymbol{x}; \boldsymbol{\theta}) = [1/K]^K$. As such, it is often preferable to enforce this as a soft-penalty function $\mathcal{P} : \mathbb{R}^K \to \mathbb{R}$ into the objective function minimized during training. Recalling Equation (7), we can relate this soft-penalty to the additional KL-divergence introduced by the label smoothing objective.

Proposition B.3. *A linear penalty (or a Lagrangian term) for the hard constraint $\boldsymbol{d}(\boldsymbol{x}) = \boldsymbol{0}$ is bounded from above and below by $\text{KL}\left(\boldsymbol{u} \| \widehat{\boldsymbol{\sigma}}\left(\boldsymbol{x}; \boldsymbol{\theta}\right)\right)$, up to additive constants*

$$\text{KL}[\boldsymbol{u} \| \widehat{\boldsymbol{\sigma}}\left(\boldsymbol{x}; \boldsymbol{\theta}\right)] - \log K \le \sum_{i=1}^{K} \frac{\boldsymbol{d}\left(\boldsymbol{x}\right)_i}{K} \le \text{KL}\left[\boldsymbol{u} \| \widehat{\boldsymbol{\sigma}}\left(\boldsymbol{x}; \boldsymbol{\theta}\right)\right]. \tag{9}$$

The proof (in Section B.2) indicates that label smoothing approximately minimizes, for a linear penalty, the constraint $\boldsymbol{d}(\boldsymbol{x}) = \boldsymbol{0}$, encouraging equality among the logits for each class to ensure that over-confidence is penalized.

Proof. Given the KL divergence

$$\mathrm{KL}\left[\boldsymbol{u}\|\widehat{\boldsymbol{\sigma}}\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right]=-\frac{1}{K}\sum_{k=1}^{K}\log P\left(\gamma_i|\boldsymbol{x};\boldsymbol{\theta}\right)+\text{const}$$

we have that

$$\mathrm{KL}\left[\boldsymbol{u}\|\widehat{\boldsymbol{\sigma}}\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right]=-\frac{1}{K}\sum_{k=1}^{K}\log\left(\frac{e^{\boldsymbol{\ell}(\boldsymbol{x};\boldsymbol{\theta})_i}}{\sum_{j=1}^{K}e^{\boldsymbol{\ell}(\boldsymbol{x};\boldsymbol{\theta})_j}}\right)+c=-\frac{1}{K}\sum_{k=1}^{K}\log\left(\sum_{j=1}^{K}e^{\boldsymbol{\ell}(\boldsymbol{x};\boldsymbol{\theta})_j}-\boldsymbol{\ell}\left(\boldsymbol{x};\boldsymbol{\theta}\right)_i\right)+c \tag{10}$$

Considering the property of the LogSumExp (LSE) function, it follows that

$$\max_j\ \boldsymbol{\ell}\left(\boldsymbol{x};\boldsymbol{\theta}\right)_j\leq\log\sum_{j=1}^{K}e^{\boldsymbol{\ell}(\boldsymbol{x};\boldsymbol{\theta})_j}\leq\max_j\ \boldsymbol{\ell}\left(\boldsymbol{x};\boldsymbol{\theta}\right)_j+\log\left(K\right)$$

and

$$\mathrm{KL}\left[\boldsymbol{u}\|\widehat{\boldsymbol{\sigma}}\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right]-\log K\leq-\frac{1}{K}\sum_{k=1}^{K}\left(\max_j\ \boldsymbol{\ell}\left(\boldsymbol{x};\boldsymbol{\theta}\right)_j-\boldsymbol{\ell}\left(\boldsymbol{x};\boldsymbol{\theta}\right)_k\right)\leq\mathrm{KL}\left[\boldsymbol{u}\|\widehat{\boldsymbol{\sigma}}\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right] \tag{11}$$

and given the definition of $\boldsymbol{d}\left(\boldsymbol{x}\right)$, then the additional penalty $\mathrm{KL}\left[\boldsymbol{u}\|\widehat{\boldsymbol{\sigma}}\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right]$ imposed by LS in addition to the standard cross-entropy loss $\mathcal{L}^{\mathrm{CE}}$ is approximately optimizing a linear penalty (or a Lagrangian) for the constraint

$$\boldsymbol{d}\left(\boldsymbol{x}\right)=\boldsymbol{0}$$

to encourage equality of the logits. □

C Complete Results

Table 2: Result of instruction-tuning various LLMs with and without label smoothing. We observe that using smoothing can reduce calibration error (ECE and RMS) for all model/dataset combinations, with minimal decrease in accuracy. Furthermore, smoothing maintains a higher entropy, indicating that the model retains in general greater uncertainty over its predictive distribution.

Model	SFT Dataset	Smoothing	MMLU-ProX				GlobalMMLU			
			Accuracy	Entropy	ECE	RMS	Accuracy	Entropy	ECE	RMS
Gemma2-2B	Base Model		0.137	1.361	0.040	0.064	0.326	1.787	0.051	0.057
Mistral-7B	Base Model		0.238	1.498	0.150	0.118	0.354	1.855	0.041	0.041
Llama-3.1-8B	Base Model		0.192	0.391	0.143	0.119	0.431	1.593	0.023	0.031
Gemma2-2B	Alpaca	0.0	0.194	0.878	0.102	0.109	0.370	1.111	0.242	0.068
		0.1	0.209	0.956	0.088	0.103	0.368	1.302	0.170	0.057
	OpenHermes	0.0	0.253	0.945	0.171	0.126	0.372	0.949	0.317	0.071
		0.1	0.253	1.324	0.087	0.099	0.370	1.673	0.044	0.054
	Tulu3Mixture	0.0	0.223	1.185	0.100	0.097	0.359	1.247	0.185	0.064
		0.1	0.222	1.246	0.059	0.087	0.360	1.717	0.043	0.049
Mistral-7B	Alpaca	0.0	0.236	0.894	0.197	0.135	0.348	1.251	0.192	0.058
		0.1	0.242	1.026	0.172	0.132	0.355	1.290	0.159	0.056
	OpenHermes	0.0	0.275	1.275	0.132	0.116	0.372	1.561	0.053	0.036
		0.1	0.273	1.360	0.113	0.113	0.374	1.699	0.043	0.030
	Tulu3Mixture	0.0	0.312	1.160	0.129	0.123	0.391	1.467	0.066	0.041
		0.1	0.319	1.297	0.107	0.117	0.394	1.568	0.046	0.035
LLama-3.1-8B	Alpaca	0.0	0.208	0.227	0.191	0.133	0.440	1.142	0.162	0.053
		0.1	0.322	0.513	0.166	0.122	0.441	1.227	0.131	0.050
	OpenHermes	0.0	0.259	0.435	0.159	0.131	0.439	1.345	0.071	0.041
		0.1	0.350	0.831	0.153	0.123	0.438	1.521	0.032	0.030
	Tulu3Mixture	0.0	0.296	0.583	0.163	0.127	0.440	1.363	0.073	0.041
		0.1	0.315	0.772	0.148	0.126	0.441	1.489	0.038	0.034

C.1 Individual Languages

The following tables contain the individual language results on both GlobalMMLU and MMLU-ProX datasets.

C.1.1 GlobalMMLU

Table 3: Results on the GlobalMMLU subset for the am language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.268	1.886	0.071	0.091
	Mistral-7B	None	0.256	1.956	0.072	0.087
	Llama-3.1-8B	None	0.292	1.784	0.055	0.085
Alpaca	Gemma2-2B	0.0	0.287	1.484	0.138	0.145
		0.1	0.284	1.538	0.120	0.138
		0.2	0.285	1.607	0.093	0.128
	Mistral-7B	0.0	0.249	1.547	0.144	0.149
		0.1	0.255	1.465	0.204	0.164
		0.2	0.245	1.247	0.321	0.191
	Llama-3.1-8B	0.0	0.305	1.502	0.124	0.135
		0.1	0.309	1.543	0.106	0.127
		0.2	0.306	1.561	0.113	0.127
OpenHermes	Gemma2-2B	0.0	0.281	1.474	0.187	0.156
		0.1	0.286	1.876	0.083	0.094
		0.2	0.285	1.874	0.083	0.094
	Mistral-7B	0.0	0.273	1.789	0.072	0.089
		0.1	0.274	1.876	0.069	0.086
		0.2	0.266	1.869	0.084	0.091
	Llama-3.1-8B	0.0	0.297	1.684	0.064	0.091
		0.1	0.305	1.821	0.046	0.086
		0.2	0.298	1.833	0.061	0.081
Tulu3Mixture	Gemma2-2B	0.0	0.278	1.529	0.148	0.144
		0.1	0.286	1.848	0.076	0.089
		0.2	0.280	1.860	0.085	0.094
	Mistral-7B	0.0	0.280	1.830	0.063	0.085
		0.1	0.279	1.881	0.055	0.083
		0.2	0.281	1.884	0.063	0.083
	Llama-3.1-8B	0.0	0.305	1.703	0.047	0.093
		0.1	0.307	1.807	0.043	0.080
		0.2	0.304	1.823	0.033	0.074

Table 4: Results on the GlobalMMLU subset for the ar language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.329	1.828	0.055	0.080
	Mistral-7B	None	0.305	1.907	0.060	0.083
	Llama-3.1-8B	None	0.453	1.604	0.025	0.067
Alpaca	Gemma2-2B	0.0	0.371	1.069	0.288	0.172
		0.1	0.367	1.274	0.207	0.149
		0.2	0.369	1.302	0.196	0.146
	Mistral-7B	0.0	0.305	1.205	0.331	0.185
		0.1	0.311	1.306	0.216	0.156
		0.2	0.272	1.371	0.176	0.145
	Llama-3.1-8B	0.0	0.447	1.085	0.194	0.142
		0.1	0.446	1.150	0.178	0.137
		0.2	0.442	1.196	0.162	0.130
OpenHermes	Gemma2-2B	0.0	0.376	0.913	0.354	0.184
		0.1	0.369	1.648	0.047	0.085
		0.2	0.369	1.652	0.045	0.086
	Mistral-7B	0.0	0.326	1.677	0.043	0.090
		0.1	0.327	1.778	0.045	0.078
		0.2	0.330	1.758	0.052	0.079
	Llama-3.1-8B	0.0	0.459	1.362	0.084	0.102
		0.1	0.446	1.525	0.027	0.072
		0.2	0.449	1.496	0.032	0.077
Tulu3Mixture	Gemma2-2B	0.0	0.358	1.142	0.259	0.165
		0.1	0.355	1.668	0.056	0.088
		0.2	0.354	1.752	0.041	0.080
	Mistral-7B	0.0	0.359	1.544	0.071	0.107
		0.1	0.360	1.645	0.050	0.090
		0.2	0.361	1.712	0.046	0.081
	Llama-3.1-8B	0.0	0.457	1.393	0.066	0.097
		0.1	0.449	1.472	0.054	0.087
		0.2	0.440	1.517	0.036	0.081

Table 5: Results on the GlobalMMLU subset for the bn language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.301	1.824	0.069	0.087
	Mistral-7B	None	0.277	1.923	0.070	0.087
	Llama-3.1-8B	None	0.375	1.690	0.027	0.074
Alpaca	Gemma2-2B	0.0	0.331	1.210	0.227	0.160
		0.1	0.335	1.413	0.144	0.140
		0.2	0.329	1.439	0.130	0.137
	Mistral-7B	0.0	0.278	1.210	0.310	0.187
		0.1	0.289	1.417	0.181	0.154
		0.2	0.257	1.465	0.153	0.145
	Llama-3.1-8B	0.0	0.384	1.250	0.199	0.147
		0.1	0.390	1.355	0.153	0.129
		0.2	0.388	1.379	0.129	0.123
OpenHermes	Gemma2-2B	0.0	0.332	0.935	0.373	0.189
		0.1	0.333	1.780	0.046	0.081
		0.2	0.332	1.775	0.052	0.079
	Mistral-7B	0.0	0.291	1.713	0.042	0.088
		0.1	0.306	1.859	0.049	0.078
		0.2	0.304	1.834	0.056	0.080
	Llama-3.1-8B	0.0	0.379	1.491	0.086	0.108
		0.1	0.377	1.666	0.029	0.073
		0.2	0.380	1.650	0.026	0.075
Tulu3Mixture	Gemma2-2B	0.0	0.322	1.199	0.222	0.159
		0.1	0.328	1.748	0.060	0.088
		0.2	0.323	1.754	0.057	0.091
	Mistral-7B	0.0	0.329	1.646	0.050	0.100
		0.1	0.325	1.750	0.050	0.083
		0.2	0.325	1.807	0.051	0.081
	Llama-3.1-8B	0.0	0.383	1.559	0.061	0.100
		0.1	0.389	1.653	0.029	0.082
		0.2	0.384	1.645	0.045	0.090

Table 6: Results on the GlobalMMLU subset for the cs language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.353	1.788	0.039	0.076
	Mistral-7B	None	0.425	1.817	0.027	0.071
	Llama-3.1-8B	None	0.479	1.523	0.024	0.065
Alpaca	Gemma2-2B	0.0	0.393	1.015	0.253	0.162
		0.1	0.393	1.260	0.163	0.138
		0.2	0.394	1.286	0.154	0.136
	Mistral-7B	0.0	0.419	1.097	0.196	0.143
		0.1	0.422	1.064	0.231	0.151
		0.2	0.366	1.108	0.285	0.166
	Llama-3.1-8B	0.0	0.488	0.994	0.194	0.141
		0.1	0.486	1.104	0.158	0.130
		0.2	0.484	1.160	0.139	0.120
OpenHermes	Gemma2-2B	0.0	0.407	0.798	0.361	0.184
		0.1	0.399	1.632	0.030	0.078
		0.2	0.395	1.650	0.033	0.077
	Mistral-7B	0.0	0.434	1.385	0.075	0.102
		0.1	0.437	1.600	0.026	0.069
		0.2	0.429	1.589	0.038	0.072
	Llama-3.1-8B	0.0	0.483	1.209	0.104	0.109
		0.1	0.483	1.412	0.038	0.076
		0.2	0.483	1.397	0.041	0.078
Tulu3Mixture	Gemma2-2B	0.0	0.384	1.197	0.180	0.146
		0.1	0.377	1.718	0.028	0.075
		0.2	0.371	1.783	0.037	0.076
	Mistral-7B	0.0	0.456	1.314	0.089	0.107
		0.1	0.446	1.438	0.058	0.091
		0.2	0.443	1.495	0.048	0.083
	Llama-3.1-8B	0.0	0.491	1.276	0.066	0.097
		0.1	0.484	1.392	0.043	0.082
		0.2	0.476	1.433	0.043	0.080

Table 7: Results on the GlobalMMLU subset for the de language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.367	1.744	0.044	0.078
	Mistral-7B	None	0.445	1.773	0.022	0.064
	Llama-3.1-8B	None	0.515	1.440	0.022	0.067
Alpaca	Gemma2-2B	0.0	0.428	0.884	0.264	0.163
		0.1	0.426	1.112	0.204	0.146
		0.2	0.425	1.138	0.204	0.143
	Mistral-7B	0.0	0.448	1.003	0.204	0.145
		0.1	0.453	1.041	0.183	0.141
		0.2	0.399	1.034	0.257	0.161
	Llama-3.1-8B	0.0	0.520	0.942	0.161	0.132
		0.1	0.517	1.068	0.125	0.117
		0.2	0.514	1.125	0.108	0.111
OpenHermes	Gemma2-2B	0.0	0.441	0.734	0.309	0.173
		0.1	0.437	1.516	0.048	0.088
		0.2	0.434	1.535	0.043	0.086
	Mistral-7B	0.0	0.490	1.333	0.052	0.092
		0.1	0.474	1.525	0.022	0.068
		0.2	0.473	1.500	0.025	0.069
	Llama-3.1-8B	0.0	0.518	1.134	0.090	0.105
		0.1	0.518	1.363	0.029	0.073
		0.2	0.521	1.351	0.029	0.073
Tulu3Mixture	Gemma2-2B	0.0	0.416	1.064	0.231	0.153
		0.1	0.408	1.655	0.035	0.080
		0.2	0.401	1.715	0.024	0.074
	Mistral-7B	0.0	0.482	1.269	0.084	0.102
		0.1	0.474	1.409	0.053	0.085
		0.2	0.468	1.472	0.050	0.081
	Llama-3.1-8B	0.0	0.516	1.147	0.091	0.104
		0.1	0.508	1.322	0.048	0.082
		0.2	0.496	1.389	0.035	0.077

Table 8: Results on the GlobalMMLU subset for the el language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.329	1.811	0.048	0.080
	Mistral-7B	None	0.305	1.936	0.034	0.069
	Llama-3.1-8B	None	0.442	1.571	0.027	0.070
Alpaca	Gemma2-2B	0.0	0.372	0.989	0.307	0.174
		0.1	0.368	1.209	0.209	0.154
		0.2	0.365	1.221	0.201	0.153
	Mistral-7B	0.0	0.306	1.224	0.291	0.182
		0.1	0.314	1.358	0.212	0.157
		0.2	0.265	1.444	0.152	0.140
	Llama-3.1-8B	0.0	0.447	1.061	0.185	0.141
		0.1	0.451	1.135	0.156	0.133
		0.2	0.444	1.173	0.151	0.129
OpenHermes	Gemma2-2B	0.0	0.379	0.797	0.380	0.189
		0.1	0.372	1.646	0.050	0.091
		0.2	0.372	1.675	0.045	0.085
	Mistral-7B	0.0	0.314	1.736	0.054	0.085
		0.1	0.302	1.802	0.067	0.085
		0.2	0.308	1.794	0.072	0.087
	Llama-3.1-8B	0.0	0.451	1.291	0.102	0.110
		0.1	0.445	1.467	0.052	0.084
		0.2	0.450	1.445	0.050	0.087
Tulu3Mixture	Gemma2-2B	0.0	0.359	1.177	0.213	0.155
		0.1	0.361	1.693	0.046	0.087
		0.2	0.358	1.768	0.047	0.081
	Mistral-7B	0.0	0.349	1.566	0.081	0.109
		0.1	0.347	1.653	0.053	0.094
		0.2	0.344	1.703	0.051	0.086
	Llama-3.1-8B	0.0	0.459	1.352	0.070	0.102
		0.1	0.449	1.447	0.059	0.090
		0.2	0.438	1.483	0.047	0.086

Table 9: Results on the GlobalMMLU subset for the en language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.461	1.552	0.032	0.069
	Mistral-7B	None	0.576	1.488	0.022	0.066
	Llama-3.1-8B	None	0.603	1.260	0.020	0.065
Alpaca	Gemma2-2B	0.0	0.511	0.768	0.249	0.156
		0.1	0.510	1.100	0.139	0.125
		0.2	0.508	1.100	0.144	0.125
	Mistral-7B	0.0	0.544	0.828	0.234	0.151
		0.1	0.551	0.891	0.158	0.128
		0.2	0.469	0.975	0.164	0.132
	Llama-3.1-8B	0.0	0.611	0.760	0.135	0.121
		0.1	0.614	0.928	0.080	0.098
		0.2	0.612	0.977	0.069	0.092
OpenHermes	Gemma2-2B	0.0	0.529	0.541	0.311	0.171
		0.1	0.523	1.309	0.045	0.086
		0.2	0.527	1.327	0.048	0.083
	Mistral-7B	0.0	0.590	0.993	0.071	0.095
		0.1	0.584	1.261	0.022	0.066
		0.2	0.582	1.236	0.017	0.060
	Llama-3.1-8B	0.0	0.627	0.904	0.074	0.095
		0.1	0.624	1.162	0.014	0.060
		0.2	0.627	1.161	0.021	0.062
Tulu3Mixture	Gemma2-2B	0.0	0.500	0.985	0.174	0.133
		0.1	0.502	1.473	0.043	0.076
		0.2	0.497	1.547	0.032	0.074
	Mistral-7B	0.0	0.578	0.967	0.110	0.107
		0.1	0.574	1.076	0.095	0.102
		0.2	0.566	1.172	0.058	0.083
	Llama-3.1-8B	0.0	0.609	0.927	0.092	0.100
		0.1	0.599	1.110	0.049	0.078
		0.2	0.594	1.206	0.019	0.064

Table 10: Results on the GlobalMMLU subset for the es language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.381	1.723	0.042	0.077
	Mistral-7B	None	0.489	1.700	0.021	0.066
	Llama-3.1-8B	None	0.525	1.410	0.023	0.064
Alpaca	Gemma2-2B	0.0	0.442	0.817	0.276	0.167
		0.1	0.444	1.108	0.193	0.143
		0.2	0.442	1.140	0.181	0.140
	Mistral-7B	0.0	0.465	0.967	0.213	0.145
		0.1	0.473	0.967	0.219	0.147
		0.2	0.412	1.041	0.255	0.159
	Llama-3.1-8B	0.0	0.535	0.915	0.151	0.127
		0.1	0.532	1.025	0.122	0.115
		0.2	0.528	1.078	0.101	0.108
OpenHermes	Gemma2-2B	0.0	0.458	0.671	0.323	0.175
		0.1	0.456	1.429	0.062	0.095
		0.2	0.448	1.455	0.051	0.089
	Mistral-7B	0.0	0.504	1.243	0.066	0.094
		0.1	0.497	1.494	0.021	0.066
		0.2	0.499	1.474	0.019	0.064
	Llama-3.1-8B	0.0	0.548	1.110	0.079	0.097
		0.1	0.543	1.332	0.019	0.065
		0.2	0.543	1.318	0.024	0.069
Tulu3Mixture	Gemma2-2B	0.0	0.437	1.029	0.219	0.148
		0.1	0.434	1.549	0.056	0.091
		0.2	0.427	1.620	0.042	0.080
	Mistral-7B	0.0	0.507	1.149	0.096	0.108
		0.1	0.508	1.347	0.050	0.085
		0.2	0.499	1.393	0.048	0.078
	Llama-3.1-8B	0.0	0.542	1.126	0.079	0.099
		0.1	0.544	1.286	0.052	0.082
		0.2	0.539	1.371	0.029	0.070

Table 11: Results on the GlobalMMLU subset for the fa language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.320	1.851	0.049	0.078
	Mistral-7B	None	0.298	1.914	0.057	0.081
	Llama-3.1-8B	None	0.446	1.605	0.026	0.066
Alpaca	Gemma2-2B	0.0	0.365	1.120	0.256	0.163
		0.1	0.366	1.303	0.178	0.146
		0.2	0.364	1.334	0.173	0.143
	Mistral-7B	0.0	0.300	1.236	0.285	0.180
		0.1	0.306	1.342	0.203	0.155
		0.2	0.263	1.392	0.163	0.144
	Llama-3.1-8B	0.0	0.446	1.112	0.161	0.133
		0.1	0.447	1.187	0.137	0.125
		0.2	0.443	1.224	0.128	0.121
OpenHermes	Gemma2-2B	0.0	0.372	0.957	0.327	0.179
		0.1	0.363	1.701	0.043	0.081
		0.2	0.360	1.684	0.047	0.083
	Mistral-7B	0.0	0.320	1.688	0.043	0.092
		0.1	0.318	1.784	0.048	0.081
		0.2	0.316	1.786	0.051	0.080
	Llama-3.1-8B	0.0	0.448	1.340	0.077	0.104
		0.1	0.442	1.510	0.035	0.078
		0.2	0.443	1.491	0.036	0.079
Tulu3Mixture	Gemma2-2B	0.0	0.354	1.240	0.199	0.151
		0.1	0.353	1.764	0.040	0.076
		0.2	0.348	1.816	0.040	0.075
	Mistral-7B	0.0	0.340	1.573	0.058	0.103
		0.1	0.337	1.641	0.053	0.095
		0.2	0.335	1.750	0.053	0.083
	Llama-3.1-8B	0.0	0.452	1.431	0.052	0.092
		0.1	0.445	1.523	0.036	0.080
		0.2	0.431	1.545	0.035	0.079

Table 12: Results on the GlobalMMLU subset for the fil language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.373	1.735	0.046	0.079
	Mistral-7B	None	0.393	1.820	0.019	0.063
	Llama-3.1-8B	None	0.454	1.558	0.028	0.070
Alpaca	Gemma2-2B	0.0	0.413	0.984	0.265	0.162
		0.1	0.414	1.250	0.165	0.136
		0.2	0.412	1.279	0.148	0.134
	Mistral-7B	0.0	0.356	1.164	0.308	0.177
		0.1	0.365	1.196	0.230	0.153
		0.2	0.303	1.244	0.198	0.146
	Llama-3.1-8B	0.0	0.460	1.077	0.168	0.134
		0.1	0.464	1.166	0.129	0.122
		0.2	0.462	1.211	0.112	0.116
OpenHermes	Gemma2-2B	0.0	0.419	0.748	0.342	0.179
		0.1	0.410	1.621	0.033	0.076
		0.2	0.412	1.610	0.029	0.077
	Mistral-7B	0.0	0.382	1.516	0.064	0.098
		0.1	0.381	1.686	0.037	0.072
		0.2	0.383	1.699	0.043	0.075
	Llama-3.1-8B	0.0	0.459	1.298	0.077	0.102
		0.1	0.459	1.463	0.031	0.076
		0.2	0.460	1.438	0.038	0.079
Tulu3Mixture	Gemma2-2B	0.0	0.405	1.133	0.214	0.149
		0.1	0.402	1.657	0.038	0.081
		0.2	0.399	1.705	0.037	0.076
	Mistral-7B	0.0	0.411	1.317	0.093	0.115
		0.1	0.416	1.440	0.073	0.102
		0.2	0.415	1.552	0.041	0.081
	Llama-3.1-8B	0.0	0.462	1.275	0.108	0.111
		0.1	0.463	1.404	0.084	0.103
		0.2	0.451	1.494	0.049	0.083

Table 13: Results on the GlobalMMLU subset for the fr language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.376	1.728	0.043	0.077
	Mistral-7B	None	0.478	1.757	0.025	0.069
	Llama-3.1-8B	None	0.525	1.446	0.022	0.066
Alpaca	Gemma2-2B	0.0	0.435	0.828	0.301	0.172
		0.1	0.430	1.097	0.225	0.149
		0.2	0.433	1.136	0.192	0.144
	Mistral-7B	0.0	0.459	0.984	0.226	0.147
		0.1	0.467	0.985	0.212	0.144
		0.2	0.419	1.044	0.234	0.155
	Llama-3.1-8B	0.0	0.538	0.919	0.154	0.128
		0.1	0.538	1.040	0.123	0.116
		0.2	0.534	1.085	0.111	0.112
OpenHermes	Gemma2-2B	0.0	0.449	0.670	0.352	0.180
		0.1	0.449	1.466	0.051	0.089
		0.2	0.443	1.496	0.044	0.087
	Mistral-7B	0.0	0.491	1.276	0.067	0.094
		0.1	0.488	1.493	0.035	0.072
		0.2	0.488	1.483	0.031	0.070
	Llama-3.1-8B	0.0	0.540	1.103	0.089	0.102
		0.1	0.538	1.347	0.027	0.068
		0.2	0.545	1.314	0.029	0.070
Tulu3Mixture	Gemma2-2B	0.0	0.425	1.081	0.185	0.145
		0.1	0.422	1.560	0.051	0.090
		0.2	0.415	1.628	0.049	0.083
	Mistral-7B	0.0	0.499	1.165	0.107	0.111
		0.1	0.495	1.327	0.067	0.094
		0.2	0.498	1.387	0.052	0.084
	Llama-3.1-8B	0.0	0.541	1.137	0.079	0.098
		0.1	0.546	1.312	0.042	0.078
		0.2	0.541	1.377	0.026	0.072

Table 14: Results on the GlobalMMLU subset for the ha language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.288	1.817	0.050	0.083
	Mistral-7B	None	0.261	1.929	0.086	0.092
	Llama-3.1-8B	None	0.347	1.759	0.037	0.075
Alpaca	Gemma2-2B	0.0	0.304	1.371	0.210	0.158
		0.1	0.303	1.463	0.184	0.147
		0.2	0.299	1.518	0.155	0.145
	Mistral-7B	0.0	0.271	1.327	0.249	0.171
		0.1	0.278	1.487	0.123	0.128
		0.2	0.255	1.546	0.072	0.111
	Llama-3.1-8B	0.0	0.355	1.386	0.167	0.142
		0.1	0.355	1.414	0.151	0.138
		0.2	0.355	1.447	0.135	0.133
OpenHermes	Gemma2-2B	0.0	0.307	1.248	0.239	0.166
		0.1	0.308	1.846	0.052	0.080
		0.2	0.306	1.851	0.059	0.080
	Mistral-7B	0.0	0.286	1.793	0.073	0.088
		0.1	0.286	1.857	0.060	0.082
		0.2	0.287	1.846	0.076	0.090
	Llama-3.1-8B	0.0	0.338	1.621	0.046	0.090
		0.1	0.336	1.704	0.029	0.073
		0.2	0.338	1.705	0.029	0.074
Tulu3Mixture	Gemma2-2B	0.0	0.307	1.545	0.100	0.127
		0.1	0.306	1.851	0.056	0.082
		0.2	0.305	1.883	0.055	0.081
	Mistral-7B	0.0	0.293	1.757	0.050	0.082
		0.1	0.292	1.807	0.069	0.085
		0.2	0.294	1.834	0.074	0.089
	Llama-3.1-8B	0.0	0.353	1.597	0.050	0.097
		0.1	0.348	1.684	0.029	0.079
		0.2	0.344	1.697	0.026	0.078

Table 15: Results on the GlobalMMLU subset for the he language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.309	1.791	0.062	0.088
	Mistral-7B	None	0.280	1.935	0.066	0.086
	Llama-3.1-8B	None	0.392	1.676	0.028	0.072
Alpaca	Gemma2-2B	0.0	0.352	1.044	0.308	0.177
		0.1	0.352	1.302	0.191	0.150
		0.2	0.346	1.319	0.180	0.148
	Mistral-7B	0.0	0.281	1.122	0.397	0.200
		0.1	0.282	1.372	0.232	0.163
		0.2	0.253	1.475	0.162	0.143
	Llama-3.1-8B	0.0	0.405	1.196	0.192	0.142
		0.1	0.409	1.274	0.142	0.129
		0.2	0.407	1.304	0.132	0.127
OpenHermes	Gemma2-2B	0.0	0.356	0.888	0.369	0.187
		0.1	0.349	1.752	0.043	0.080
		0.2	0.345	1.761	0.049	0.081
	Mistral-7B	0.0	0.304	1.799	0.053	0.083
		0.1	0.300	1.856	0.062	0.082
		0.2	0.304	1.841	0.062	0.083
	Llama-3.1-8B	0.0	0.395	1.467	0.066	0.103
		0.1	0.393	1.620	0.036	0.077
		0.2	0.398	1.598	0.036	0.082
Tulu3Mixture	Gemma2-2B	0.0	0.338	1.229	0.239	0.159
		0.1	0.329	1.782	0.048	0.079
		0.2	0.328	1.840	0.054	0.080
	Mistral-7B	0.0	0.316	1.782	0.050	0.080
		0.1	0.311	1.839	0.057	0.079
		0.2	0.316	1.882	0.045	0.071
	Llama-3.1-8B	0.0	0.411	1.490	0.061	0.099
		0.1	0.406	1.574	0.041	0.087
		0.2	0.396	1.588	0.056	0.090

Table 16: Results on the GlobalMMLU subset for the hi language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.312	1.817	0.055	0.085
	Mistral-7B	None	0.287	1.912	0.067	0.084
	Llama-3.1-8B	None	0.433	1.701	0.019	0.064
Alpaca	Gemma2-2B	0.0	0.363	1.036	0.304	0.175
		0.1	0.355	1.280	0.208	0.153
		0.2	0.357	1.299	0.192	0.150
	Mistral-7B	0.0	0.290	1.183	0.329	0.188
		0.1	0.299	1.389	0.172	0.152
		0.2	0.260	1.423	0.162	0.145
	Llama-3.1-8B	0.0	0.430	1.203	0.145	0.130
		0.1	0.429	1.260	0.137	0.124
		0.2	0.425	1.284	0.131	0.122
OpenHermes	Gemma2-2B	0.0	0.363	0.751	0.417	0.198
		0.1	0.362	1.655	0.056	0.092
		0.2	0.356	1.666	0.045	0.088
	Mistral-7B	0.0	0.304	1.736	0.062	0.080
		0.1	0.307	1.835	0.056	0.079
		0.2	0.304	1.803	0.036	0.082
	Llama-3.1-8B	0.0	0.427	1.424	0.068	0.101
		0.1	0.424	1.589	0.035	0.077
		0.2	0.426	1.565	0.033	0.078
Tulu3Mixture	Gemma2-2B	0.0	0.344	1.130	0.246	0.162
		0.1	0.340	1.712	0.045	0.090
		0.2	0.330	1.730	0.043	0.092
	Mistral-7B	0.0	0.335	1.662	0.067	0.107
		0.1	0.330	1.754	0.027	0.076
		0.2	0.322	1.795	0.051	0.077
	Llama-3.1-8B	0.0	0.430	1.494	0.053	0.091
		0.1	0.419	1.588	0.026	0.077
		0.2	0.411	1.600	0.031	0.078

Table 17: Results on the GlobalMMLU subset for the id language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.351	1.815	0.041	0.076
	Mistral-7B	None	0.397	1.819	0.029	0.068
	Llama-3.1-8B	None	0.492	1.548	0.023	0.065
Alpaca	Gemma2-2B	0.0	0.417	1.049	0.253	0.160
		0.1	0.411	1.255	0.182	0.141
		0.2	0.411	1.284	0.168	0.137
	Mistral-7B	0.0	0.387	1.166	0.272	0.169
		0.1	0.393	1.157	0.208	0.148
		0.2	0.322	1.174	0.192	0.145
	Llama-3.1-8B	0.0	0.495	1.036	0.149	0.127
		0.1	0.499	1.156	0.114	0.112
		0.2	0.496	1.195	0.097	0.107
OpenHermes	Gemma2-2B	0.0	0.427	0.869	0.293	0.169
		0.1	0.429	1.628	0.030	0.074
		0.2	0.424	1.615	0.023	0.076
	Mistral-7B	0.0	0.412	1.473	0.056	0.097
		0.1	0.413	1.646	0.022	0.065
		0.2	0.410	1.645	0.020	0.067
	Llama-3.1-8B	0.0	0.505	1.242	0.073	0.097
		0.1	0.501	1.418	0.028	0.072
		0.2	0.499	1.397	0.027	0.074
Tulu3Mixture	Gemma2-2B	0.0	0.407	1.279	0.135	0.130
		0.1	0.405	1.676	0.034	0.077
		0.2	0.404	1.744	0.032	0.070
	Mistral-7B	0.0	0.445	1.280	0.094	0.113
		0.1	0.440	1.382	0.080	0.105
		0.2	0.440	1.506	0.045	0.084
	Llama-3.1-8B	0.0	0.497	1.275	0.069	0.094
		0.1	0.492	1.434	0.027	0.072
		0.2	0.484	1.479	0.021	0.069

Table 18: Results on the GlobalMMLU subset for the ig language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.289	1.774	0.068	0.093
	Mistral-7B	None	0.267	1.927	0.080	0.090
	Llama-3.1-8B	None	0.356	1.732	0.033	0.076
Alpaca	Gemma2-2B	0.0	0.310	1.300	0.229	0.164
		0.1	0.311	1.422	0.185	0.152
		0.2	0.307	1.480	0.151	0.143
	Mistral-7B	0.0	0.271	1.533	0.279	0.175
		0.1	0.273	1.465	0.141	0.135
		0.2	0.255	1.316	0.082	0.117
	Llama-3.1-8B	0.0	0.363	1.333	0.163	0.137
		0.1	0.364	1.373	0.136	0.131
		0.2	0.363	1.391	0.134	0.128
OpenHermes	Gemma2-2B	0.0	0.306	1.135	0.295	0.176
		0.1	0.302	1.799	0.067	0.087
		0.2	0.304	1.798	0.067	0.089
	Mistral-7B	0.0	0.286	1.769	0.070	0.089
		0.1	0.285	1.840	0.072	0.087
		0.2	0.286	1.829	0.068	0.084
	Llama-3.1-8B	0.0	0.349	1.596	0.051	0.092
		0.1	0.345	1.713	0.044	0.077
		0.2	0.353	1.707	0.036	0.078
Tulu3Mixture	Gemma2-2B	0.0	0.299	1.419	0.148	0.144
		0.1	0.300	1.817	0.050	0.083
		0.2	0.296	1.846	0.058	0.083
	Mistral-7B	0.0	0.305	1.675	0.052	0.094
		0.1	0.307	1.734	0.049	0.091
		0.2	0.300	1.772	0.047	0.097
	Llama-3.1-8B	0.0	0.356	1.480	0.110	0.120
		0.1	0.355	1.579	0.074	0.099
		0.2	0.350	1.628	0.047	0.088

Table 19: Results on the GlobalMMLU subset for the it language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.344	1.668	0.043	0.082
	Mistral-7B	None	0.463	1.764	0.018	0.065
	Llama-3.1-8B	None	0.521	1.461	0.020	0.062
Alpaca	Gemma2-2B	0.0	0.428	0.881	0.276	0.166
		0.1	0.423	1.166	0.173	0.140
		0.2	0.426	1.195	0.167	0.137
	Mistral-7B	0.0	0.445	1.022	0.268	0.164
		0.1	0.450	1.018	0.235	0.152
		0.2	0.382	1.072	0.218	0.148
	Llama-3.1-8B	0.0	0.523	0.945	0.162	0.129
		0.1	0.519	1.050	0.123	0.117
		0.2	0.519	1.095	0.115	0.112
OpenHermes	Gemma2-2B	0.0	0.441	0.705	0.350	0.180
		0.1	0.442	1.460	0.072	0.098
		0.2	0.439	1.488	0.064	0.092
	Mistral-7B	0.0	0.480	1.255	0.077	0.101
		0.1	0.484	1.502	0.033	0.070
		0.2	0.483	1.486	0.025	0.068
	Llama-3.1-8B	0.0	0.536	1.154	0.076	0.098
		0.1	0.533	1.366	0.018	0.067
		0.2	0.534	1.344	0.023	0.070
Tulu3Mixture	Gemma2-2B	0.0	0.425	1.100	0.193	0.143
		0.1	0.422	1.587	0.049	0.085
		0.2	0.317	1.666	0.037	0.077
	Mistral-7B	0.0	0.496	1.171	0.097	0.114
		0.1	0.494	1.318	0.064	0.095
		0.2	0.481	1.395	0.050	0.084
	Llama-3.1-8B	0.0	0.533	1.193	0.068	0.094
		0.1	0.534	1.338	0.031	0.075
		0.2	0.526	1.413	0.023	0.069

Table 20: Results on the GlobalMMLU subset for the ja language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.330	1.865	0.039	0.074
	Mistral-7B	None	0.373	1.867	0.017	0.065
	Llama-3.1-8B	None	0.461	1.459	0.031	0.076
Alpaca	Gemma2-2B	0.0	0.388	1.174	0.190	0.150
		0.1	0.387	1.264	0.172	0.142
		0.2	0.388	1.293	0.163	0.139
	Mistral-7B	0.0	0.364	1.163	0.307	0.175
		0.1	0.370	1.260	0.214	0.152
		0.2	0.316	1.248	0.200	0.151
	Llama-3.1-8B	0.0	0.477	1.007	0.170	0.136
		0.1	0.479	1.111	0.139	0.125
		0.2	0.474	1.180	0.111	0.115
OpenHermes	Gemma2-2B	0.0	0.398	1.086	0.223	0.158
		0.1	0.392	1.658	0.034	0.078
		0.2	0.396	1.660	0.034	0.079
	Mistral-7B	0.0	0.401	1.432	0.072	0.106
		0.1	0.399	1.524	0.053	0.089
		0.2	0.401	1.502	0.050	0.093
	Llama-3.1-8B	0.0	0.473	1.135	0.117	0.120
		0.1	0.472	1.367	0.052	0.085
		0.2	0.475	1.363	0.053	0.089
Tulu3Mixture	Gemma2-2B	0.0	0.372	1.271	0.163	0.142
		0.1	0.360	1.673	0.044	0.088
		0.2	0.358	1.757	0.037	0.080
	Mistral-7B	0.0	0.406	1.373	0.113	0.119
		0.1	0.413	1.441	0.087	0.109
		0.2	0.403	1.525	0.049	0.094
	Llama-3.1-8B	0.0	0.472	1.173	0.103	0.117
		0.1	0.474	1.356	0.060	0.093
		0.2	0.465	1.420	0.058	0.090

Table 21: Results on the GlobalMMLU subset for the ko language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.323	1.846	0.054	0.080
	Mistral-7B	None	0.385	1.836	0.022	0.065
	Llama-3.1-8B	None	0.447	1.462	0.033	0.077
Alpaca	Gemma2-2B	0.0	0.368	1.058	0.276	0.168
		0.1	0.367	1.226	0.202	0.154
		0.2	0.365	1.257	0.188	0.150
	Mistral-7B	0.0	0.362	1.136	0.254	0.162
		0.1	0.367	1.136	0.249	0.159
		0.2	0.328	1.130	0.299	0.173
	Llama-3.1-8B	0.0	0.465	1.046	0.174	0.135
		0.1	0.472	1.186	0.120	0.117
		0.2	0.466	1.247	0.102	0.109
OpenHermes	Gemma2-2B	0.0	0.376	0.903	0.339	0.180
		0.1	0.380	1.658	0.044	0.083
		0.2	0.372	1.666	0.050	0.085
	Mistral-7B	0.0	0.388	1.418	0.080	0.111
		0.1	0.383	1.585	0.048	0.086
		0.2	0.385	1.560	0.050	0.089
	Llama-3.1-8B	0.0	0.468	1.148	0.135	0.123
		0.1	0.466	1.375	0.056	0.091
		0.2	0.465	1.379	0.057	0.091
Tulu3Mixture	Gemma2-2B	0.0	0.354	1.088	0.258	0.165
		0.1	0.350	1.701	0.051	0.090
		0.2	0.342	1.783	0.044	0.080
	Mistral-7B	0.0	0.401	1.306	0.113	0.125
		0.1	0.402	1.426	0.097	0.110
		0.2	0.398	1.548	0.061	0.093
	Llama-3.1-8B	0.0	0.460	1.196	0.141	0.123
		0.1	0.466	1.366	0.088	0.101
		0.2	0.460	1.418	0.080	0.098

Table 22: Results on the GlobalMMLU subset for the ky language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.288	1.834	0.067	0.088
	Mistral-7B	None	0.274	1.936	0.066	0.083
	Llama-3.1-8B	None	0.362	1.556	0.078	0.101
Alpaca	Gemma2-2B	0.0	0.307	1.383	0.159	0.149
		0.1	0.309	1.494	0.124	0.138
		0.2	0.302	1.546	0.112	0.132
	Mistral-7B	0.0	0.280	1.585	0.298	0.179
		0.1	0.291	1.490	0.120	0.133
		0.2	0.264	1.275	0.068	0.113
	Llama-3.1-8B	0.0	0.390	1.188	0.210	0.148
		0.1	0.392	1.307	0.150	0.131
		0.2	0.387	1.361	0.122	0.122
OpenHermes	Gemma2-2B	0.0	0.296	1.345	0.207	0.157
		0.1	0.299	1.849	0.049	0.079
		0.2	0.296	1.859	0.054	0.081
	Mistral-7B	0.0	0.299	1.778	0.057	0.086
		0.1	0.293	1.852	0.070	0.086
		0.2	0.289	1.847	0.074	0.088
	Llama-3.1-8B	0.0	0.380	1.427	0.086	0.111
		0.1	0.379	1.598	0.043	0.082
		0.2	0.381	1.595	0.042	0.085
Tulu3Mixture	Gemma2-2B	0.0	0.293	1.415	0.153	0.145
		0.1	0.288	1.817	0.055	0.083
		0.2	0.287	1.868	0.063	0.085
	Mistral-7B	0.0	0.343	1.699	0.041	0.084
		0.1	0.339	1.783	0.047	0.077
		0.2	0.338	1.811	0.036	0.073
	Llama-3.1-8B	0.0	0.377	1.394	0.119	0.122
		0.1	0.378	1.563	0.049	0.093
		0.2	0.370	1.592	0.044	0.090

Table 23: Results on the GlobalMMLU subset for the lt language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.317	1.792	0.057	0.083
	Mistral-7B	None	0.295	1.919	0.060	0.082
	Llama-3.1-8B	None	0.389	1.636	0.026	0.076
Alpaca	Gemma2-2B	0.0	0.356	1.159	0.233	0.159
		0.1	0.354	1.356	0.151	0.139
		0.2	0.352	1.404	0.131	0.135
	Mistral-7B	0.0	0.295	1.171	0.344	0.190
		0.1	0.309	1.379	0.175	0.147
		0.2	0.257	1.440	0.136	0.135
	Llama-3.1-8B	0.0	0.411	1.182	0.200	0.145
		0.1	0.409	1.291	0.155	0.129
		0.2	0.408	1.336	0.140	0.123
OpenHermes	Gemma2-2B	0.0	0.358	0.958	0.337	0.180
		0.1	0.357	1.716	0.040	0.078
		0.2	0.350	1.731	0.037	0.079
	Mistral-7B	0.0	0.321	1.753	0.069	0.086
		0.1	0.314	1.852	0.056	0.079
		0.2	0.306	1.838	0.049	0.080
	Llama-3.1-8B	0.0	0.398	1.424	0.096	0.109
		0.1	0.393	1.600	0.029	0.074
		0.2	0.401	1.604	0.031	0.074
Tulu3Mixture	Gemma2-2B	0.0	0.347	1.254	0.202	0.153
		0.1	0.340	1.723	0.046	0.083
		0.2	0.330	1.795	0.043	0.077
	Mistral-7B	0.0	0.356	1.605	0.061	0.096
		0.1	0.352	1.694	0.042	0.081
		0.2	0.344	1.716	0.052	0.080
	Llama-3.1-8B	0.0	0.405	1.425	0.096	0.106
		0.1	0.401	1.560	0.049	0.084
		0.2	0.390	1.612	0.027	0.076

Table 24: Results on the GlobalMMLU subset for the mg language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.289	1.792	0.073	0.093
	Mistral-7B	None	0.263	1.927	0.090	0.094
	Llama-3.1-8B	None	0.334	1.785	0.041	0.074
Alpaca	Gemma2-2B	0.0	0.311	1.416	0.176	0.151
		0.1	0.305	1.489	0.149	0.145
		0.2	0.301	1.557	0.120	0.138
	Mistral-7B	0.0	0.274	1.355	0.244	0.168
		0.1	0.279	1.470	0.125	0.131
		0.2	0.263	1.544	0.064	0.113
	Llama-3.1-8B	0.0	0.341	1.435	0.142	0.130
		0.1	0.342	1.464	0.115	0.124
		0.2	0.340	1.482	0.110	0.121
OpenHermes	Gemma2-2B	0.0	0.302	1.236	0.274	0.173
		0.1	0.298	1.812	0.068	0.088
		0.2	0.293	1.813	0.076	0.092
	Mistral-7B	0.0	0.297	1.805	0.069	0.085
		0.1	0.290	1.849	0.058	0.084
		0.2	0.282	1.831	0.083	0.090
	Llama-3.1-8B	0.0	0.330	1.619	0.052	0.092
		0.1	0.332	1.737	0.051	0.079
		0.2	0.324	1.742	0.050	0.078
Tulu3Mixture	Gemma2-2B	0.0	0.295	1.425	0.178	0.145
		0.1	0.296	1.767	0.068	0.090
		0.2	0.295	1.802	0.071	0.090
	Mistral-7B	0.0	0.321	1.638	0.060	0.097
		0.1	0.320	1.719	0.030	0.081
		0.2	0.312	1.744	0.051	0.080
	Llama-3.1-8B	0.0	0.349	1.547	0.105	0.126
		0.1	0.345	1.637	0.074	0.118
		0.2	0.335	1.669	0.059	0.101

Table 25: Results on the GlobalMMLU subset for the ms language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.341	1.820	0.045	0.076
	Mistral-7B	None	0.377	1.847	0.023	0.063
	Llama-3.1-8B	None	0.467	1.597	0.021	0.063
Alpaca	Gemma2-2B	0.0	0.407	1.087	0.245	0.158
		0.1	0.400	1.295	0.164	0.139
		0.2	0.400	1.326	0.151	0.135
	Mistral-7B	0.0	0.362	1.222	0.267	0.170
		0.1	0.369	1.210	0.182	0.145
		0.2	0.311	1.216	0.209	0.149
	Llama-3.1-8B	0.0	0.475	1.083	0.169	0.132
		0.1	0.477	1.176	0.143	0.121
		0.2	0.471	1.218	0.120	0.114
OpenHermes	Gemma2-2B	0.0	0.408	0.912	0.305	0.172
		0.1	0.407	1.657	0.025	0.075
		0.2	0.406	1.638	0.028	0.078
	Mistral-7B	0.0	0.387	1.535	0.064	0.096
		0.1	0.385	1.687	0.029	0.069
		0.2	0.386	1.691	0.033	0.071
	Llama-3.1-8B	0.0	0.480	1.314	0.075	0.097
		0.1	0.474	1.491	0.031	0.072
		0.2	0.477	1.466	0.029	0.071
Tulu3Mixture	Gemma2-2B	0.0	0.390	1.297	0.139	0.134
		0.1	0.388	1.697	0.037	0.078
		0.2	0.384	1.762	0.032	0.071
	Mistral-7B	0.0	0.415	1.360	0.093	0.114
		0.1	0.419	1.425	0.079	0.105
		0.2	0.420	1.543	0.050	0.087
	Llama-3.1-8B	0.0	0.464	1.349	0.073	0.096
		0.1	0.469	1.488	0.032	0.078
		0.2	0.458	1.528	0.028	0.074

Table 26: Results on the GlobalMMLU subset for the ne language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.296	1.795	0.067	0.090
	Mistral-7B	None	0.265	1.917	0.085	0.092
	Llama-3.1-8B	None	0.391	1.752	0.031	0.069
Alpaca	Gemma2-2B	0.0	0.332	1.133	0.272	0.170
		0.1	0.328	1.306	0.199	0.155
		0.2	0.323	1.339	0.187	0.151
	Mistral-7B	0.0	0.273	1.166	0.345	0.192
		0.1	0.275	1.442	0.168	0.152
		0.2	0.254	1.490	0.133	0.142
	Llama-3.1-8B	0.0	0.394	1.304	0.162	0.137
		0.1	0.391	1.326	0.155	0.132
		0.2	0.389	1.357	0.143	0.129
OpenHermes	Gemma2-2B	0.0	0.331	0.998	0.355	0.186
		0.1	0.326	1.740	0.034	0.082
		0.2	0.324	1.751	0.039	0.081
	Mistral-7B	0.0	0.287	1.764	0.061	0.082
		0.1	0.297	1.855	0.041	0.086
		0.2	0.290	1.823	0.077	0.088
	Llama-3.1-8B	0.0	0.393	1.522	0.053	0.094
		0.1	0.387	1.671	0.025	0.070
		0.2	0.392	1.649	0.026	0.076
Tulu3Mixture	Gemma2-2B	0.0	0.323	1.161	0.248	0.163
		0.1	0.322	1.767	0.051	0.082
		0.2	0.312	1.789	0.056	0.086
	Mistral-7B	0.0	0.320	1.632	0.075	0.110
		0.1	0.317	1.732	0.033	0.085
		0.2	0.314	1.771	0.029	0.071
	Llama-3.1-8B	0.0	0.401	1.526	0.058	0.098
		0.1	0.403	1.625	0.042	0.083
		0.2	0.393	1.725	0.052	0.088

Table 27: Results on the GlobalMMLU subset for the nl language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.372	1.706	0.044	0.081
	Mistral-7B	None	0.441	1.771	0.024	0.066
	Llama-3.1-8B	None	0.502	1.484	0.021	0.066
Alpaca	Gemma2-2B	0.0	0.418	0.856	0.299	0.171
		0.1	0.418	1.109	0.207	0.149
		0.2	0.413	1.136	0.191	0.146
	Mistral-7B	0.0	0.433	1.054	0.252	0.162
		0.1	0.439	1.043	0.203	0.144
		0.2	0.372	1.092	0.182	0.142
	Llama-3.1-8B	0.0	0.514	0.986	0.140	0.125
		0.1	0.515	1.088	0.113	0.114
		0.2	0.511	1.134	0.104	0.110
OpenHermes	Gemma2-2B	0.0	0.429	0.705	0.358	0.183
		0.1	0.424	1.541	0.051	0.086
		0.2	0.422	1.556	0.052	0.085
	Mistral-7B	0.0	0.459	1.289	0.095	0.103
		0.1	0.457	1.536	0.036	0.070
		0.2	0.456	1.525	0.033	0.069
	Llama-3.1-8B	0.0	0.521	1.186	0.077	0.098
		0.1	0.511	1.392	0.028	0.070
		0.2	0.513	1.380	0.023	0.068
Tulu3Mixture	Gemma2-2B	0.0	0.408	1.083	0.208	0.151
		0.1	0.408	1.679	0.029	0.074
		0.2	0.399	1.740	0.026	0.073
	Mistral-7B	0.0	0.477	1.256	0.097	0.107
		0.1	0.474	1.374	0.064	0.092
		0.2	0.471	1.456	0.042	0.079
	Llama-3.1-8B	0.0	0.512	1.200	0.081	0.098
		0.1	0.510	1.365	0.055	0.081
		0.1	0.500	1.465	0.065	0.091

Table 28: Results on the GlobalMMLU subset for the ny language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.286	1.798	0.061	0.087
	Mistral-7B	None	0.273	1.931	0.073	0.088
	Llama-3.1-8B	None	0.324	1.829	0.035	0.073
Alpaca	Gemma2-2B	0.0	0.301	1.422	0.155	0.146
		0.1	0.299	1.477	0.141	0.138
		0.2	0.297	1.563	0.099	0.127
	Mistral-7B	0.0	0.275	1.199	0.328	0.186
		0.1	0.277	1.470	0.143	0.133
		0.2	0.253	1.551	0.068	0.115
	Llama-3.1-8B	0.0	0.330	1.525	0.092	0.118
		0.1	0.331	1.534	0.087	0.115
		0.2	0.327	1.552	0.072	0.113
OpenHermes	Gemma2-2B	0.0	0.301	1.291	0.231	0.164
		0.1	0.291	1.819	0.064	0.087
		0.2	0.290	1.813	0.078	0.090
	Mistral-7B	0.0	0.298	1.794	0.073	0.087
		0.1	0.297	1.853	0.061	0.080
		0.2	0.293	1.836	0.062	0.083
	Llama-3.1-8B	0.0	0.319	1.686	0.057	0.083
		0.1	0.313	1.787	0.041	0.086
		0.2	0.316	1.783	0.056	0.083
Tulu3Mixture	Gemma2-2B	0.0	0.297	1.592	0.074	0.117
		0.1	0.292	1.881	0.054	0.078
		0.2	0.293	1.898	0.057	0.083
	Mistral-7B	0.0	0.314	1.718	0.051	0.088
		0.1	0.310	1.798	0.044	0.079
		0.2	0.306	1.800	0.050	0.078
	Llama-3.1-8B	0.0	0.329	1.647	0.050	0.095
		0.1	0.328	1.697	0.040	0.082
		0.2	0.318	1.897	0.038	0.079

Table 29: Results on the GlobalMMLU subset for the pl language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.340	1.772	0.057	0.084
	Mistral-7B	None	0.421	1.816	0.022	0.067
	Llama-3.1-8B	None	0.466	1.542	0.027	0.068
Alpaca	Gemma2-2B	0.0	0.393	0.945	0.301	0.174
		0.1	0.392	1.234	0.185	0.144
		0.2	0.392	1.257	0.175	0.141
	Mistral-7B	0.0	0.414	1.103	0.263	0.162
		0.1	0.424	1.082	0.198	0.146
		0.2	0.366	1.120	0.188	0.143
	Llama-3.1-8B	0.0	0.477	1.005	0.199	0.144
		0.1	0.475	1.126	0.169	0.134
		0.2	0.473	1.177	0.147	0.127
OpenHermes	Gemma2-2B	0.0	0.403	0.732	0.393	0.191
		0.1	0.401	1.613	0.041	0.084
		0.2	0.396	1.633	0.042	0.083
	Mistral-7B	0.0	0.437	1.412	0.063	0.095
		0.1	0.435	1.613	0.031	0.072
		0.2	0.431	1.604	0.032	0.071
	Llama-3.1-8B	0.0	0.474	1.241	0.095	0.105
		0.1	0.474	1.443	0.034	0.076
		0.2	0.475	1.422	0.031	0.077
Tulu3Mixture	Gemma2-2B	0.0	0.382	1.126	0.212	0.153
		0.1	0.382	1.688	0.030	0.078
		0.2	0.379	1.696	0.031	0.076
	Mistral-7B	0.0	0.452	1.317	0.086	0.108
		0.1	0.449	1.416	0.065	0.094
		0.2	0.444	1.497	0.049	0.082
	Llama-3.1-8B	0.0	0.478	1.297	0.074	0.100
		0.1	0.473	1.432	0.040	0.079
		0.2	0.476	1.482	0.038	0.084

Table 30: Results on the GlobalMMLU subset for the pt language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.365	1.716	0.038	0.078
	Mistral-7B	None	0.471	1.744	0.023	0.067
	Llama-3.1-8B	None	0.522	1.434	0.025	0.064
Alpaca	Gemma2-2B	0.0	0.436	0.881	0.271	0.163
		0.1	0.435	1.152	0.187	0.141
		0.2	0.432	1.188	0.169	0.137
	Mistral-7B	0.0	0.448	1.022	0.293	0.170
		0.1	0.458	1.021	0.217	0.147
		0.2	0.381	1.055	0.206	0.143
	Llama-3.1-8B	0.0	0.531	0.916	0.154	0.129
		0.1	0.532	1.034	0.121	0.114
		0.2	0.528	1.078	0.107	0.111
OpenHermes	Gemma2-2B	0.0	0.444	0.672	0.343	0.179
		0.1	0.435	1.427	0.065	0.099
		0.2	0.433	1.458	0.066	0.098
	Mistral-7B	0.0	0.482	1.301	0.069	0.093
		0.1	0.479	1.533	0.025	0.066
		0.2	0.471	1.504	0.029	0.069
	Llama-3.1-8B	0.0	0.538	1.094	0.086	0.103
		0.1	0.533	1.322	0.030	0.069
		0.2	0.538	1.303	0.050	0.077
Tulu3Mixture	Gemma2-2B	0.0	0.426	1.011	0.234	0.153
		0.1	0.429	1.525	0.052	0.092
		0.2	0.419	1.732	0.049	0.082
	Mistral-7B	0.0	0.500	1.170	0.098	0.109
		0.1	0.492	1.383	0.050	0.081
		0.2	0.487	1.424	0.038	0.076
	Llama-3.1-8B	0.0	0.532	1.099	0.095	0.107
		0.1	0.528	1.285	0.047	0.083
		0.2	0.526	1.334	0.037	0.077

Table 31: Results on the GlobalMMLU subset for the ro language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.353	1.757	0.040	0.076
	Mistral-7B	None	0.429	1.805	0.019	0.067
	Llama-3.1-8B	None	0.480	1.516	0.028	0.066
Alpaca	Gemma2-2B	0.0	0.405	0.943	0.275	0.165
		0.1	0.405	1.220	0.177	0.141
		0.2	0.401	1.242	0.170	0.139
	Mistral-7B	0.0	0.416	1.084	0.194	0.145
		0.1	0.424	1.102	0.198	0.143
		0.2	0.365	1.139	0.263	0.162
	Llama-3.1-8B	0.0	0.496	0.990	0.163	0.132
		0.1	0.499	1.080	0.138	0.122
		0.2	0.497	1.131	0.120	0.117
OpenHermes	Gemma2-2B	0.0	0.415	0.791	0.346	0.180
		0.1	0.404	1.604	0.040	0.081
		0.2	0.402	1.620	0.039	0.083
	Mistral-7B	0.0	0.440	1.390	0.059	0.094
		0.1	0.435	1.593	0.035	0.073
		0.2	0.435	1.589	0.029	0.071
	Llama-3.1-8B	0.0	0.492	1.235	0.094	0.103
		0.1	0.490	1.435	0.033	0.073
		0.2	0.496	1.414	0.032	0.074
Tulu3Mixture	Gemma2-2B	0.0	0.393	1.205	0.194	0.143
		0.1	0.392	1.683	0.035	0.078
		0.1	0.385	1.712	0.037	0.082
	Mistral-7B	0.0	0.450	1.260	0.113	0.115
		0.1	0.449	1.413	0.076	0.099
		0.2	0.439	1.485	0.039	0.079
	Llama-3.1-8B	0.0	0.497	1.234	0.089	0.102
		0.1	0.501	1.380	0.068	0.086
		0.1	0.493	1.380	0.057	0.090

Table 32: Results on the GlobalMMLU subset for the ru language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.342	1.753	0.048	0.076
	Mistral-7B	None	0.435	1.779	0.023	0.070
	Llama-3.1-8B	None	0.488	1.458	0.019	0.067
Alpaca	Gemma2-2B	0.0	0.399	0.941	0.270	0.168
		0.1	0.390	1.213	0.186	0.146
		0.2	0.393	1.246	0.174	0.143
	Mistral-7B	0.0	0.427	1.016	0.251	0.156
		0.1	0.432	1.062	0.220	0.148
		0.2	0.377	1.054	0.296	0.170
	Llama-3.1-8B	0.0	0.497	0.954	0.177	0.138
		0.1	0.500	1.083	0.140	0.123
		0.2	0.500	1.145	0.115	0.115
OpenHermes	Gemma2-2B	0.0	0.401	0.737	0.394	0.190
		0.1	0.397	1.611	0.054	0.088
		0.2	0.398	1.629	0.045	0.085
	Mistral-7B	0.0	0.454	1.309	0.085	0.106
		0.1	0.448	1.495	0.037	0.077
		0.2	0.450	1.492	0.039	0.078
	Llama-3.1-8B	0.0	0.500	1.148	0.106	0.113
		0.1	0.498	1.365	0.040	0.079
		0.2	0.501	1.352	0.042	0.081
Tulu3Mixture	Gemma2-2B	0.0	0.386	1.043	0.268	0.165
		0.1	0.383	1.680	0.036	0.080
		0.2	0.373	1.724	0.046	0.088
	Mistral-7B	0.0	0.466	1.203	0.114	0.117
		0.1	0.458	1.329	0.082	0.100
		0.2	0.463	1.411	0.056	0.092
	Llama-3.1-8B	0.0	0.498	1.201	0.086	0.106
		0.1	0.499	1.346	0.047	0.084
		0.2	0.496	1.464	0.042	0.087

Table 33: Results on the GlobalMMLU subset for the si language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.284	1.860	0.067	0.088
	Mistral-7B	None	0.250	1.926	0.113	0.107
	Llama-3.1-8B	None	0.342	1.760	0.034	0.073
Alpaca	Gemma2-2B	0.0	0.297	1.378	0.196	0.157
		0.1	0.293	1.474	0.162	0.148
		0.2	0.293	1.521	0.145	0.144
	Mistral-7B	0.0	0.256	1.266	0.280	0.185
		0.1	0.262	1.478	0.174	0.151
		0.2	0.250	1.522	0.168	0.146
	Llama-3.1-8B	0.0	0.348	1.367	0.149	0.138
		0.1	0.352	1.440	0.116	0.126
		0.2	0.350	1.456	0.110	0.125
OpenHermes	Gemma2-2B	0.0	0.289	1.333	0.221	0.161
		0.1	0.293	1.825	0.054	0.082
		0.2	0.294	1.839	0.056	0.082
	Mistral-7B	0.0	0.275	1.785	0.074	0.092
		0.1	0.276	1.864	0.068	0.087
		0.2	0.273	1.843	0.085	0.091
	Llama-3.1-8B	0.0	0.343	1.594	0.062	0.105
		0.1	0.340	1.751	0.037	0.076
		0.2	0.345	1.734	0.039	0.079
Tulu3Mixture	Gemma2-2B	0.0	0.297	1.399	0.161	0.145
		0.1	0.291	1.755	0.066	0.094
		0.2	0.294	1.823	0.061	0.090
	Mistral-7B	0.0	0.303	1.764	0.060	0.088
		0.1	0.298	1.820	0.059	0.083
		0.2	0.301	1.838	0.061	0.083
	Llama-3.1-8B	0.0	0.350	1.596	0.064	0.106
		0.1	0.350	1.740	0.031	0.078
		0.2	0.346	1.826	0.034	0.077

Table 34: Results on the GlobalMMLU subset for the sn language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.285	1.821	0.065	0.087
	Mistral-7B	None	0.281	1.928	0.063	0.081
	Llama-3.1-8B	None	0.328	1.799	0.039	0.072
Alpaca	Gemma2-2B	0.0	0.307	1.453	0.149	0.142
		0.1	0.310	1.503	0.138	0.137
		0.2	0.307	1.575	0.094	0.127
	Mistral-7B	0.0	0.284	1.228	0.348	0.188
		0.1	0.282	1.438	0.158	0.139
		0.2	0.259	1.531	0.081	0.118
	Llama-3.1-8B	0.0	0.337	1.475	0.136	0.134
		0.1	0.335	1.482	0.127	0.131
		0.2	0.336	1.503	0.114	0.125
OpenHermes	Gemma2-2B	0.0	0.308	1.289	0.243	0.164
		0.1	0.305	1.818	0.057	0.083
		0.2	0.301	1.814	0.066	0.088
	Mistral-7B	0.0	0.300	1.769	0.059	0.085
		0.1	0.302	1.828	0.054	0.079
		0.2	0.298	1.810	0.065	0.083
	Llama-3.1-8B	0.0	0.329	1.648	0.046	0.090
		0.1	0.334	1.755	0.042	0.079
		0.2	0.324	1.756	0.049	0.079
Tulu3Mixture	Gemma2-2B	0.0	0.298	1.627	0.064	0.109
		0.1	0.292	1.894	0.055	0.079
		0.2	0.295	1.903	0.054	0.080
	Mistral-7B	0.0	0.326	1.681	0.056	0.095
		0.1	0.329	1.747	0.048	0.085
		0.2	0.328	1.766	0.043	0.080
	Llama-3.1-8B	0.0	0.344	1.645	0.049	0.094
		0.1	0.342	1.695	0.042	0.083
		0.2	0.346	1.706	0.038	0.083

Table 35: Results on the GlobalMMLU subset for the so language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.280	1.873	0.070	0.089
	Mistral-7B	None	0.279	1.946	0.054	0.080
	Llama-3.1-8B	None	0.317	1.781	0.041	0.076
Alpaca	Gemma2-2B	0.0	0.294	1.493	0.121	0.141
		0.1	0.297	1.556	0.101	0.130
		0.2	0.290	1.620	0.080	0.120
	Mistral-7B	0.0	0.272	1.367	0.236	0.169
		0.1	0.276	1.528	0.092	0.125
		0.2	0.254	1.579	0.074	0.111
	Llama-3.1-8B	0.0	0.321	1.479	0.117	0.131
		0.1	0.323	1.509	0.102	0.124
		0.2	0.324	1.537	0.086	0.118
OpenHermes	Gemma2-2B	0.0	0.291	1.397	0.188	0.151
		0.1	0.305	1.869	0.067	0.087
		0.2	0.285	1.868	0.075	0.091
	Mistral-7B	0.0	0.281	1.779	0.074	0.089
		0.1	0.286	1.844	0.063	0.088
		0.2	0.289	1.833	0.073	0.087
	Llama-3.1-8B	0.0	0.314	1.656	0.062	0.094
		0.1	0.311	1.764	0.064	0.085
		0.2	0.313	1.776	0.064	0.084
Tulu3Mixture	Gemma2-2B	0.0	0.289	1.678	0.070	0.111
		0.1	0.282	1.868	0.058	0.082
		0.2	0.283	1.897	0.055	0.080
	Mistral-7B	0.0	0.306	1.734	0.060	0.087
		0.1	0.305	1.793	0.052	0.083
		0.2	0.302	1.800	0.047	0.082
	Llama-3.1-8B	0.0	0.334	1.549	0.067	0.105
		0.1	0.332	1.675	0.044	0.085
		0.2	0.330	1.715	0.035	0.076

Table 36: Results on the GlobalMMLU subset for the sr language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.324	1.717	0.063	0.093
	Mistral-7B	None	0.404	1.826	0.013	0.062
	Llama-3.1-8B	None	0.432	1.539	0.035	0.075
Alpaca	Gemma2-2B	0.0	0.360	1.048	0.292	0.172
		0.1	0.357	1.273	0.192	0.150
		0.2	0.358	1.304	0.182	0.148
	Mistral-7B	0.0	0.401	1.103	0.293	0.173
		0.1	0.405	1.105	0.223	0.151
		0.2	0.333	1.146	0.183	0.144
	Llama-3.1-8B	0.0	0.453	1.066	0.187	0.141
		0.1	0.454	1.156	0.151	0.130
		0.2	0.449	1.206	0.136	0.124
OpenHermes	Gemma2-2B	0.0	0.358	0.899	0.376	0.190
		0.1	0.357	1.736	0.043	0.079
		0.2	0.354	1.750	0.045	0.081
	Mistral-7B	0.0	0.415	1.424	0.077	0.103
		0.1	0.418	1.621	0.030	0.070
		0.2	0.415	1.614	0.029	0.073
	Llama-3.1-8B	0.0	0.445	1.301	0.110	0.114
		0.1	0.447	1.482	0.043	0.080
		0.2	0.445	1.480	0.040	0.081
Tulu3Mixture	Gemma2-2B	0.0	0.349	1.193	0.243	0.160
		0.1	0.347	1.758	0.042	0.074
		0.2	0.344	1.827	0.045	0.078
	Mistral-7B	0.0	0.428	1.342	0.098	0.111
		0.1	0.430	1.511	0.054	0.089
		0.2	0.427	1.575	0.041	0.078
	Llama-3.1-8B	0.0	0.446	1.307	0.090	0.110
		0.1	0.446	1.424	0.066	0.094
		0.2	0.436	1.489	0.051	0.086

Table 37: Results on the GlobalMMLU subset for the sv language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.359	1.730	0.051	0.082
	Mistral-7B	None	0.428	1.779	0.026	0.068
	Llama-3.1-8B	None	0.497	1.525	0.016	0.067
Alpaca	Gemma2-2B	0.0	0.414	0.929	0.282	0.167
		0.1	0.406	1.154	0.197	0.147
		0.2	0.407	1.185	0.184	0.143
	Mistral-7B	0.0	0.425	1.046	0.219	0.149
		0.1	0.433	1.069	0.218	0.148
		0.2	0.377	1.082	0.270	0.163
	Llama-3.1-8B	0.0	0.502	1.010	0.151	0.129
		0.1	0.498	1.105	0.124	0.117
		0.2	0.496	1.155	0.112	0.113
OpenHermes	Gemma2-2B	0.0	0.418	0.767	0.362	0.184
		0.1	0.413	1.577	0.046	0.087
		0.2	0.411	1.594	0.041	0.082
	Mistral-7B	0.0	0.452	1.412	0.063	0.092
		0.1	0.449	1.602	0.019	0.066
		0.2	0.443	1.603	0.032	0.070
	Llama-3.1-8B	0.0	0.500	1.232	0.077	0.101
		0.1	0.497	1.433	0.024	0.068
		0.2	0.503	1.414	0.026	0.070
Tulu3Mixture	Gemma2-2B	0.0	0.395	1.145	0.206	0.148
		0.1	0.394	1.694	0.037	0.076
		0.2	0.386	1.751	0.038	0.077
	Mistral-7B	0.0	0.463	1.334	0.079	0.102
		0.1	0.465	1.466	0.040	0.080
		0.2	0.459	1.523	0.023	0.071
	Llama-3.1-8B	0.0	0.497	1.257	0.070	0.098
		0.1	0.495	1.403	0.036	0.079
		0.2	0.484	1.450	0.031	0.073

Table 38: Results on the GlobalMMLU subset for the sw language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.323	1.816	0.045	0.078
	Mistral-7B	None	0.280	1.943	0.056	0.080
	Llama-3.1-8B	None	0.356	1.731	0.039	0.075
Alpaca	Gemma2-2B	0.0	0.353	1.203	0.218	0.157
		0.1	0.356	1.347	0.187	0.145
		0.2	0.352	1.388	0.171	0.141
	Mistral-7B	0.0	0.280	1.241	0.295	0.182
		0.1	0.280	1.455	0.153	0.138
		0.2	0.255	1.516	0.097	0.121
	Llama-3.1-8B	0.0	0.373	1.307	0.156	0.137
		0.1	0.376	1.351	0.143	0.132
		0.2	0.375	1.380	0.132	0.128
OpenHermes	Gemma2-2B	0.0	0.356	1.042	0.311	0.175
		0.1	0.349	1.732	0.044	0.081
		0.2	0.345	1.731	0.047	0.082
	Mistral-7B	0.0	0.295	1.792	0.078	0.089
		0.1	0.297	1.864	0.064	0.083
		0.2	0.289	1.852	0.051	0.080
	Llama-3.1-8B	0.0	0.355	1.525	0.062	0.100
		0.1	0.356	1.682	0.037	0.078
		0.2	0.359	1.682	0.038	0.076
Tulu3Mixture	Gemma2-2B	0.0	0.340	1.421	0.146	0.133
		0.1	0.340	1.802	0.049	0.079
		0.2	0.338	1.823	0.049	0.080
	Mistral-7B	0.0	0.307	1.696	0.049	0.095
		0.1	0.311	1.769	0.039	0.085
		0.2	0.317	1.800	0.046	0.079
	Llama-3.1-8B	0.0	0.362	1.536	0.072	0.104
		0.1	0.365	1.631	0.045	0.086
		0.2	0.359	1.662	0.031	0.080

Table 39: Results on the GlobalMMLU subset for the te language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.288	1.823	0.082	0.094
	Mistral-7B	None	0.250	1.920	0.111	0.105
	Llama-3.1-8B	None	0.356	1.758	0.042	0.075
Alpaca	Gemma2-2B	0.0	0.319	1.211	0.261	0.168
		0.1	0.321	1.409	0.160	0.144
		0.2	0.318	1.424	0.149	0.142
	Mistral-7B	0.0	0.254	1.229	0.291	0.187
		0.1	0.259	1.474	0.164	0.155
		0.2	0.248	1.515	0.152	0.149
	Llama-3.1-8B	0.0	0.360	1.339	0.160	0.137
		0.1	0.364	1.397	0.126	0.127
		0.2	0.363	1.406	0.122	0.125
OpenHermes	Gemma2-2B	0.0	0.317	1.031	0.363	0.189
		0.1	0.319	1.776	0.032	0.084
		0.2	0.321	1.770	0.043	0.084
	Mistral-7B	0.0	0.275	1.788	0.065	0.087
		0.1	0.285	1.885	0.060	0.080
		0.2	0.273	1.871	0.085	0.092
	Llama-3.1-8B	0.0	0.352	1.567	0.066	0.103
		0.1	0.349	1.712	0.038	0.074
		0.2	0.353	1.690	0.030	0.077
Tulu3Mixture	Gemma2-2B	0.0	0.316	1.267	0.206	0.152
		0.1	0.311	1.757	0.058	0.086
		0.2	0.305	1.779	0.060	0.086
	Mistral-7B	0.0	0.303	1.768	0.048	0.086
		0.1	0.302	1.823	0.026	0.075
		0.2	0.296	1.859	0.072	0.085
	Llama-3.1-8B	0.0	0.360	1.691	0.039	0.084
		0.1	0.363	1.780	0.021	0.069
		0.2	0.357	1.749	0.024	0.075

Table 40: Results on the GlobalMMLU subset for the tr language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.328	1.830	0.052	0.078
	Mistral-7B	None	0.341	1.885	0.032	0.069
	Llama-3.1-8B	None	0.452	1.539	0.028	0.071
Alpaca	Gemma2-2B	0.0	0.373	1.141	0.214	0.155
		0.1	0.372	1.332	0.154	0.137
		0.2	0.371	1.363	0.137	0.133
	Mistral-7B	0.0	0.337	1.229	0.297	0.176
		0.1	0.350	1.252	0.232	0.156
		0.2	0.286	1.310	0.192	0.148
	Llama-3.1-8B	0.0	0.463	1.076	0.189	0.138
		0.1	0.465	1.180	0.156	0.129
		0.2	0.462	1.234	0.139	0.123
OpenHermes	Gemma2-2B	0.0	0.376	0.966	0.311	0.174
		0.1	0.375	1.690	0.038	0.078
		0.2	0.375	1.693	0.037	0.076
	Mistral-7B	0.0	0.361	1.643	0.044	0.085
		0.1	0.365	1.739	0.044	0.076
		0.2	0.360	1.739	0.049	0.078
	Llama-3.1-8B	0.0	0.470	1.230	0.106	0.114
		0.1	0.464	1.430	0.048	0.081
		0.2	0.463	1.419	0.048	0.085
Tulu3Mixture	Gemma2-2B	0.0	0.361	1.261	0.172	0.143
		0.1	0.353	1.673	0.055	0.092
		0.2	0.350	1.745	0.043	0.080
	Mistral-7B	0.0	0.398	1.517	0.070	0.102
		0.1	0.399	1.649	0.044	0.080
		0.2	0.391	1.710	0.030	0.072
	Llama-3.1-8B	0.0	0.455	1.285	0.094	0.109
		0.1	0.458	1.424	0.061	0.093
		0.2	0.451	1.480	0.046	0.086

Table 41: Results on the GlobalMMLU subset for the uk language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.331	1.758	0.051	0.082
	Mistral-7B	None	0.423	1.794	0.021	0.065
	Llama-3.1-8B	None	0.460	1.499	0.022	0.070
Alpaca	Gemma2-2B	0.0	0.385	0.963	0.297	0.171
		0.1	0.378	1.237	0.186	0.147
		0.2	0.378	1.265	0.173	0.143
	Mistral-7B	0.0	0.414	1.057	0.299	0.172
		0.1	0.424	1.043	0.245	0.154
		0.2	0.364	1.081	0.212	0.148
	Llama-3.1-8B	0.0	0.477	1.013	0.177	0.136
		0.1	0.480	1.134	0.143	0.123
		0.2	0.473	1.193	0.125	0.117
OpenHermes	Gemma2-2B	0.0	0.388	0.760	0.395	0.191
		0.1	0.379	1.680	0.042	0.077
		0.2	0.377	1.695	0.039	0.078
	Mistral-7B	0.0	0.447	1.370	0.065	0.097
		0.1	0.442	1.546	0.026	0.071
		0.2	0.441	1.549	0.032	0.074
	Llama-3.1-8B	0.0	0.470	1.206	0.127	0.120
		0.1	0.469	1.421	0.042	0.084
		0.2	0.473	1.407	0.043	0.084
Tulu3Mixture	Gemma2-2B	0.0	0.371	1.085	0.242	0.160
		0.1	0.366	1.715	0.031	0.072
		0.2	0.356	1.792	0.028	0.069
	Mistral-7B	0.0	0.454	1.246	0.102	0.115
		0.1	0.452	1.384	0.066	0.096
		0.2	0.449	1.461	0.039	0.083
	Llama-3.1-8B	0.0	0.467	1.246	0.104	0.109
		0.1	0.470	1.390	0.056	0.087
		0.2	0.461	1.440	0.044	0.084

Table 42: Results on the GlobalMMLU subset for the vi language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.299	1.752	0.044	0.081
	Mistral-7B	None	0.365	1.870	0.020	0.067
	Llama-3.1-8B	None	0.473	1.523	0.020	0.067
Alpaca	Gemma2-2B	0.0	0.384	1.036	0.282	0.169
		0.1	0.387	1.274	0.179	0.146
		0.2	0.382	1.258	0.186	0.147
	Mistral-7B	0.0	0.346	1.212	0.283	0.171
		0.1	0.351	1.210	0.248	0.161
		0.2	0.308	1.232	0.230	0.159
	Llama-3.1-8B	0.0	0.487	1.025	0.160	0.133
		0.1	0.487	1.127	0.132	0.122
		0.2	0.482	1.166	0.115	0.118
OpenHermes	Gemma2-2B	0.0	0.399	0.856	0.348	0.183
		0.1	0.398	1.632	0.038	0.083
		0.2	0.398	1.663	0.031	0.078
	Mistral-7B	0.0	0.356	1.546	0.070	0.104
		0.1	0.351	1.697	0.049	0.081
		0.2	0.351	1.697	0.048	0.079
	Llama-3.1-8B	0.0	0.491	1.171	0.111	0.114
		0.1	0.493	1.351	0.050	0.088
		0.2	0.488	1.318	0.067	0.095
Tulu3Mixture	Gemma2-2B	0.0	0.352	1.137	0.265	0.164
		0.1	0.351	1.677	0.029	0.084
		0.2	0.341	1.697	0.029	0.086
	Mistral-7B	0.0	0.388	1.443	0.084	0.110
		0.1	0.386	1.502	0.077	0.101
		0.2	0.376	1.620	0.055	0.084
	Llama-3.1-8B	0.0	0.482	1.252	0.095	0.108
		0.1	0.484	1.377	0.066	0.090
		0.2	0.471	1.397	0.058	0.088

Table 43: Results on the GlobalMMLU subset for the yo language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.276	1.796	0.068	0.093
	Mistral-7B	None	0.265	1.945	0.069	0.087
	Llama-3.1-8B	None	0.308	1.809	0.052	0.079
Alpaca	Gemma2-2B	0.0	0.289	1.450	0.141	0.146
		0.1	0.286	1.485	0.149	0.141
		0.2	0.288	1.556	0.105	0.132
	Mistral-7B	0.0	0.266	1.434	0.176	0.158
		0.1	0.268	1.552	0.075	0.119
		0.2	0.253	1.604	0.075	0.110
	Llama-3.1-8B	0.0	0.320	1.532	0.113	0.124
		0.1	0.319	1.553	0.098	0.118
		0.2	0.316	1.566	0.082	0.116
OpenHermes	Gemma2-2B	0.0	0.296	1.304	0.240	0.166
		0.1	0.290	1.841	0.065	0.086
		0.2	0.290	1.841	0.071	0.090
	Mistral-7B	0.0	0.289	1.792	0.064	0.087
		0.1	0.285	1.866	0.062	0.084
		0.2	0.283	1.847	0.071	0.086
	Llama-3.1-8B	0.0	0.310	1.686	0.054	0.087
		0.1	0.311	1.760	0.065	0.085
		0.2	0.310	1.775	0.056	0.081
Tulu3Mixture	Gemma2-2B	0.0	0.290	1.502	0.110	0.135
		0.1	0.292	1.793	0.053	0.088
		0.2	0.289	1.787	0.048	0.089
	Mistral-7B	0.0	0.313	1.680	0.056	0.094
		0.1	0.313	1.737	0.049	0.089
		0.2	0.316	1.789	0.049	0.079
	Llama-3.1-8B	0.0	0.323	1.611	0.057	0.102
		0.1	0.328	1.660	0.049	0.094
		0.2	0.320	1.683	0.040	0.089

Table 44: Results on the GlobalMMLU subset for the zh language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.343	1.881	0.027	0.069
	Mistral-7B	None	0.410	1.781	0.021	0.066
	Llama-3.1-8B	None	0.491	1.378	0.036	0.077
Alpaca	Gemma2-2B	0.0	0.408	1.126	0.216	0.151
		0.1	0.408	1.244	0.169	0.139
		0.2	0.407	1.278	0.157	0.137
	Mistral-7B	0.0	0.400	1.042	0.299	0.172
		0.1	0.407	1.076	0.252	0.159
		0.2	0.348	1.112	0.265	0.163
	Llama-3.1-8B	0.0	0.500	0.928	0.172	0.135
		0.1	0.503	1.060	0.123	0.118
		0.2	0.500	1.125	0.112	0.113
OpenHermes	Gemma2-2B	0.0	0.419	1.019	0.263	0.161
		0.1	0.415	1.620	0.037	0.080
		0.2	0.413	1.629	0.039	0.081
	Mistral-7B	0.0	0.430	1.313	0.108	0.115
		0.1	0.427	1.468	0.048	0.089
		0.2	0.426	1.441	0.067	0.094
	Llama-3.1-8B	0.0	0.509	1.064	0.123	0.118
		0.1	0.502	1.278	0.057	0.089
		0.2	0.507	1.257	0.063	0.094
Tulu3Mixture	Gemma2-2B	0.0	0.388	1.195	0.206	0.149
		0.1	0.385	1.693	0.038	0.081
		0.2	0.373	1.769	0.034	0.071
	Mistral-7B	0.0	0.445	1.212	0.122	0.124
		0.1	0.447	1.261	0.108	0.119
		0.2	0.446	1.403	0.065	0.101
	Llama-3.1-8B	0.0	0.506	1.112	0.109	0.114
		0.1	0.507	1.296	0.060	0.090
		0.2	0.492	1.340	0.051	0.087

C.1.2 MMLU-ProX

Table 45: Results on the MMLU-ProX subset for the af language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.115	3.098	0.085	0.095
	Mistral-7B	None	0.176	2.997	0.066	0.083
	Llama-3.1-8B	None	0.122	1.566	0.227	0.163
Alpaca	Gemma2-2B	0.0	0.159	2.866	0.091	0.110
		0.1	0.164	2.826	0.059	0.090
		0.2	0.158	2.637	0.063	0.091
	Mistral-7B	0.0	0.182	2.114	0.125	0.150
		0.1	0.195	2.188	0.086	0.129
		0.2	0.176	2.258	0.073	0.121
	Llama-3.1-8B	0.0	0.133	1.169	0.442	0.212
		0.1	0.182	1.859	0.114	0.138
		0.2	0.183	1.953	0.090	0.127
OpenHermes	Gemma2-2B	0.0	0.168	2.323	0.104	0.133
		0.1	0.165	3.073	0.057	0.081
		0.2	0.165	3.115	0.045	0.075
	Mistral-7B	0.0	0.181	2.630	0.089	0.096
		0.1	0.198	2.756	0.086	0.096
		0.2	0.188	2.847	0.083	0.093
	Llama-3.1-8B	0.0	0.146	1.522	0.293	0.181
		0.1	0.199	2.651	0.106	0.105
		0.2	0.202	2.706	0.096	0.100
Tulu3Mixture	Gemma2-2B	0.0	0.138	2.770	0.090	0.099
		0.1	0.163	3.152	0.077	0.077
		0.2	0.145	3.140	0.050	0.077
	Mistral-7B	0.0	0.163	2.817	0.103	0.101
		0.1	0.196	2.972	0.047	0.076
		0.2	0.179	2.914	0.069	0.087
	Llama-3.1-8B	0.0	0.150	1.736	0.171	0.149
		0.1	0.161	2.211	0.101	0.110
		0.2	0.149	2.225	0.112	0.113

Table 46: Results on the MMLU-ProX subset for the ar language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.123	3.147	0.060	0.083
	Mistral-7B	None	0.158	3.042	0.074	0.088
	Llama-3.1-8B	None	0.126	1.530	0.251	0.167
Alpaca	Gemma2-2B	0.0	0.150	2.787	0.089	0.105
		0.1	0.151	2.911	0.077	0.095
		0.2	0.148	2.926	0.076	0.094
	Mistral-7B	0.0	0.156	2.329	0.081	0.127
		0.1	0.166	2.277	0.083	0.135
		0.2	0.148	2.148	0.133	0.160
	Llama-3.1-8B	0.0	0.140	1.172	0.427	0.209
		0.1	0.179	1.716	0.197	0.161
		0.2	0.184	1.848	0.127	0.140
OpenHermes	Gemma2-2B	0.0	0.155	2.289	0.086	0.133
		0.1	0.151	3.034	0.078	0.091
		0.2	0.153	3.066	0.068	0.086
	Mistral-7B	0.0	0.168	2.669	0.104	0.102
		0.1	0.177	2.800	0.099	0.102
		0.2	0.163	2.910	0.093	0.098
	Llama-3.1-8B	0.0	0.141	1.391	0.407	0.208
		0.1	0.194	2.636	0.103	0.104
		0.2	0.201	2.687	0.092	0.098
Tulu3Mixture	Gemma2-2B	0.0	0.152	2.673	0.100	0.109
		0.1	0.146	3.073	0.061	0.082
		0.2	0.139	3.121	0.065	0.084
	Mistral-7B	0.0	0.176	2.805	0.072	0.091
		0.1	0.172	2.980	0.062	0.083
		0.2	0.160	2.987	0.088	0.095
	Llama-3.1-8B	0.0	0.173	1.987	0.075	0.112
		0.1	0.174	2.427	0.046	0.083
		0.2	0.160	2.395	0.035	0.084

Table 47: Results on the MMLU-ProX subset for the bn language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.099	3.156	0.090	0.099
	Mistral-7B	None	0.137	3.029	0.097	0.100
	Llama-3.1-8B	None	0.121	1.610	0.197	0.154
Alpaca	Gemma2-2B	0.0	0.139	2.839	0.088	0.102
		0.1	0.135	2.941	0.079	0.094
		0.2	0.129	2.922	0.094	0.101
	Mistral-7B	0.0	0.140	2.415	0.081	0.129
		0.1	0.144	2.388	0.083	0.134
		0.2	0.132	2.238	0.129	0.160
	Llama-3.1-8B	0.0	0.128	1.300	0.385	0.203
		0.1	0.163	2.065	0.089	0.125
		0.2	0.166	2.151	0.091	0.117
OpenHermes	Gemma2-2B	0.0	0.136	2.374	0.093	0.128
		0.1	0.143	3.120	0.071	0.086
		0.2	0.140	3.140	0.058	0.078
	Mistral-7B	0.0	0.160	2.774	0.109	0.104
		0.1	0.154	2.899	0.107	0.104
		0.2	0.142	2.952	0.107	0.103
	Llama-3.1-8B	0.0	0.133	1.652	0.345	0.195
		0.1	0.170	2.825	0.100	0.102
		0.2	0.176	2.841	0.087	0.096
Tulu3Mixture	Gemma2-2B	0.0	0.122	2.737	0.128	0.117
		0.1	0.132	3.110	0.101	0.101
		0.2	0.124	3.122	0.091	0.098
	Mistral-7B	0.0	0.114	2.923	0.177	0.132
		0.1	0.161	2.935	0.081	0.093
		0.2	0.136	2.976	0.125	0.113
	Llama-3.1-8B	0.0	0.152	2.176	0.069	0.109
		0.1	0.150	2.639	0.111	0.106
		0.2	0.130	2.587	0.108	0.109

Table 48: Results on the MMLU-ProX subset for the cs language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.123	3.079	0.084	0.094
	Mistral-7B	None	0.206	2.996	0.046	0.075
	Llama-3.1-8B	None	0.132	1.539	0.243	0.165
Alpaca	Gemma2-2B	0.0	0.167	2.637	0.072	0.113
		0.1	0.171	2.834	0.049	0.085
		0.2	0.168	2.822	0.047	0.085
	Mistral-7B	0.0	0.211	2.089	0.111	0.140
		0.1	0.221	2.173	0.106	0.134
		0.2	0.205	2.218	0.088	0.131
	Llama-3.1-8B	0.0	0.141	1.157	0.426	0.209
		0.1	0.193	1.748	0.164	0.153
		0.2	0.195	1.870	0.109	0.136
OpenHermes	Gemma2-2B	0.0	0.177	2.312	0.088	0.129
		0.1	0.174	3.043	0.056	0.078
		0.2	0.173	3.085	0.047	0.075
	Mistral-7B	0.0	0.223	2.430	0.070	0.095
		0.1	0.219	2.647	0.069	0.092
		0.2	0.213	2.775	0.067	0.086
	Llama-3.1-8B	0.0	0.163	1.394	0.418	0.209
		0.1	0.216	2.557	0.099	0.101
		0.2	0.223	2.612	0.085	0.095
Tulu3Mixture	Gemma2-2B	0.0	0.166	2.649	0.090	0.104
		0.1	0.162	3.078	0.054	0.078
		0.2	0.151	3.138	0.056	0.079
	Mistral-7B	0.0	0.217	2.803	0.100	0.102
		0.1	0.213	2.793	0.068	0.087
		0.2	0.206	2.633	0.049	0.082
	Llama-3.1-8B	0.0	0.192	1.813	0.131	0.133
		0.1	0.202	2.247	0.039	0.092
		0.2	0.183	2.278	0.041	0.092

Table 49: Results on the MMLU-ProX subset for the de language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.123	3.109	0.078	0.090
	Mistral-7B	None	0.210	2.979	0.048	0.076
	Llama-3.1-8B	None	0.144	1.603	0.228	0.163
Alpaca	Gemma2-2B	0.0	0.173	2.646	0.073	0.106
		0.1	0.174	2.825	0.056	0.088
		0.2	0.166	2.840	0.060	0.088
	Mistral-7B	0.0	0.213	2.030	0.120	0.142
		0.1	0.223	2.035	0.108	0.135
		0.2	0.206	2.117	0.081	0.127
	Llama-3.1-8B	0.0	0.153	1.172	0.426	0.210
		0.1	0.215	1.820	0.113	0.136
		0.2	0.218	1.925	0.087	0.123
OpenHermes	Gemma2-2B	0.0	0.189	2.291	0.111	0.133
		0.1	0.185	3.032	0.048	0.076
		0.2	0.182	3.065	0.041	0.072
	Mistral-7B	0.0	0.235	2.438	0.061	0.091
		0.1	0.225	2.614	0.068	0.087
		0.2	0.229	2.696	0.057	0.082
	Llama-3.1-8B	0.0	0.185	1.523	0.285	0.178
		0.1	0.234	2.531	0.088	0.097
		0.2	0.238	2.576	0.081	0.094
Tulu3Mixture	Gemma2-2B	0.0	0.175	2.689	0.080	0.100
		0.1	0.172	3.094	0.044	0.074
		0.2	0.168	3.141	0.051	0.077
	Mistral-7B	0.0	0.221	2.654	0.069	0.086
		0.1	0.211	2.798	0.049	0.076
		0.2	0.199	2.841	0.044	0.081
	Llama-3.1-8B	0.0	0.203	1.790	0.129	0.134
		0.1	0.208	2.212	0.061	0.096
		0.2	0.190	2.236	0.069	0.099

Table 50: Results on the MMLU-ProX subset for the en language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.155	3.053	0.064	0.085
	Mistral-7B	None	0.268	2.870	0.022	0.063
	Llama-3.1-8B	None	0.175	1.571	0.238	0.161
Alpaca	Gemma2-2B	0.0	0.200	2.420	0.093	0.119
		0.1	0.198	2.694	0.063	0.096
		0.2	0.199	2.708	0.057	0.096
	Mistral-7B	0.0	0.254	1.910	0.116	0.138
		0.1	0.268	1.940	0.108	0.130
		0.2	0.244	1.900	0.138	0.142
	Llama-3.1-8B	0.0	0.177	1.144	0.408	0.205
		0.1	0.238	1.751	0.115	0.130
		0.2	0.241	1.829	0.089	0.120
OpenHermes	Gemma2-2B	0.0	0.214	1.950	0.147	0.154
		0.1	0.220	2.905	0.028	0.067
		0.2	0.218	2.940	0.028	0.070
	Mistral-7B	0.0	0.296	2.125	0.064	0.100
		0.1	0.293	2.407	0.064	0.088
		0.2	0.283	2.492	0.064	0.086
	Llama-3.1-8B	0.0	0.216	1.449	0.278	0.174
		0.1	0.282	2.421	0.078	0.091
		0.2	0.287	2.446	0.067	0.088
Tulu3Mixture	Gemma2-2B	0.0	0.201	2.494	0.090	0.107
		0.1	0.202	3.015	0.050	0.076
		0.2	0.192	3.081	0.052	0.077
	Mistral-7B	0.0	0.264	2.468	0.061	0.085
		0.1	0.265	2.625	0.036	0.077
		0.2	0.248	2.677	0.045	0.091
	Llama-3.1-8B	0.0	0.222	1.500	0.239	0.162
		0.1	0.233	1.962	0.146	0.128
		0.2	0.207	1.974	0.163	0.134

Table 51: Results on the MMLU-ProX subset for the es language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.128	3.115	0.071	0.086
	Mistral-7B	None	0.229	2.948	0.034	0.069
	Llama-3.1-8B	None	0.141	1.600	0.219	0.159
Alpaca	Gemma2-2B	0.0	0.182	2.574	0.075	0.112
		0.1	0.180	2.779	0.060	0.095
		0.2	0.172	2.788	0.067	0.095
	Mistral-7B	0.0	0.217	1.997	0.124	0.143
		0.1	0.232	2.008	0.109	0.138
		0.2	0.212	2.005	0.113	0.135
	Llama-3.1-8B	0.0	0.147	1.170	0.414	0.206
		0.1	0.212	1.773	0.133	0.138
		0.2	0.217	1.890	0.089	0.119
OpenHermes	Gemma2-2B	0.0	0.193	2.162	0.133	0.147
		0.1	0.193	2.972	0.042	0.072
		0.2	0.190	3.007	0.037	0.070
	Mistral-7B	0.0	0.246	2.328	0.065	0.096
		0.1	0.245	2.601	0.063	0.084
		0.2	0.244	2.702	0.057	0.080
	Llama-3.1-8B	0.0	0.171	1.468	0.316	0.186
		0.1	0.235	2.519	0.101	0.102
		0.2	0.243	2.566	0.086	0.096
Tulu3Mixture	Gemma2-2B	0.0	0.176	2.654	0.087	0.103
		0.1	0.170	3.009	0.058	0.081
		0.2	0.155	3.081	0.069	0.085
	Mistral-7B	0.0	0.231	2.547	0.085	0.096
		0.1	0.223	2.746	0.047	0.086
		0.2	0.207	2.783	0.057	0.081
	Llama-3.1-8B	0.0	0.201	1.800	0.108	0.123
		0.1	0.212	2.210	0.064	0.093
		0.2	0.194	2.242	0.052	0.088

Table 52: Results on the MMLU-ProX subset for the fr language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.116	3.129	0.077	0.090
	Mistral-7B	None	0.221	2.955	0.042	0.073
	Llama-3.1-8B	None	0.153	1.672	0.178	0.145
Alpaca	Gemma2-2B	0.0	0.169	2.628	0.083	0.112
		0.1	0.171	2.804	0.066	0.095
		0.2	0.168	2.821	0.066	0.096
	Mistral-7B	0.0	0.219	2.017	0.106	0.138
		0.1	0.233	2.005	0.114	0.135
		0.2	0.213	2.015	0.120	0.141
	Llama-3.1-8B	0.0	0.170	1.247	0.374	0.198
		0.1	0.223	1.833	0.109	0.131
		0.2	0.226	1.937	0.078	0.117
OpenHermes	Gemma2-2B	0.0	0.187	2.255	0.124	0.139
		0.1	0.187	3.007	0.044	0.073
		0.2	0.190	3.044	0.034	0.070
	Mistral-7B	0.0	0.242	2.350	0.062	0.097
		0.1	0.242	2.584	0.065	0.087
		0.2	0.239	2.687	0.065	0.085
	Llama-3.1-8B	0.0	0.196	1.573	0.300	0.182
		0.1	0.240	2.577	0.088	0.096
		0.2	0.246	2.592	0.078	0.091
Tulu3Mixture	Gemma2-2B	0.0	0.176	2.705	0.082	0.098
		0.1	0.169	3.061	0.054	0.078
		0.2	0.156	3.107	0.061	0.082
	Mistral-7B	0.0	0.230	2.557	0.043	0.092
		0.1	0.220	2.728	0.042	0.078
		0.2	0.205	2.767	0.067	0.086
	Llama-3.1-8B	0.0	0.226	1.900	0.099	0.119
		0.1	0.223	2.350	0.052	0.084
		0.2	0.204	2.337	0.027	0.080

Table 53: Results on the MMLU-ProX subset for the hi language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.102	3.155	0.090	0.099
	Mistral-7B	None	0.148	3.006	0.089	0.096
	Llama-3.1-8B	None	0.121	1.544	0.256	0.169
Alpaca	Gemma2-2B	0.0	0.148	2.733	0.070	0.112
		0.1	0.148	2.879	0.061	0.094
		0.2	0.144	2.869	0.067	0.097
	Mistral-7B	0.0	0.153	2.399	0.133	0.159
		0.1	0.160	2.372	0.083	0.131
		0.2	0.140	2.170	0.081	0.127
	Llama-3.1-8B	0.0	0.136	1.272	0.380	0.201
		0.1	0.185	1.926	0.105	0.136
		0.2	0.186	2.038	0.087	0.124
OpenHermes	Gemma2-2B	0.0	0.148	2.227	0.099	0.142
		0.1	0.145	3.036	0.076	0.089
		0.2	0.143	3.067	0.070	0.087
	Mistral-7B	0.0	0.168	2.687	0.111	0.107
		0.1	0.161	2.832	0.107	0.104
		0.2	0.160	2.895	0.103	0.102
	Llama-3.1-8B	0.0	0.151	1.686	0.272	0.176
		0.1	0.194	2.707	0.095	0.100
		0.2	0.197	2.734	0.086	0.094
Tulu3Mixture	Gemma2-2B	0.0	0.140	2.647	0.122	0.116
		0.1	0.133	3.137	0.078	0.089
		0.2	0.127	3.143	0.081	0.091
	Mistral-7B	0.0	0.164	2.905	0.115	0.109
		0.1	0.158	2.980	0.101	0.102
		0.2	0.151	2.999	0.076	0.089
	Llama-3.1-8B	0.0	0.164	2.144	0.055	0.101
		0.1	0.159	2.398	0.031	0.082
		0.2	0.149	2.452	0.048	0.086

Table 54: Results on the MMLU-ProX subset for the hu language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.116	3.149	0.067	0.088
	Mistral-7B	None	0.176	2.978	0.092	0.097
	Llama-3.1-8B	None	0.129	1.507	0.244	0.169
Alpaca	Gemma2-2B	0.0	0.159	2.724	0.072	0.104
		0.1	0.164	2.924	0.050	0.081
		0.2	0.157	2.927	0.054	0.083
	Mistral-7B	0.0	0.197	2.250	0.089	0.131
		0.1	0.211	2.173	0.079	0.124
		0.2	0.198	2.184	0.062	0.116
	Llama-3.1-8B	0.0	0.131	1.083	0.467	0.217
		0.1	0.185	1.705	0.179	0.158
		0.2	0.187	1.824	0.126	0.143
OpenHermes	Gemma2-2B	0.0	0.157	2.415	0.107	0.128
		0.1	0.158	3.104	0.053	0.078
		0.2	0.159	3.135	0.044	0.073
	Mistral-7B	0.0	0.213	2.551	0.086	0.095
		0.1	0.206	2.728	0.077	0.094
		0.2	0.198	2.842	0.074	0.088
	Llama-3.1-8B	0.0	0.155	1.386	0.374	0.200
		0.1	0.201	2.520	0.106	0.104
		0.2	0.206	2.573	0.099	0.102
Tulu3Mixture	Gemma2-2B	0.0	0.152	2.750	0.089	0.100
		0.1	0.149	3.099	0.062	0.084
		0.2	0.133	3.124	0.065	0.084
	Mistral-7B	0.0	0.203	2.764	0.056	0.082
		0.1	0.205	2.939	0.049	0.078
		0.2	0.183	2.963	0.073	0.088
	Llama-3.1-8B	0.0	0.164	1.722	0.154	0.148
		0.1	0.175	2.192	0.058	0.102
		0.2	0.162	2.239	0.063	0.100

Table 55: Results on the MMLU-ProX subset for the id language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.127	3.113	0.075	0.088
	Mistral-7B	None	0.200	2.997	0.046	0.076
	Llama-3.1-8B	None	0.143	1.709	0.185	0.149
Alpaca	Gemma2-2B	0.0	0.170	2.675	0.068	0.105
		0.1	0.164	2.816	0.063	0.092
		0.2	0.166	2.818	0.063	0.092
	Mistral-7B	0.0	0.197	2.158	0.112	0.142
		0.1	0.209	2.134	0.087	0.127
		0.2	0.183	2.117	0.075	0.125
	Llama-3.1-8B	0.0	0.157	1.267	0.368	0.199
		0.1	0.212	1.931	0.086	0.126
		0.2	0.210	2.023	0.079	0.114
OpenHermes	Gemma2-2B	0.0	0.177	2.298	0.109	0.132
		0.1	0.181	3.037	0.045	0.073
		0.2	0.180	3.073	0.038	0.071
	Mistral-7B	0.0	0.211	2.452	0.072	0.093
		0.1	0.209	2.712	0.081	0.092
		0.2	0.204	2.796	0.079	0.092
	Llama-3.1-8B	0.0	0.177	1.572	0.353	0.196
		0.1	0.229	2.599	0.090	0.100
		0.2	0.237	2.641	0.075	0.091
Tulu3Mixture	Gemma2-2B	0.0	0.169	2.705	0.094	0.101
		0.1	0.160	3.050	0.065	0.083
		0.2	0.154	3.108	0.067	0.084
	Mistral-7B	0.0	0.213	2.561	0.119	0.110
		0.1	0.208	2.744	0.060	0.097
		0.2	0.184	2.742	0.063	0.088
	Llama-3.1-8B	0.0	0.203	1.954	0.090	0.118
		0.1	0.200	2.350	0.021	0.074
		0.2	0.185	2.351	0.031	0.081

Table 56: Results on the MMLU-ProX subset for the it language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.124	3.118	0.076	0.089
	Mistral-7B	None	0.219	2.984	0.039	0.072
	Llama-3.1-8B	None	0.129	1.556	0.302	0.182
Alpaca	Gemma2-2B	0.0	0.173	2.599	0.082	0.112
		0.1	0.169	2.791	0.063	0.093
		0.2	0.170	2.815	0.056	0.090
	Mistral-7B	0.0	0.214	2.053	0.096	0.132
		0.1	0.226	2.056	0.098	0.129
		0.2	0.207	2.051	0.117	0.140
	Llama-3.1-8B	0.0	0.147	1.201	0.420	0.209
		0.1	0.211	1.842	0.095	0.127
		0.2	0.214	1.923	0.085	0.118
OpenHermes	Gemma2-2B	0.0	0.183	2.279	0.109	0.137
		0.1	0.184	2.989	0.046	0.075
		0.2	0.184	3.028	0.039	0.071
	Mistral-7B	0.0	0.239	2.282	0.063	0.097
		0.1	0.244	2.565	0.074	0.090
		0.2	0.234	2.687	0.063	0.084
	Llama-3.1-8B	0.0	0.166	1.519	0.279	0.175
		0.1	0.232	2.517	0.096	0.101
		0.2	0.233	2.557	0.093	0.101
Tulu3Mixture	Gemma2-2B	0.0	0.162	2.731	0.087	0.099
		0.1	0.160	3.072	0.056	0.080
		0.2	0.152	3.123	0.059	0.081
	Mistral-7B	0.0	0.229	2.592	0.104	0.105
		0.1	0.226	2.717	0.082	0.095
		0.2	0.215	2.765	0.047	0.086
	Llama-3.1-8B	0.0	0.185	1.765	0.159	0.141
		0.1	0.193	2.149	0.134	0.120
		0.2	0.176	2.189	0.116	0.116

Table 57: Results on the MMLU-ProX subset for the ja language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.116	3.194	0.056	0.079
	Mistral-7B	None	0.196	3.016	0.044	0.073
	Llama-3.1-8B	None	0.141	1.618	0.216	0.157
Alpaca	Gemma2-2B	0.0	0.159	2.845	0.062	0.096
		0.1	0.164	2.910	0.059	0.089
		0.2	0.157	2.901	0.062	0.090
	Mistral-7B	0.0	0.189	2.185	0.074	0.125
		0.1	0.197	2.143	0.089	0.130
		0.2	0.182	2.088	0.126	0.145
	Llama-3.1-8B	0.0	0.144	1.136	0.439	0.212
		0.1	0.187	1.714	0.191	0.158
		0.2	0.190	1.858	0.120	0.136
OpenHermes	Gemma2-2B	0.0	0.171	2.547	0.081	0.115
		0.1	0.168	3.022	0.059	0.079
		0.2	0.164	3.056	0.055	0.079
	Mistral-7B	0.0	0.208	2.465	0.078	0.097
		0.1	0.206	2.587	0.091	0.099
		0.2	0.205	2.694	0.088	0.097
	Llama-3.1-8B	0.0	0.162	1.424	0.380	0.201
		0.1	0.206	2.540	0.100	0.103
		0.2	0.209	2.593	0.093	0.100
Tulu3Mixture	Gemma2-2B	0.0	0.159	2.881	0.077	0.095
		0.1	0.155	3.082	0.060	0.084
		0.2	0.143	3.123	0.063	0.083
	Mistral-7B	0.0	0.189	2.735	0.069	0.092
		0.1	0.182	2.866	0.071	0.089
		0.2	0.172	2.896	0.090	0.097
	Llama-3.1-8B	0.0	0.185	1.887	0.097	0.122
		0.1	0.185	2.307	0.034	0.089
		0.2	0.163	2.288	0.056	0.097

Table 58: Results on the MMLU-ProX subset for the ko language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.113	3.165	0.064	0.085
	Mistral-7B	None	0.197	3.053	0.032	0.067
	Llama-3.1-8B	None	0.136	1.658	0.188	0.152
Alpaca	Gemma2-2B	0.0	0.159	2.785	0.066	0.101
		0.1	0.155	2.912	0.068	0.093
		0.2	0.154	2.907	0.063	0.091
	Mistral-7B	0.0	0.189	2.169	0.115	0.144
		0.1	0.197	2.155	0.088	0.131
		0.2	0.180	2.116	0.086	0.130
	Llama-3.1-8B	0.0	0.146	1.218	0.400	0.205
		0.1	0.187	1.825	0.128	0.144
		0.2	0.192	1.961	0.095	0.127
OpenHermes	Gemma2-2B	0.0	0.165	2.403	0.096	0.125
		0.1	0.165	3.005	0.076	0.090
		0.2	0.156	3.034	0.068	0.086
	Mistral-7B	0.0	0.208	2.531	0.084	0.095
		0.1	0.212	2.727	0.071	0.093
		0.2	0.187	2.810	0.089	0.096
	Llama-3.1-8B	0.0	0.161	1.543	0.329	0.189
		0.1	0.199	2.580	0.096	0.100
		0.2	0.206	2.639	0.085	0.095
Tulu3Mixture	Gemma2-2B	0.0	0.148	2.611	0.122	0.115
		0.1	0.148	3.108	0.066	0.084
		0.2	0.143	3.147	0.060	0.082
	Mistral-7B	0.0	0.191	2.755	0.102	0.102
		0.1	0.189	2.917	0.067	0.086
		0.2	0.175	2.934	0.058	0.086
	Llama-3.1-8B	0.0	0.185	1.967	0.085	0.122
		0.1	0.179	2.265	0.075	0.104
		0.2	0.163	2.328	0.051	0.095

Table 59: Results on the MMLU-ProX subset for the mr language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.110	3.163	0.074	0.091
	Mistral-7B	None	0.140	2.991	0.105	0.104
	Llama-3.1-8B	None	0.119	1.571	0.230	0.162
Alpaca	Gemma2-2B	0.0	0.140	2.769	0.086	0.113
		0.1	0.146	2.870	0.067	0.099
		0.2	0.141	2.843	0.068	0.101
	Mistral-7B	0.0	0.147	2.433	0.095	0.138
		0.1	0.152	2.394	0.089	0.138
		0.2	0.139	2.100	0.167	0.173
	Llama-3.1-8B	0.0	0.135	1.278	0.386	0.204
		0.1	0.178	1.916	0.114	0.142
		0.2	0.178	2.018	0.088	0.129
OpenHermes	Gemma2-2B	0.0	0.139	2.443	0.090	0.121
		0.1	0.135	3.067	0.080	0.091
		0.2	0.136	3.094	0.070	0.086
	Mistral-7B	0.0	0.166	2.762	0.109	0.104
		0.1	0.158	2.858	0.103	0.102
		0.2	0.145	2.924	0.112	0.105
	Llama-3.1-8B	0.0	0.142	1.646	0.319	0.188
		0.1	0.185	2.713	0.107	0.105
		0.2	0.191	2.741	0.091	0.097
Tulu3Mixture	Gemma2-2B	0.0	0.131	2.788	0.124	0.113
		0.1	0.130	3.119	0.078	0.090
		0.2	0.128	3.122	0.083	0.091
	Mistral-7B	0.0	0.167	2.918	0.107	0.104
		0.1	0.155	3.019	0.085	0.094
		0.2	0.143	3.037	0.076	0.089
	Llama-3.1-8B	0.0	0.163	2.194	0.082	0.097
		0.1	0.166	2.496	0.049	0.099
		0.2	0.154	2.523	0.072	0.093

Table 60: Results on the MMLU-ProX subset for the ne language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.108	3.174	0.069	0.089
	Mistral-7B	None	0.141	2.997	0.105	0.104
	Llama-3.1-8B	None	0.123	1.764	0.128	0.132
Alpaca	Gemma2-2B	0.0	0.145	2.764	0.074	0.110
		0.1	0.145	2.893	0.064	0.096
		0.2	0.140	2.866	0.067	0.096
	Mistral-7B	0.0	0.148	2.429	0.091	0.134
		0.1	0.149	2.406	0.090	0.135
		0.2	0.141	2.158	0.148	0.167
	Llama-3.1-8B	0.0	0.139	1.360	0.345	0.195
		0.1	0.180	2.012	0.091	0.130
		0.2	0.180	2.124	0.078	0.117
OpenHermes	Gemma2-2B	0.0	0.141	2.412	0.084	0.122
		0.1	0.134	3.066	0.082	0.093
		0.2	0.136	3.095	0.071	0.087
	Mistral-7B	0.0	0.160	2.743	0.117	0.108
		0.1	0.152	2.858	0.109	0.104
		0.2	0.146	2.936	0.107	0.103
	Llama-3.1-8B	0.0	0.155	1.939	0.156	0.139
		0.1	0.186	2.793	0.093	0.098
		0.2	0.188	2.814	0.086	0.095
Tulu3Mixture	Gemma2-2B	0.0	0.137	2.670	0.120	0.115
		0.1	0.127	3.132	0.063	0.082
		0.2	0.120	3.151	0.068	0.085
	Mistral-7B	0.0	0.173	2.904	0.117	0.108
		0.1	0.174	3.000	0.091	0.097
		0.2	0.162	2.996	0.066	0.087
	Llama-3.1-8B	0.0	0.175	2.458	0.108	0.108
		0.1	0.171	2.524	0.056	0.085
		0.2	0.159	2.530	0.067	0.089

Table 61: Results on the MMLU-ProX subset for the pt language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.127	3.145	0.069	0.086
	Mistral-7B	None	0.220	2.966	0.044	0.074
	Llama-3.1-8B	None	0.147	1.580	0.237	0.163
Alpaca	Gemma2-2B	0.0	0.174	2.631	0.081	0.106
		0.1	0.175	2.813	0.060	0.091
		0.2	0.170	2.852	0.060	0.087
	Mistral-7B	0.0	0.216	2.055	0.092	0.132
		0.1	0.231	2.068	0.092	0.127
		0.2	0.213	2.047	0.114	0.138
	Llama-3.1-8B	0.0	0.150	1.195	0.404	0.204
		0.1	0.211	1.785	0.129	0.139
		0.2	0.214	1.878	0.087	0.124
OpenHermes	Gemma2-2B	0.0	0.184	2.153	0.131	0.145
		0.1	0.187	2.990	0.050	0.076
		0.2	0.185	3.031	0.045	0.071
	Mistral-7B	0.0	0.240	2.364	0.066	0.096
		0.1	0.239	2.615	0.068	0.088
		0.2	0.228	2.715	0.064	0.084
	Llama-3.1-8B	0.0	0.176	1.409	0.348	0.194
		0.1	0.235	2.496	0.097	0.101
		0.2	0.240	2.545	0.089	0.098
Tulu3Mixture	Gemma2-2B	0.0	0.173	2.665	0.081	0.098
		0.1	0.169	3.026	0.057	0.080
		0.2	0.157	3.090	0.066	0.085
	Mistral-7B	0.0	0.227	2.569	0.116	0.110
		0.1	0.226	2.746	0.053	0.088
		0.2	0.208	2.735	0.077	0.091
	Llama-3.1-8B	0.0	0.207	1.794	0.130	0.131
		0.1	0.214	2.237	0.051	0.090
		0.2	0.200	2.273	0.044	0.086

Table 62: Results on the MMLU-ProX subset for the rj language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.111	3.137	0.077	0.089
	Mistral-7B	None	0.211	2.983	0.040	0.072
	Llama-3.1-8B	None	0.141	1.614	0.192	0.152
Alpaca	Gemma2-2B	0.0	0.165	2.707	0.069	0.104
		0.1	0.166	2.862	0.055	0.087
		0.2	0.162	2.870	0.062	0.091
	Mistral-7B	0.0	0.213	2.079	0.145	0.151
		0.1	0.219	1.981	0.094	0.132
		0.2	0.199	1.953	0.120	0.139
	Llama-3.1-8B	0.0	0.149	1.191	0.402	0.204
		0.1	0.206	1.804	0.125	0.139
		0.2	0.209	1.918	0.088	0.123
OpenHermes	Gemma2-2B	0.0	0.169	2.210	0.087	0.134
		0.1	0.167	3.062	0.061	0.083
		0.2	0.164	3.093	0.052	0.077
	Mistral-7B	0.0	0.241	2.402	0.071	0.094
		0.1	0.232	2.613	0.060	0.089
		0.2	0.221	2.731	0.062	0.083
	Llama-3.1-8B	0.0	0.172	1.566	0.298	0.182
		0.1	0.226	2.566	0.086	0.097
		0.2	0.232	2.604	0.080	0.092
Tulu3Mixture	Gemma2-2B	0.0	0.161	2.642	0.097	0.107
		0.1	0.156	3.086	0.062	0.084
		0.2	0.136	3.134	0.067	0.085
	Mistral-7B	0.0	0.225	2.592	0.048	0.090
		0.1	0.219	2.731	0.057	0.084
		0.2	0.198	2.771	0.086	0.096
	Llama-3.1-8B	0.0	0.198	1.858	0.111	0.127
		0.1	0.203	2.240	0.042	0.092
		0.2	0.184	2.270	0.049	0.093

Table 63: Results on the MMLU-ProX subset for the sr language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.116	3.103	0.085	0.095
	Mistral-7B	None	0.189	3.012	0.058	0.081
	Llama-3.1-8B	None	0.131	1.630	0.213	0.160
Alpaca	Gemma2-2B	0.0	0.160	2.635	0.085	0.113
		0.1	0.163	2.818	0.055	0.091
		0.2	0.162	2.816	0.050	0.092
	Mistral-7B	0.0	0.199	2.192	0.129	0.147
		0.1	0.212	2.122	0.078	0.124
		0.2	0.191	2.081	0.094	0.130
	Llama-3.1-8B	0.0	0.136	1.188	0.422	0.210
		0.1	0.191	1.834	0.125	0.143
		0.2	0.192	1.934	0.094	0.130
OpenHermes	Gemma2-2B	0.0	0.168	2.366	0.097	0.131
		0.1	0.163	3.065	0.056	0.080
		0.2	0.167	3.105	0.043	0.071
	Mistral-7B	0.0	0.218	2.479	0.079	0.092
		0.1	0.219	2.714	0.066	0.089
		0.2	0.202	2.829	0.068	0.086
	Llama-3.1-8B	0.0	0.151	1.503	0.346	0.194
		0.1	0.197	2.616	0.106	0.104
		0.2	0.204	2.668	0.091	0.099
Tulu3Mixture	Gemma2-2B	0.0	0.160	2.732	0.082	0.101
		0.1	0.156	3.101	0.053	0.077
		0.2	0.137	3.147	0.056	0.080
	Mistral-7B	0.0	0.210	2.691	0.082	0.093
		0.1	0.204	2.835	0.052	0.084
		0.2	0.170	2.796	0.118	0.109
	Llama-3.1-8B	0.0	0.174	1.882	0.111	0.129
		0.1	0.178	2.256	0.093	0.107
		0.2	0.159	2.302	0.074	0.098

Table 64: Results on the MMLU-ProX subset for the sw language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.109	3.134	0.085	0.093
	Mistral-7B	None	0.141	3.027	0.092	0.098
	Llama-3.1-8B	None	0.110	1.531	0.270	0.174
Alpaca	Gemma2-2B	0.0	0.146	2.804	0.089	0.103
		0.1	0.144	2.905	0.080	0.094
		0.2	0.138	2.905	0.086	0.096
	Mistral-7B	0.0	0.148	2.482	0.118	0.156
		0.1	0.162	2.454	0.066	0.113
		0.2	0.147	2.199	0.085	0.113
	Llama-3.1-8B	0.0	0.115	1.154	0.462	0.216
		0.1	0.151	1.825	0.144	0.151
		0.2	0.153	1.953	0.103	0.136
OpenHermes	Gemma2-2B	0.0	0.149	2.521	0.079	0.114
		0.1	0.150	3.113	0.057	0.080
		0.2	0.148	3.138	0.055	0.078
	Mistral-7B	0.0	0.162	2.799	0.108	0.104
		0.1	0.158	2.894	0.105	0.102
		0.2	0.153	2.989	0.096	0.098
	Llama-3.1-8B	0.0	0.122	1.482	0.321	0.188
		0.1	0.153	2.724	0.135	0.117
		0.2	0.154	2.786	0.132	0.114
Tulu3Mixture	Gemma2-2B	0.0	0.142	2.829	0.101	0.103
		0.1	0.142	3.115	0.070	0.086
		0.2	0.134	3.147	0.072	0.087
	Mistral-7B	0.0	0.168	2.937	0.072	0.088
		0.1	0.160	3.036	0.070	0.086
		0.2	0.137	3.049	0.110	0.105
	Llama-3.1-8B	0.0	0.118	1.779	0.244	0.160
		0.1	0.111	2.122	0.208	0.159
		0.2	0.104	2.175	0.232	0.155

Table 65: Results on the MMLU-ProX subset for the te language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.099	3.156	0.092	0.102
	Mistral-7B	None	0.140	3.000	0.110	0.108
	Llama-3.1-8B	None	0.112	1.520	0.256	0.170
Alpaca	Gemma2-2B	0.0	0.137	2.803	0.089	0.108
		0.1	0.133	2.916	0.080	0.096
		0.2	0.128	2.903	0.085	0.100
	Mistral-7B	0.0	0.141	2.406	0.107	0.148
		0.1	0.153	2.414	0.099	0.142
		0.2	0.142	2.244	0.143	0.168
	Llama-3.1-8B	0.0	0.122	1.240	0.418	0.209
		0.1	0.163	1.964	0.103	0.134
		0.2	0.163	2.060	0.096	0.125
OpenHermes	Gemma2-2B	0.0	0.134	2.477	0.101	0.125
		0.1	0.135	3.103	0.072	0.087
		0.2	0.140	3.127	0.060	0.081
	Mistral-7B	0.0	0.150	2.778	0.126	0.112
		0.1	0.151	2.871	0.105	0.101
		0.2	0.139	2.989	0.097	0.100
	Llama-3.1-8B	0.0	0.123	1.523	0.338	0.192
		0.1	0.160	2.716	0.124	0.113
		0.2	0.165	2.756	0.111	0.107
Tulu3Mixture	Gemma2-2B	0.0	0.128	2.833	0.111	0.109
		0.1	0.128	3.150	0.067	0.085
		0.2	0.125	3.160	0.068	0.085
	Mistral-7B	0.0	0.157	2.984	0.119	0.110
		0.1	0.141	3.019	0.075	0.089
		0.2	0.131	3.017	0.137	0.116
	Llama-3.1-8B	0.0	0.135	2.043	0.076	0.111
		0.1	0.140	2.560	0.061	0.095
		0.2	0.121	2.524	0.058	0.088

Table 66: Results on the MMLU-ProX subset for the th language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.108	3.174	0.070	0.088
	Mistral-7B	None	0.152	3.027	0.075	0.090
	Llama-3.1-8B	None	0.126	1.529	0.228	0.162
Alpaca	Gemma2-2B	0.0	0.143	2.780	0.086	0.107
		0.1	0.140	2.899	0.085	0.100
		0.2	0.137	2.919	0.082	0.098
	Mistral-7B	0.0	0.142	2.166	0.138	0.165
		0.1	0.154	2.369	0.082	0.134
		0.2	0.140	2.448	0.089	0.125
	Llama-3.1-8B	0.0	0.128	1.171	0.434	0.210
		0.1	0.172	1.768	0.164	0.154
		0.2	0.177	1.908	0.104	0.135
OpenHermes	Gemma2-2B	0.0	0.149	2.279	0.086	0.134
		0.1	0.143	3.079	0.075	0.089
		0.2	0.146	3.105	0.062	0.081
	Mistral-7B	0.0	0.162	2.768	0.104	0.103
		0.1	0.160	2.881	0.104	0.102
		0.2	0.157	2.960	0.094	0.097
	Llama-3.1-8B	0.0	0.137	1.512	0.352	0.194
		0.1	0.192	2.654	0.101	0.104
		0.2	0.198	2.707	0.091	0.097
Tulu3Mixture	Gemma2-2B	0.0	0.147	2.761	0.085	0.103
		0.1	0.144	3.098	0.061	0.083
		0.2	0.140	3.131	0.067	0.086
	Mistral-7B	0.0	0.157	2.913	0.111	0.106
		0.1	0.154	3.014	0.072	0.087
		0.2	0.138	3.012	0.080	0.092
	Llama-3.1-8B	0.0	0.160	1.890	0.098	0.119
		0.1	0.157	2.282	0.077	0.100
		0.2	0.148	2.337	0.051	0.091

Table 67: Results on the MMLU-ProX subset for the uk language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.109	3.143	0.076	0.090
	Mistral-7B	None	0.201	2.989	0.047	0.073
	Llama-3.1-8B	None	0.138	1.639	0.175	0.147
Alpaca	Gemma2-2B	0.0	0.161	2.692	0.076	0.108
		0.1	0.161	2.858	0.059	0.090
		0.2	0.163	2.877	0.053	0.086
	Mistral-7B	0.0	0.210	2.109	0.090	0.131
		0.1	0.220	2.008	0.117	0.138
		0.2	0.201	1.992	0.136	0.148
	Llama-3.1-8B	0.0	0.150	1.218	0.389	0.202
		0.1	0.202	1.838	0.111	0.137
		0.2	0.205	1.953	0.084	0.121
OpenHermes	Gemma2-2B	0.0	0.165	2.244	0.085	0.133
		0.1	0.162	3.074	0.061	0.082
		0.2	0.161	3.107	0.050	0.076
	Mistral-7B	0.0	0.226	2.431	0.063	0.093
		0.1	0.225	2.635	0.071	0.088
		0.2	0.223	2.760	0.067	0.086
	Llama-3.1-8B	0.0	0.164	1.599	0.272	0.176
		0.1	0.214	2.590	0.095	0.100
		0.2	0.222	2.633	0.086	0.096
Tulu3Mixture	Gemma2-2B	0.0	0.155	2.634	0.097	0.106
		0.1	0.152	3.103	0.062	0.083
		0.2	0.146	3.147	0.062	0.083
	Mistral-7B	0.0	0.223	2.607	0.062	0.088
		0.1	0.214	2.769	0.061	0.084
		0.2	0.196	2.796	0.099	0.103
	Llama-3.1-8B	0.0	0.190	1.933	0.089	0.116
		0.1	0.194	2.220	0.065	0.098
		0.2	0.175	2.271	0.055	0.092

Table 68: Results on the MMLU-ProX subset for the ur language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.105	3.156	0.072	0.088
	Mistral-7B	None	0.145	3.012	0.102	0.103
	Llama-3.1-8B	None	0.120	1.459	0.282	0.177
Alpaca	Gemma2-2B	0.0	0.143	2.805	0.082	0.106
		0.1	0.143	2.893	0.075	0.098
		0.2	0.141	2.884	0.073	0.098
	Mistral-7B	0.0	0.147	2.360	0.130	0.159
		0.1	0.155	2.339	0.087	0.139
		0.2	0.140	2.220	0.098	0.136
	Llama-3.1-8B	0.0	0.133	1.175	0.438	0.213
		0.1	0.169	1.724	0.199	0.163
		0.2	0.172	1.832	0.147	0.146
OpenHermes	Gemma2-2B	0.0	0.147	2.458	0.082	0.118
		0.1	0.149	3.062	0.082	0.093
		0.2	0.141	3.095	0.069	0.085
	Mistral-7B	0.0	0.171	2.687	0.116	0.109
		0.1	0.166	2.823	0.110	0.105
		0.2	0.157	2.936	0.098	0.099
	Llama-3.1-8B	0.0	0.137	1.414	0.401	0.205
		0.1	0.186	2.669	0.107	0.104
		0.2	0.190	2.710	0.100	0.101
Tulu3Mixture	Gemma2-2B	0.0	0.137	2.827	0.100	0.105
		0.1	0.136	3.139	0.062	0.081
		0.2	0.133	3.163	0.061	0.081
	Mistral-7B	0.0	0.163	2.882	0.123	0.111
		0.1	0.158	2.986	0.086	0.095
		0.2	0.139	2.995	0.089	0.096
	Llama-3.1-8B	0.0	0.155	1.879	0.112	0.132
		0.1	0.157	2.294	0.032	0.089
		0.2	0.145	2.360	0.039	0.085

Table 69: Results on the MMLU-ProX subset for the vi language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.102	3.026	0.130	0.117
	Mistral-7B	None	0.191	3.003	0.045	0.072
	Llama-3.1-8B	None	0.141	1.667	0.169	0.146
Alpaca	Gemma2-2B	0.0	0.156	2.636	0.110	0.113
		0.1	0.156	2.850	0.079	0.099
		0.2	0.153	2.864	0.082	0.097
	Mistral-7B	0.0	0.186	2.198	0.077	0.125
		0.1	0.197	2.169	0.078	0.127
		0.2	0.179	2.193	0.087	0.132
	Llama-3.1-8B	0.0	0.157	1.245	0.383	0.201
		0.1	0.214	1.862	0.105	0.131
		0.2	0.216	1.991	0.080	0.116
OpenHermes	Gemma2-2B	0.0	0.173	2.413	0.104	0.125
		0.1	0.173	3.074	0.045	0.074
		0.2	0.175	3.113	0.032	0.066
	Mistral-7B	0.0	0.191	2.577	0.100	0.098
		0.1	0.200	2.759	0.093	0.098
		0.2	0.185	2.854	0.086	0.095
	Llama-3.1-8B	0.0	0.181	1.616	0.277	0.177
		0.1	0.233	2.555	0.080	0.093
		0.2	0.237	2.579	0.069	0.088
Tulu3Mixture	Gemma2-2B	0.0	0.148	3.054	0.089	0.095
		0.1	0.144	3.120	0.088	0.095
		0.2	0.138	3.266	0.066	0.102
	Mistral-7B	0.0	0.198	2.709	0.117	0.109
		0.1	0.188	2.758	0.093	0.099
		0.2	0.175	2.841	0.049	0.083
	Llama-3.1-8B	0.0	0.209	2.095	0.070	0.108
		0.1	0.209	2.436	0.050	0.085
		0.2	0.194	2.476	0.059	0.089

Table 70: Results on the MMLU-ProX subset for the wo language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.103	3.150	0.087	0.094
	Mistral-7B	None	0.141	3.038	0.087	0.095
	Llama-3.1-8B	None	0.113	1.637	0.169	0.149
Alpaca	Gemma2-2B	0.0	0.136	2.968	0.090	0.100
		0.1	0.129	3.036	0.075	0.089
		0.2	0.130	3.051	0.074	0.088
	Mistral-7B	0.0	0.150	2.511	0.101	0.117
		0.1	0.163	2.454	0.077	0.110
		0.2	0.151	2.291	0.085	0.139
	Llama-3.1-8B	0.0	0.116	1.190	0.435	0.211
		0.1	0.141	1.921	0.113	0.141
		0.2	0.145	2.098	0.084	0.126
OpenHermes	Gemma2-2B	0.0	0.135	2.678	0.090	0.102
		0.1	0.131	3.154	0.068	0.086
		0.2	0.131	3.186	0.055	0.077
	Mistral-7B	0.0	0.161	2.789	0.113	0.107
		0.1	0.160	2.916	0.097	0.099
		0.2	0.157	2.998	0.090	0.095
	Llama-3.1-8B	0.0	0.119	1.613	0.305	0.184
		0.1	0.148	2.800	0.143	0.119
		0.2	0.149	2.869	0.131	0.114
Tulu3Mixture	Gemma2-2B	0.0	0.137	2.963	0.095	0.099
		0.1	0.131	3.162	0.068	0.085
		0.2	0.126	3.189	0.063	0.084
	Mistral-7B	0.0	0.152	2.880	0.094	0.097
		0.1	0.153	3.060	0.067	0.085
		0.2	0.133	3.018	0.113	0.105
	Llama-3.1-8B	0.0	0.118	1.909	0.094	0.125
		0.1	0.111	2.360	0.056	0.092
		0.2	0.102	2.364	0.067	0.093

Table 71: Results on the MMLU-ProX subset for the yo language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.106	3.111	0.099	0.100
	Mistral-7B	None	0.143	3.035	0.085	0.093
	Llama-3.1-8B	None	0.108	1.666	0.165	0.147
Alpaca	Gemma2-2B	0.0	0.134	2.914	0.099	0.104
		0.1	0.131	3.004	0.076	0.090
		0.2	0.128	2.989	0.090	0.097
	Mistral-7B	0.0	0.143	2.503	0.108	0.121
		0.1	0.153	2.505	0.093	0.115
		0.2	0.143	2.380	0.088	0.133
	Llama-3.1-8B	0.0	0.114	1.263	0.409	0.208
		0.1	0.147	1.999	0.103	0.135
		0.2	0.147	2.101	0.096	0.127
OpenHermes	Gemma2-2B	0.0	0.125	2.625	0.080	0.101
		0.1	0.131	3.143	0.078	0.091
		0.2	0.125	3.174	0.068	0.084
	Mistral-7B	0.0	0.148	2.792	0.109	0.104
		0.1	0.162	2.897	0.100	0.100
		0.2	0.141	2.985	0.096	0.098
	Llama-3.1-8B	0.0	0.123	1.730	0.228	0.164
		0.1	0.147	2.842	0.136	0.116
		0.2	0.153	2.894	0.121	0.110
Tulu3Mixture	Gemma2-2B	0.0	0.122	2.834	0.123	0.113
		0.1	0.118	3.096	0.083	0.092
		0.2	0.114	3.115	0.085	0.094
	Mistral-7B	0.0	0.155	2.888	0.088	0.096
		0.1	0.158	3.034	0.065	0.085
		0.2	0.140	3.027	0.104	0.103
	Llama-3.1-8B	0.0	0.115	1.998	0.081	0.121
		0.1	0.115	2.523	0.082	0.099
		0.2	0.101	2.378	0.073	0.100

Table 72: Results on the MMLU-ProX subset for the zh language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.111	3.203	0.056	0.081
	Mistral-7B	None	0.202	3.003	0.043	0.073
	Llama-3.1-8B	None	0.155	1.789	0.102	0.123
Alpaca	Gemma2-2B	0.0	0.171	2.869	0.079	0.090
		0.1	0.174	2.907	0.062	0.089
		0.2	0.172	2.903	0.063	0.090
	Mistral-7B	0.0	0.206	2.038	0.109	0.136
		0.1	0.214	2.049	0.104	0.135
		0.2	0.193	2.095	0.104	0.138
	Llama-3.1-8B	0.0	0.148	1.133	0.437	0.212
		0.1	0.200	1.689	0.202	0.157
		0.2	0.202	1.804	0.139	0.140
OpenHermes	Gemma2-2B	0.0	0.173	2.445	0.097	0.123
		0.1	0.174	2.993	0.056	0.080
		0.2	0.171	3.026	0.048	0.076
	Mistral-7B	0.0	0.208	2.426	0.077	0.095
		0.1	0.226	2.607	0.065	0.093
		0.2	0.204	2.669	0.062	0.092
	Llama-3.1-8B	0.0	0.188	1.670	0.256	0.169
		0.1	0.232	2.526	0.076	0.092
		0.2	0.236	2.535	0.070	0.090
Tulu3Mixture	Gemma2-2B	0.0	0.144	2.787	0.089	0.102
		0.1	0.148	3.007	0.074	0.090
		0.2	0.140	3.064	0.075	0.090
	Mistral-7B	0.0	0.202	2.655	0.039	0.088
		0.1	0.211	2.786	0.044	0.079
		0.2	0.203	2.861	0.070	0.087
	Llama-3.1-8B	0.0	0.208	2.026	0.087	0.116
		0.1	0.214	2.426	0.050	0.086
		0.2	0.204	2.393	0.033	0.083

Table 73: Results on the MMLU-ProX subset for the zu language by model, SFT dataset, and label smoothing.

SFT Dataset	Base	Smoothing	Accuracy	Entropy	ECE	RMS
Base Model	Gemma2-2B	None	0.107	3.143	0.084	0.094
	Mistral-7B	None	0.144	3.030	0.089	0.096
	Llama-3.1-8B	None	0.106	1.505	0.253	0.171
Alpaca	Gemma2-2B	0.0	0.125	2.899	0.118	0.112
		0.1	0.129	2.982	0.094	0.100
		0.2	0.126	3.007	0.095	0.099
	Mistral-7B	0.0	0.145	2.527	0.099	0.116
		0.1	0.156	2.499	0.075	0.112
		0.2	0.143	2.265	0.104	0.148
	Llama-3.1-8B	0.0	0.111	1.114	0.476	0.218
		0.1	0.134	1.811	0.151	0.154
		0.2	0.137	1.941	0.108	0.141
OpenHermes	Gemma2-2B	0.0	0.134	2.734	0.094	0.103
		0.1	0.132	3.169	0.061	0.082
		0.2	0.133	3.191	0.051	0.075
	Mistral-7B	0.0	0.143	2.797	0.112	0.106
		0.1	0.159	2.895	0.110	0.104
		0.2	0.146	2.983	0.102	0.101
	Llama-3.1-8B	0.0	0.112	1.473	0.310	0.185
		0.1	0.139	2.735	0.152	0.123
		0.2	0.146	2.821	0.140	0.119
Tulu3Mixture	Gemma2-2B	0.0	0.117	2.982	0.103	0.102
		0.1	0.132	3.181	0.061	0.082
		0.2	0.119	3.211	0.053	0.079
	Mistral-7B	0.0	0.144	2.936	0.094	0.097
		0.1	0.167	3.022	0.066	0.084
		0.2	0.156	3.045	0.061	0.081
	Llama-3.1-8B	0.0	0.101	1.687	0.229	0.166
		0.1	0.113	2.236	0.097	0.112
		0.2	0.108	2.221	0.153	0.129

D Individual Reliability Plots

D.1 GlobalMMLU

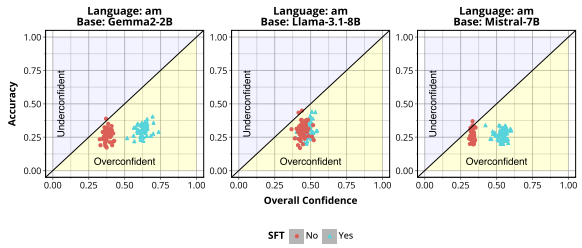

Figure 3: Reliability diagrams for the **GlobalMMLU** dataset for the am language.

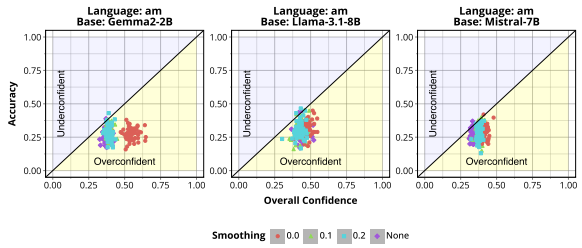

Figure 4: Reliability diagrams for the **GlobalMMLU** dataset for the am language after instruction-tuning on the **Tulu3Mixture** dataset.

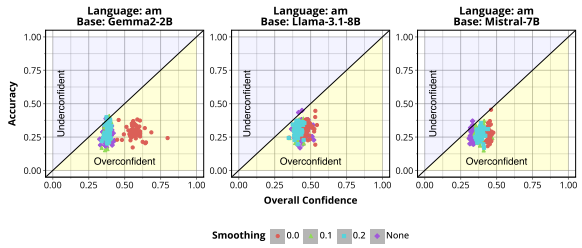

Figure 5: Reliability diagrams for the **GlobalMMLU** dataset for the am language after instruction-tuning on the **OpenHermes** dataset.

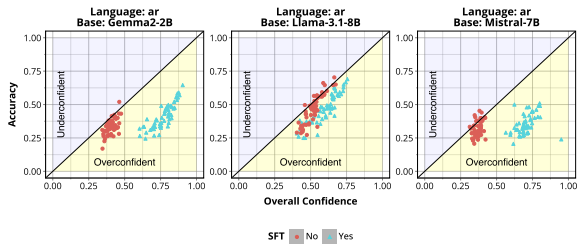

Figure 6: Reliability diagrams for the **GlobalMMLU** dataset for the ar language.

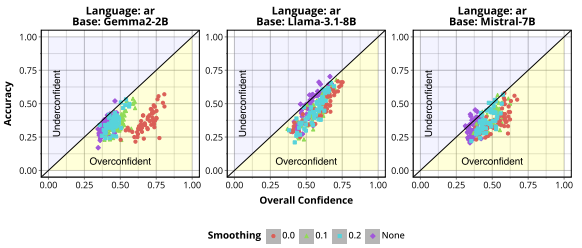

Figure 7: Reliability diagrams for the **GlobalMMLU** dataset for the ar language after instruction-tuning on the **Tulu3Mixture** dataset.

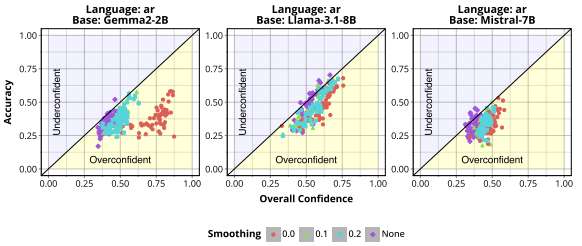

Figure 8: Reliability diagrams for the **GlobalMMLU** dataset for the ar language after instruction-tuning on the **OpenHermes** dataset.

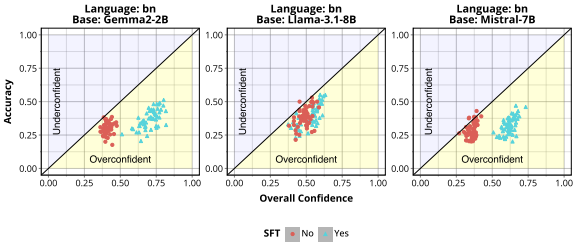

Figure 9: Reliability diagrams for the **GlobalMMLU** dataset for the bn language.

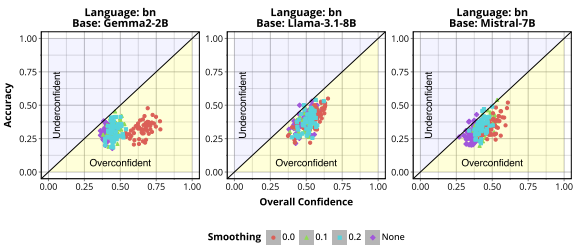

Figure 10: Reliability diagrams for the **GlobalMMLU** dataset for the bn language after instruction-tuning on the **Tulu3Mixture** dataset.

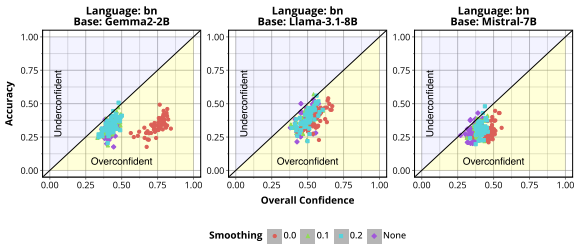

Figure 11: Reliability diagrams for the **GlobalMMLU** dataset for the bn language after instruction-tuning on the **OpenHermes** dataset.

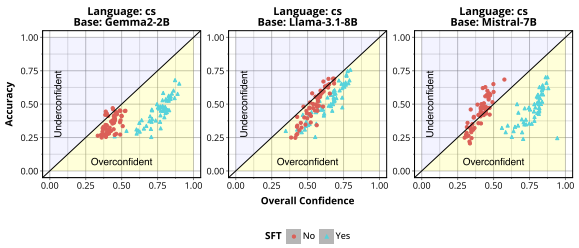

Figure 12: Reliability diagrams for the **GlobalMMLU** dataset for the cs language.

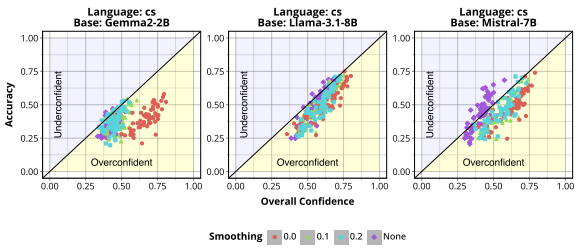

Figure 13: Reliability diagrams for the **GlobalMMLU** dataset for the cs language after instruction-tuning on the **Tulu3Mixture** dataset.

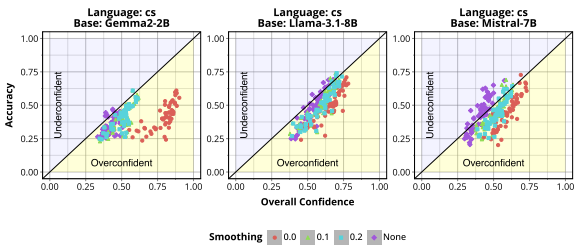

Figure 14: Reliability diagrams for the **GlobalMMLU** dataset for the cs language after instruction-tuning on the **OpenHermes** dataset.

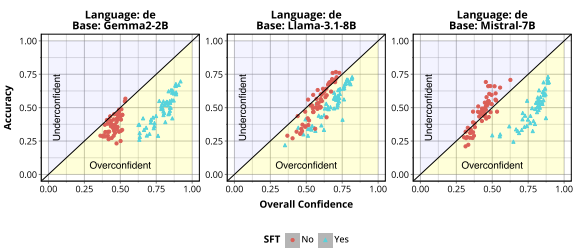

Figure 15: Reliability diagrams for the **GlobalMMLU** dataset for the de language.

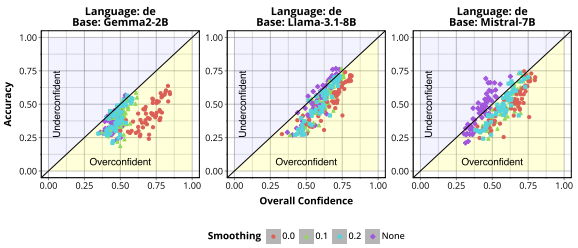

Figure 16: Reliability diagrams for the **GlobalMMLU** dataset for the de language after instruction-tuning on the **Tulu3Mixture** dataset.

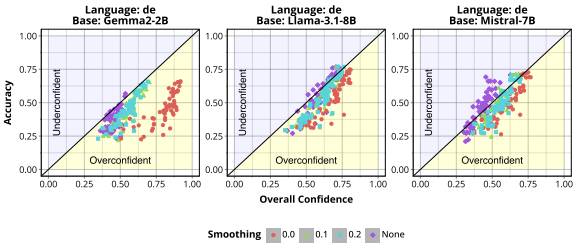

Figure 17: Reliability diagrams for the **GlobalMMLU** dataset for the de language after instruction-tuning on the **OpenHermes** dataset.

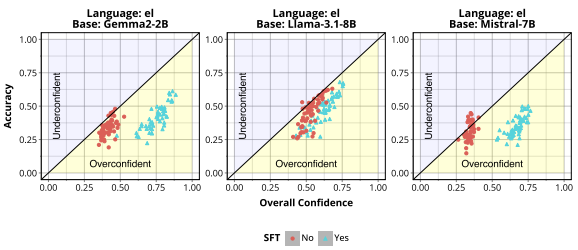

Figure 18: Reliability diagrams for the **GlobalMMLU** dataset for the el language.

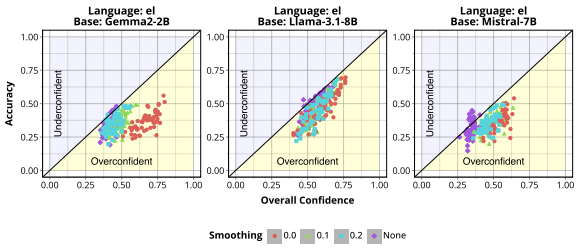

Figure 19: Reliability diagrams for the **GlobalMMLU** dataset for the el language after instruction-tuning on the **Tulu3Mixture** dataset.

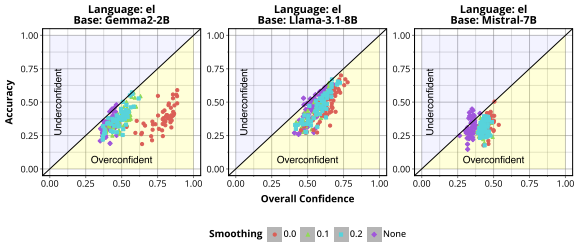

Figure 20: Reliability diagrams for the **GlobalMMLU** dataset for the el language after instruction-tuning on the **OpenHermes** dataset.

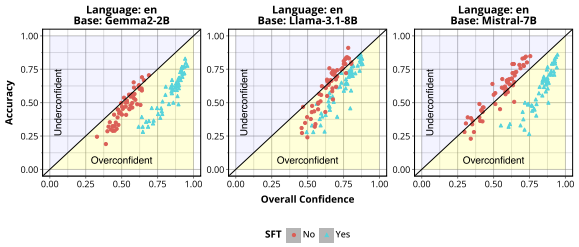

Figure 21: Reliability diagrams for the **GlobalMMLU** dataset for the en language.

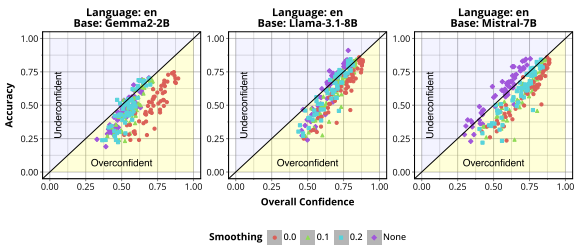

Figure 22: Reliability diagrams for the **GlobalMMLU** dataset for the en language after instruction-tuning on the **Tulu3Mixture** dataset.

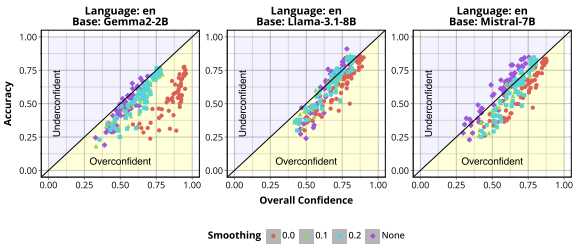

Figure 23: Reliability diagrams for the **GlobalMMLU** dataset for the en language after instruction-tuning on the **OpenHermes** dataset.

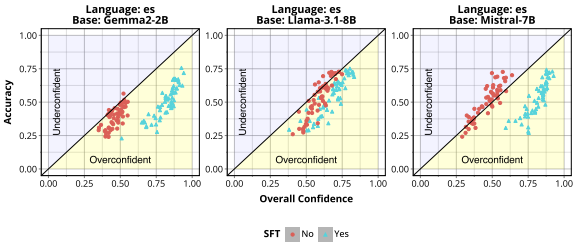

Figure 24: Reliability diagrams for the **GlobalMMLU** dataset for the es language.

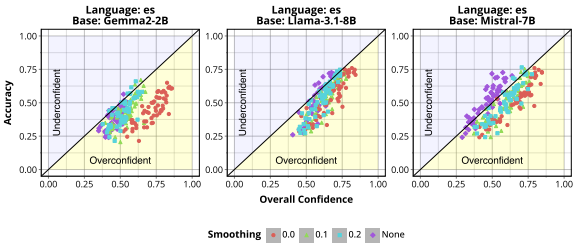

Figure 25: Reliability diagrams for the **GlobalMMLU** dataset for the es language after instruction-tuning on the **Tulu3Mixture** dataset.

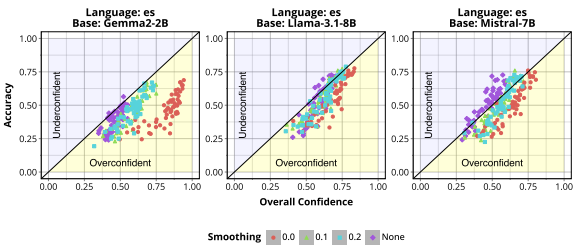

Figure 26: Reliability diagrams for the **GlobalMMLU** dataset for the es language after instruction-tuning on the **OpenHermes** dataset.

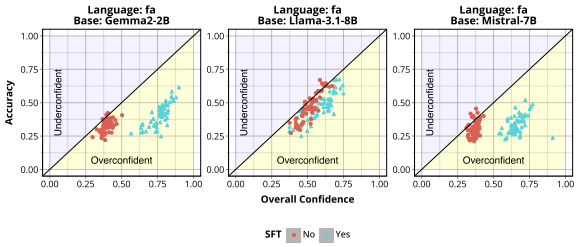

Figure 27: Reliability diagrams for the **GlobalMMLU** dataset for the `fa` language.

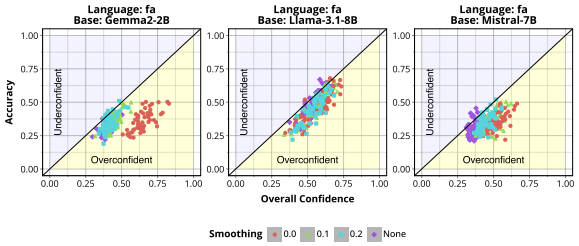

Figure 28: Reliability diagrams for the **GlobalMMLU** dataset for the `fa` language after instruction-tuning on the **Tulu3Mixture** dataset.

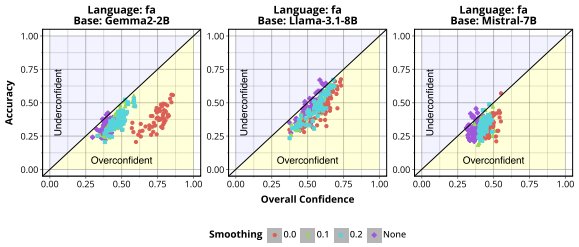

Figure 29: Reliability diagrams for the **GlobalMMLU** dataset for the `fa` language after instruction-tuning on the **OpenHermes** dataset.

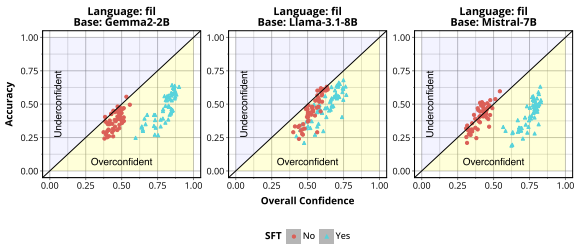

Figure 30: Reliability diagrams for the **GlobalMMLU** dataset for the `fil` language.

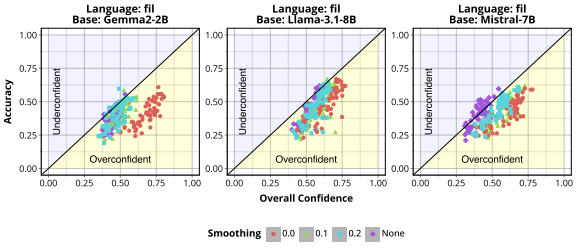

Figure 31: Reliability diagrams for the **GlobalMMLU** dataset for the `fil` language after instruction-tuning on the **Tulu3Mixture** dataset.

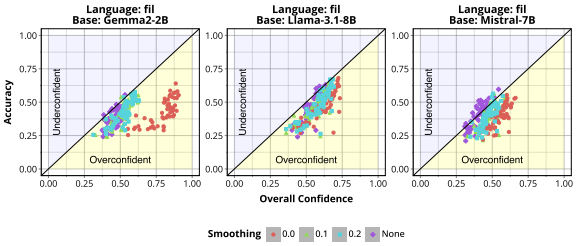

Figure 32: Reliability diagrams for the **GlobalMMLU** dataset for the `fil` language after instruction-tuning on the **OpenHermes** dataset.

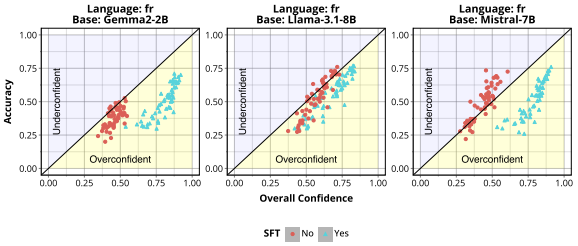

Figure 33: Reliability diagrams for the **GlobalMMLU** dataset for the `fr` language.

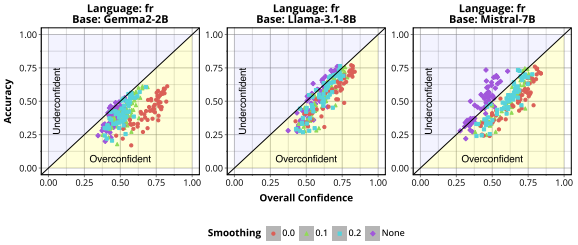

Figure 34: Reliability diagrams for the **GlobalMMLU** dataset for the `fr` language after instruction-tuning on the **Tulu3Mixture** dataset.

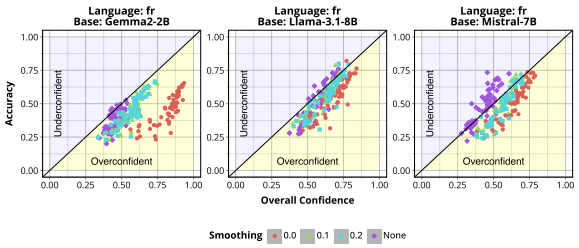

Figure 35: Reliability diagrams for the **GlobalMMLU** dataset for the fr language after instruction-tuning on the **OpenHermes** dataset.

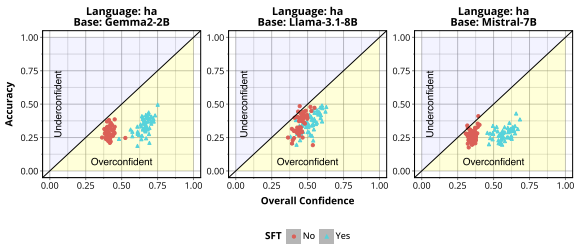

Figure 36: Reliability diagrams for the **GlobalMMLU** dataset for the ha language.

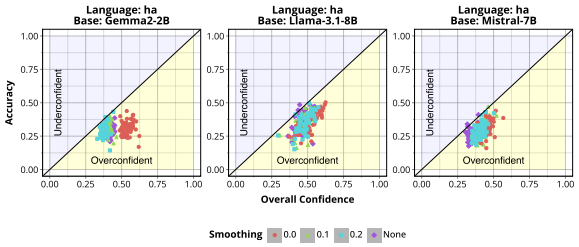

Figure 37: Reliability diagrams for the **GlobalMMLU** dataset for the ha language after instruction-tuning on the **Tulu3Mixture** dataset.

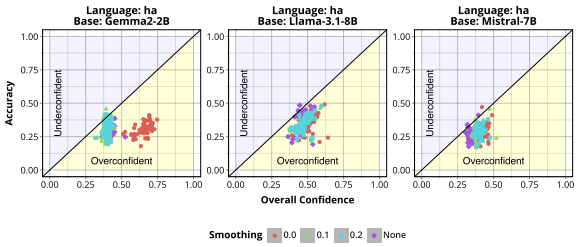

Figure 38: Reliability diagrams for the **GlobalMMLU** dataset for the ha language after instruction-tuning on the **OpenHermes** dataset.

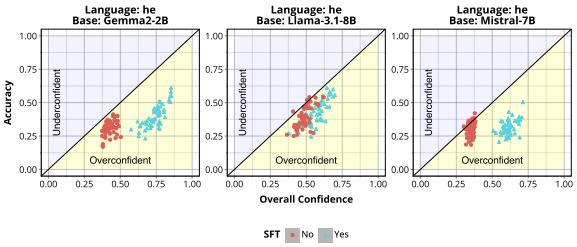

Figure 39: Reliability diagrams for the **GlobalMMLU** dataset for the he language.

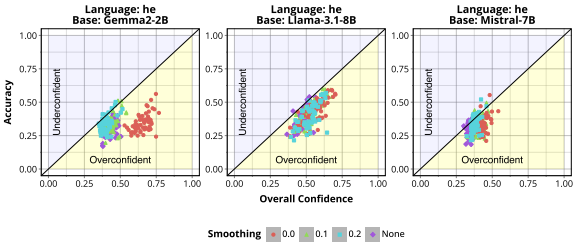

Figure 40: Reliability diagrams for the **GlobalMMLU** dataset for the he language after instruction-tuning on the **Tulu3Mixture** dataset.

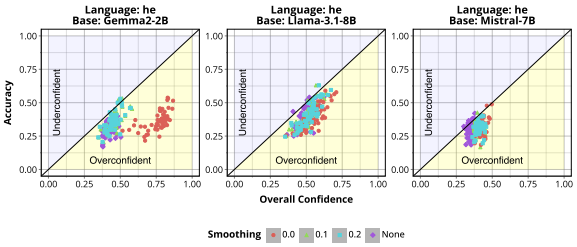

Figure 41: Reliability diagrams for the **GlobalMMLU** dataset for the he language after instruction-tuning on the **OpenHermes** dataset.

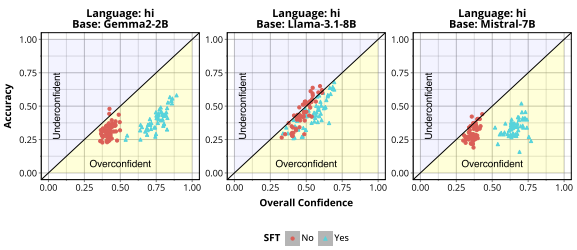

Figure 42: Reliability diagrams for the **GlobalMMLU** dataset for the hi language.

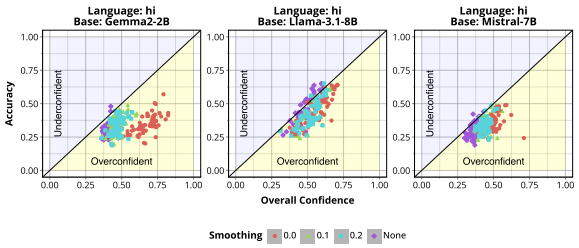

Figure 43: Reliability diagrams for the **GlobalMMLU** dataset for the hi language after instruction-tuning on the **Tulu3Mixture** dataset.

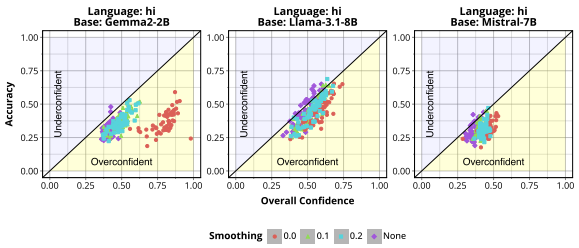

Figure 44: Reliability diagrams for the **GlobalMMLU** dataset for the hi language after instruction-tuning on the **OpenHermes** dataset.

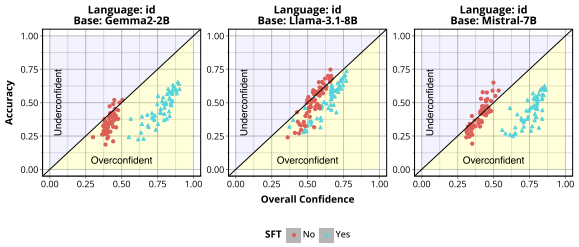

Figure 45: Reliability diagrams for the **GlobalMMLU** dataset for the id language.

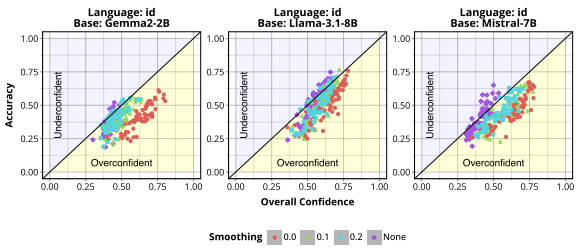

Figure 46: Reliability diagrams for the **GlobalMMLU** dataset for the id language after instruction-tuning on the **Tulu3Mixture** dataset.

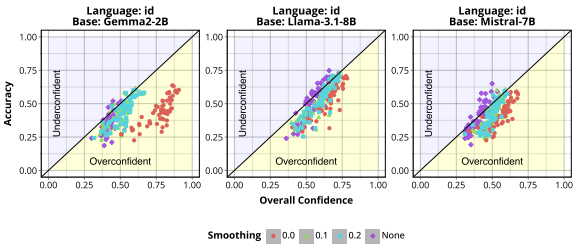

Figure 47: Reliability diagrams for the **GlobalMMLU** dataset for the id language after instruction-tuning on the **OpenHermes** dataset.

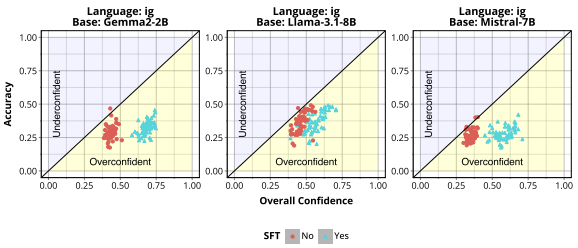

Figure 48: Reliability diagrams for the **GlobalMMLU** dataset for the ig language.

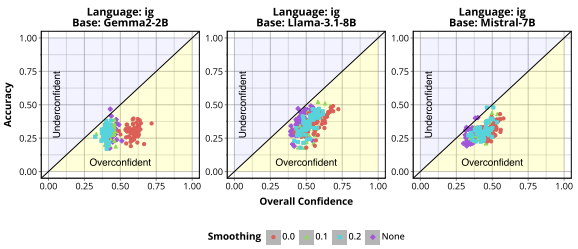

Figure 49: Reliability diagrams for the **GlobalMMLU** dataset for the ig language after instruction-tuning on the **Tulu3Mixture** dataset.

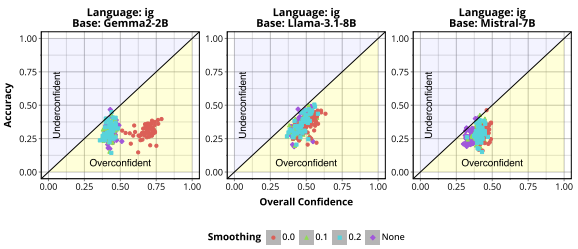

Figure 50: Reliability diagrams for the **GlobalMMLU** dataset for the ig language after instruction-tuning on the **OpenHermes** dataset.

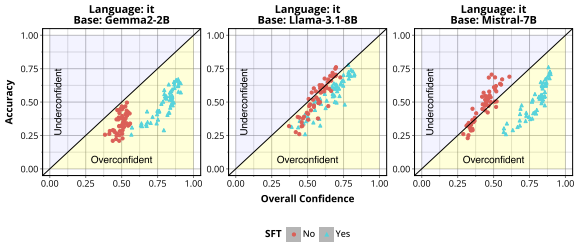

Figure 51: Reliability diagrams for the **GlobalMMLU** dataset for the `it` language.

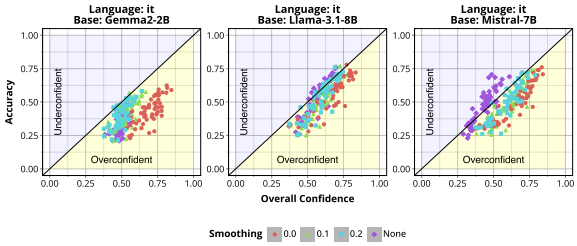

Figure 52: Reliability diagrams for the **GlobalMMLU** dataset for the `it` language after instruction-tuning on the **Tulu3Mixture** dataset.

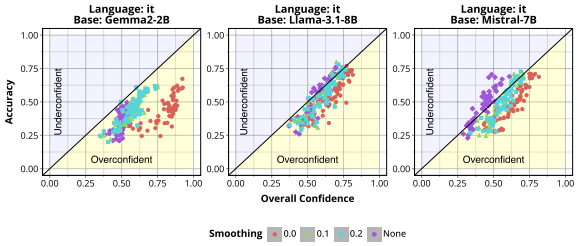

Figure 53: Reliability diagrams for the **GlobalMMLU** dataset for the `it` language after instruction-tuning on the **OpenHermes** dataset.

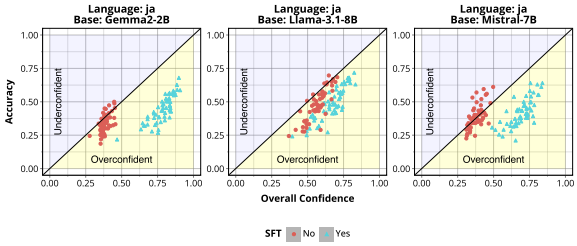

Figure 54: Reliability diagrams for the **GlobalMMLU** dataset for the `ja` language.

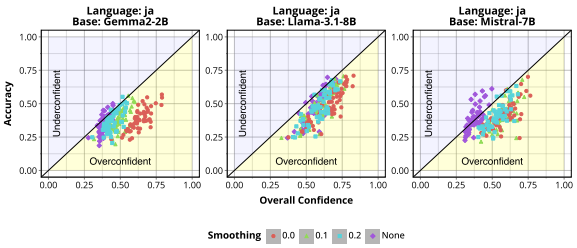

Figure 55: Reliability diagrams for the **GlobalMMLU** dataset for the `ja` language after instruction-tuning on the **Tulu3Mixture** dataset.

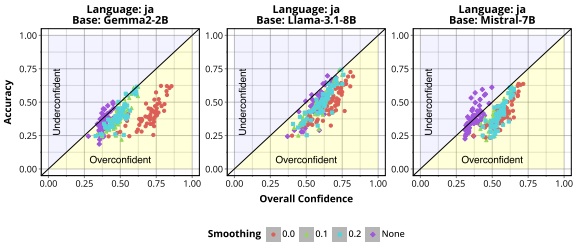

Figure 56: Reliability diagrams for the **GlobalMMLU** dataset for the `ja` language after instruction-tuning on the **OpenHermes** dataset.

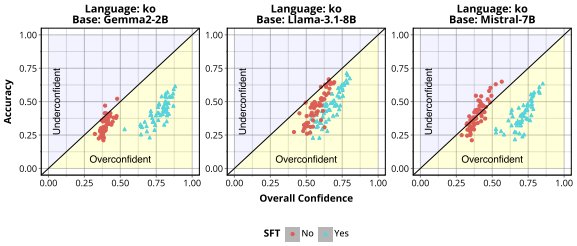

Figure 57: Reliability diagrams for the **GlobalMMLU** dataset for the `ko` language.

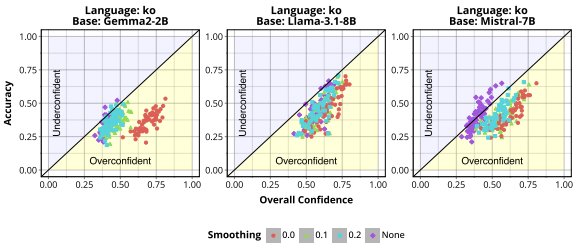

Figure 58: Reliability diagrams for the **GlobalMMLU** dataset for the `ko` language after instruction-tuning on the **Tulu3Mixture** dataset.

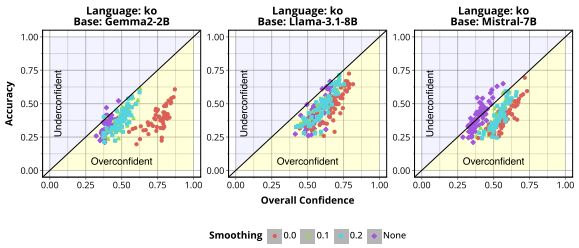

Figure 59: Reliability diagrams for the **GlobalMMLU** dataset for the ko language after instruction-tuning on the **OpenHermes** dataset.

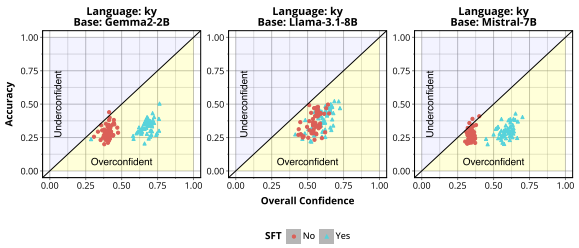

Figure 60: Reliability diagrams for the **GlobalMMLU** dataset for the ky language.

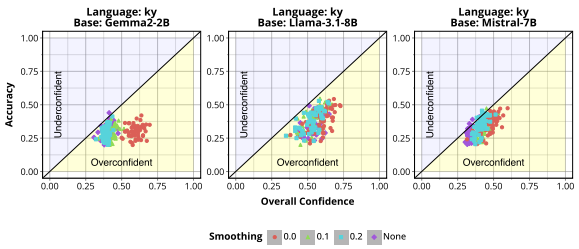

Figure 61: Reliability diagrams for the **GlobalMMLU** dataset for the ky language after instruction-tuning on the **Tulu3Mixture** dataset.

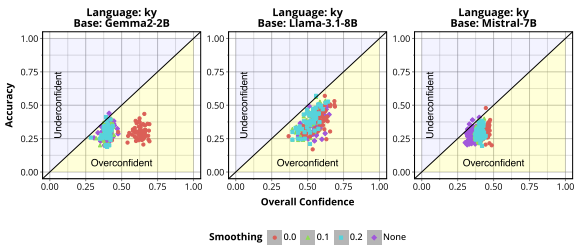

Figure 62: Reliability diagrams for the **GlobalMMLU** dataset for the ky language after instruction-tuning on the **OpenHermes** dataset.

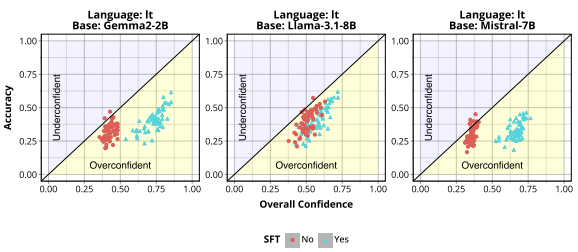

Figure 63: Reliability diagrams for the **GlobalMMLU** dataset for the lt language.

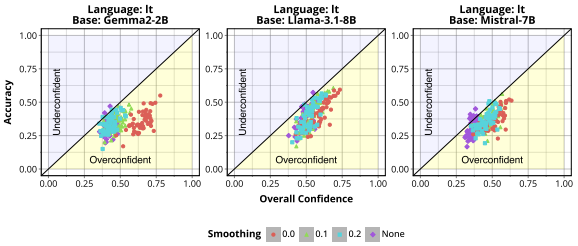

Figure 64: Reliability diagrams for the **GlobalMMLU** dataset for the lt language after instruction-tuning on the **Tulu3Mixture** dataset.

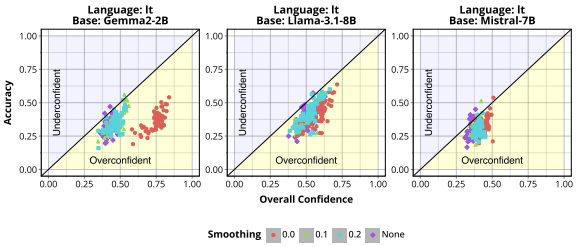

Figure 65: Reliability diagrams for the **GlobalMMLU** dataset for the lt language after instruction-tuning on the **OpenHermes** dataset.

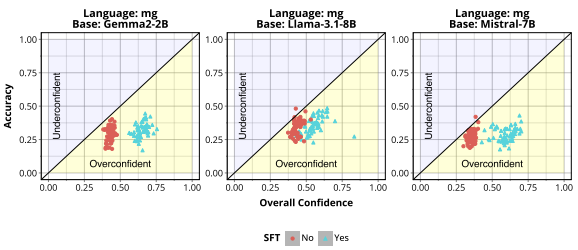

Figure 66: Reliability diagrams for the **GlobalMMLU** dataset for the mg language.

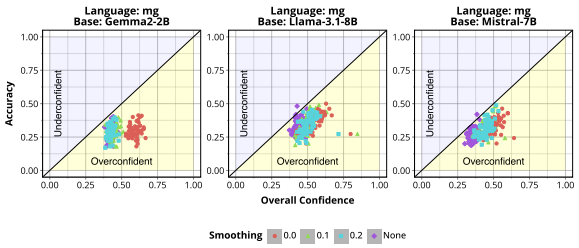

Figure 67: Reliability diagrams for the **GlobalMMLU** dataset for the mg language after instruction-tuning on the **Tulu3Mixture** dataset.

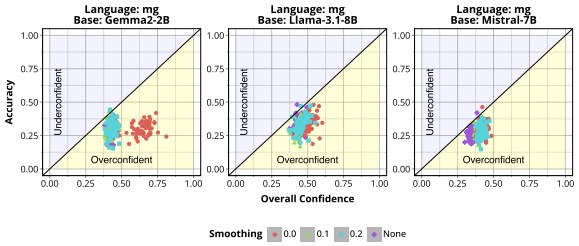

Figure 68: Reliability diagrams for the **GlobalMMLU** dataset for the mg language after instruction-tuning on the **OpenHermes** dataset.

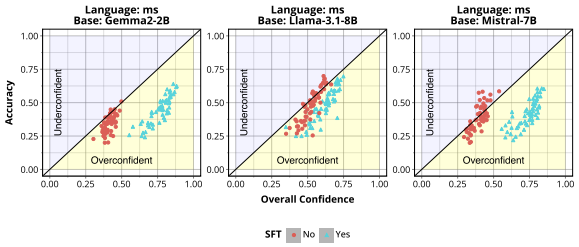

Figure 69: Reliability diagrams for the **GlobalMMLU** dataset for the ms language.

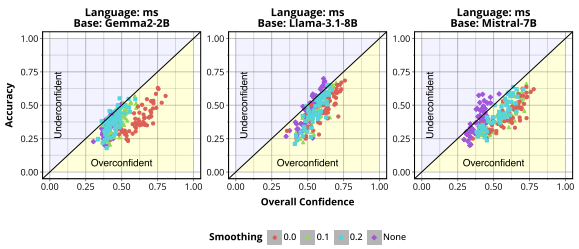

Figure 70: Reliability diagrams for the **GlobalMMLU** dataset for the ms language after instruction-tuning on the **Tulu3Mixture** dataset.

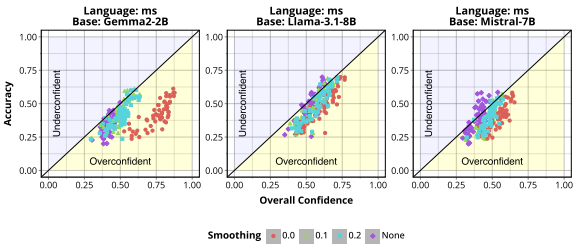

Figure 71: Reliability diagrams for the **GlobalMMLU** dataset for the ms language after instruction-tuning on the **OpenHermes** dataset.

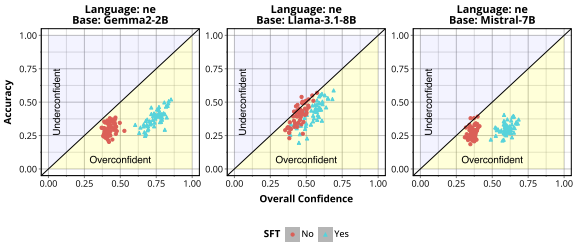

Figure 72: Reliability diagrams for the **GlobalMMLU** dataset for the ne language.

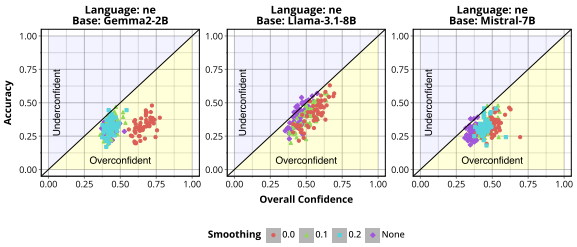

Figure 73: Reliability diagrams for the **GlobalMMLU** dataset for the ne language after instruction-tuning on the **Tulu3Mixture** dataset.

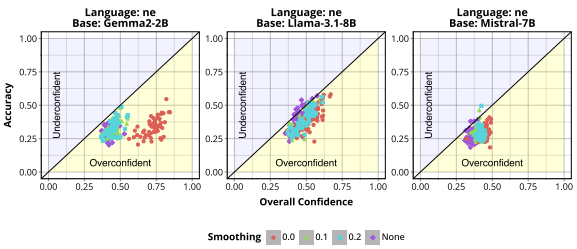

Figure 74: Reliability diagrams for the **GlobalMMLU** dataset for the ne language after instruction-tuning on the **OpenHermes** dataset.

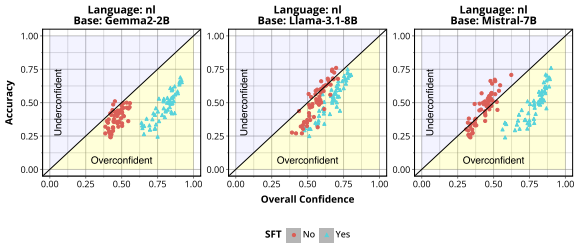

Figure 75: Reliability diagrams for the **GlobalMMLU** dataset for the nl language.

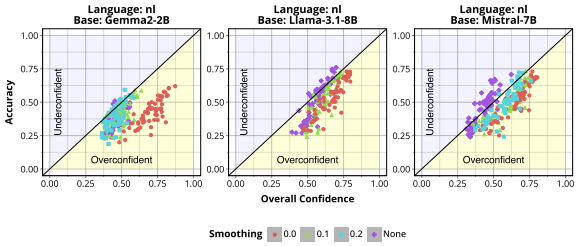

Figure 76: Reliability diagrams for the **GlobalMMLU** dataset for the nl language after instruction-tuning on the **Tulu3Mixture** dataset.

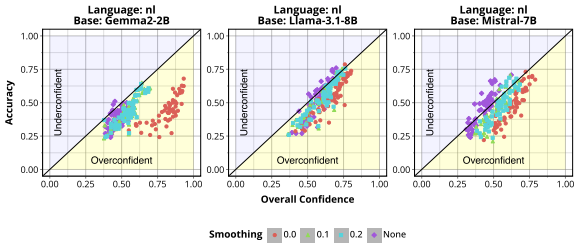

Figure 77: Reliability diagrams for the **GlobalMMLU** dataset for the nl language after instruction-tuning on the **OpenHermes** dataset.

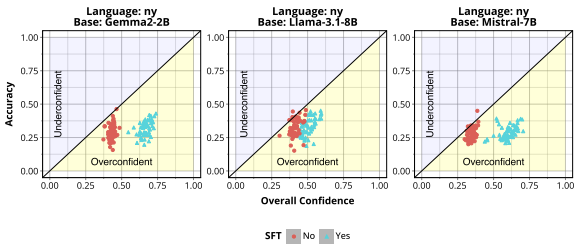

Figure 78: Reliability diagrams for the **GlobalMMLU** dataset for the ny language.

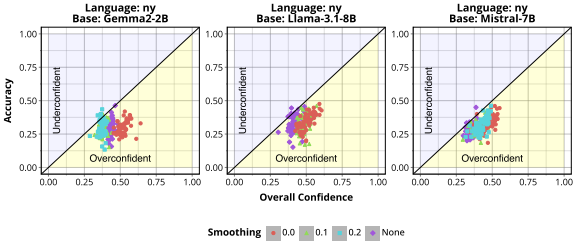

Figure 79: Reliability diagrams for the **GlobalMMLU** dataset for the ny language after instruction-tuning on the **Tulu3Mixture** dataset.

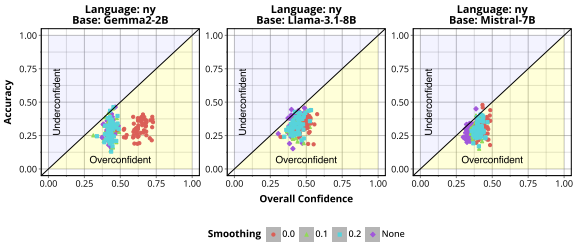

Figure 80: Reliability diagrams for the **GlobalMMLU** dataset for the ny language after instruction-tuning on the **OpenHermes** dataset.

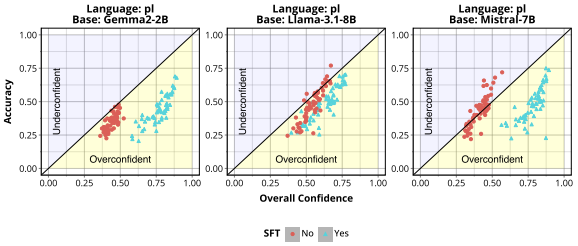

Figure 81: Reliability diagrams for the **GlobalMMLU** dataset for the pl language.

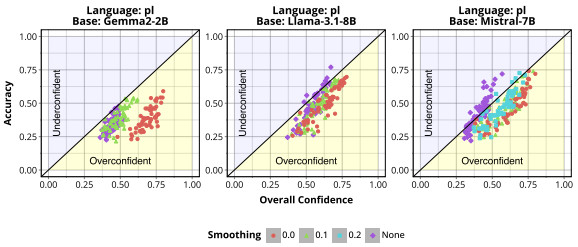

Figure 82: Reliability diagrams for the **GlobalMMLU** dataset for the pl language after instruction-tuning on the **Tulu3Mixture** dataset.

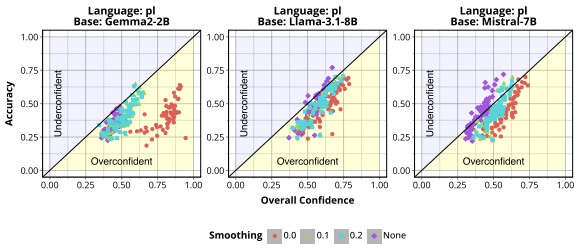

Figure 83: Reliability diagrams for the **GlobalMMLU** dataset for the pl language after instruction-tuning on the **OpenHermes** dataset.

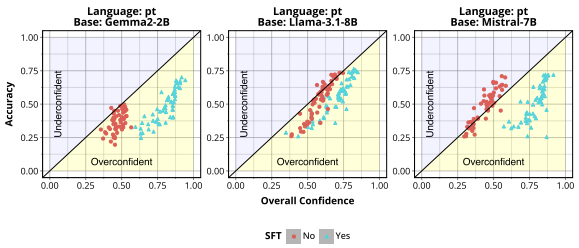

Figure 84: Reliability diagrams for the **GlobalMMLU** dataset for the pt language.

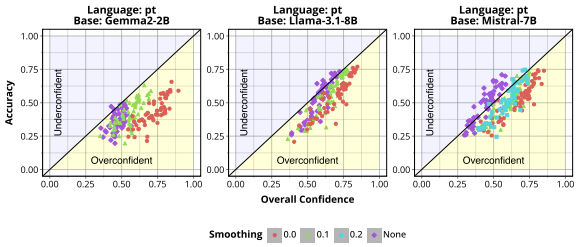

Figure 85: Reliability diagrams for the **GlobalMMLU** dataset for the pt language after instruction-tuning on the **Tulu3Mixture** dataset.

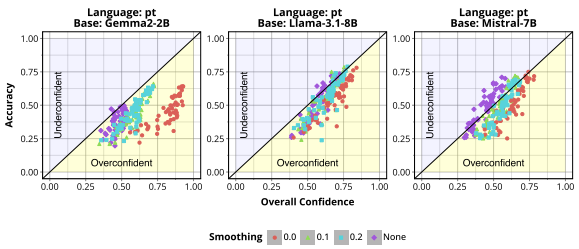

Figure 86: Reliability diagrams for the **GlobalMMLU** dataset for the pt language after instruction-tuning on the **OpenHermes** dataset.

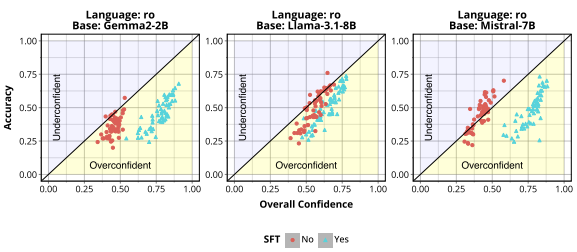

Figure 87: Reliability diagrams for the **GlobalMMLU** dataset for the ro language.

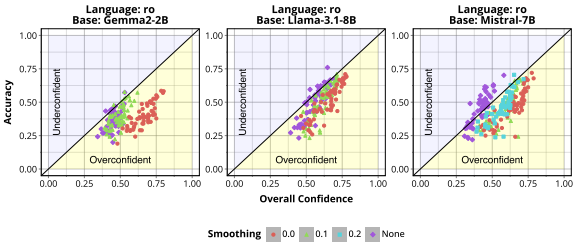

Figure 88: Reliability diagrams for the **GlobalMMLU** dataset for the ro language after instruction-tuning on the **Tulu3Mixture** dataset.

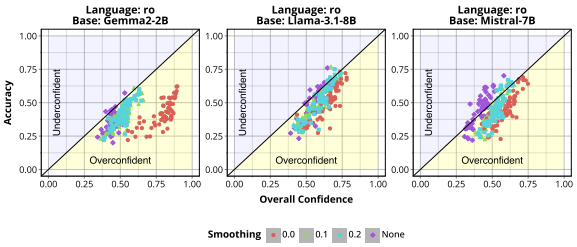

Figure 89: Reliability diagrams for the **GlobalMMLU** dataset for the ro language after instruction-tuning on the **OpenHermes** dataset.

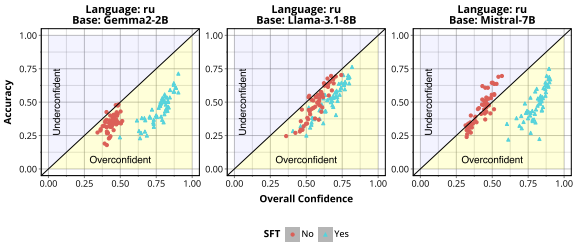

Figure 90: Reliability diagrams for the **GlobalMMLU** dataset for the ru language.

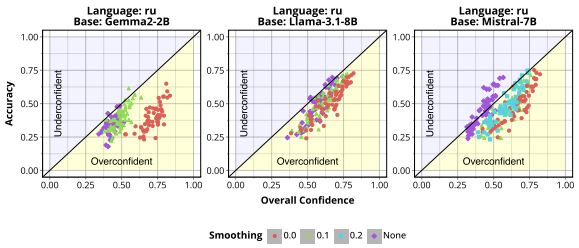

Figure 91: Reliability diagrams for the **GlobalMMLU** dataset for the ru language after instruction-tuning on the **Tulu3Mixture** dataset.

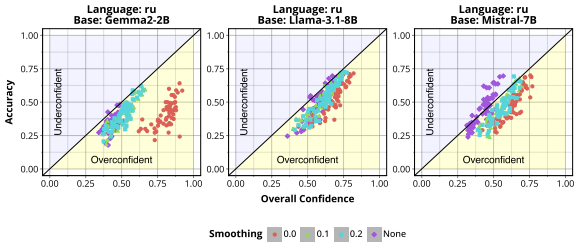

Figure 92: Reliability diagrams for the **GlobalMMLU** dataset for the ru language after instruction-tuning on the **OpenHermes** dataset.

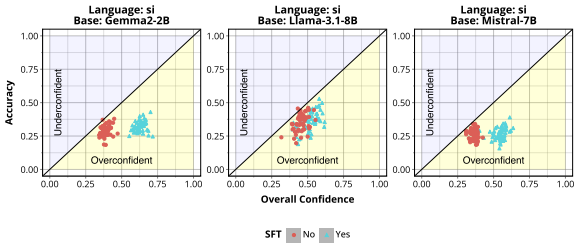

Figure 93: Reliability diagrams for the **GlobalMMLU** dataset for the si language.

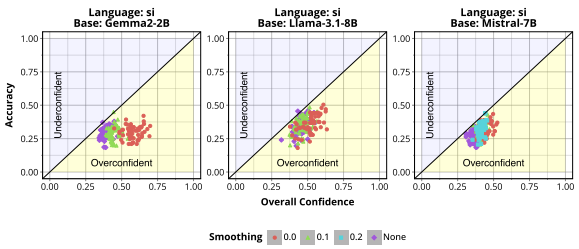

Figure 94: Reliability diagrams for the **GlobalMMLU** dataset for the si language after instruction-tuning on the **Tulu3Mixture** dataset.

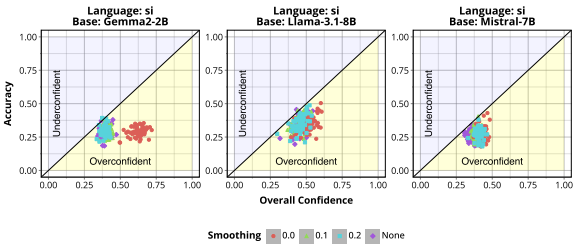

Figure 95: Reliability diagrams for the **GlobalMMLU** dataset for the si language after instruction-tuning on the **OpenHermes** dataset.

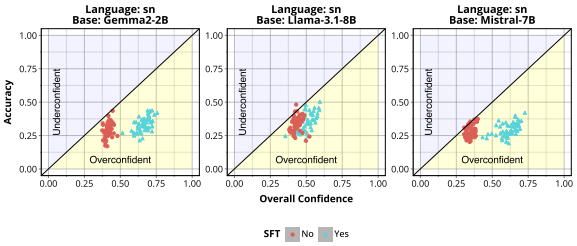

Figure 96: Reliability diagrams for the **GlobalMMLU** dataset for the sn language.

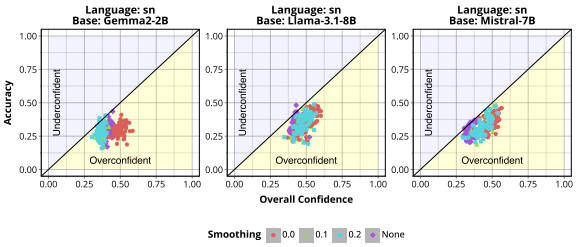

Figure 97: Reliability diagrams for the **GlobalMMLU** dataset for the sn language after instruction-tuning on the **Tulu3Mixture** dataset.

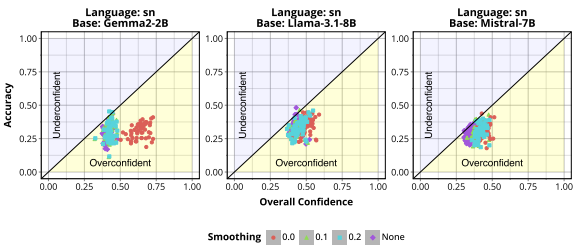

Figure 98: Reliability diagrams for the **GlobalMMLU** dataset for the sn language after instruction-tuning on the **OpenHermes** dataset.

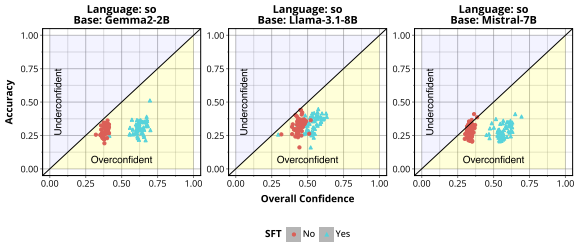

Figure 99: Reliability diagrams for the **GlobalMMLU** dataset for the so language.

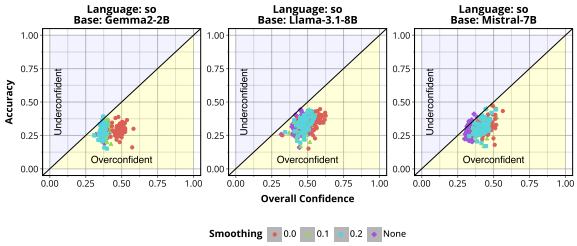

Figure 100: Reliability diagrams for the **GlobalMMLU** dataset for the so language after instruction-tuning on the **Tulu3Mixture** dataset.

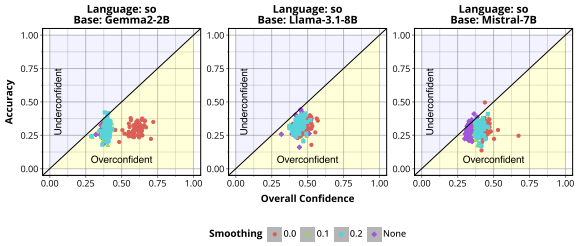

Figure 101: Reliability diagrams for the **GlobalMMLU** dataset for the so language after instruction-tuning on the **OpenHermes** dataset.

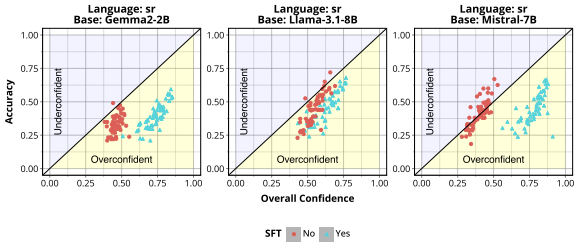

Figure 102: Reliability diagrams for the **GlobalMMLU** dataset for the sr language.

D.2 MMLU-ProX

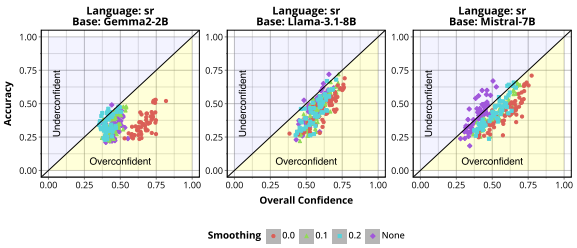

Figure 103: Reliability diagrams for the **GlobalMMLU** dataset for the sr language after instruction-tuning on the **Tulu3Mixture** dataset.

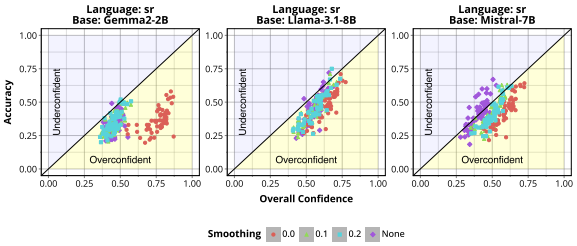

Figure 104: Reliability diagrams for the **GlobalMMLU** dataset for the sr language after instruction-tuning on the **OpenHermes** dataset.

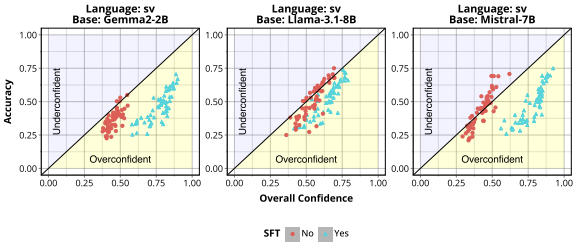

Figure 105: Reliability diagrams for the **GlobalMMLU** dataset for the sv language.

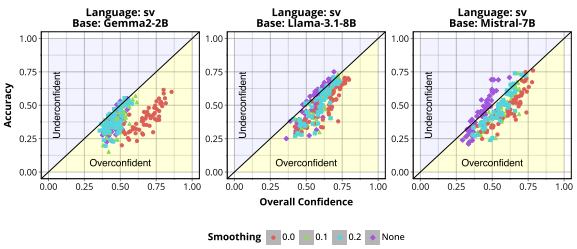

Figure 106: Reliability diagrams for the **GlobalMMLU** dataset for the sv language after instruction-tuning on the **Tulu3Mixture** dataset.

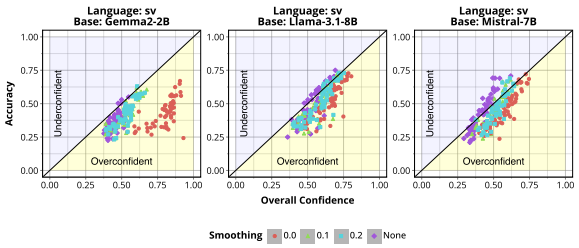

Figure 107: Reliability diagrams for the **GlobalMMLU** dataset for the sv language after instruction-tuning on the **OpenHermes** dataset.

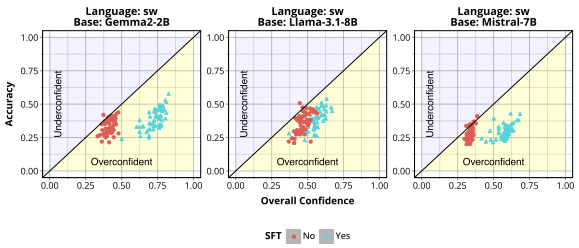

Figure 108: Reliability diagrams for the **GlobalMMLU** dataset for the sw language.

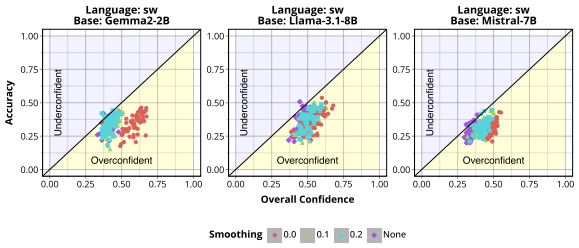

Figure 109: Reliability diagrams for the **GlobalMMLU** dataset for the sw language after instruction-tuning on the **Tulu3Mixture** dataset.

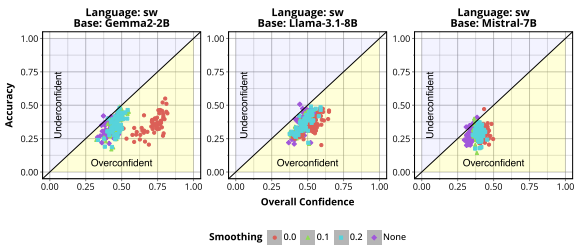

Figure 110: Reliability diagrams for the **GlobalMMLU** dataset for the sw language after instruction-tuning on the **OpenHermes** dataset.

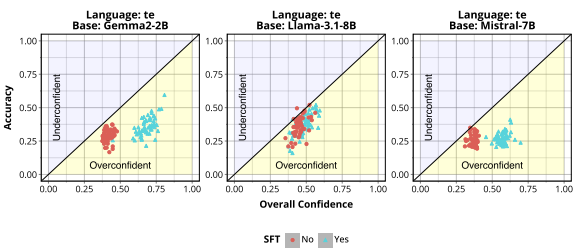

Figure 111: Reliability diagrams for the **GlobalMMLU** dataset for the te language.

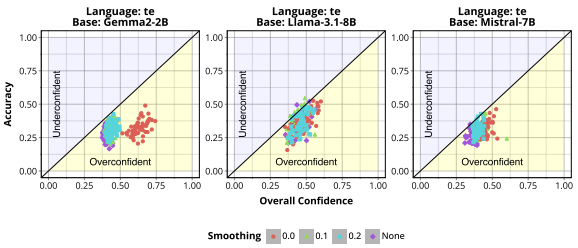

Figure 112: Reliability diagrams for the **GlobalMMLU** dataset for the te language after instruction-tuning on the **Tulu3Mixture** dataset.

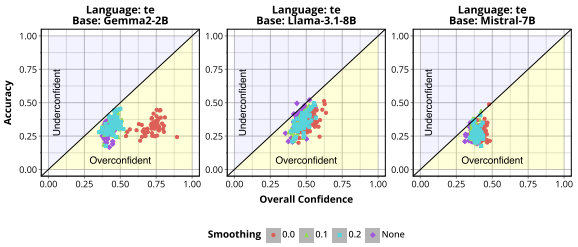

Figure 113: Reliability diagrams for the **GlobalMMLU** dataset for the te language after instruction-tuning on the **OpenHermes** dataset.

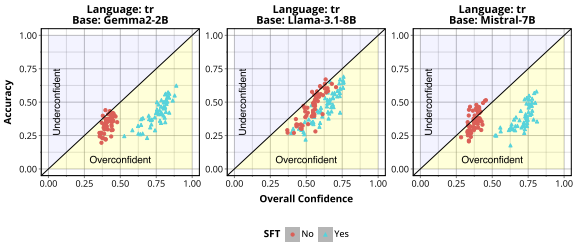

Figure 114: Reliability diagrams for the **GlobalMMLU** dataset for the tr language.

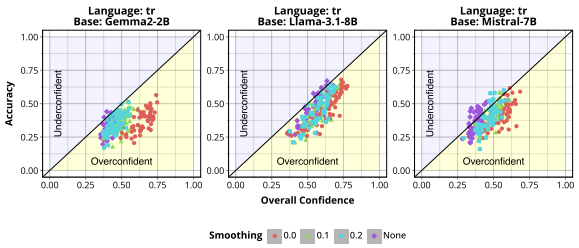

Figure 115: Reliability diagrams for the **GlobalMMLU** dataset for the `tr` language after instruction-tuning on the **Tulu3Mixture** dataset.

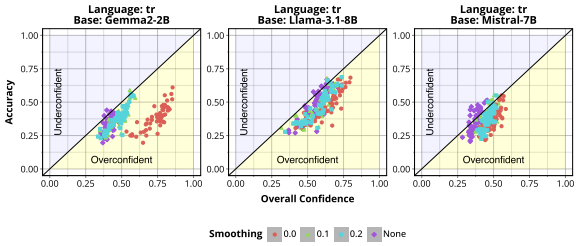

Figure 116: Reliability diagrams for the **GlobalMMLU** dataset for the `tr` language after instruction-tuning on the **OpenHermes** dataset.

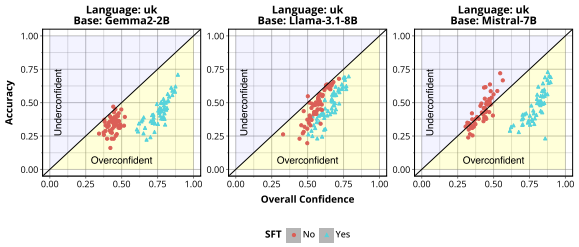

Figure 117: Reliability diagrams for the **GlobalMMLU** dataset for the uk language.

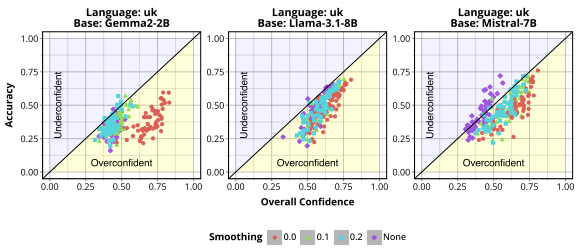

Figure 118: Reliability diagrams for the **GlobalMMLU** dataset for the uk language after instruction-tuning on the **Tulu3Mixture** dataset.

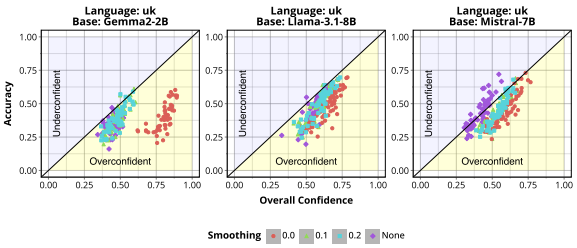

Figure 119: Reliability diagrams for the **GlobalMMLU** dataset for the uk language after instruction-tuning on the **OpenHermes** dataset.

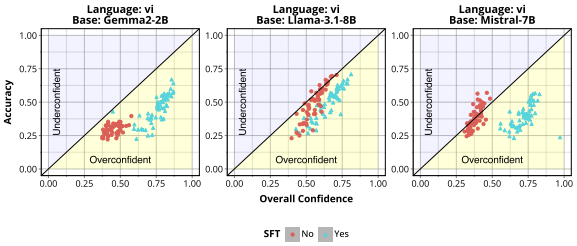

Figure 120: Reliability diagrams for the **GlobalMMLU** dataset for the `vi` language.

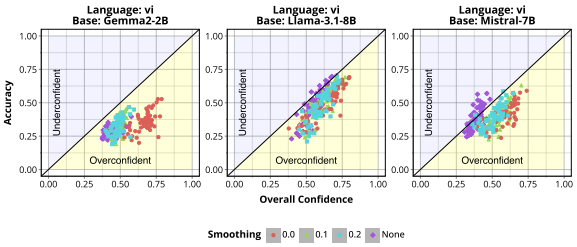

Figure 121: Reliability diagrams for the **GlobalMMLU** dataset for the `vi` language after instruction-tuning on the **Tulu3Mixture** dataset.

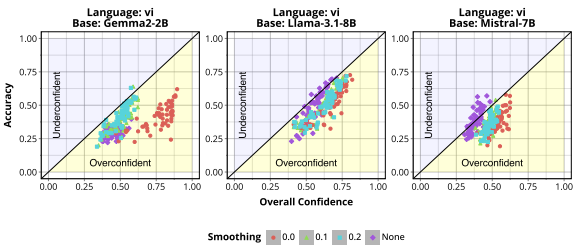

Figure 122: Reliability diagrams for the **GlobalMMLU** dataset for the `vi` language after instruction-tuning on the **OpenHermes** dataset.

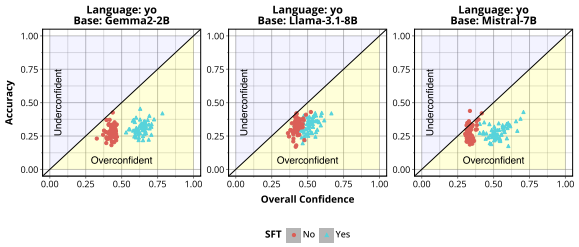

Figure 123: Reliability diagrams for the **GlobalMMLU** dataset for the yo language.

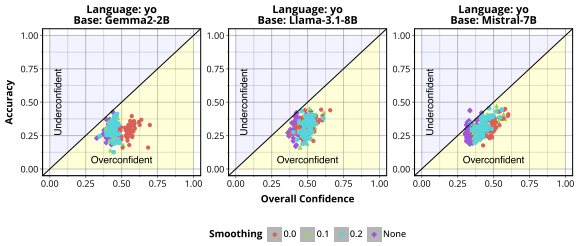

Figure 124: Reliability diagrams for the **GlobalMMLU** dataset for the yo language after instruction-tuning on the **Tulu3Mixture** dataset.

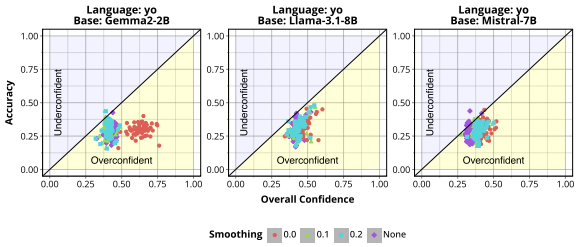

Figure 125: Reliability diagrams for the **GlobalMMLU** dataset for the yo language after instruction-tuning on the **OpenHermes** dataset.

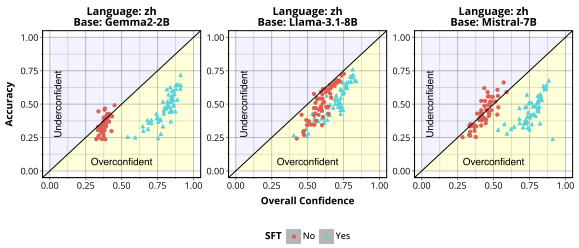

Figure 126: Reliability diagrams for the **GlobalMMLU** dataset for the zh language.

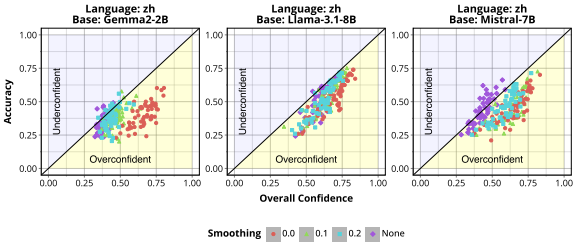

Figure 127: Reliability diagrams for the **GlobalMMLU** dataset for the zh language after instruction-tuning on the **Tulu3Mixture** dataset.

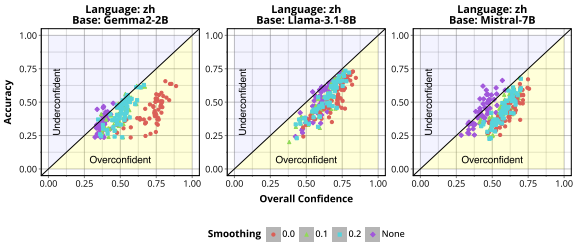

Figure 128: Reliability diagrams for the **GlobalMMLU** dataset for the zh language after instruction-tuning on the **OpenHermes** dataset.

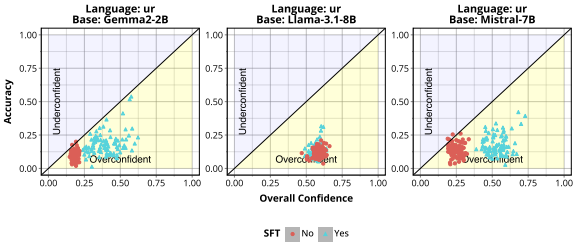

Figure 129: Reliability diagrams for the **MMLU-ProX** dataset for the ur language.

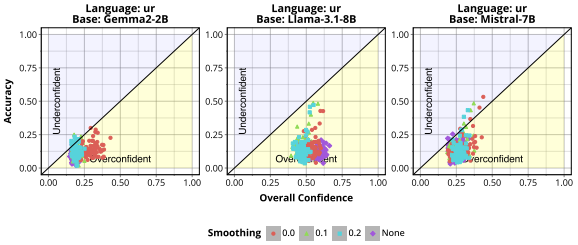

Figure 130: Reliability diagrams for the **MMLU-ProX** dataset for the ur language after instruction-tuning on the **Tulu3Mixture** dataset.

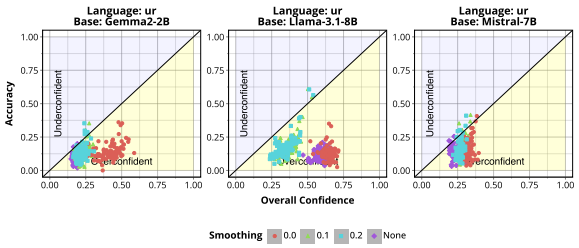

Figure 131: Reliability diagrams for the **MMLU-ProX** dataset for the ur language after instruction-tuning on the **OpenHermes** dataset.

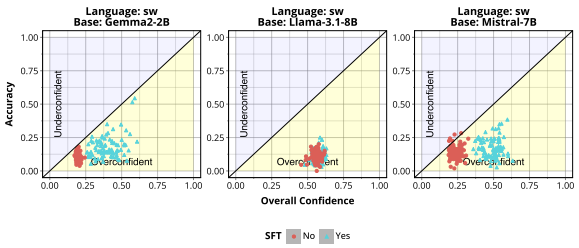

Figure 132: Reliability diagrams for the **MMLU-ProX** dataset for the sw language.

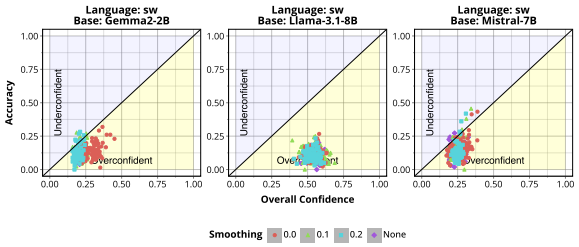

Figure 133: Reliability diagrams for the **MMLU-ProX** dataset for the sw language after instruction-tuning on the **Tulu3Mixture** dataset.

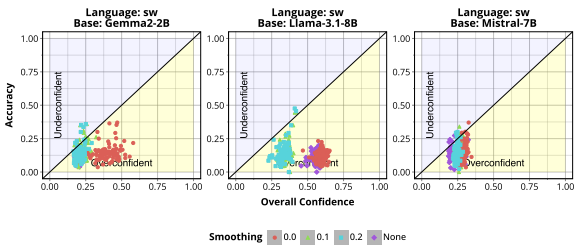

Figure 134: Reliability diagrams for the **MMLU-ProX** dataset for the sw language after instruction-tuning on the **OpenHermes** dataset.

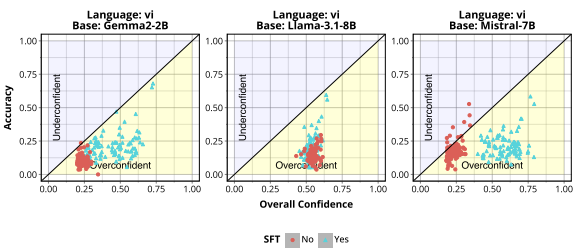

Figure 135: Reliability diagrams for the **MMLU-ProX** dataset for the vi language.

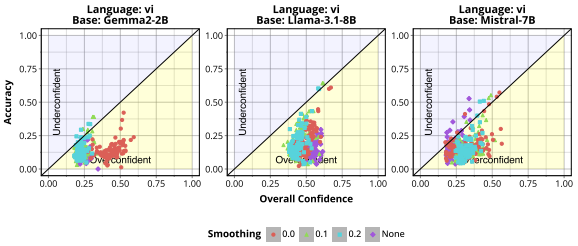

Figure 136: Reliability diagrams for the **MMLU-ProX** dataset for the vi language after instruction-tuning on the **Tulu3Mixture** dataset.

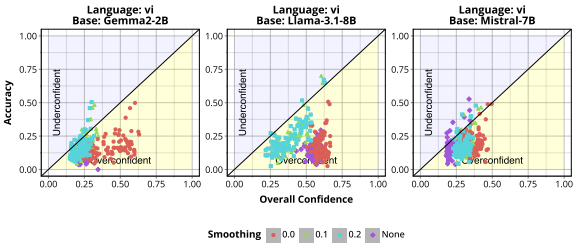

Figure 137: Reliability diagrams for the **MMLU-ProX** dataset for the vi language after instruction-tuning on the **OpenHermes** dataset.

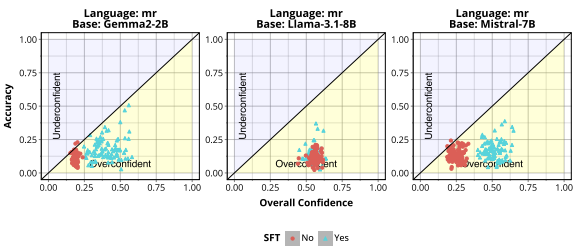

Figure 138: Reliability diagrams for the **MMLU-ProX** dataset for the mr language.

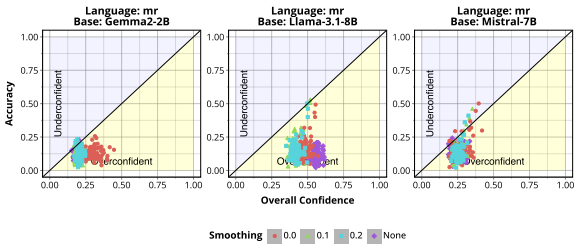

Figure 139: Reliability diagrams for the **MMLU-ProX** dataset for the mr language after instruction-tuning on the **Tulu3Mixture** dataset.

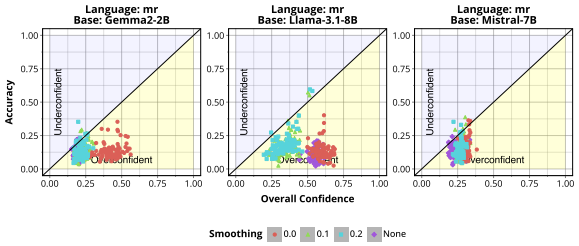

Figure 140: Reliability diagrams for the **MMLU-ProX** dataset for the mr language after instruction-tuning on the **OpenHermes** dataset.

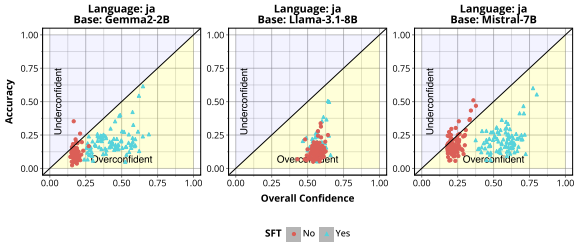

Figure 141: Reliability diagrams for the **MMLU-ProX** dataset for the ja language.

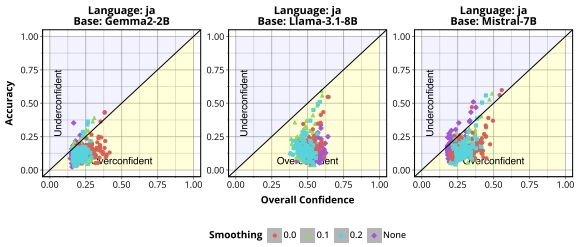

Figure 142: Reliability diagrams for the **MMLU-ProX** dataset for the ja language after instruction-tuning on the **Tulu3Mixture** dataset.

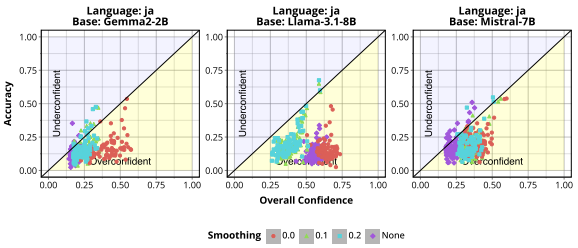

Figure 143: Reliability diagrams for the **MMLU-ProX** dataset for the ja language after instruction-tuning on the **OpenHermes** dataset.

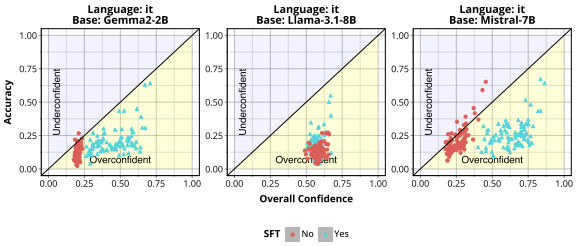

Figure 144: Reliability diagrams for the **MMLU-ProX** dataset for the it language.

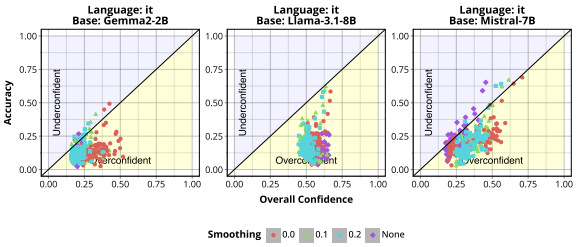

Figure 145: Reliability diagrams for the **MMLU-ProX** dataset for the it language after instruction-tuning on the **Tulu3Mixture** dataset.

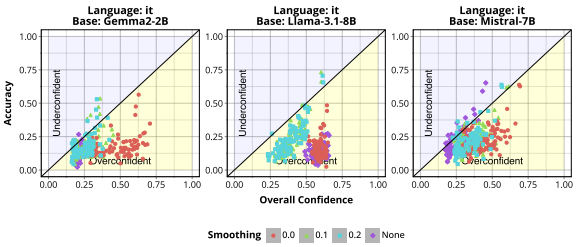

Figure 146: Reliability diagrams for the **MMLU-ProX** dataset for the it language after instruction-tuning on the **OpenHermes** dataset.

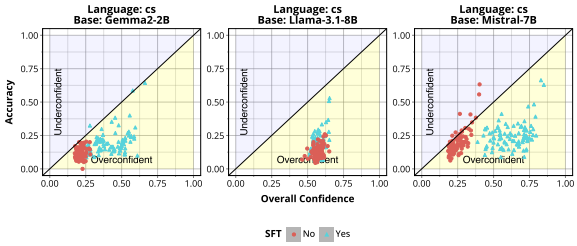

Figure 147: Reliability diagrams for the **MMLU-ProX** dataset for the cs language.

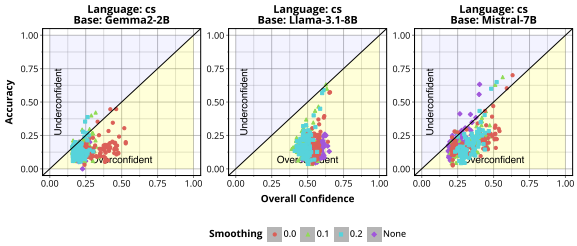

Figure 148: Reliability diagrams for the **MMLU-ProX** dataset for the cs language after instruction-tuning on the **Tulu3Mixture** dataset.

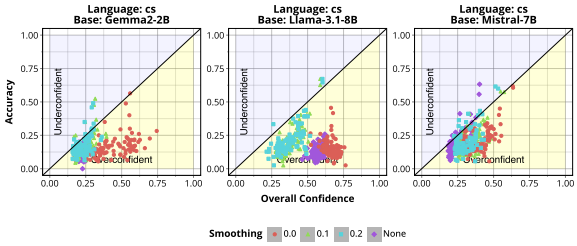

Figure 149: Reliability diagrams for the **MMLU-ProX** dataset for the cs language after instruction-tuning on the **OpenHermes** dataset.

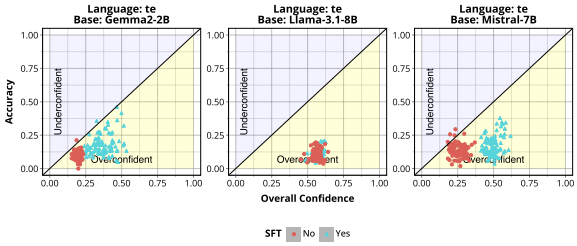

Figure 150: Reliability diagrams for the **MMLU-ProX** dataset for the te language.

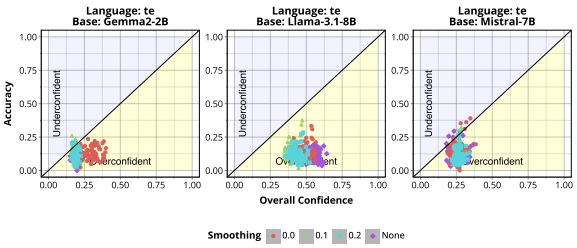

Figure 151: Reliability diagrams for the **MMLU-ProX** dataset for the te language after instruction-tuning on the **Tulu3Mixture** dataset.

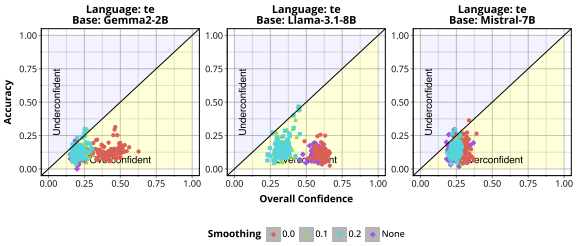

Figure 152: Reliability diagrams for the **MMLU-ProX** dataset for the te language after instruction-tuning on the **OpenHermes** dataset.

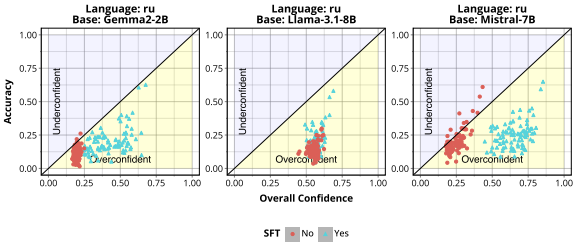

Figure 153: Reliability diagrams for the **MMLU-ProX** dataset for the ru language.

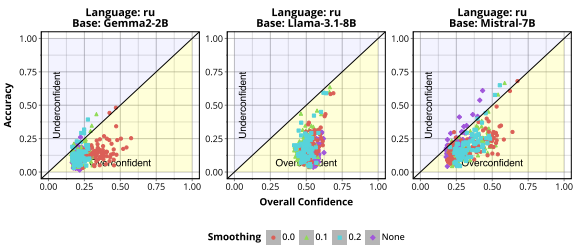

Figure 154: Reliability diagrams for the **MMLU-ProX** dataset for the ru language after instruction-tuning on the **Tulu3Mixture** dataset.

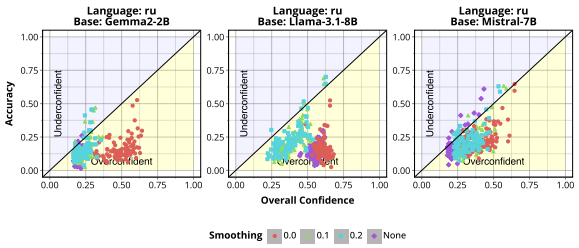

Figure 155: Reliability diagrams for the **MMLU-ProX** dataset for the ru language after instruction-tuning on the **OpenHermes** dataset.

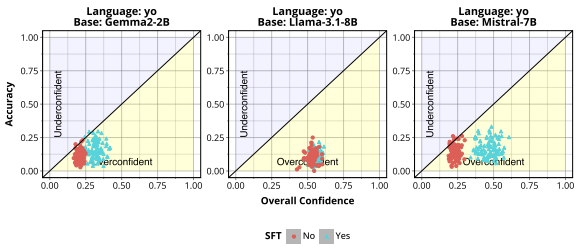

Figure 156: Reliability diagrams for the **MMLU-ProX** dataset for the yo language.

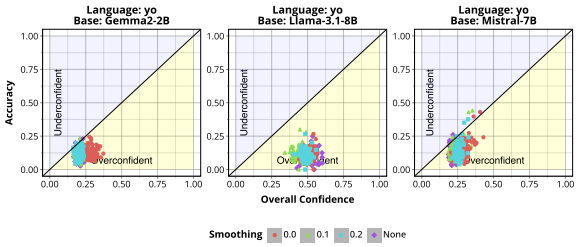

Figure 157: Reliability diagrams for the **MMLU-ProX** dataset for the yo language after instruction-tuning on the **Tulu3Mixture** dataset.

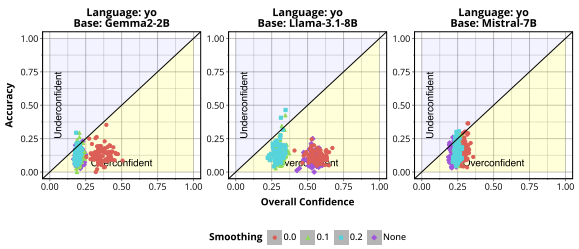

Figure 158: Reliability diagrams for the **MMLU-ProX** dataset for the yo language after instruction-tuning on the **OpenHermes** dataset.

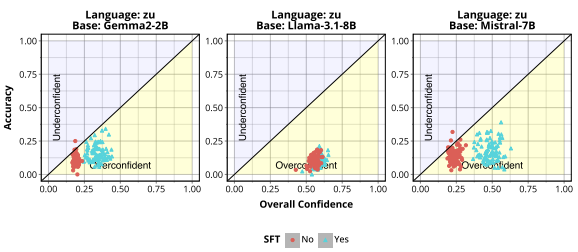

Figure 159: Reliability diagrams for the **MMLU-ProX** dataset for the zu language.

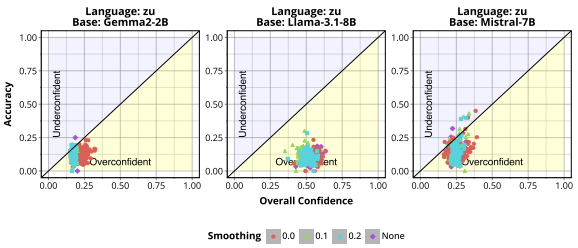

Figure 160: Reliability diagrams for the **MMLU-ProX** dataset for the zu language after instruction-tuning on the **Tulu3Mixture** dataset.

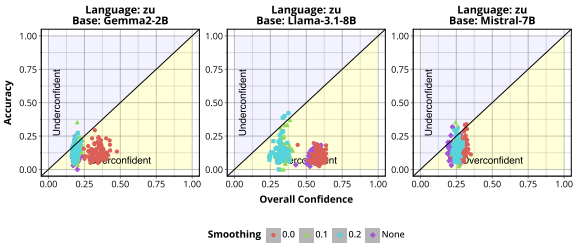

Figure 161: Reliability diagrams for the **MMLU-ProX** dataset for the zu language after instruction-tuning on the **OpenHermes** dataset.

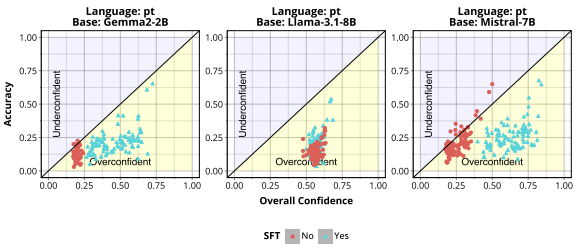

Figure 162: Reliability diagrams for the **MMLU-ProX** dataset for the pt language.

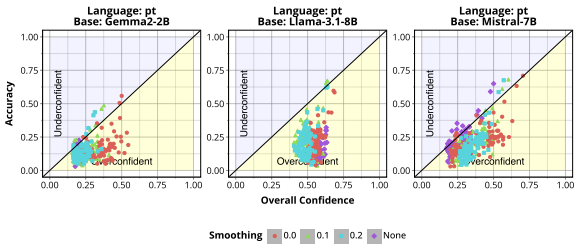

Figure 163: Reliability diagrams for the **MMLU-ProX** dataset for the pt language after instruction-tuning on the **Tulu3Mixture** dataset.

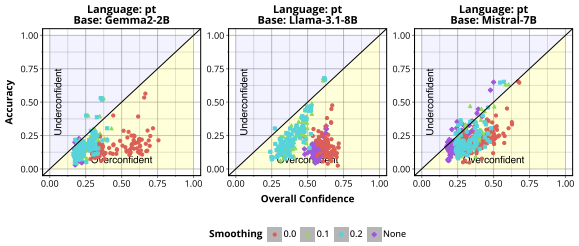

Figure 164: Reliability diagrams for the **MMLU-ProX** dataset for the pt language after instruction-tuning on the **OpenHermes** dataset.

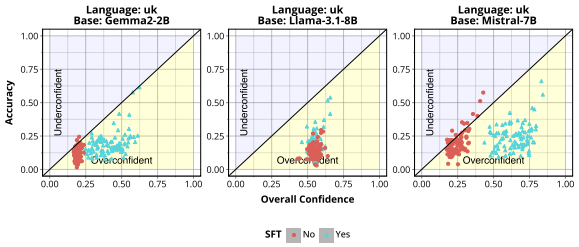

Figure 165: Reliability diagrams for the **MMLU-ProX** dataset for the uk language.

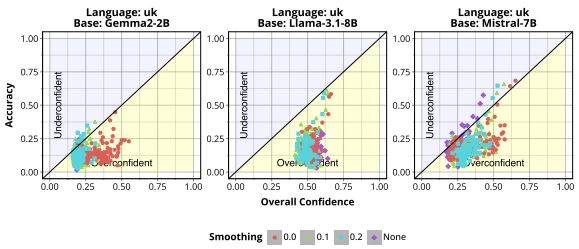

Figure 166: Reliability diagrams for the **MMLU-ProX** dataset for the uk language after instruction-tuning on the **Tulu3Mixture** dataset.

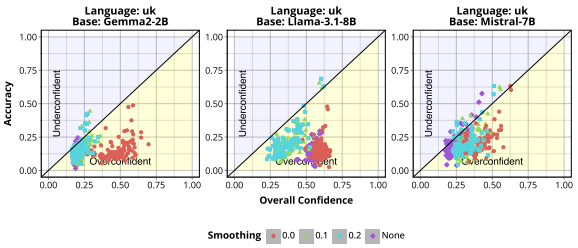

Figure 167: Reliability diagrams for the **MMLU-ProX** dataset for the uk language after instruction-tuning on the **OpenHermes** dataset.

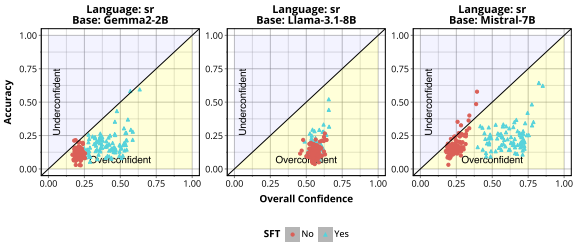

Figure 168: Reliability diagrams for the **MMLU-ProX** dataset for the sr language.

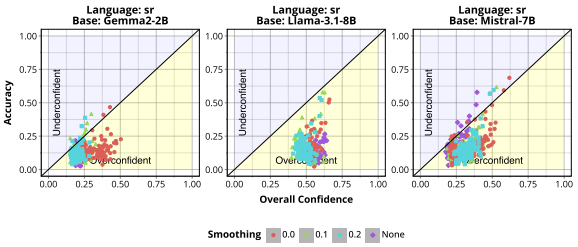

Figure 169: Reliability diagrams for the **MMLU-ProX** dataset for the sr language after instruction-tuning on the **Tulu3Mixture** dataset.

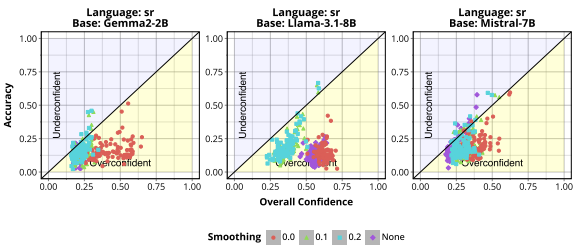

Figure 170: Reliability diagrams for the **MMLU-ProX** dataset for the sr language after instruction-tuning on the **OpenHermes** dataset.

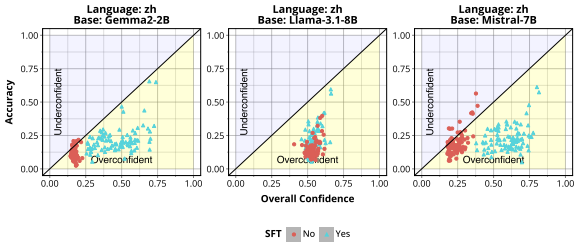

Figure 171: Reliability diagrams for the **MMLU-ProX** dataset for the zh language.

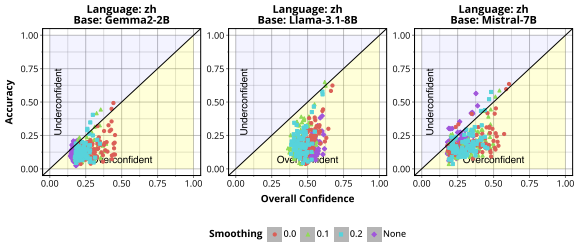

Figure 172: Reliability diagrams for the **MMLU-ProX** dataset for the zh language after instruction-tuning on the **Tulu3Mixture** dataset.

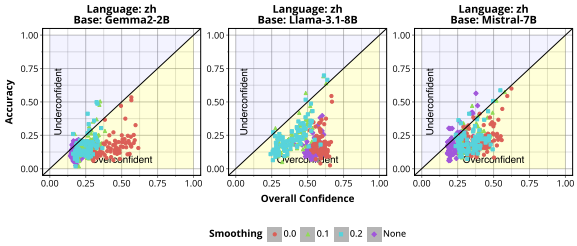

Figure 173: Reliability diagrams for the **MMLU-ProX** dataset for the zh language after instruction-tuning on the **OpenHermes** dataset.

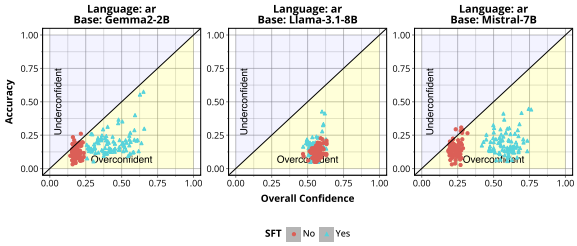

Figure 174: Reliability diagrams for the **MMLU-ProX** dataset for the ar language.

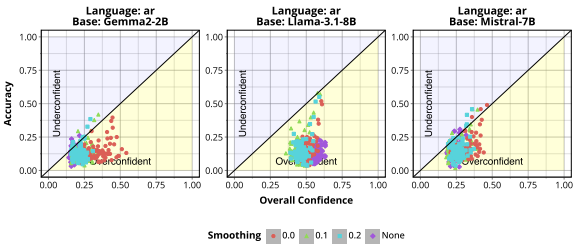

Figure 175: Reliability diagrams for the **MMLU-ProX** dataset for the ar language after instruction-tuning on the **Tulu3Mixture** dataset.

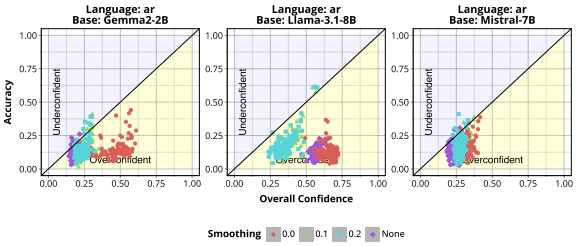

Figure 176: Reliability diagrams for the **MMLU-ProX** dataset for the ar language after instruction-tuning on the **OpenHermes** dataset.

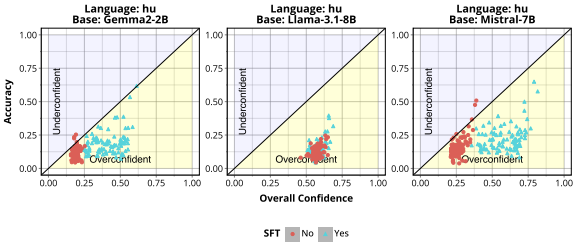

Figure 177: Reliability diagrams for the **MMLU-ProX** dataset for the hu language.

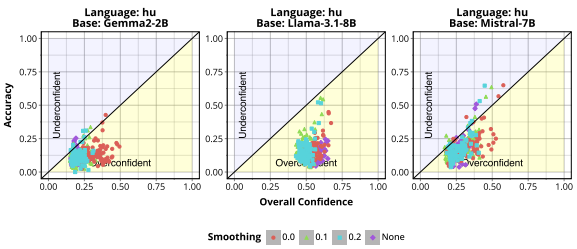

Figure 178: Reliability diagrams for the **MMLU-ProX** dataset for the hu language after instruction-tuning on the **Tulu3Mixture** dataset.

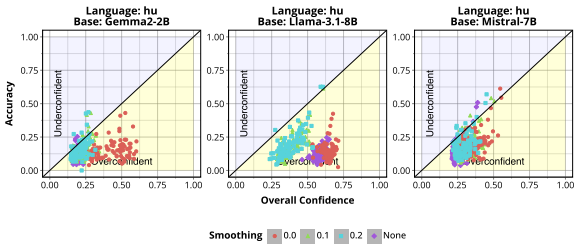

Figure 179: Reliability diagrams for the **MMLU-ProX** dataset for the hu language after instruction-tuning on the **OpenHermes** dataset.

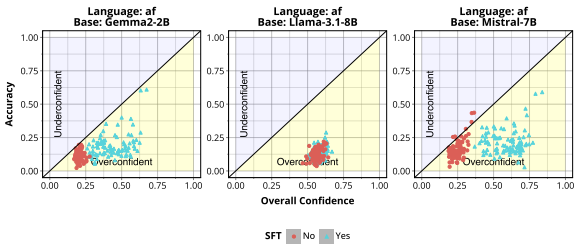

Figure 180: Reliability diagrams for the **MMLU-ProX** dataset for the af language.

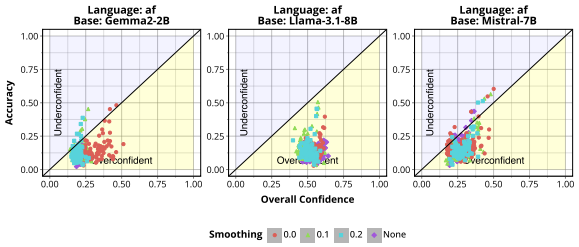

Figure 181: Reliability diagrams for the **MMLU-ProX** dataset for the af language after instruction-tuning on the **Tulu3Mixture** dataset.

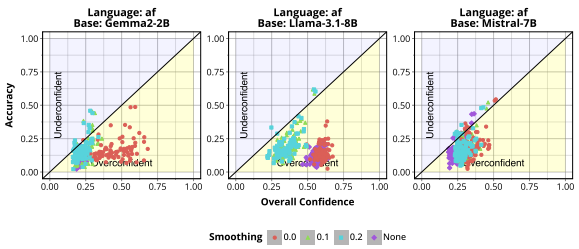

Figure 182: Reliability diagrams for the **MMLU-ProX** dataset for the af language after instruction-tuning on the **OpenHermes** dataset.

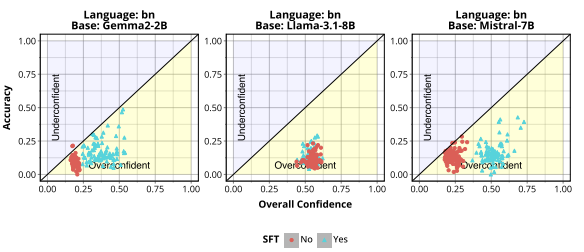

Figure 183: Reliability diagrams for the **MMLU-ProX** dataset for the bn language.

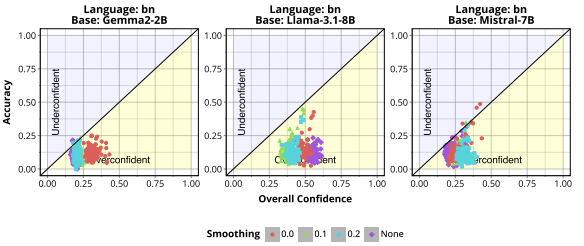

Figure 184: Reliability diagrams for the **MMLU-ProX** dataset for the bn language after instruction-tuning on the **Tulu3Mixture** dataset.

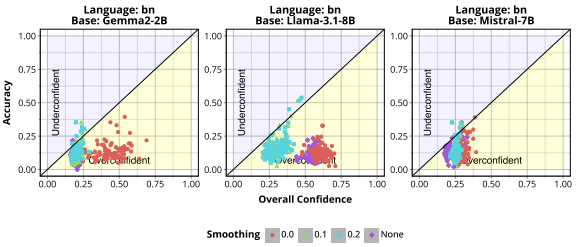

Figure 185: Reliability diagrams for the **MMLU-ProX** dataset for the bn language after instruction-tuning on the **OpenHermes** dataset.

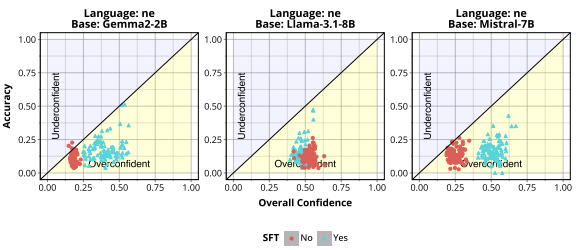

Figure 186: Reliability diagrams for the **MMLU-ProX** dataset for the ne language.

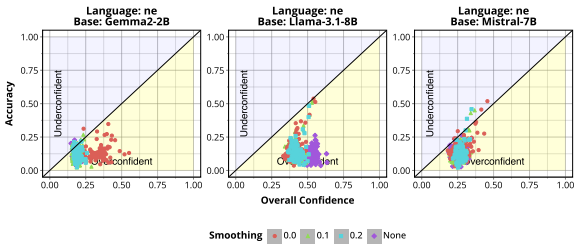

Figure 187: Reliability diagrams for the **MMLU-ProX** dataset for the ne language after instruction-tuning on the **Tulu3Mixture** dataset.

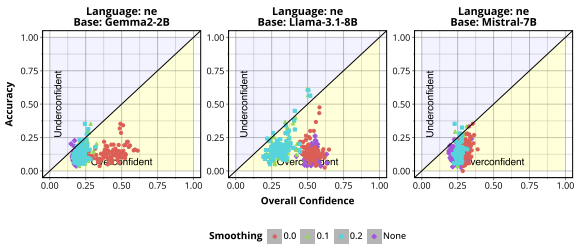

Figure 188: Reliability diagrams for the **MMLU-ProX** dataset for the ne language after instruction-tuning on the **OpenHermes** dataset.

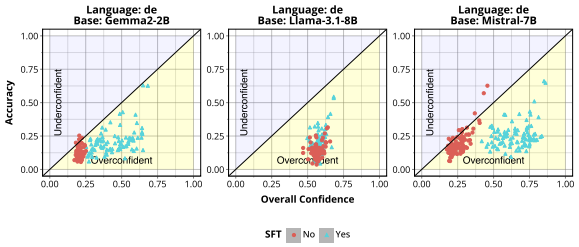

Figure 189: Reliability diagrams for the **MMLU-ProX** dataset for the de language.

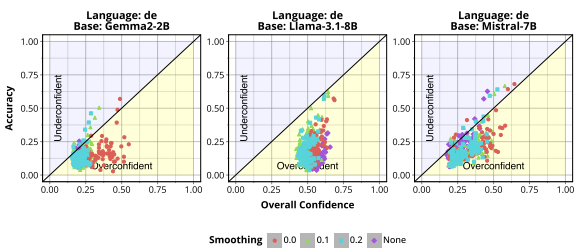

Figure 190: Reliability diagrams for the **MMLU-ProX** dataset for the de language after instruction-tuning on the **Tulu3Mixture** dataset.

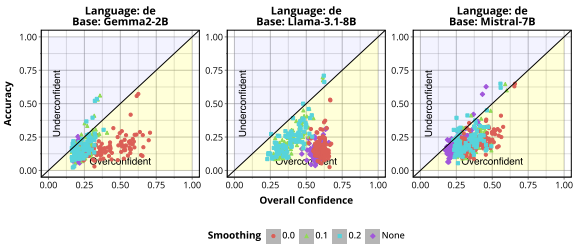

Figure 191: Reliability diagrams for the **MMLU-ProX** dataset for the de language after instruction-tuning on the **OpenHermes** dataset.

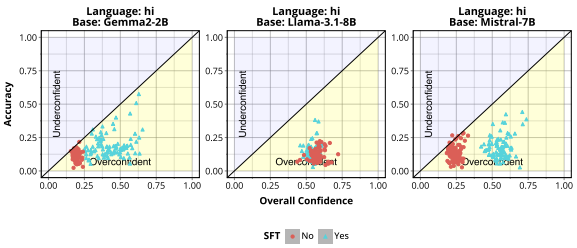

Figure 192: Reliability diagrams for the **MMLU-ProX** dataset for the hi language.

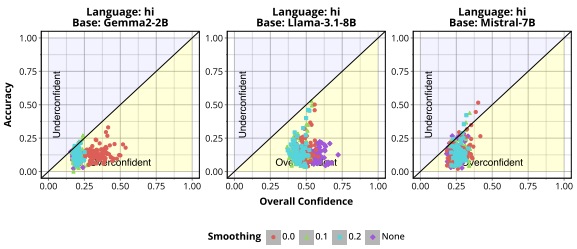

Figure 193: Reliability diagrams for the **MMLU-ProX** dataset for the hi language after instruction-tuning on the **Tulu3Mixture** dataset.

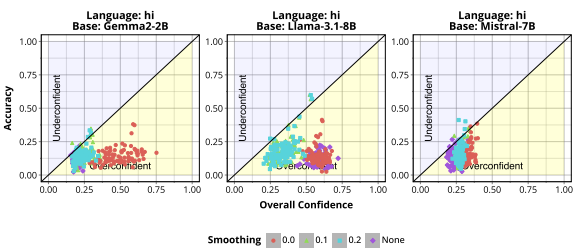

Figure 194: Reliability diagrams for the **MMLU-ProX** dataset for the hi language after instruction-tuning on the **OpenHermes** dataset.

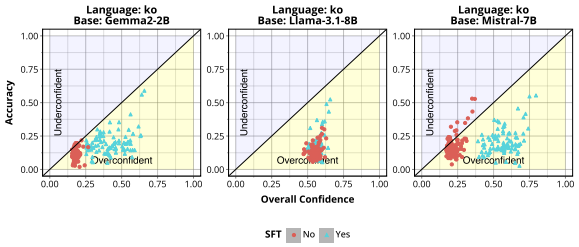

Figure 195: Reliability diagrams for the **MMLU-ProX** dataset for the ko language.

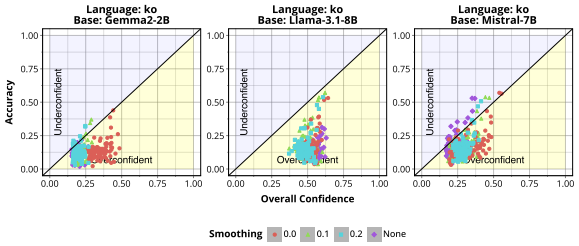

Figure 196: Reliability diagrams for the **MMLU-ProX** dataset for the ko language after instruction-tuning on the **Tulu3Mixture** dataset.

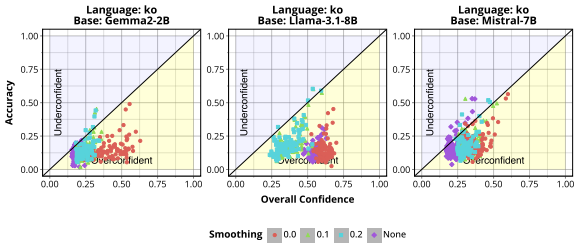

Figure 197: Reliability diagrams for the **MMLU-ProX** dataset for the ko language after instruction-tuning on the **OpenHermes** dataset.

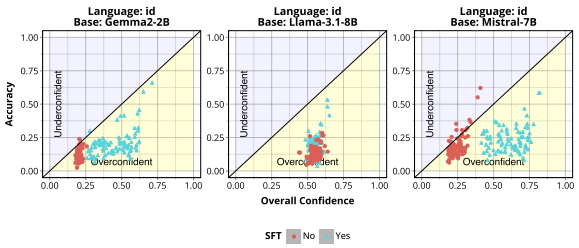

Figure 198: Reliability diagrams for the **MMLU-ProX** dataset for the `id` language.

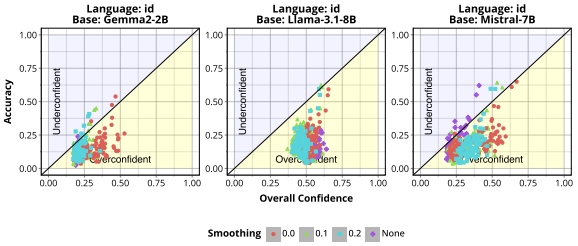

Figure 199: Reliability diagrams for the **MMLU-ProX** dataset for the `id` language after instruction-tuning on the **Tulu3Mixture** dataset.

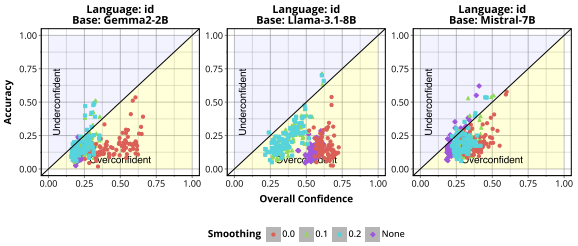

Figure 200: Reliability diagrams for the **MMLU-ProX** dataset for the `id` language after instruction-tuning on the **OpenHermes** dataset.

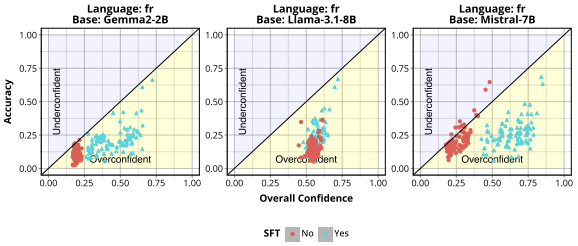

Figure 201: Reliability diagrams for the **MMLU-ProX** dataset for the `fr` language.

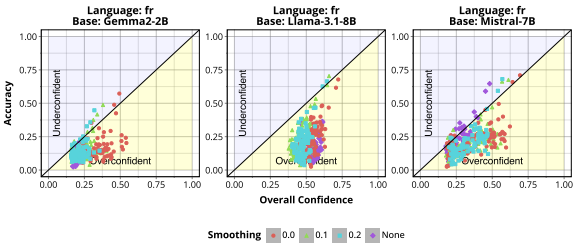

Figure 202: Reliability diagrams for the **MMLU-ProX** dataset for the `fr` language after instruction-tuning on the **Tulu3Mixture** dataset.

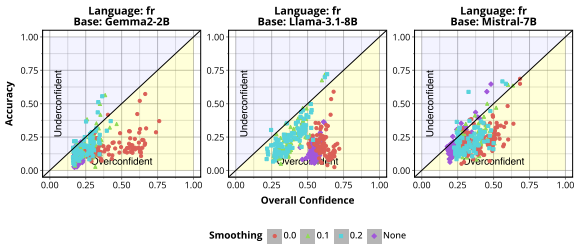

Figure 203: Reliability diagrams for the **MMLU-ProX** dataset for the `fr` language after instruction-tuning on the **OpenHermes** dataset.

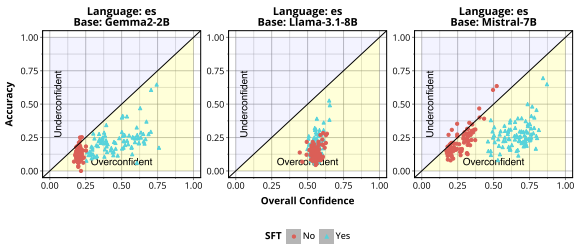

Figure 204: Reliability diagrams for the **MMLU-ProX** dataset for the `es` language.

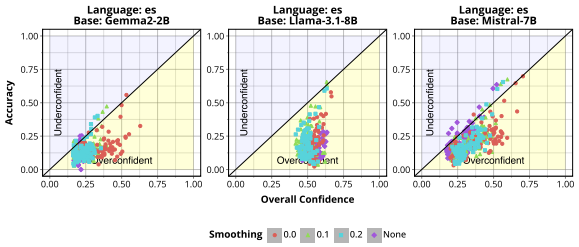

Figure 205: Reliability diagrams for the **MMLU-ProX** dataset for the `es` language after instruction-tuning on the **Tulu3Mixture** dataset.

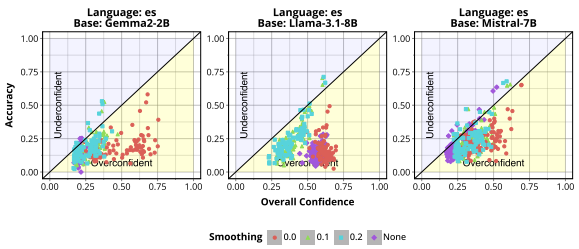

Figure 206: Reliability diagrams for the **MMLU-ProX** dataset for the `es` language after instruction-tuning on the **OpenHermes** dataset.

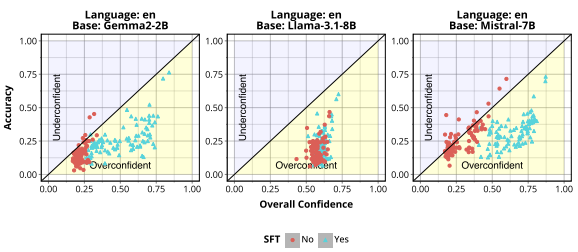

Figure 207: Reliability diagrams for the **MMLU-ProX** dataset for the `en` language.

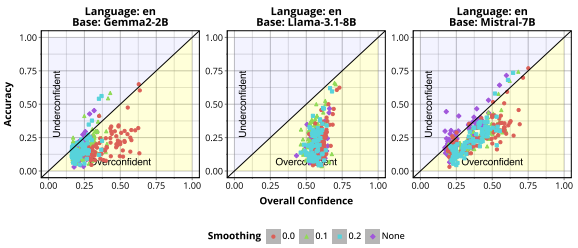

Figure 208: Reliability diagrams for the **MMLU-ProX** dataset for the `en` language after instruction-tuning on the **Tulu3Mixture** dataset.

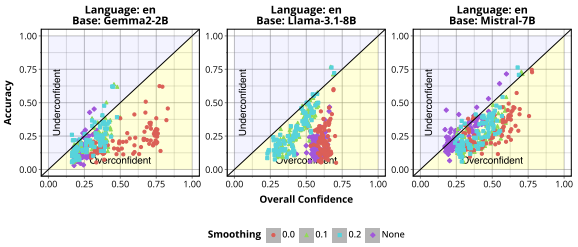

Figure 209: Reliability diagrams for the **MMLU-ProX** dataset for the `en` language after instruction-tuning on the **OpenHermes** dataset.

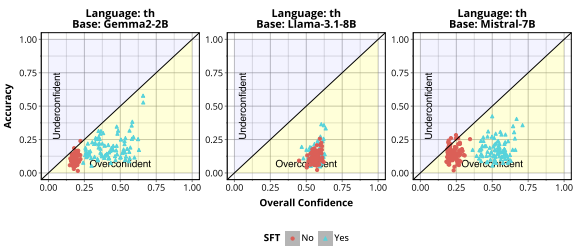

Figure 210: Reliability diagrams for the **MMLU-ProX** dataset for the `th` language.

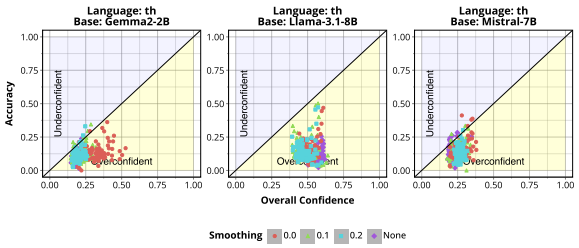

Figure 211: Reliability diagrams for the **MMLU-ProX** dataset for the th language after instruction-tuning on the **Tulu3Mixture** dataset.

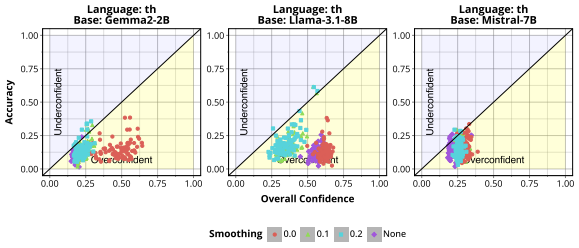

Figure 212: Reliability diagrams for the **MMLU-ProX** dataset for the th language after instruction-tuning on the **OpenHermes** dataset.

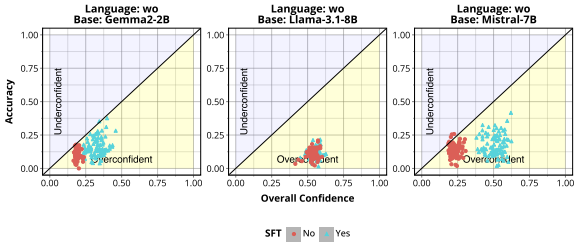

Figure 213: Reliability diagrams for the **MMLU-ProX** dataset for the wo language.

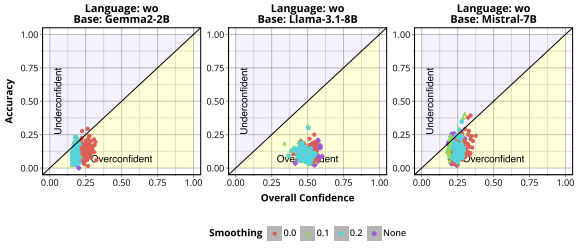

Figure 214: Reliability diagrams for the **MMLU-ProX** dataset for the wo language after instruction-tuning on the **Tulu3Mixture** dataset.

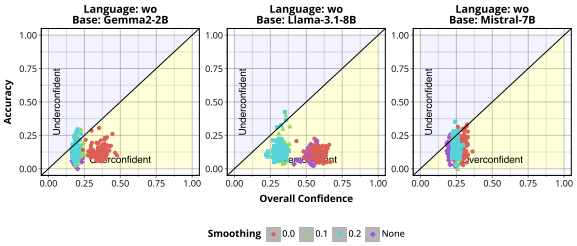

Figure 215: Reliability diagrams for the **MMLU-ProX** dataset for the wo language after instruction-tuning on the **OpenHermes** dataset.