探究语言模型指令微调的多语言校准效应

Jerry Huang[♣♠♡*] **PengLu**[♠] **Qiuhao Zeng**[♢▽] **Yusuke Iwasawa**[♡] **Yutaka Matsuo**[♡]
SarathChandar[♣✠■] **Edison Marrese-Taylor**[♡★†] **IreneLi**[♡††♣]Mila - Quebec 人工智能研究所 [♠]蒙特利尔大学 [♡]东京大学 [♢]西安大略大学[▽]Vector 研究所 [✠]蒙特利尔理工学院 [■]CIFAR 人工智能讲席 [★]AIST

[†]**通讯作者：** irene.li@weblab.t.u-tokyo.ac.jp

arXiv:2601.01362v1 [cs.CL] 4 Jan 2026

摘要

确保深度学习模型的预测不确定性得到良好校准，对于维持其可信度和可靠性至关重要。然而，尽管基础模型研究不断取得进展，大语言模型（LLM）与其校准之间的关系仍是一个有待研究的领域。在这项工作中，我们关注多语言环境下大语言模型校准方面的一个关键空白，旨在更好地理解数据稀缺可能如何导致不同的校准效应，以及常用技术如何应用于这些环境。我们对两个多语言基准测试进行了分析，二者分别涵盖 29 种和 42 种语言。结果表明，即使对于低资源语言，在高资源语言监督微调数据集上进行指令微调后，模型置信度也会显著提高。然而，准确率的提升微乎其微，甚至毫无提升，从而导致校准失准，这凸显了标准监督微调在多语言场景中的一个关键缺陷。此外，我们观察到，使用标签平滑是缓解这一问题的一种合理方法，而且无需低资源监督微调数据，就能在所有语言中保持更好的校准。总体而言，这凸显了在训练和微调大语言模型时考虑多语言因素的重要性，以提升其在下游应用中的可靠性和公平性。

1 引言

在构建能够遵循自然语言指令的模型方面，人们已取得巨大进展（Ouyang 等人，2022；Sanh 等人，2022；Chung 等人，2024年）。这些模型通过使用在大量数据上预训练的大语言模型，以及能帮助它们学会以类似人类的方式互动的高质量数据集来构建（Bach 等人，2022；Wang 等人，2022、2023a）。然而，这类模型已表现出预测过度自信的倾向（Zhao 等，2021；Jiang 等，2021；Xiong 等，2024），引发了人们对其用于高风险决策场景的担忧。对于神经网络而言，这类现象并不新鲜；研究一再表明，神经网络会出现过度自信的预测，并高估其预测正确的似然（Guo 等人，2017；Szegedy 等人，2016；Müller 等人，2019；Naeini 等人，2015；Minderer 等人，2021）。为改善这一问题，人们提出了温度缩放（Guo 等人，2017）和标签平滑（Müller 等人，2019）等方法，效果各异；这也推动了更多研究，以确保预测与置信度保持一致（Lin 等人，2017；Mukhoti 等人，2020；Pereyra 等人，2017；Liu 等人，2022）。

关于大语言模型校准的研究仍然有限（Zhao 等人，2021年），但研究表明，就校准而言，它们的表现与原型网络相似（Huang 等人，2025a）。然而，大语言模型在应用方面仍有其独特之处，尤其是它们能够应用于资源稀缺程度各异的不同语言。这引出了一个有趣的问题：

用于校准大语言模型的解决方案能否同时作用于多种语言？

在这项工作中，我们对这一问题给出了初步的实证回答。我们的工作和贡献总结如下：

a) 我们在一系列多语言分类任务上进行评估，考察大语言模型的下游校准在不同语言之间是否保持不变。

b) 有趣的是，所有语言都呈现出一个有趣的特征：经过指令微调的模型校准效果不如其基础模型。然而，另一个观察结果涉及未用于指令微调的语言：这些语言的置信度

[*]作为访问学生在东京大学完成的工作。
[†]共同指导。

有所提高，但准确率并未提升，导致过度自信加剧。

c) 仅在指令调优数据上使用标签平滑，与不使用平滑的模型相比，我们发现校准效果更好，且准确率仅略有下降。

总的来说，我们的工作展示了一种简单而有效的解决方案，可提升多语言语言模型的校准效果，并有望为改进其鲁棒性和可靠性、以适用于下游应用开辟道路。

2 相关工作

不确定性校准。 不确定性校准（Brier，1950年；Murphy，1972；DeGroot 和 Fienberg，1983）旨在使针对不同输入得出的预测概率与这些输入对应的预期准确率相匹配。用于衡量此类属性的常用指标包括期望校准误差（ECE）（Naeini 等人，2015）。其他提出的指标包括均方根（Hendrycks 等人，2019年）以及静态/自适应校准误差（SCE/ACE）（Nixon 等人，2019年），这些指标提供了互补的视角，从而能够更全面地评估不确定性校准情况。

然而，从统计学角度来看，大语言模型的校准仍未得到充分探索。Zhao 等人(2021年)首次指出，模型普遍缺乏校准能力。一种简单的解决方法是通过额外的无内容上下文样本（即输入为“N/A”的示例）和校准参数来提示模型。Huang 等人(2025a)则指出，指令微调本身可能导致模型相较于基础模型的校准能力显著下降，并提出在训练期间使用标签平滑作为解决方案。

多语言大语言模型。 多语言自然语言处理已大幅超越最初以英语为中心的范式，以应对世界的语言多样性。近期的大语言模型（Grattafiori 等人，2024年；Yang 等人，2024年；OpenAI，2023年；Gemma 团队，2024年）利用涵盖数十至数百种语言的大规模预训练数据集，展现出卓越的多语言能力。然而，研究表明，这些系统仍面临持续存在的挑战，尤其是模型在内部处理非英语输入时会采用类似英语的表征，以及高资源语言与低资源语言之间始终存在的性能差距（Zhong 等人，2024年）。

3 方法论

模型与设置。为了检验我们的假设，我们训练并测试了 Mistral-7B（Jiang 等人，2023年）、Llama3.1-8B（Grattafiori 等人，2024年）和 Gemma2-2B（Gemma 团队，2024年）模型。我们使用 Alpaca（Dubois 等人，2023年）、Tulu3Mixture（Wang 等人，2023a）和 OpenHermes（Teknium，2023年）数据集进行微调，并参考 Wang 等人（2023a）的建议。我们采用 AdamW优化器，通过网格搜索选择学习率 {5e-6、1e-5、5e-5、1e-4}，在训练前 2% 的过程中进行线性预热，批量大小设为 128，丢弃率设为 0.1。

任务。 为公平评估多种语言的校准情况，我们使用 **MMLU-ProX**数据集（Xuan 等人，2025年）和 **GlobalMMLU**（Singh 等人，2024年），这两个综合性基准测试涵盖了多种语言（分别为 29 种和 42 种），并基于仅包含英语的 MMLU 数据集构建（Hendrycks 等人，2021年）。每种语言版本均包含 12k 和 14k 道相同的问题，从而能够更直接地进行跨语言比较。对于每个问题，系统会提供一个输入，模型必须从 K 个不同的候选答案中进行选择（**MMLU-ProX**最多有 10 个，**GlobalMMLU**最多有 4 个）。具体而言，我们会计算候选生成结果的困惑度，并将其归一化为不同选项上的概率分布，再用于分类。

4 结果

4.1 跨语言校准

我们首先试图回答的问题是：大语言模型校准是否只与大语言模型训练数据中出现的语言有关。为此，我们使用若干开放模型及其指令微调变体，并在多种语言上进行评估，开展初步验证。结果如图 1 所示。有趣的是，我们在所有语言中都观察到一种一致的现象：指令微调会使所有语言的校准效果变差，包括那些不太可能出现在微调数据集中的语言。

4.2 考虑过度自信

微调后出现了过度自信问题，这使我们能够进一步提出一个问题：如何在不

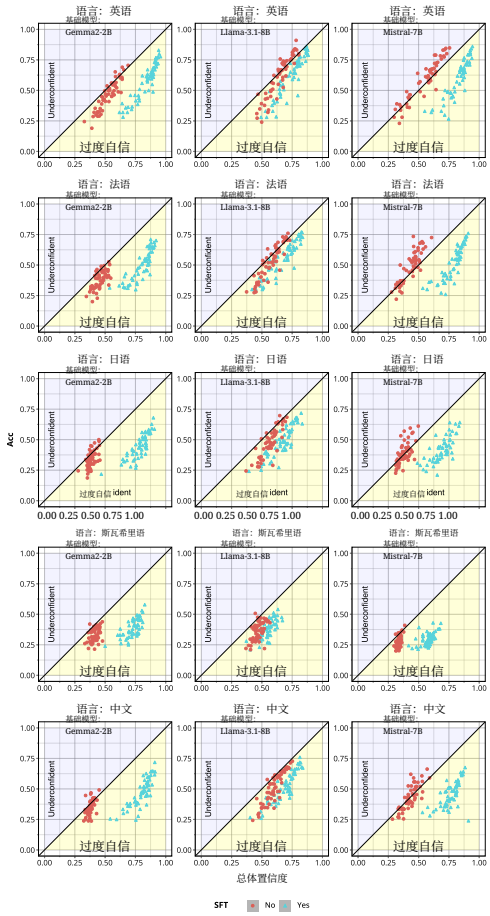


图1：比较**基础模型**（红色）与指令微调（蓝色）模型在各种语言上的表现，评测基准为**GlobalMMLU**。与直线的偏离表示置信不足或过度自信；模型在所有语言中都表现出日益严重的过度自信。

降低性能的情况下？为此，我们可以考虑将标签平滑作为解决方案。

标签平滑（LS）已被证明是一种很有前景的方法，可用于防止模型变得Szegedy 等人、2016年；Müller 等人，2019年）过度自信，或者在提供的标签中存在噪声时Lukasik 等人，2020年；Wei 等人，2022b；Lu 等人，2023年）。考虑一个参数化为 $\boldsymbol{\theta}$、用于输出条件分布 $P(\cdot|\boldsymbol{x};\boldsymbol{\theta})$ 的模型，其定义在一个标签集上。模型通常通过最小化数据集 $\mathcal{D}=\{\boldsymbol{x}_n,y_n\}_{n=1}^N$ 上的交叉熵（CE）损失进行训练。

从未知分布中采样 $p(\boldsymbol{x},\quad y)$，

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

其中，δ_i^j 是克罗内克δ，而 $\widehat{\boldsymbol{\sigma}}(\boldsymbol{x};\boldsymbol{\theta})\in[0,1]^K$ 是输出分布。标签平滑将原始分布与离散均匀分布 $\mathcal{U}=[1/K]^K\in\mathbb{R}^K$ 按平滑率 $\beta\in[0,1]$进行混合。因此，标签平滑可以理解为一种正则化方法，使输出标签趋向于均匀分布；它通过鼓励较高的熵来防止过拟合，因此对于抑制过度自信尤其有效。这一论断在第 B.1 节中有所说明。此外，这也可以视为一个约束优化问题，它通过鼓励各类别的 logit 相等来确保对过度自信进行惩罚（第 B.2 节）。

标签平滑有助于多语言校准吗？一个自然的问题是，标签平滑是否足以帮助实现整体多语言校准。表1 显示，使用平滑的模型在两项任务上的平均校准效果都显著改善，而准确率的下降幅度仍然很小（≤ 0准确率仅下降 0.005 个百分点）。在单个语言上的结果也进一步印证了这一点（第 C节），凸显了这种方法在维持更稳健、可靠的语言模型方面的通用性。然而，过高程度的平滑也可能导致准确率有所下降，因此需要以固定或自适应的方式妥善处理这一超参数。

图2 以约鲁巴语为例进行了具体说明，该语言未出现在微调数据中；虽然准确率没有变化（没有垂直位移），置信度却有所提高（向右水平移动），导致校准偏差加剧。标签平滑明显缓解了这种过度自信，同时没有影响准确率。

5 讨论

为什么监督微调会导致校准失准？Oh 等人（2024年）从分布外（OOD）泛化的角度提出了一种观点。为了更好地

表1：对采用或未采用标签平滑的各种大语言模型进行指令微调的结果。对于所有模型/数据集组合，使用平滑都能降低校准误差（期望校准误差和均方根），同时准确率仅略有下降。表现较好的设置以**粗体** 标示；当差异具有统计显著性时（使用学生 t-检验），还会进一步斜体标示 。

模型	监督微调数据集	平滑	MMLU-ProX 准确率	MMLU-ProX 熵	MMLU-ProX ECE	MMLU-ProX 均方根	GlobalMMLU 准确率	GlobalMMLU 熵	GlobalMMLU ECE	GlobalMMLU RMS
Gemma2-2B	基础模型		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	基础模型		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	**基础模型**		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***

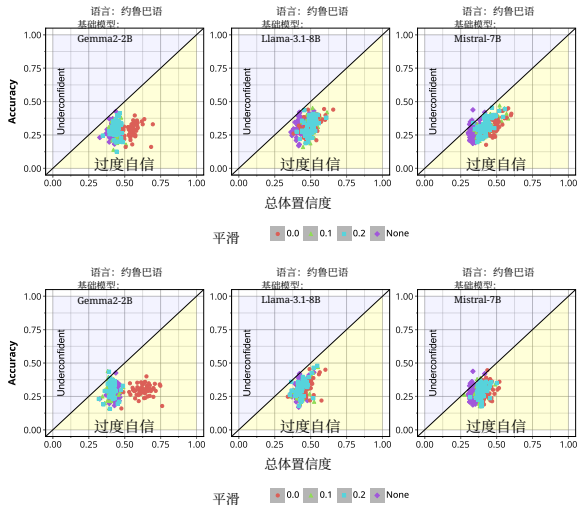


图 2：指令微调后，GlobalMMLU 的 yo 划分在 Tulu3Mixture（上）和 OpenHermes（下）上的可靠性图。

理解这一点，可以将监督微调数据视为由一个分布内（ID）数据集构成，而用于测试泛化和校准的下游数据集则构成一个分布外数据集。在这种设定下，最终分类器（对于自回归大语言模型而言，即语言建模头）的准确率和校准能力能否同时保持，与特征嵌入的多样性直接相关。具体而言，他们表明，该界限取决于协方差矩阵的最小奇异值，这意味着，随着一组学习到的特征嵌入（在此语境中即提示词的嵌入）之间的相互依赖性降低后，校准误差和分类误差都可以得到最小化。以往研究表明，微调能够显著降低此类特征的多样性（Mukhoti 等人，2024年；Kumar 等人，2022；Huh 等人，2024年），这也解释了为什么监督微调会显著损害校准效果。

关于大语言模型的差异置信度度量。除了模型的统计属性之外，大语言模型还具有一个额外维度，即能够直接通过提示询问其内部的置信度度量，这被称为口头表达的置信度（Geng 等人，2024年）。然而，此类方法通常依赖大语言模型生成一种内部化的置信度度量（Tao 等人，2024年），例如一个原始数值 $c \in [0, 1]$，但人们仍不清楚这类响应究竟是否真正衡量了模型置信度，或这类数值是否值得信任。总体而言，这促使我们进一步理解大语言模型中的置信度/不确定性校准，既要从统计角度出发，也要从面向用户的角度出发。

6 结论

通过对各种大语言模型开展受控调优实验，我们发现了一个令人惊讶的现象：即使在调优过程中从未接触过低资源语言，大语言模型在这些语言上的预测仍会显著失准。具体而言，

模型表现出一种内在倾向，即变得过度自信，即使性能没有提升也是如此；这一特定情形使标签平滑能够作为一种通用解决方案。总而言之，我们的研究揭示，语言模型可能以传统评估无法察觉的细微方式受到不利的下游影响。尽管存在缓解这些问题的可能方案，但研究也凸显出，我们需要更审慎地考虑如何使用数据集进行训练，以及如何充分评估模型。

局限性

缺乏明确的解决方案

作为一项实证研究，我们的工作仅考察了因输入语言不同而产生的各种校准现象是否存在。因此，我们没有提供适用于任意语言的模型校准整体解决方案。

伦理考量

本文提出了一种提升大词汇量语言模型校准效果的方法。我们预计，本研究的发现可能带来的社会影响或伦理问题微乎其微。

致谢

Jerry Huang 获得了加拿大自然科学与工程研究委员会（NSERC）加拿大研究生奖学金（编号 589326）以及日本学术振兴会（JSPS）日本研究博士后奖学金（编号 SP25408）的资助。Sarath Chandar 获得了加拿大CIFAR人工智能讲席、加拿大终身机器学习研究讲席以及加拿大自然科学与工程研究委员会探索基金的资助。作者感谢 Yufei Cui 的讨论，这些讨论促成了本工作的最初方向。本研究的完成部分得益于魁北克计算[1], 加拿大数字研究联盟（DRAC）[2] 和人工智能桥接云基础设施（ABCI）[3]提供的计算资源。

参考文献

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、Gun-jan Chhablani、Han Wang、Jason Alan Fries 及其他 8 位作者。2022 年。Promptsource：用于自然语言提示的集成开发环境和存储库。载于 第 *60* 届计算语言学协会年会论文集，*ACL 2022*——系统演示，爱尔兰都柏林，*2022* 年 *5* 月 *22* 日至 *27* 日，第 93–104 页。

Nematollah Kayhan Batmanghelich、Gerald T. Quon、Alex Kulesza、Manolis Kellis、Polina Golland 和 Luke Bornn。2014年。利用变分行列式点过程提升稀疏性的多样性。

Dimitri P. Bertsekas。1999年。非线性规划 *ng*.

Glenn W. Brier. 1950年. 以概率形式表示的预测验证。《月刊天气评论》，78(1):1–3。

Ta-Chung Chi、Ting-Han Fan 和 Alexander Rudnicky。2024年。注意力对齐和灵活的位置嵌入可改善 Transformer 长度外推。载于《计算语言学协会论文集：*NAACL 2024*》，墨西哥，墨西哥城，*2024*年*6*月*16*日至*21*日，第132–148页。

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 等另外 16位作者。2024年。扩展指令微调语言模型。《机器学习研究期刊》，25:70:1–70:53。

Tri Dao。2024年。FlashAttention-2：通过改进并行性和工作划分加速注意力机制。收录于第十二届国际学习表征会议，*ICLR 2024*，维也纳，奥地利，*2024*年*5*月*7*日至*11*日。

Tri Dao、Daniel Y. Fu、Stefano Ermon、Atri Rudra 和 Christopher Ré。2022。FlashAttention：一种快速且内存高效的精确注意力机制，具备 IO 感知能力。收录于神经信息处理系统进展第*35*卷：*2022*年神经信息处理系统年会，*NeurIPS 2022*，美国路易斯安那州新奥尔良，*2022*年*11*月*28*日至*12*月*9*日。

Morris H. DeGroot 和 Stephen E. Fienberg。1983。预测者的比较与评估。《皇家统计学会期刊》。*D*辑（统计学家），32(1):12–22。

Yann Dubois、Chen Xuechen Li、Rohan Taori、Tianyi Zhang、Ishaan Gulrajani、Jimmy Ba、Carlos Guestrin、Percy Liang 和 Tatsunori B. Hashimoto。2023。Alpacafarm：一种用于研究从人类反馈中学习的方法的仿真框架。收录于神经信息处理系统进展第*36*卷：*2023*年神经信息处理系统年会，*NeurIPS 2023*，美国路易斯安那州新奥尔良，*2023*年*12*月*10*日至*16*日。

Gemma 团队。2024年。Gemma 2：提升实用规模的开放语言模型。

Jiahui Geng、Fengyu Cai、Yuxia Wang、Heinz Koeppl、Preslav Nakov 和 Iryna Gurevych。2024年。大语言模型置信度估计与校准综述。载于《*2024*年北美计算语言学协会分会会议论文集：人类语言技术》（第*1*卷：长篇论文），*NAACL 2024*，墨西哥城，墨西哥，*2024*年*6*月*16*日至*21*日，第6577–6595页。计算语言学协会。

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

Asish Ghoshal、Xilun Chen、Sonal Gupta、Luke Zettlemoyer 和 Yashar Mehdad。2021年。利用低秩自适应标签平滑学习更优的结构化表征。载于第九届国际学习表征会议，*ICLR 2021*，线上会议，奥地利，*2021*年*5*月*3*日至*7*日。

Aaron Grattafiori 和另外 1 位作者。2024年。Llama 3 模型群。

Chuan Guo、Geoff Pleiss、Yu Sun 和 Kilian Q. Wein- berger。2017年。现代神经网络的校准。载于第 *34* 届国际机器学习会议，*ICML 2017*，澳大利亚新南威尔士州悉尼，*2017*年 *8* 月 *6—11* 日，机器学习研究论文集第 70 卷机器学习研究论文集，第 1321—1330 页。

Dan Hendrycks、Collin Burns、Steven Basart、Andy Zou、Mantas Mazeika、Dawn Song 和 Jacob Stein- hardt。2021年。衡量大规模多任务语言理解。载于 第 *9* 届国际学习表征会议，*ICLR 2021*，线上会议，奥地利，*2021*年 *5* 月 *3—7* 日。

Dan Hendrycks、Mantas Mazeika 和 Thomas G. Di- etterich。2019年。利用离群值暴露进行深度异常检测。载于第 *7* 届国际学习表征会议，*ICLR 2019*，美国路易斯安那州新奥尔良，*2019*年 *5* 月 *6—9* 日。

Pin-Lun Hsu、Yun Dai、Vignesh Kothapalli、Qingquan Song、Shao Tang、Siyu Zhu、Steven Shimizu、Shivam Sahni、Haowen Ning 和 Yanning Chen。2024年。Liger 内核：用于大语言模型训练的高效 Triton 内核。

Jerry Huang. 2025年。 长序列模型对长序列的建模效果如何？比较架构归纳偏置对长上下文能力的影响。载于第*31*届国际计算语言学会议论文集，第29–39页，阿布扎比，阿联酋。计算语言学协会。Jerry Huang、Peng Lu 和 Qiuhao Zeng。2025a。 经过校准的语言模型，以及如何使用标签平滑找到它们。载于第*42*届国际机器学习会议论文集，*ICML 2025*，加拿大不列颠哥伦比亚省温哥华，*2025*年*7*月*13*日至*19*日，机器学习研究论文集。Jerry Huang、Prasanna Parthasarathi、Mehdi Reza-gholizadeh 和 Sarath Chandar。2024a。 利用大语言模型实现推理加速的上下文感知助手选择。载于*2024*年自然语言处理实证方法会议论文集，第5817–5830页，美国佛罗里达州迈阿密。计算语言学协会。Jerry Huang、Prasanna Parthasarathi、Mehdi Reza-gholizadeh 和 Sarath Chandar。2024b。 迈向持续学习大语言模型的实用工具使用。预印本，arXiv:2404.09339。Jerry Huang、Prasanna Parthasarathi、Mehdi Reza-gholizadeh、Boxing Chen 和 Sarath Chandar。2025b。 机器蛇会像电羊一样做梦吗？探究架构归纳偏置对幻觉的影响。载于计算语言学协会论文集：*ACL 2025* 年度会议论文集，第1075–1096页，奥地利维也纳。计算语言学协会。

Minyoung Huh、Brian Cheung、Tongzhou Wang 和 Phillip Isola。2024年。立场：柏拉图式表征假说。载于第四十一届国际机器学习会议（*ICML 2024*），奥地利维也纳，*2024*年*7*月*21*日至*27*日。

Like Hui、Mikhail Belkin 和 Stephen Wright。2023年。用 squentropy 减少损失。载于 国际机器学习会议（*ICML 2023*），*2023*年*7*月*23*日至*29*日，美国夏威夷檀香山，第 202 卷，出自 机器学习研究论文集，第 14114–14131 页。

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 和 William El Sayed。2023年。Mistral 7B。

Zhengbao Jiang、Jun Araki、Haibo Ding 和 Graham Neubig。2021年。我们如何知道语言模型知道什么？关于用于问答的语言模型校准。计算语言学协会汇刊，9：962–977。

Andrew Kerr、Duane Merrill、Julien Demouth 和 John Tran。2017。Cutlass：CUDA C++ 中的快速线性代数。Ananya Kumar、Aditi Raghunathan、Robbie Matthew Jones、Tengyu Ma 和 Percy Liang。2022。微调可能会扭曲预训练特征，并导致分布外表现不佳。发表于第十届国际学习表征会议，*ICLR 2022*，线上会议，*2022*年*4*月*25*日至*29*日。Will LeVine、Benjamin Pikus、Pranav Raja 和 Fernando Amat Gil。2023。实现大型视觉语言模型零样本推理的校准。Tsung-Yi Lin、Priya Goyal、Ross B. Girshick、Kaiming He 和 Piotr Dollár。2017。用于密集目标检测的焦点损失。发表于*IEEE*国际计算机视觉会议，*ICCV 2017*，意大利威尼斯，*2017*年*10*月*22*日至*29*日，第2999–3007页。Bingyuan Liu、Ismail Ben Ayed、Adrian Galdran 和 Jose Dolz。2022。关键在于边界：基于边界的标签平滑用于网络校准。发表于*IEEE/CVF*计算机视觉与模式识别会议，*CVPR 2022*，美国路易斯安那州新奥尔良，*2022*年*6*月*18*日至*24*日，第80–88页。Peng Lu、Ahmad Rashid、Ivan Kobyzev、Mehdi Rezagholizadeh 和 Philippe Langlais。2023。LABO：通过双层优化实现最优标签正则化的学习。发表于计算语言学协会论文集

计算语言学协会：*ACL 2023*，加拿大多伦多，*2023*年*7*月*9*日至*14*日，第5759–5774页。

Michal Lukasik、Srinadh Bhojanapalli、Aditya Krishna Menon 和 Sanjiv Kumar。2020年。标签平滑能否缓解标签噪声？载于第*37*届国际机器学习大会（*ICML 2020*），*2020*年*7*月*13*日至*18*日，线上会议，机器学习研究论文集第119卷，第6448–6458页。

Matthias Minderer、Josip Djolonga、Rob Romijnders、Frances Hubis、Xiaohua Zhai、Neil Houlsby、Dustin Tran 和 Mario Lucic。2021年。重新审视现代神经网络的校准。载于神经信息处理系统进展第*34*卷：*2021*年神经信息处理系统年会（*NeurIPS 2021*），*2021*年*12*月*6*日至*14*日，线上，第15682–15694页。

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 及其他35人。2025年。SHADES：迈向对大语言模型中刻板印象的多语言评估。载于*2025*年美洲分会计算语言学协会会议论文集：人类语言技术（第*1*卷：长篇论文），第11995–12041页，阿尔伯克基，新墨西哥州。计算语言学协会。

Jishnu Mukhoti、Yarin Gal、Philip Torr 和 Puneet K. Dokania。2024年。微调可能会损害你的基础模型；保留特征或许是解决方案。机器学习研究汇刊，2024年。

Jishnu Mukhoti、Viveka Kulharia、Amartya Sanyal、Stuart Golodetz、Philip H. S. Torr 和 Puneet K. Dokania。2020年。使用焦点损失校准深度神经网络。载于神经信息处理系统进展第*33*卷：*2020*年神经信息处理系统年会，*NeurIPS 2020*，*2020*年*12*月*6*日至*12*日，线上。

Rafael Müller、Simon Kornblith 和 Geoffrey E. Hinton。2019年。标签平滑何时有帮助？载于神经信息处理系统进展第*32*卷：*2019*年神经信息处理系统年会，*NeurIPS 2019*，*2019*年*12*月*8*日至*14*日，温哥华，不列颠哥伦比亚省，加拿大，第4696–4705页。

Allan H. Murphy。1972年。概率评分的标量与向量划分：第一部分：双状态情形。《应用气象学与气候学杂志》，11(2):273–282。

Mahdi Pakdaman Naeini、Gregory F. Cooper 和 Milos Hauskrecht。2015年。使用贝叶斯分箱获得校准良好的概率。载于第二十九届 *AAAI* 人工智能大会论文集，*2015*年*1*月*25*日至*30*日，奥斯汀，得克萨斯州，美国，第2901–2907页。

Jeremy Nixon、Michael W. Dusenberry、Linchuan Zhang、Ghassen Jerfel 和 Dustin Tran。2019年。衡量深度学习中的校准。载于*IEEE* 计算机视觉与模式识别会议研讨会，*CVPR 2019*研讨会，*2019*年*6*月*16*日至*20*日，美国加利福尼亚州长滩，第38–41页。

Changdae Oh、Hyesu Lim、Mijoo Kim、Dongyoon Han、Sangdoo Yun、Jaegul Choo、Alexander G Hauptmann、Zhi-Qi Cheng 和 Kyungwoo Song。2024年。迈向视觉语言模型的校准的鲁棒微调。发表于神经信息处理系统进展 *38*：神经信息处理系统年度会议 *2024*，*NeurIPS 2024*，*2024*年*12*月*10*日至*15*日，加拿大不列颠哥伦比亚省温哥华。

OpenAI。2023年。GPT-4技术报告。*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 和 Ryan Lowe。2022年。通过人类反馈训练语言模型遵循指令。发表于神经信息处理系统进展 *35*：神经信息处理系统年度会议 *2022*，*NeurIPS 2022*，美国路易斯安那州新奥尔良，*2022*年*11*月*28*日至*12*月*9*日。

Gabriel Pereyra、George Tucker、Jan Chorowski、Lukasz Kaiser 和 Geoffrey E. Hinton。2017年。通过惩罚置信输出分布来正则化神经网络。载于第五届国际学习表征会议（*ICLR 2017*）论文集，法国土伦，*2017*年*4*月*24*日至*26*日，研讨会专题。

Gabriele Prato、Jerry Huang、Prasanna Parthasarathi、Shagun Sodhani 和 Sarath Chandar。2023年。EpiK-eval：将语言模型作为认知模型进行评估。载于*2023*年自然语言处理实证方法会议论文集，第9523–9557页，新加坡。计算语言学协会。

Gabriele Prato、Jerry Huang、Prasanna Parthasarathi、Shagun Sodhani 和 Sarath Chandar。2024年。大型语言模型知道自己知道多少吗？载于*2024*年自然语言处理实证方法会议论文集，第6054–6070页，美国佛罗里达州迈阿密。计算语言学协会。

PyTorch。2024年。torchtune：PyTorch 的微调库。

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、Debajy- oti Datta，以及另外21位作者。2022年。多任务提示训练实现零样本任务泛化。载于*The*

第十届国际学习表征会议，*ICLR 2022*，线上会议，*2022*年*4*月*25*日至*29*日。

Shivalika Singh、Angelika Romanou、Clémentine Fourrier、David I. Adelani、Jian Gang Ngui、Daniel Vila-Suero、Peerat Limkonchotiwat、Kelly Marchi- sio、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 等。2024年。《全球MMLU：理解并应对多语言评估中的文化和语言偏见》。

Christian Szegedy、Vincent Vanhoucke、Sergey Ioffe、Jonathon Shlens 和 Zbigniew Wojna。2016年。重新思考用于计算机视觉的 Inception 架构。载于*2016*年 *IEEE*计算机视觉与模式识别会议，*CVPR 2016*，拉斯维加斯，内华达州，美国，*2016*年*6*月*27*日至*30*日，第 2818–2826 页。

Shuchang Tao、Liuyi Yao、Hanxing Ding、Yuexiang Xie、Qi Cao、Fei Sun、Jinyang Gao、Huawei Shen 和 Bolin Ding。2024年。何时信任大语言模型：使置信度与回答质量相匹配。载于计算语言学协会论文集，*ACL 2024*，泰国曼谷及线上会议，*2024*年*8*月*11*日至*16*日，第5984–5996页。计算语言学协会。

Teknium。2023年。OpenHermes 2.5：面向通用型大语言模型助手的合成数据开放数据集。

Sunil Thulasidasan、Gopinath Chennupati、Jeff A. Bilmes、Tanmoy Bhattacharya 和 Sarah Michalak。2019年。关于 Mixup 训练：改进深度神经网络的校准和预测不确定性。载于 神经信息处理系统进展 *32*：*2019*年神经信息处理系统年会，*NeurIPS 2019*，*2019*年*12*月*8*日至*14*日，加拿大不列颠哥伦比亚省温哥华，第13888–13899页。

Philippe Tillet、Hsiang-Tsung Kung与David D. Cox。2019年。Triton：一种用于分块神经网络计算的中间语言与编译器。载于第*3*届*ACM SIGPLAN*机器学习与编程语言国际研讨会论文集，*MAPL@PLDI 2019*，美国亚利桑那州凤凰城，*2019*年*6*月*22*日，第10–19页。

Xinyu Wang、Linrui Ma、Jerry Huang、Peng Lu、Prasanna Parthasarathi、Xiao-Wen Chang、Boxing Chen与Yufei Cui。2025年。Resona：利用检索改进线性递归模型中的上下文复制。载于第二届语言建模会议。

Yizhong Wang、Hamish Ivison、Pradeep Dasigi、Jack Hessel、Tushar Khot、Khyathi Raghavi Chandu、David Wadden、Kelsey MacMillan、Noah A. Smith、Iz Beltagy 和 Hannaneh Hajishirzi。2023a。骆驼能走多远？探索开放资源上的指令微调现状。载于神经信息处理进展 第*36*卷：神经信息处理系统年会，*NeurIPS 2023*，美国路易斯安那州新奥尔良，*2023*年*12*月*10*日至*16*日。

Yizhong Wang、Yeganeh Kordi、Swaroop Mishra、Alisa Liu、Noah A. Smith、Daniel Khashabi与Hannaneh Hajishirzi。2023b。自我指令：利用自生成指令对齐语言模型。载于第*61*届计算语言学协会年会论文集（第*1*卷：长篇论文），*ACL 2023*，加拿大多伦多，*2023*年*7*月*9*日至*14*日，第13484–13508页。

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 Puro- hit、Ishani Mondal、Jacob Anderson、Kirby Kuznia、Kri ma Doshi、Kuntal Kumar Pal及另外16位作者。2022。Super-naturalinstructions：通过声明式指令实现自然语言处理任务的泛化。在*2022*年自然语言处理实证方法会议论文集（*EMNLP 2022*），阿拉伯联合酋长国阿布扎比，*2022*年*12*月*7*日至*11*日，第5085–5109页。

Jason Wei、Maarten Bosma、Vincent Y. Zhao、Kelvin Guu、Adams Wei Yu、Brian Lester、Nan Du、Andrew M. Dai和Quoc V. Le。2022a。微调语言模型是零样本学习器。在第十届国际学习表征会议，*ICLR 2022*，线上会议，*2022*年*4*月*25*日至*29*日。

Jiaheng Wei、Hangyu Liu、Tongliang Liu、Gang Niu、Masashi Sugiyama 和 Yang Liu。2022b。《平滑还是不平滑？标签平滑遇上噪声标签》。载于国际机器学习会议，*ICML 2022*，*2022*年*7*月*17*日至*23*日，巴尔的摩，马里兰州，美国，机器学习研究论文集第162卷，第23589–23614页。

Erik Wijmans、Brody Huval、Alexander Hertzberg、Vladlen Koltun 和 Philipp Krähenbühl。2025年。《削减大词汇量语言模型中的损失》。载于第十三届国际学习表征会议，*ICLR 2025*，新加坡，*2025*年*4*月*24*日至*28*日。

Ronald J. Williams 和 David Zipser。1989年。《一种持续运行的全循环神经网络学习算法》。*Neural Comput.*，1(2)：270–280。

Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pier- ric Cistac、Tim Rault、Rémi Louf、Morgan Funtow- icz 和 Jamie Brew。2019年。Huggingface 的 Transformers：最先进的自然语言处理。abs/1910.03771。

Miao Xiong、Zhiyuan Hu、Xinyang Lu、Yifei Li、Jie Fu、Junxian He 和 Bryan Hooi。2024年。大语言模型能否表达自身的不确定性？一项实证评估

大语言模型中的置信度 elicitation。发表于第十二届国际学习表征会议（*The Twelfth International Conference on Learning Representations*），*ICLR 2024*，奥地利维也纳，*2024*年*5*月*7*日至*11*日。

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Yun Xing, Junjue Wang, Huitao Li, Xin Li, Kunyu Yu, Nan Liu, Qingyu Chen, Dou- glas Teodoro, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo 和 Irene Li。2025年。Mmlu-prox：用于高级大语言模型评估的多语言基准。*CoRR*，arXiv:2503.10497。

An Yang、BaosongYang、BeichenZhang、BinyuanHui、Bo Zheng、Bowen Yu、Chengyuan Li、Dayiheng Liu、Fei Huang、Haoran Wei 和其他 1 位作者。2024年。Qwen2.5 技术报告。

QIUHAO Zeng、Jerry Huang、Peng Lu、Gezheng Xu、Boxing Chen、Charles Ling 和 Boyu Wang。2025年。ZETA：利用 z 阶曲线实现高效的 top-k 注意力。载于第十三届国际学习表征会议。

Shuangfei Zhai、Tatiana Likhomanenko、Etai Littwin、Dan Busbridge、Jason Ramapuram、Yizhe Zhang、Jiatao Gu 和 Joshua M. Susskind。2023年。通过防止注意力熵坍塌稳定 Transformer 训练。载于国际机器学习会议，*ICML 2023*，*2023*年*7*月*23*日至*29*日，美国夏威夷州檀香山，第202卷，机器学习研究论文集，第40770–40803页。

Chang-Bin Zhang、Peng-Tao Jiang、Qibin Hou、Yun-chao Wei、Qi Han、Zhen Li 和 Ming-Ming Cheng。2021年。深入探究标签平滑。*IEEE* 图像处理汇刊，30：5984–5996。

Zhilu Zhang 和 Mert R. Sabuncu。2020年。自蒸馏作为实例特定标签平滑。在神经信息处理系统进展*33*：神经信息处理系统年会*2020*，*NeurIPS 2020*，*2020*年*12*月*6*日至*12*日，线上。

Zihao Zhao、Eric Wallace、Shi Feng、Dan Klein 和 Sameer Singh。2021 年。使用前先校准：提升语言模型的少样本性能。载于第*38*届国际机器学习会议论文集（*ICML 2021*），*2021*年*7*月*18*日至*24*日，线上会议，第139卷，机器学习研究论文集，第12697–12706页。

Chengzhi Zhong、Fei Cheng、Qianying Liu、J unfeng Jiang、Zhen Wan、Chenhui Chu、Yugo Murawaki 和 Sadao Kurohashi。2024年。超越以英语为中心的大语言模型：多语言语言模型用什么语言思考?

Łukasz Rajkowski. 2019年. 高斯狄利克雷过程混合模型中的最大后验划分分析。*BayesianAnalysis*，14(2):477 – 494。

A 扩展相关研究

A.1 不确定性校准

在 K 类分类设置中，令 $\mathcal{X} \in \mathbb{R}^D$ 和 $\mathcal{Y} \in \{\gamma_k\}_{k=1}^K$ 分别表示输入空间和标签空间。令 f 为分类器，$f(\widehat{y}|\boldsymbol{x}) = \widehat{c}$ 为其置信度，即对应于其预测 $\widehat{y}$的 K 个维度中的最大概率。当一个模型完全校准时，

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

模型校准可以表示为 $\mathbb{E}\left[|P(\widehat{y} = y|\widehat{c} = c) - c|\right]$。

期望校准误差。常用的校准指标包括期望校准误差（ECE）（Naeini 等人，2015），该指标将 N 个样本的置信度分数划分为 M 均匀置信度分箱 $\{B_m\}_{m=1}^M$，并对各分箱误差进行加权求和。

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

多分类与静态校准误差。静态校准误差（SCE）是期望校准误差在多分类情形下对每个概率进行计算的一种简单扩展。SCE 针对每个类别概率分别对预测进行分箱，计算各分箱内的校准误差，并对各分箱取平均:

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

这里，acc(m, k) 和 conf(m, k) 分别表示类别标签为 k 的分箱 m 的准确率和置信度；$|B_{mk}|$ 是类别标签为 k 的分箱 b 中的预测数量。

自适应性与自适应校准误差。 自适应校准区间的设计源于区间选择中的偏差—方差权衡，这表明，为了获得整体校准误差的最佳估计，该指标应聚焦于预测所在的区域（而较少关注预测数量较少的区域）。由此产生了自适应校准误差（ACE），它采用自适应方案来划分区间，使每个区间都包含相同数量的预测。

具体而言，ACE 的输入包括预测 P（通常来自 Softmax）、正确标签以及范围数 R。

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

此处，acc(r, k) 和 conf(r, k) 分别表示类别标签为 k 的自适应校准区间 r 的准确率和置信度。校准区间 r 由排序并经过阈值处理后的预测中的第 $\lfloor N/R \rfloor$ 个索引定义。

均方根与平均绝对偏差校准误差。 均方根校准误差衡量置信度水平下置信度与准确率之间期望平方差的平方根。平均绝对值（MAD）采用了类似的公式，但对较大的置信度与准确率偏差惩罚较轻；它是均方根校准误差的下界。

为经验估计这些失准度量，将 N 样本再次划分为M 个区间。这里，区间并非等距划分，因为置信度值的分布并不均匀，而是动态变化的。具体而言，均方根校准误差采用以下数值稳定公式进行估计

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

类似地，MAD 校准误差由于使用绝对差而非平方差，属于不适当评分规则，其估计公式为

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

B 证明

B.1 标签平滑命题的证明

注释 B.1. 标签平滑可以理解为对输出标签上的均匀分布进行正则化，通过鼓励更高的熵来防止过拟合，因此在抑制过度自信方面尤其有效。

证明。 沿用第 4.2 节中的符号，考虑平滑率 $\beta\in[0,1]$ 和交叉熵损失。若使用标签平滑，则损失变为

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

□

B.2 约束命题的证明

为了更好地理解标签平滑的潜在影响，可以从约束优化的角度重新表述其作用，其中约束由正则化惩罚项施加（Bertsekas，1999年）。首先，定义

定义 B.2. 对于**logit 距离**向量 $\boldsymbol{x}$， $\boldsymbol{d}(\boldsymbol{x})$，有

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

确保模型不会高估某个特定类别的一种方法，是将其设为硬约束，从而使所有类别的 logit 相等，并产生Softmax输出 $\boldsymbol{o}=f(\boldsymbol{x};\ \boldsymbol{\theta})=[1/K]^K$。因此，通常更可取的做法是将其作为软惩罚函数 $\mathcal{P}:\ \mathbb{R}^K\to\mathbb{R}$ 加入训练期间最小化的目标函数中。回顾公式（7），我们可以将这种软惩罚与标签平滑目标函数引入的额外 KL 散度联系起来。

命题 B.3。对于硬约束 $\boldsymbol{d}(\boldsymbol{x})=\mathbf{0}$，线性惩罚项（或拉格朗日项）有上界和下界，分别为KL（$\boldsymbol{u}\|\widehat{\boldsymbol{\sigma}}(\boldsymbol{x};\ \boldsymbol{\theta})$），相差加性常数

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

证明（见第B.2节）表明，对于线性惩罚项，标签平滑近似地最小化约束 $\boldsymbol{d}(\boldsymbol{x})=\mathbf{0}$，促使每个类别的 logit 相等，从而确保对过度自信施加惩罚。

证明。给定 KL散度

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

我们有

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

考虑到 LogSumExp（LSE）函数的性质，可得

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

and

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

并且根据 $\boldsymbol{d}(\boldsymbol{x})$ 的定义，LS 除标准交叉熵损失 $\mathcal{L}^{\mathrm{CE}}$ 外额外施加的惩罚项 $\mathrm{KL}\left[\boldsymbol{u} \| \widehat{\boldsymbol{\sigma}}(\boldsymbol{x} ; \boldsymbol{\theta})\right]$，近似于针对以下约束优化线性惩罚项（或拉格朗日函数）

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

以促进各个 logit 相等。 □

C 完整结果

表2：对各种大语言模型进行有无标签平滑的指令微调结果。我们观察到，在所有模型/数据集组合中，使用平滑都能降低校准误差（期望校准误差和均方根），而准确率仅略有下降。此外，平滑能保持较高的熵，这表明模型对其预测分布总体上仍保有更大的不确定性。

模型	监督微调数据集	平滑	MMLU-ProX				GlobalMMLU			
			准确率	熵	ECE	均方根	准确率	熵	ECE	RMS
Gemma2-2B	基础模型		0.137	1.361	0.040	0.064	0.326	1.787	0.051	0.057
Mistral-7B	基础模型		0.238	1.498	0.150	0.118	0.354	1.855	0.041	0.041
Llama-3.1-8B	基础模型		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
Gemma2-2B	Alpaca	0.1	0.209	0.956	0.088	0.103	0.368	1.302	0.170	0.057
Gemma2-2B	OpenHermes	0.0	0.253	0.945	0.171	0.126	0.372	0.949	0.317	0.071
Gemma2-2B	OpenHermes	0.1	0.253	1.324	0.087	0.099	0.370	1.673	0.044	0.054
Gemma2-2B	Tulu3Mixture	0.0	0.223	1.185	0.100	0.097	0.359	1.247	0.185	0.064
Gemma2-2B	Tulu3Mixture	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
Mistral-7B	Alpaca	0.1	0.242	1.026	0.172	0.132	0.355	1.290	0.159	0.056
Mistral-7B	OpenHermes	0.0	0.275	1.275	0.132	0.116	0.372	1.561	0.053	0.036
Mistral-7B	OpenHermes	0.1	0.273	1.360	0.113	0.113	0.374	1.699	0.043	0.030
Mistral-7B	Tulu3Mixture	0.0	0.312	1.160	0.129	0.123	0.391	1.467	0.066	0.041
Mistral-7B	Tulu3Mixture	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
LLama-3.1-8B	Alpaca	0.1	0.322	0.513	0.166	0.122	0.441	1.227	0.131	0.050
LLama-3.1-8B	OpenHermes	0.0	0.259	0.435	0.159	0.131	0.439	1.345	0.071	0.041
LLama-3.1-8B	OpenHermes	0.1	0.350	0.831	0.153	0.123	0.438	1.521	0.032	0.030
LLama-3.1-8B	Tulu3Mixture	0.0	0.296	0.583	0.163	0.127	0.440	1.363	0.073	0.041
LLama-3.1-8B	Tulu3Mixture	0.1	0.315	0.772	0.148	0.126	0.441	1.489	0.038	0.034

C.1 单个语言

下表列出了 GlobalMMLU 和 MMLU-ProX 数据集上的单个语言结果。

C.1.1 GlobalMMLU

表 3：按阿姆哈拉语、模型、监督微调数据集和平滑方式划分的 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 4：不同模型、监督微调数据集和平滑方式在阿拉伯语 GlobalMMLU 子集上的结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 5：按模型、监督微调数据集和标签平滑划分的孟加拉语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 6：捷克语中不同模型、监督微调数据集和平滑方式在 GlobalMMLU 子集上的结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表7：德语模型、监督微调数据和平滑在 GlobalMMLU 子集上的结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 8：按模型、监督微调数据集和平滑方式划分的希腊语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 9：按模型、监督微调数据集和标签平滑方式划分的英语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表10：按模型、监督微调数据集和标签平滑划分的西班牙语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表11：按模型、监督微调数据集和平滑方式划分的波斯语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表12：按模型、监督微调数据集和平滑方式划分的菲律宾语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表13：按模型、监督微调数据集和平滑方式划分的法语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表14：按模型、监督微调数据集和平滑方法划分的豪萨语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表15：按模型、监督微调数据集和平滑方式划分的希伯来语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表16：按模型、监督微调数据集和平滑方式划分的印地语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表17：不同模型、监督微调数据集和平滑方式在印尼语 GlobalMMLU 子集上的结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表18：按模型、监督微调数据集和标签平滑划分的伊博语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表19：按模型、监督微调数据集和平滑方法划分的意大利语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表20：按模型、监督微调数据集和平滑方式划分的日语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表21：按模型、监督微调数据集和平滑方式划分的韩语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表22：吉尔吉斯语中不同模型、监督微调数据集和平滑方法在 GlobalMMLU 子集上的结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表23：不同模型、监督微调数据集和平滑方式在立陶宛语 GlobalMMLU 子集上的结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表24：按模型、监督微调数据集和平滑方式划分的 mg 语言 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表25：按模型、监督微调数据集和平滑方式划分的马来语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表26：按模型、监督微调数据集和平滑方式划分的尼泊尔语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表27：按模型、监督微调数据集和平滑方式划分的荷兰语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表28：按模型、监督微调数据集和平滑方法划分的奇切瓦语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表29：不同模型、监督微调数据集和标签平滑设置下，在GlobalMMLU 子集的pl语言上的结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 30：不同模型、监督微调数据集和标签平滑设置下，葡萄牙语 GlobalMMLU 子集上的结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 31：按模型、监督微调数据集和平滑方式划分的罗马尼亚语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表32：按模型、监督微调数据集和标签平滑划分的俄语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表33：按模型、监督微调数据集和平滑方式划分的僧伽罗语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表34：按模型、监督微调数据集和平滑方式划分的sn语言 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 35：按模型、监督微调数据集和平滑方式划分的索马里语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 36：按模型、监督微调数据集和平滑方式划分的塞尔维亚语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 37：按模型、监督微调数据集和平滑方式划分的瑞典语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 38：斯瓦希里语模型在 GlobalMMLU 子集上的结果，按监督微调数据集和平滑方式划分。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表39：按模型、监督微调数据集和平滑方式划分的泰卢固语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 40：按模型、监督微调数据集和平滑方式划分的土耳其语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 41：按模型、监督微调数据集和标签平滑方式划分的乌克兰语 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 42：不同模型、监督微调数据集和平滑设置下越南语 GlobalMMLU 子集的结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 43：约鲁巴语在 GlobalMMLU 子集上的结果，按模型、监督微调数据集和标签平滑分类。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表44：按模型、监督微调数据集和标签平滑划分的中文语言 GlobalMMLU 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 45：af 语言在不同模型、监督微调数据集和平滑方式下的 MMLU-ProX 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表46：阿拉伯语模型在 MMLU-ProX 子集上的结果，按监督微调数据集和平滑方式划分。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表47：按模型、监督微调数据集和标签平滑划分的孟加拉语 MMLU-ProX 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 48：按模型、监督微调数据集和平滑方式划分的捷克语 MMLU-ProX 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 49：德语模型、监督微调数据集和平滑设置下 MMLU-ProX 子集上的结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 51：按模型、监督微调数据集和平滑方法划分的西班牙语 MMLU-ProX 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表50：按模型、监督微调数据集和平滑方式划分的英语 MMLU-ProX 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表52：按模型、监督微调数据集和平滑方式划分的 MMLU-ProX 法语子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 53：按模型、监督微调数据集和平滑方法划分的印地语 MMLU-ProX 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 54：匈牙利语在不同模型、监督微调数据集和平滑方式下的 MMLU-ProX 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 55：按模型、监督微调数据集和标签平滑划分的印尼语 MMLU-ProX 子集上的结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 56：按模型、监督微调数据集和标签平滑划分的意大利语 MMLU-ProX 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表57：日语语言按模型、监督微调数据集和平滑方式划分的 MMLU-ProX 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表59：按模型、监督微调数据集和平滑方式划分的马拉地语 MMLU-ProX 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 58：韩语模型、监督微调数据集和平滑设置下 MMLU-ProX 子集的结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 60：按模型、监督微调数据集和平滑方式划分的尼泊尔语 MMLU-ProX 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 61：葡萄牙语模型在不同监督微调数据集和平滑设置下的 MMLU-ProX 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表62：rj语言模型在 MMLU-ProX 子集上的结果，按监督微调数据集和平滑方式分类。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 63：按模型、监督微调数据集和标签平滑划分的塞尔维亚语 MMLU-ProX 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表64：按模型、监督微调数据集和标签平滑划分的斯瓦希里语 MMLU-ProX 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表65：泰卢固语中按模型、监督微调数据集和平滑方式划分的 MMLU-ProX 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 66：泰语在不同模型、监督微调数据集和平滑设置下的 MMLU-ProX 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 67：按模型、监督微调数据集和平滑方式划分的乌克兰语 MMLU-ProX 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 68：按模型、监督微调数据集和平滑方法划分的乌尔都语 MMLU-ProX 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表 69：越南语在不同模型、监督微调数据集和平滑方式下的 MMLU-ProX 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表70：wo 语言在不同模型、监督微调数据集和平滑方式下的 MMLU-ProX 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表71：按模型、监督微调数据集和平滑方式划分的约鲁巴语 MMLU-ProX 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表72：按模型、监督微调数据集和标签平滑划分的中文语言 MMLU-ProX 子集结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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

表73：祖鲁语的 MMLU-ProX 子集在不同模型、监督微调数据集和平滑方式下的结果。

监督微调数据集	Base	平滑	准确率	熵	ECE	RMS
基础模型	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 个体可靠性图

D.1 GlobalMMLU

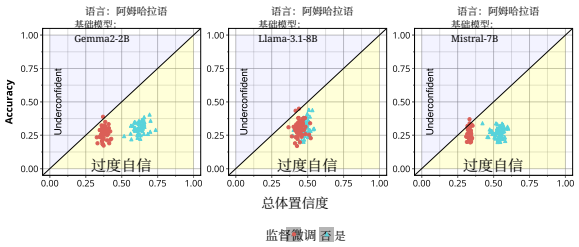


图 3：阿姆哈拉语数据集的**GlobalMMLU**可靠性图。

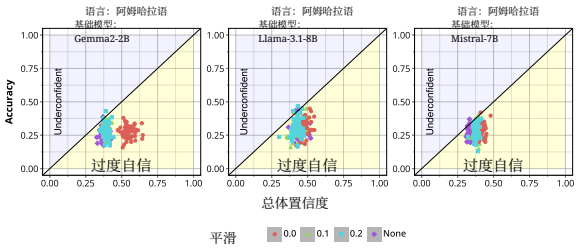


图 4：在**Tulu3Mixture** 数据集上进行指令微调后，阿姆哈拉语数据集的**GlobalMMLU**可靠性图。

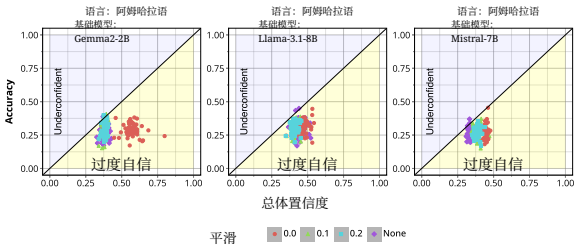


图5：在 **OpenHermes** 数据集上进行指令微调后，阿姆哈拉语在**GlobalMMLU**数据集上的可靠性图。

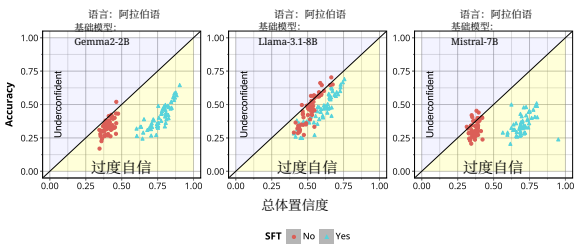


图6：阿拉伯语在**GlobalMMLU**数据集上的可靠性图。

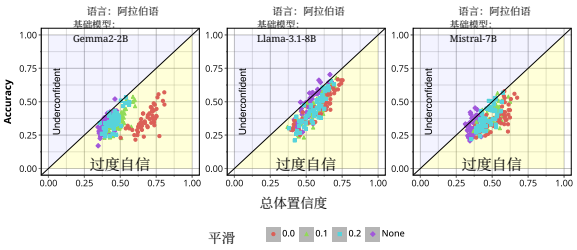


图7：在 **Tulu3Mixture** 数据集上进行指令微调后，阿拉伯语在**GlobalMMLU**数据集上的可靠性图。

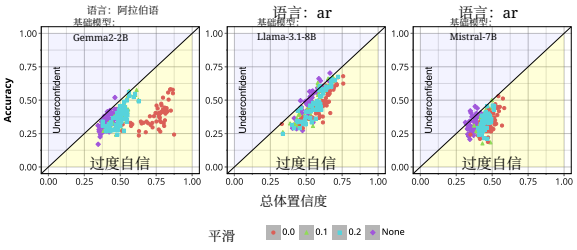


图8：在 **OpenHermes** 数据集上进行指令微调后，阿拉伯语在**GlobalMMLU**数据集上的可靠性图。

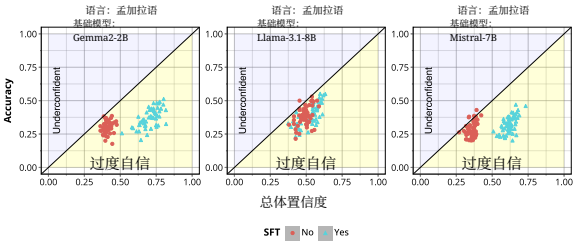


图9：孟加拉语在**GlobalMMLU**数据集上的可靠性图。

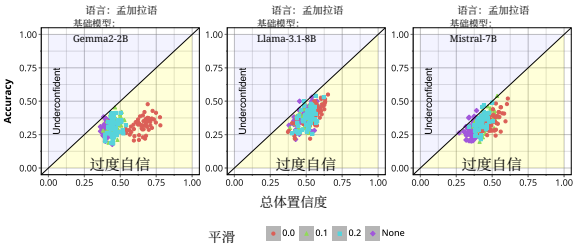


图10：在 **Tulu3Mixture** 数据集上进行指令微调后，孟加拉语在**GlobalMMLU**数据集上的可靠性图。

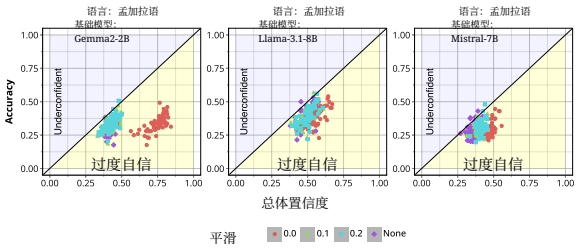


图11：在 **GlobalMMLU**数据集上，孟加拉语在使用 **OpenHermes** 数据集进行指令微调后的可靠性图。

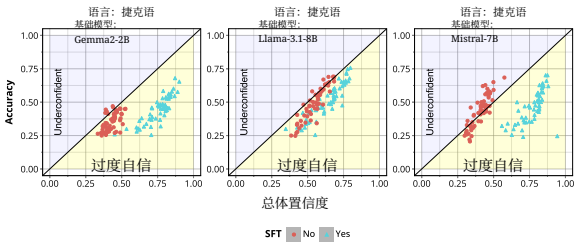


图12：**GlobalMMLU**数据集上捷克语的可靠性图。

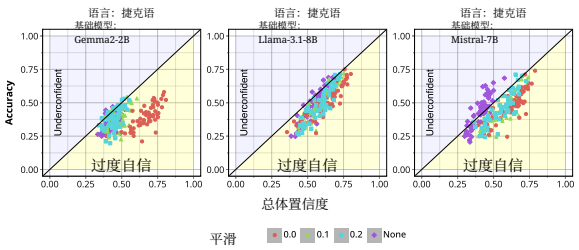


图13：在 **GlobalMMLU**数据集上，使用**Tulu3Mixture** 数据集进行指令微调后的捷克语可靠性图。

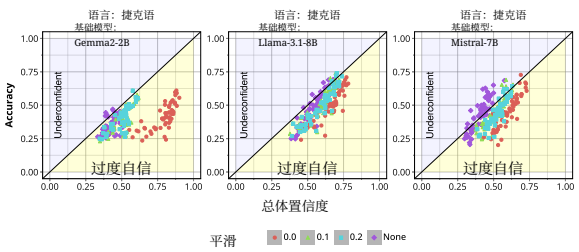


图14：在 **GlobalMMLU**数据集上，使用 **OpenHermes** 数据集进行指令微调后的捷克语可靠性图。

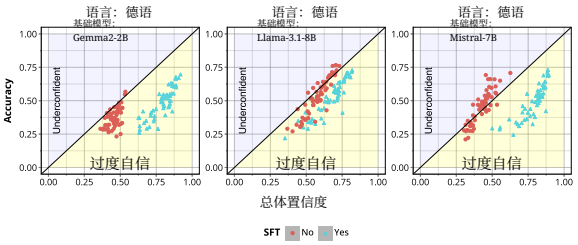


图15：**GlobalMMLU**数据集上德语的可靠性图。

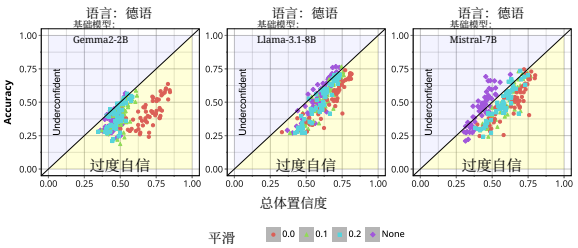


图 16：在 **Tulu3Mixture** 数据集上进行指令微调后，德语在**GlobalMMLU**数据集上的可靠性图。

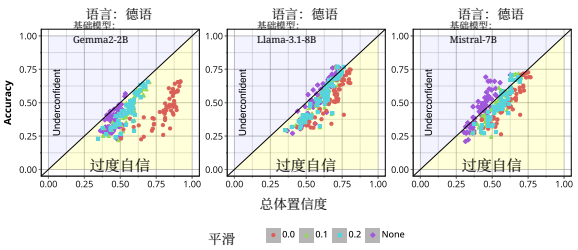


图 17：在 **OpenHermes** 数据集上进行指令微调后，德语在**GlobalMMLU**数据集上的可靠性图。

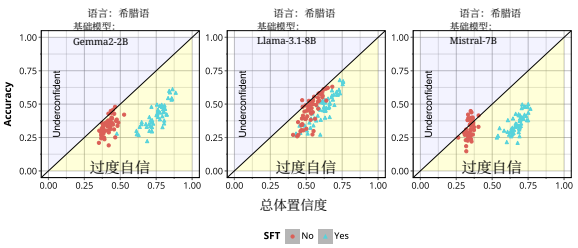


图 18：希腊语在**GlobalMMLU**数据集上的可靠性图。

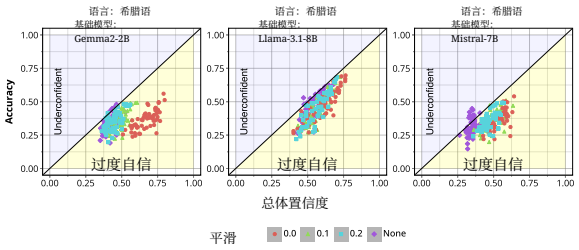


图19：在**GlobalMMLU**数据集的希腊语部分上，使用**Tulu3Mixture**数据集进行指令微调后的可靠性图。

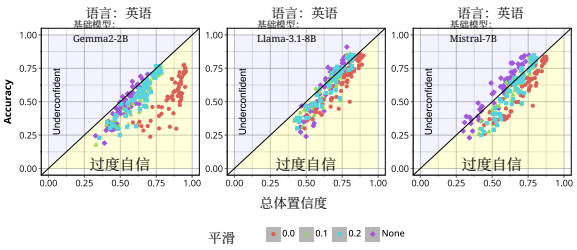


图23：在**OpenHermes**数据集上进行指令微调后，英语版**GlobalMMLU**数据集的可靠性图。

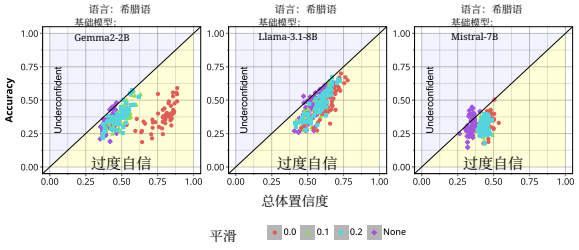


图20：在**GlobalMMLU**数据集的希腊语部分上，使用**OpenHermes**数据集进行指令微调后的可靠性图。

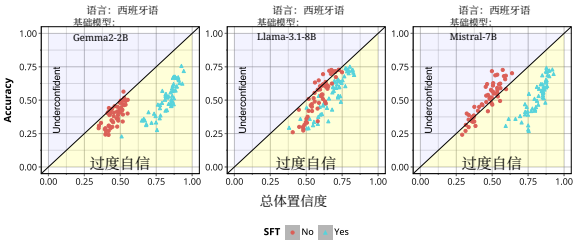


图24：西班牙语版**GlobalMMLU**数据集的可靠性图。

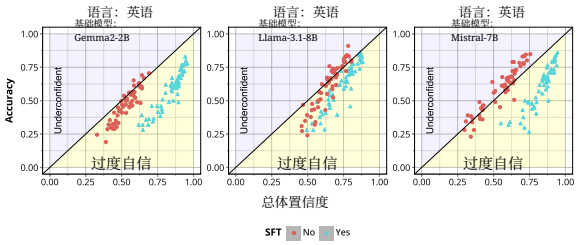


图21：**GlobalMMLU**数据集的英语部分的可靠性图。

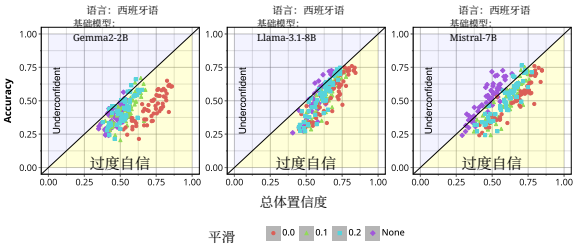


图25：在**Tulu3Mixture**数据集上进行指令微调后，西班牙语版**GlobalMMLU**数据集的可靠性图。

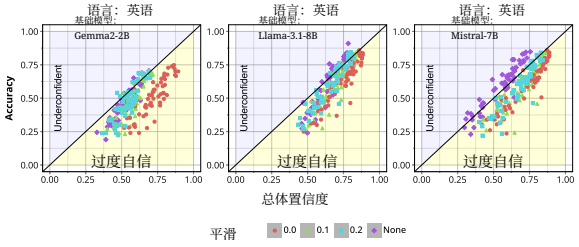


图22：在**GlobalMMLU**数据集的英语部分上，使用**Tulu3Mixture**数据集进行指令微调后的可靠性图。

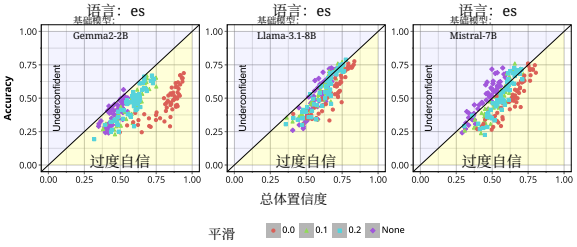


图26：在**OpenHermes**数据集上进行指令微调后，西班牙语版**GlobalMMLU**数据集的可靠性图。

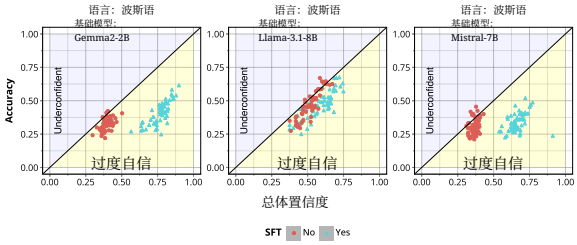


图 27：波斯语的 **GlobalMMLU**数据集的可靠性图。

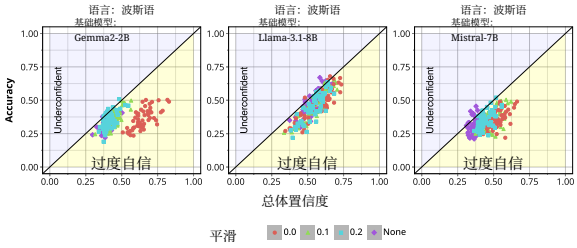


图 28：在**Tulu3Mixture** 数据集上进行指令微调后，波斯语的 **GlobalMMLU**数据集的可靠性图。

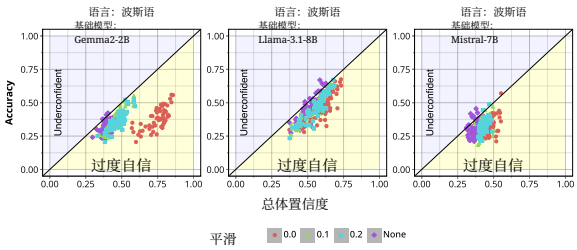


图 29：在 **OpenHermes** 数据集上进行指令微调后，波斯语的 **GlobalMMLU**数据集的可靠性图。

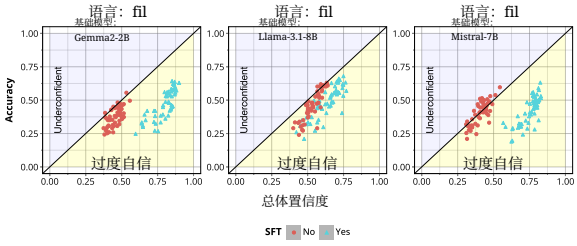


图 30：菲律宾语的 **GlobalMMLU**数据集的可靠性图。

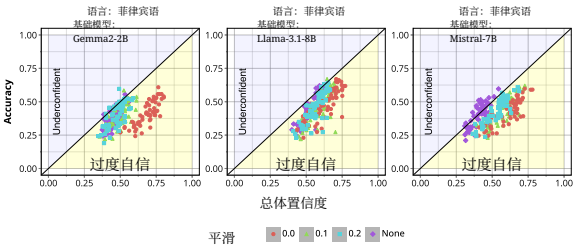


图 31：在 **Tulu3Mixture** 数据集上进行指令微调后，菲律宾语的 **GlobalMMLU**数据集的可靠性图。

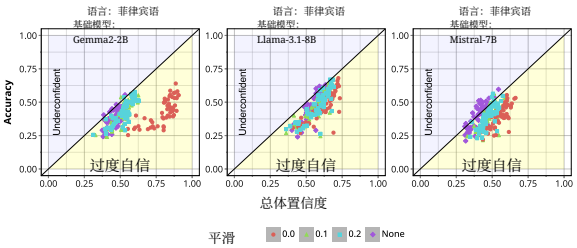


图 32：在 **OpenHermes** 数据集上进行指令微调后，菲律宾语在**GlobalMMLU**数据集上的可靠性图。

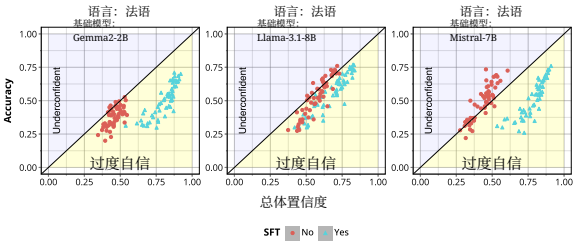


图 33：法语在**GlobalMMLU**数据集上的可靠性图。

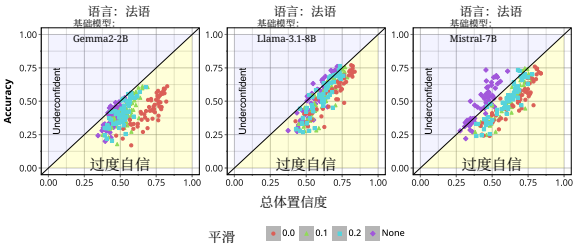


图 34：在**Tulu3Mixture** 数据集上进行指令微调后，法语在**GlobalMMLU**数据集上的可靠性图。

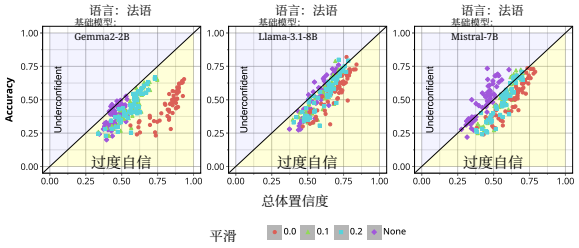


图35：在 **OpenHermes** 数据集上进行指令微调后，**GlobalMMLU**数据集的法语可靠性图。

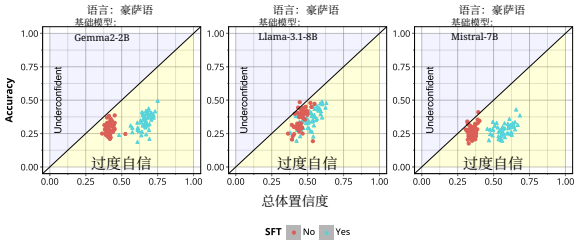


图36：**GlobalMMLU**数据集的豪萨语可靠性图。

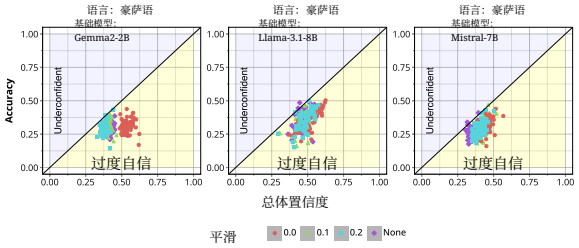


图37：在**Tulu3Mixture** 数据集上进行指令微调后，**GlobalMMLU**数据集的豪萨语可靠性图。

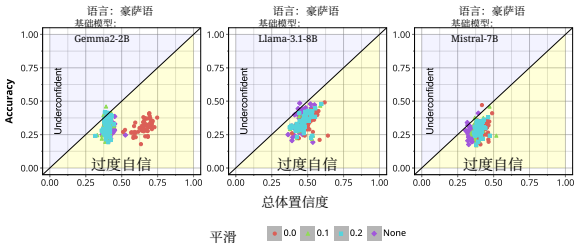


图38：在 **OpenHermes** 数据集上进行指令微调后，**GlobalMMLU**数据集的豪萨语可靠性图。

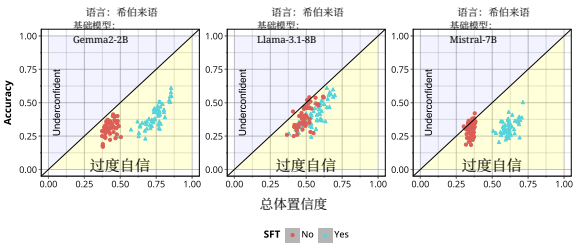


图39：**GlobalMMLU**数据集的希伯来语可靠性图。

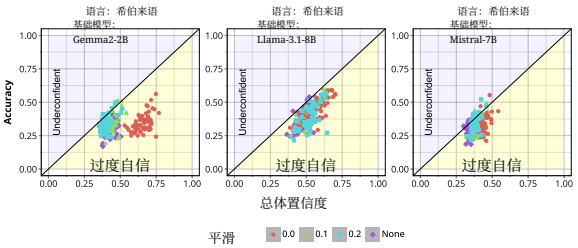


图40：在**Tulu3Mixture** 数据集上进行指令微调后，希伯来语在**GlobalMMLU**数据集上的可靠性图。

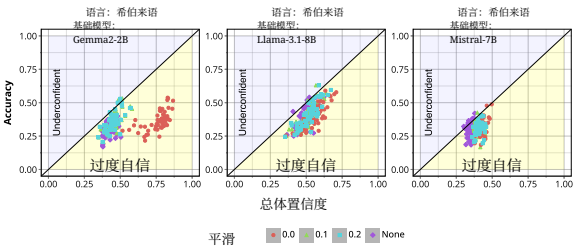


图41：在**OpenHermes** 数据集上进行指令微调后，希伯来语在**GlobalMMLU**数据集上的可靠性图。

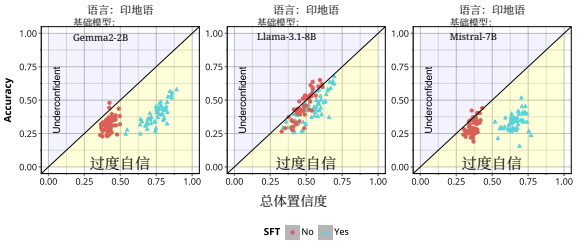


图42：印地语在**GlobalMMLU**数据集上的可靠性图。

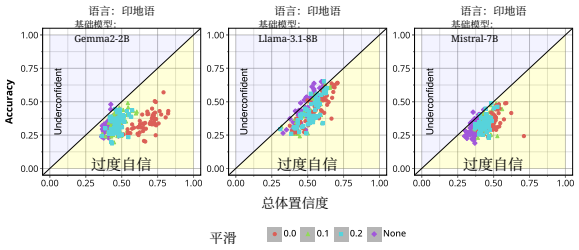


图 43：在 **GlobalMMLU**数据集上，印地语经过 **Tulu3Mixture** 数据集指令微调后的可靠性图。

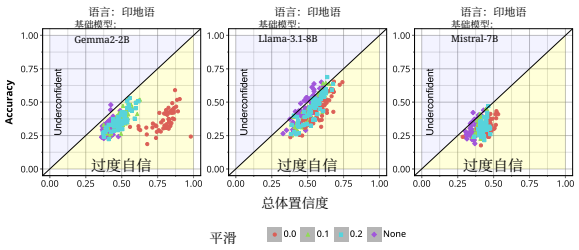


图 44：在 **GlobalMMLU**数据集上，印地语经过 **OpenHermes** 数据集指令微调后的可靠性图。

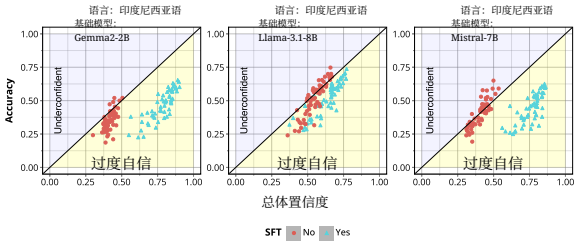


图 45：**GlobalMMLU**数据集上的印尼语可靠性图。

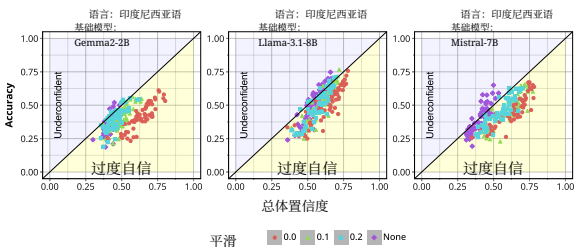


图 46：在 **GlobalMMLU**数据集上，印尼语经过 **Tulu3Mixture** 数据集指令微调后的可靠性图。

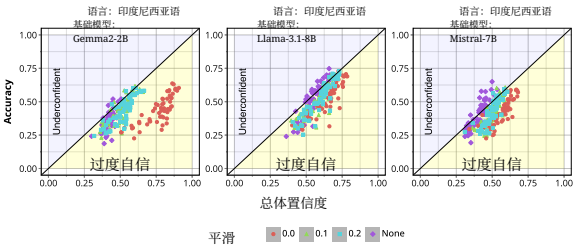


图47：在 **GlobalMMLU**数据集上，印度尼西亚语的可靠性图；在 **OpenHermes** 数据集上进行指令微调后。

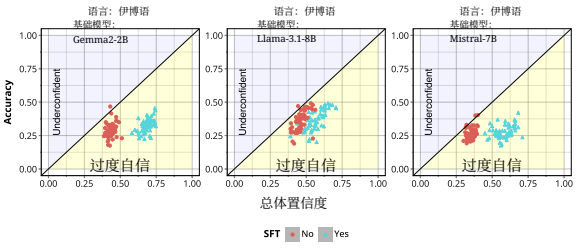


图48：**GlobalMMLU**数据集上伊博语的可靠性图。

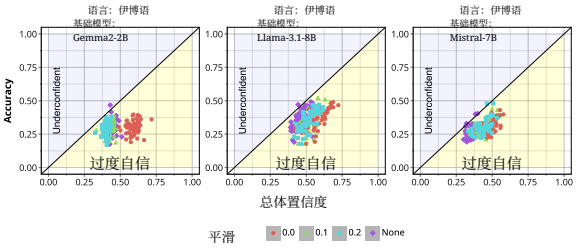


图49：在 **GlobalMMLU**数据集上，伊博语的可靠性图；在 **Tulu3Mixture** 数据集上进行指令微调后。

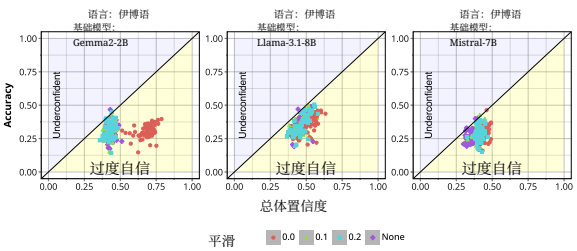


图50：在 **GlobalMMLU**数据集上，伊博语的可靠性图；在 **OpenHermes** 数据集上进行指令微调后。

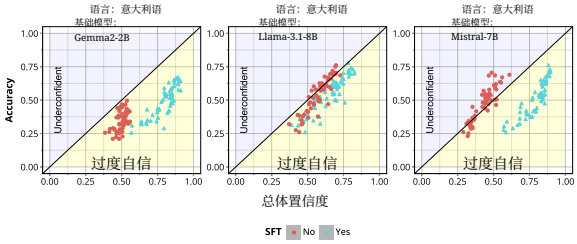


图 51：意大利语 GlobalMMLU数据集的可靠性图。

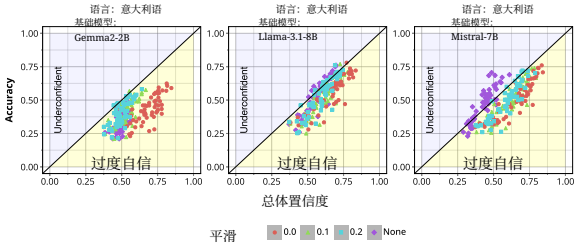


图 52：在 Tulu3Mixture 数据集上进行指令微调后，意大利语 GlobalMMLU数据集的可靠性图。

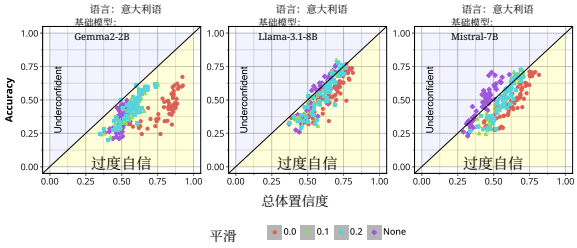


图 53：在 OpenHermes 数据集上进行指令微调后，意大利语 GlobalMMLU数据集的可靠性图。

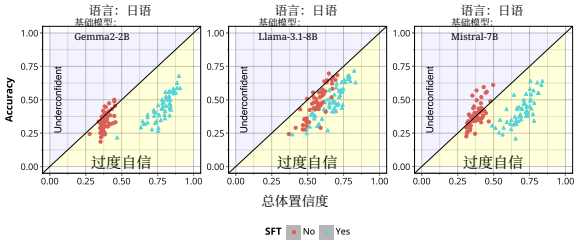


图 54：日语 GlobalMMLU数据集的可靠性图。

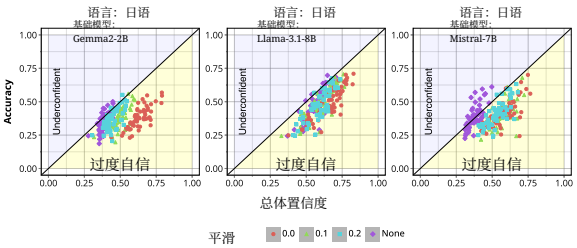


图 55：在 Tulu3Mixture 数据集上进行指令微调后，日语 GlobalMMLU数据集的可靠性图。

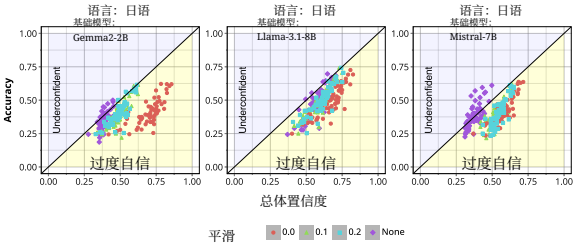


图 56：在 **OpenHermes** 数据集上进行指令微调后，**GlobalMMLU**数据集的日语可靠性图。

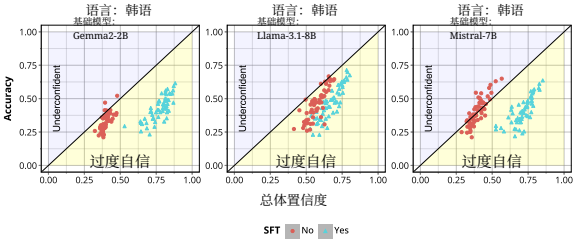


图 57：**GlobalMMLU**数据集的韩语可靠性图。

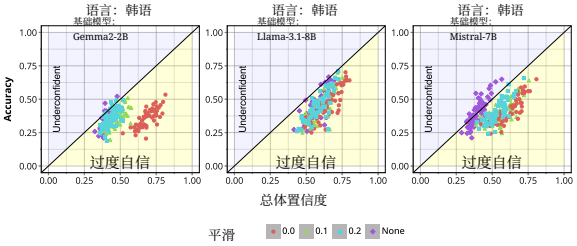


图 58：在 **Tulu3Mixture** 数据集上进行指令微调后，**GlobalMMLU**数据集的韩语可靠性图。

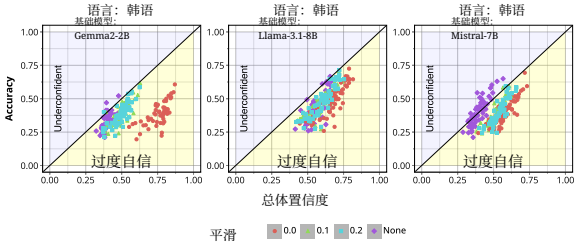


图59：**GlobalMMLU**数据集中韩语的可靠性图，基于**OpenHermes** 数据集进行指令微调后。

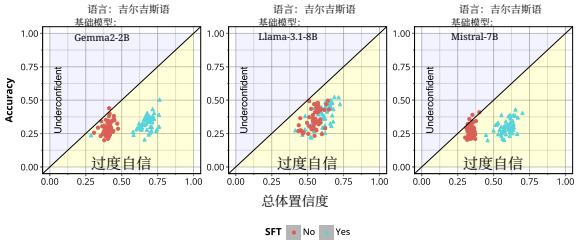


图60：**GlobalMMLU**数据集中吉尔吉斯语的可靠性图。

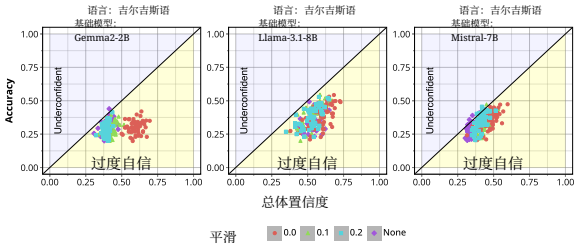


图61：**GlobalMMLU**数据集中吉尔吉斯语的可靠性图，基于**Tulu3Mixture** 数据集进行指令微调后。

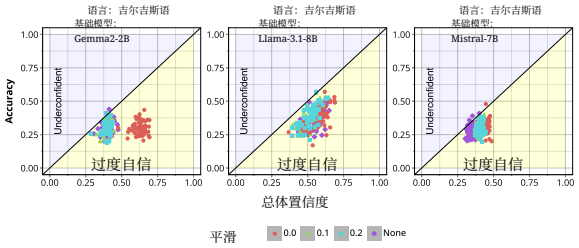


图62：**GlobalMMLU**数据集中吉尔吉斯语的可靠性图，基于**OpenHermes** 数据集进行指令微调后。

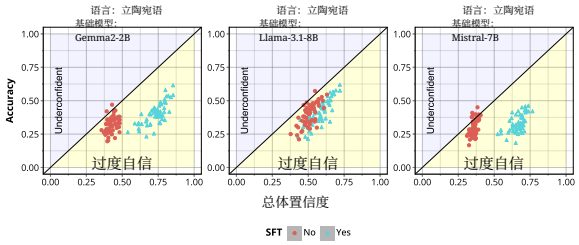


图63：**GlobalMMLU**数据集中立陶宛语的可靠性图。

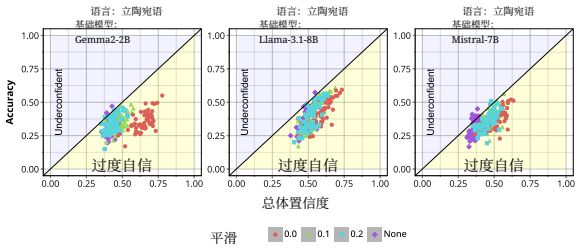


图 64：在 **Tulu3Mixture** 数据集上进行指令微调后，立陶宛语的 **GlobalMMLU**数据集的可靠性图。

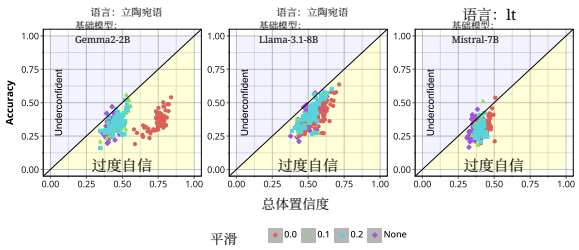


图 65：在 **OpenHermes** 数据集上进行指令微调后，立陶宛语的 **GlobalMMLU**数据集的可靠性图。

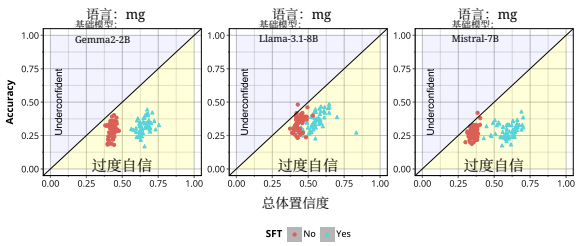


图 66：mg 语言的 **GlobalMMLU**数据集的可靠性图。

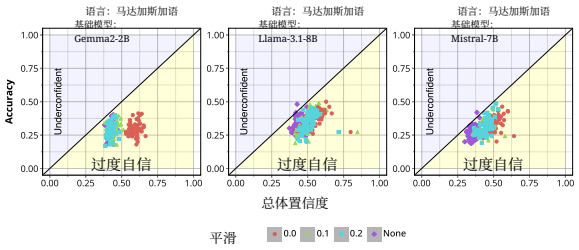


图67：在**GlobalMMLU**数据集上，mg 语言经过**Tulu3Mixture** 数据集指令微调后的可靠性图。

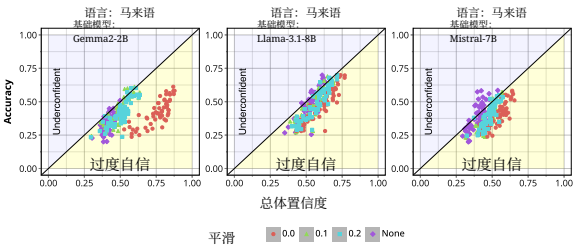


图71：**GlobalMMLU**数据集马来语的可靠性图，在OpenHermes 数据集上进行指令微调后。

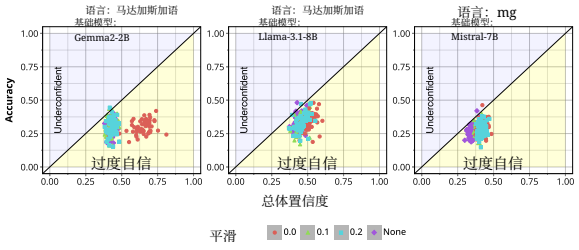


图68：在**GlobalMMLU**数据集上，mg 语言经过**OpenHermes** 数据集指令微调后的可靠性图。

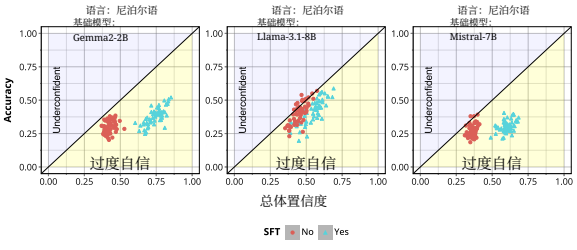


图72：**GlobalMMLU**数据集尼泊尔语的可靠性图。

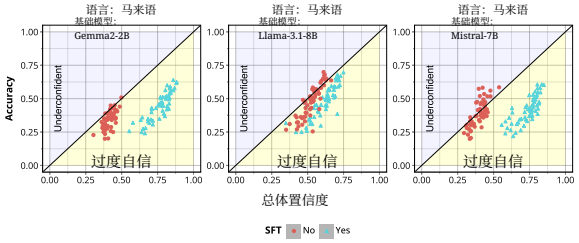


图69：**GlobalMMLU**数据集上 ms 语言的可靠性图。

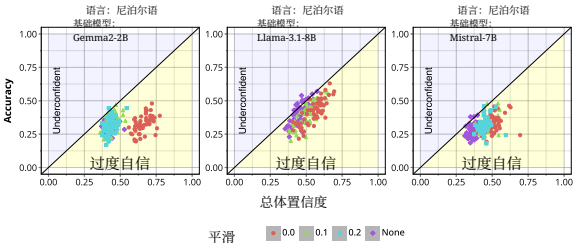


图73：**GlobalMMLU**数据集尼泊尔语的可靠性图，在**Tulu3Mixture** 数据集上进行指令微调后。

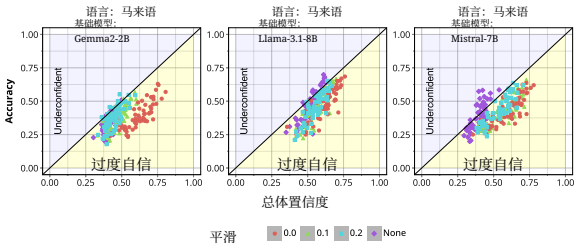


图70：在**GlobalMMLU**数据集上，ms 语言经过**Tulu3Mixture** 数据集指令微调后的可靠性图。

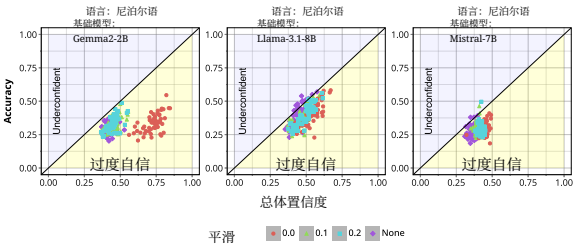


图74：**GlobalMMLU**数据集尼泊尔语的可靠性图，在**OpenHermes** 数据集上进行指令微调后。

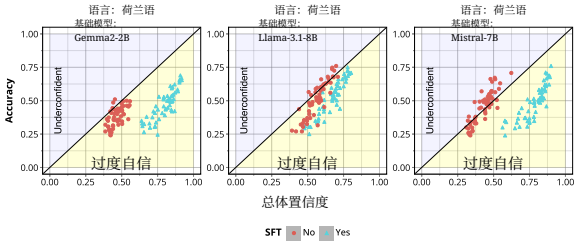


图75：荷兰语 GlobalMMLU 数据集的可靠性图。

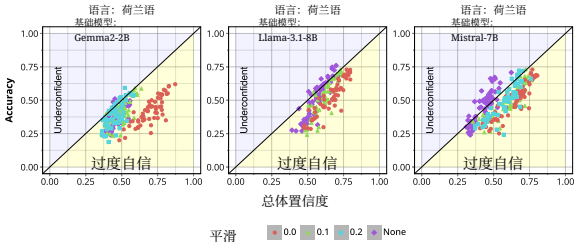


图76：使用 Tulu3Mixture 数据集进行指令微调后，荷兰语 GlobalMMLU 数据集的可靠性图。

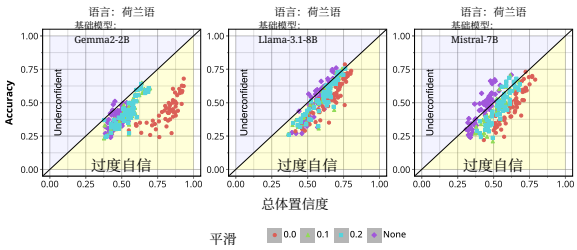


图77：使用 OpenHermes 数据集进行指令微调后，荷兰语 GlobalMMLU 数据集的可靠性图。

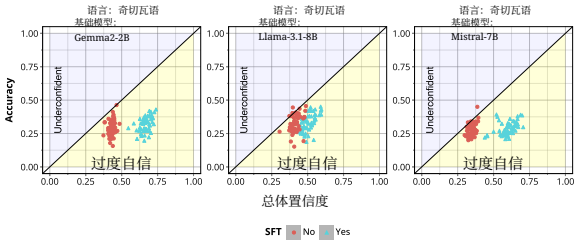


图78：ny语言 GlobalMMLU 数据集的可靠性图。

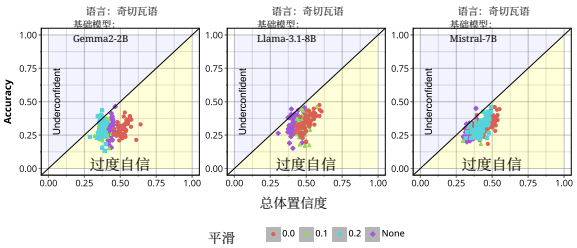


图79：使用 Tulu3Mixture 数据集进行指令微调后，ny语言 GlobalMMLU 数据集的可靠性图。

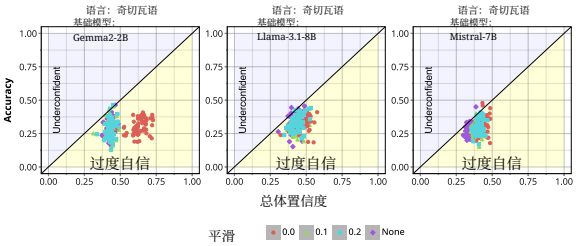


图80：在 **OpenHermes** 数据集上进行指令微调后，**GlobalMMLU**数据集的 ny语言可靠性图。

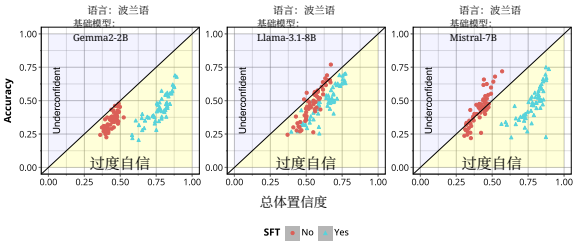


图81：**GlobalMMLU**数据集的 pl语言可靠性图。

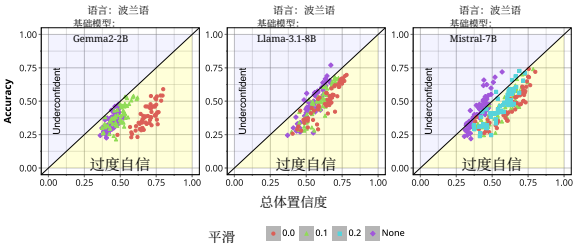


图82：在 **Tulu3Mixture** 数据集上进行指令微调后，**GlobalMMLU**数据集的 pl语言可靠性图。

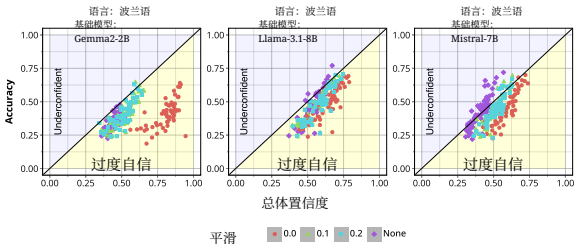


图83：在**GlobalMMLU**数据集的波兰语上，使用**OpenHermes**数据集进行指令微调后的可靠性图。

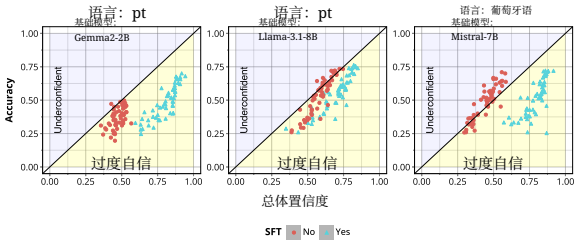


图84：**GlobalMMLU**数据集葡萄牙语的可靠性图。

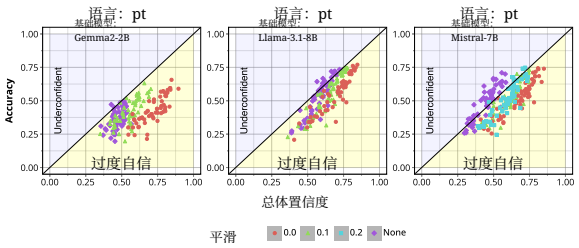


图85：在**GlobalMMLU**数据集的葡萄牙语上，使用**Tulu3Mixture**数据集进行指令微调后的可靠性图。

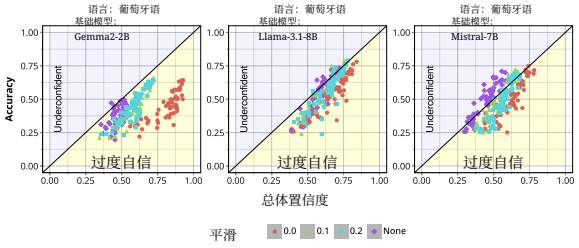


图86：在**GlobalMMLU**数据集的葡萄牙语上，使用**OpenHermes**数据集进行指令微调后的可靠性图。

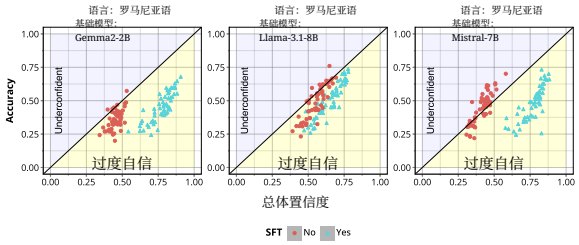


图87：**GlobalMMLU**数据集罗马尼亚语的可靠性图。

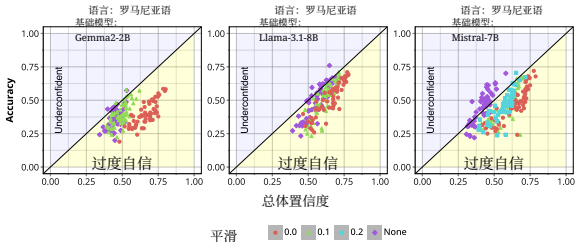


图88：在**Tulu3Mixture**数据集上进行指令微调后，罗马尼亚语在**GlobalMMLU**数据集上的可靠性图。

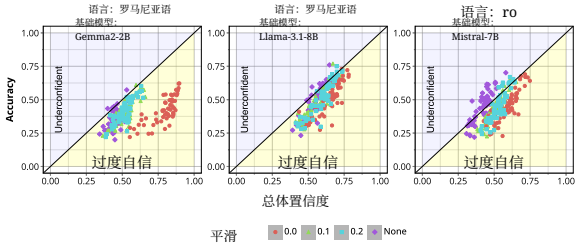


图89：在**OpenHermes**数据集上进行指令微调后，罗马尼亚语在**GlobalMMLU**数据集上的可靠性图。

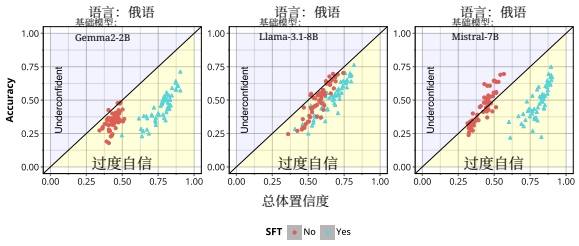


图90：俄语在**GlobalMMLU**数据集上的可靠性图。

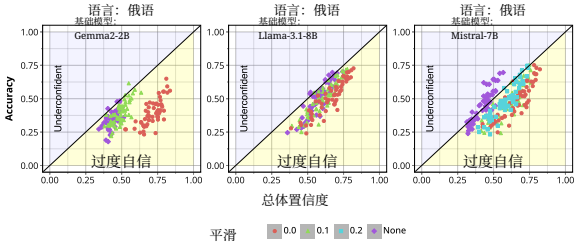


图91：在**Tulu3Mixture** 数据集上进行指令微调后，俄语在**GlobalMMLU**数据集上的可靠性图。

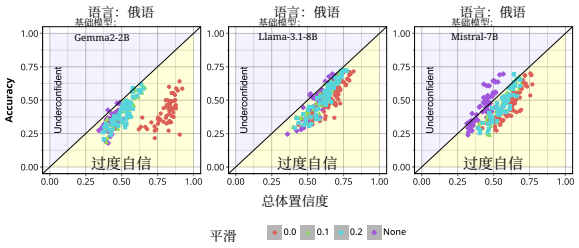


图92：在**OpenHermes** 数据集上进行指令微调后，俄语在**GlobalMMLU**数据集上的可靠性图。

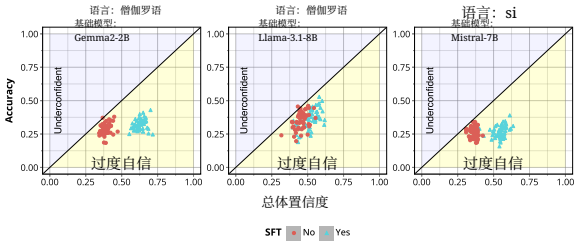


图93：僧伽罗语在**GlobalMMLU**数据集上的可靠性图。

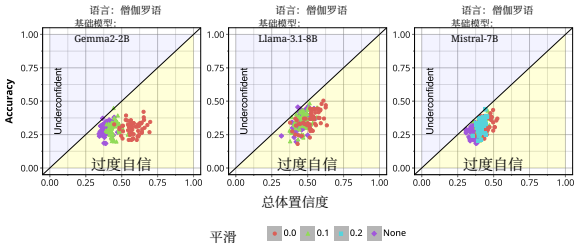


图94：在**Tulu3Mixture** 数据集上进行指令微调后，僧伽罗语在**GlobalMMLU**数据集上的可靠性图。

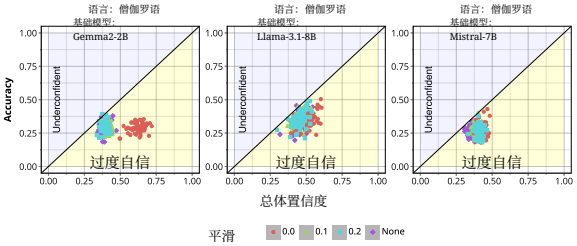


图95：僧伽罗语的**GlobalMMLU**数据集在使用**OpenHermes** 数据集进行指令微调后的可靠性图。

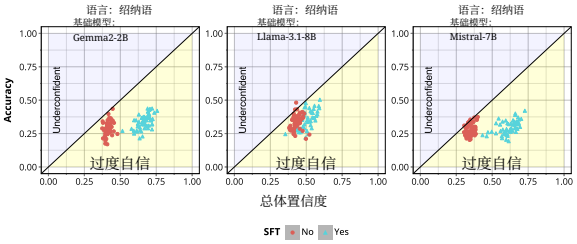


图96：sn语言的**GlobalMMLU**数据集的可靠性图。

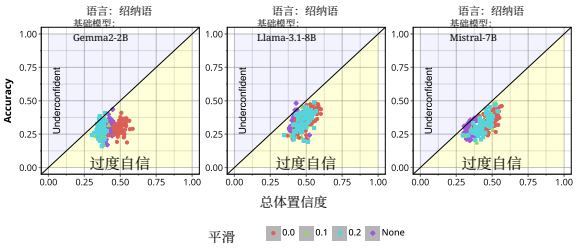


图97：sn语言的**GlobalMMLU**数据集在使用**Tulu3Mixture** 数据集进行指令微调后的可靠性图。

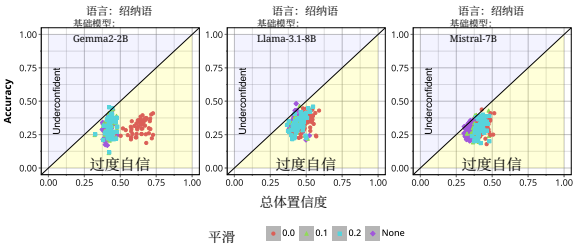


图98：sn语言的**GlobalMMLU**数据集在使用**OpenHermes** 数据集进行指令微调后的可靠性图。

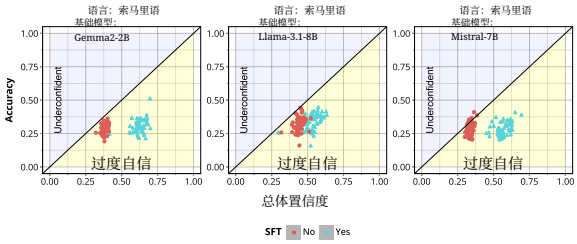


图99：索马里语 GlobalMMLU 数据集的可靠性图。

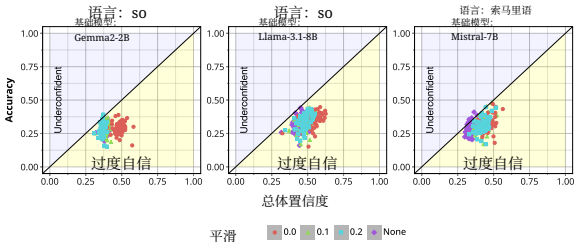


图100：在 Tulu3Mixture 数据集上进行指令微调后，索马里语 GlobalMMLU 数据集的可靠性图。

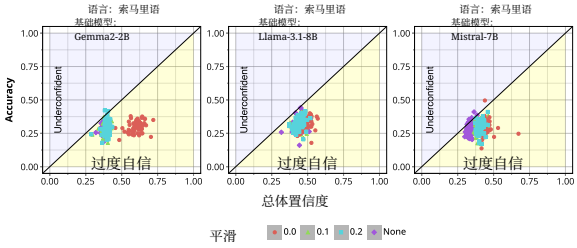


图101：在 OpenHermes 数据集上进行指令微调后，索马里语 GlobalMMLU 数据集的可靠性图。

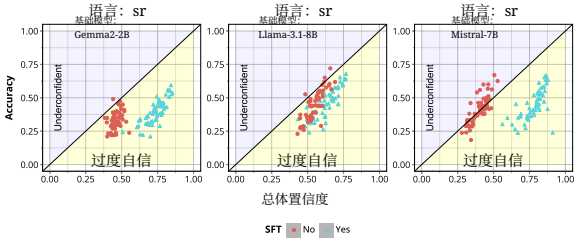


图102：塞尔维亚语 GlobalMMLU 数据集的可靠性图。

D.2 MMLU-ProX

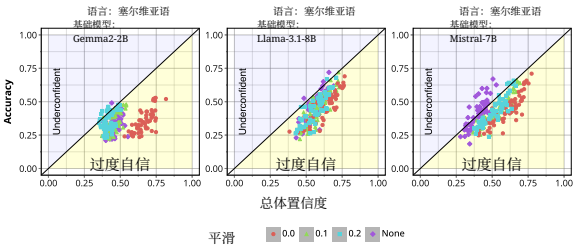


图103：塞尔维亚语在**GlobalMMLU**数据集上的可靠性图，该数据集经过**Tulu3Mixture** 数据集上的指令微调。

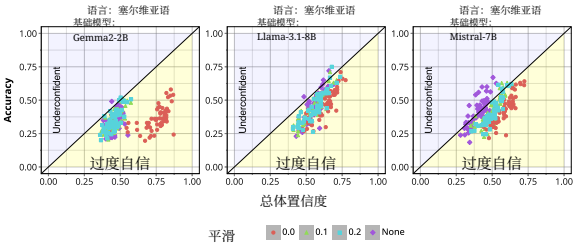


图104：塞尔维亚语在**GlobalMMLU**数据集上的可靠性图，该数据集经过**OpenHermes** 数据集上的指令微调。

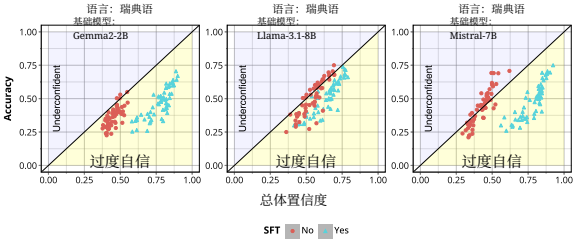


图105：瑞典语在**GlobalMMLU**数据集上的可靠性图。

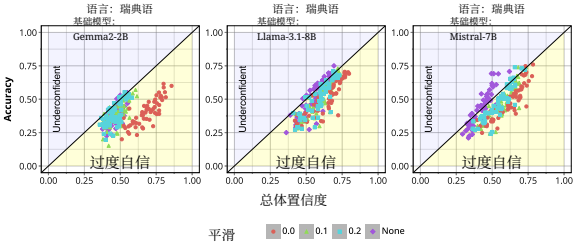


图106：瑞典语在**GlobalMMLU**数据集上的可靠性图，该数据集经过**Tulu3Mixture** 数据集上的指令微调。

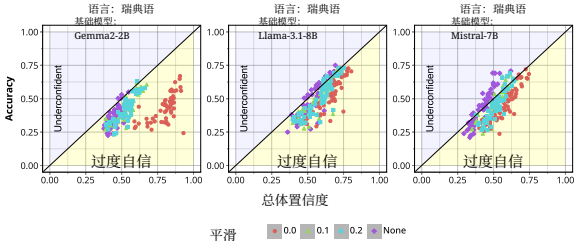


图107：在**OpenHermes** 数据集上进行指令微调后，瑞典语的**GlobalMMLU**数据集可靠性图。

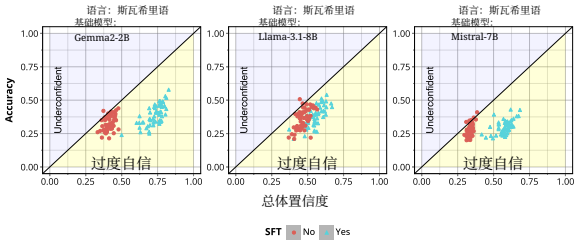


图108：斯瓦希里语的**GlobalMMLU**数据集可靠性图。

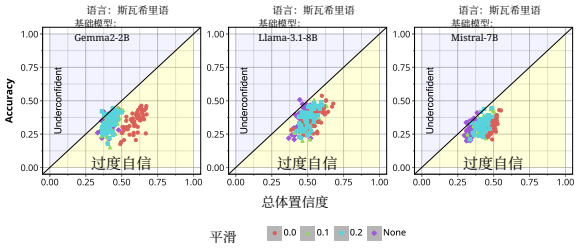


图109：在**Tulu3Mixture** 数据集上进行指令微调后，斯瓦希里语的**GlobalMMLU**数据集可靠性图。

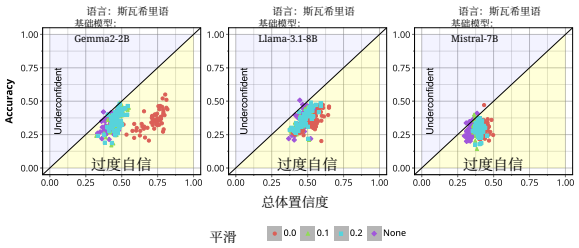


图110：在**OpenHermes** 数据集上进行指令微调后，斯瓦希里语的**GlobalMMLU**数据集可靠性图。

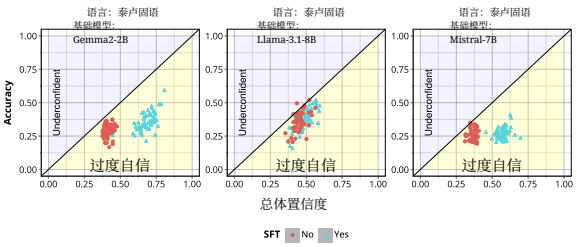


图111：泰卢固语的**GlobalMMLU**数据集可靠性图。

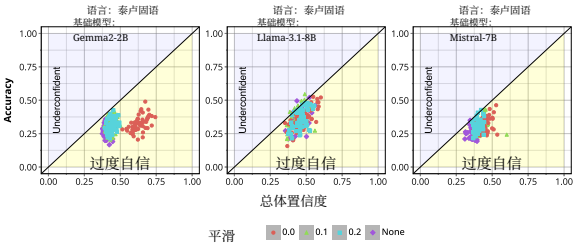


图112：在**Tulu3Mixture** 数据集上进行指令微调后，**GlobalMMLU**数据集在泰卢固语上的可靠性图。

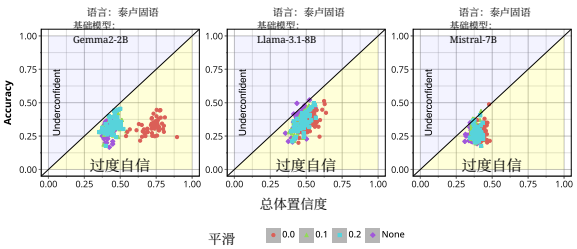


图113：在**OpenHermes** 数据集上进行指令微调后，**GlobalMMLU**数据集在泰卢固语上的可靠性图。

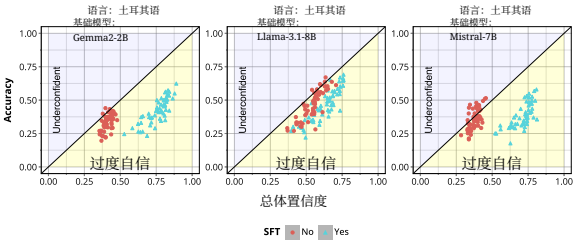


图114：**GlobalMMLU**数据集在土耳其语上的可靠性图。

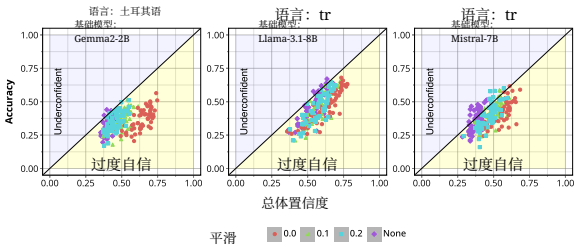


图115：在**GlobalMMLU**数据集的土耳其语部分上，使用**Tulu3Mixture** 数据集进行指令微调后的可靠性图。

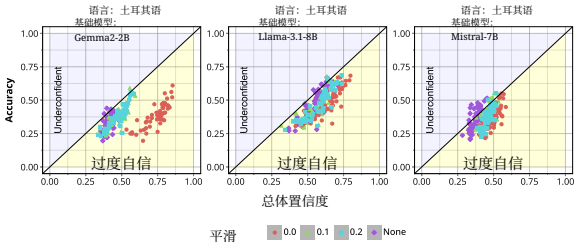


图116：在**GlobalMMLU**数据集的土耳其语部分上，使用**OpenHermes** 数据集进行指令微调后的可靠性图。

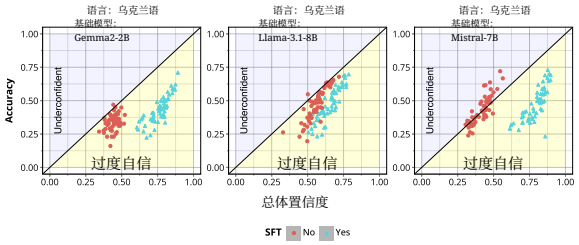


图117：**GlobalMMLU**数据集中乌克兰语部分的可靠性图。

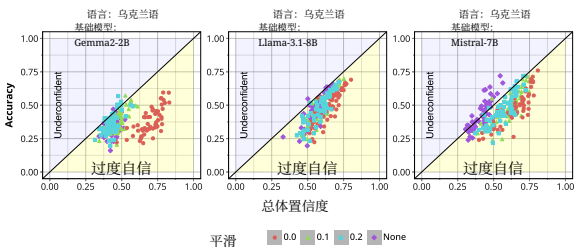


图118：在**GlobalMMLU**数据集的乌克兰语部分上，使用**Tulu3Mixture** 数据集进行指令微调后的可靠性图。

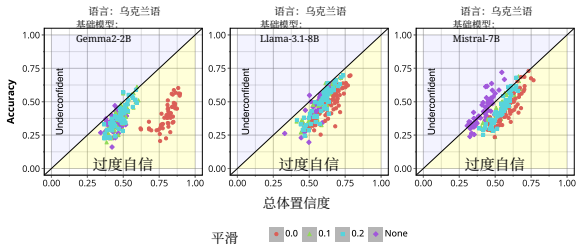


图119：在 **OpenHermes** 数据集上进行指令微调后，**GlobalMMLU**数据集的乌克兰语可靠性图。

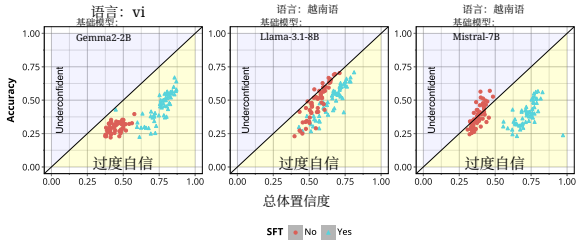


图120：**GlobalMMLU**数据集的越南语可靠性图。

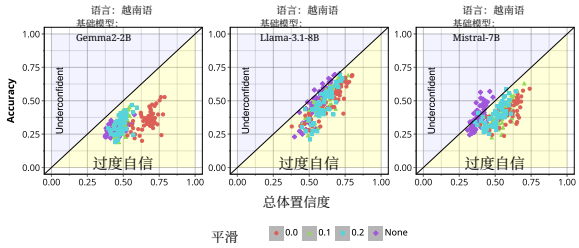


图121：在 **Tulu3Mixture** 数据集上进行指令微调后，**GlobalMMLU**数据集的越南语可靠性图。

图118：在**GlobalMMLU**数据集的乌克兰语部分上，使用**Tulu3Mixture** 数据集进行指令微调后的可靠性图。

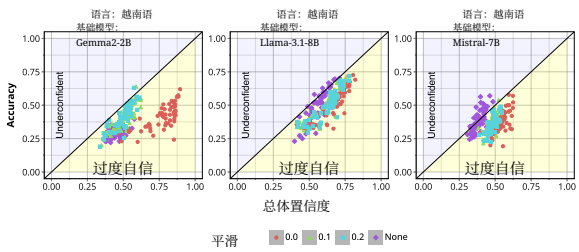


图122：在 **OpenHermes** 数据集上进行指令微调后，**GlobalMMLU**数据集的越南语可靠性图。

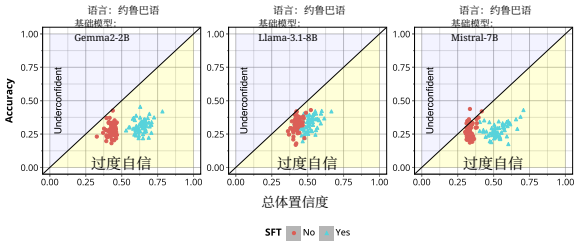
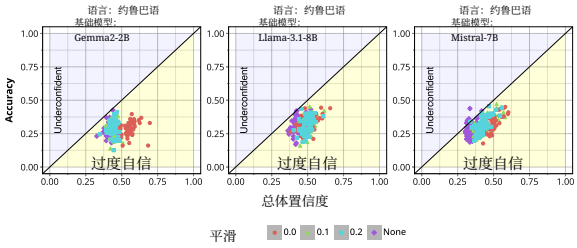
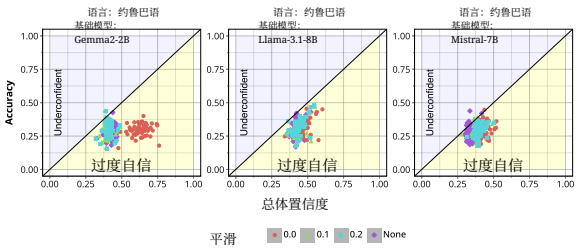


图123：约鲁巴语的**GlobalMMLU**数据集可靠性图。



在**Tulu3Mixture** 数据集上进行指令微调后，约鲁巴语的**GlobalMMLU**数据集可靠性图。



在**OpenHermes** 数据集上进行指令微调后，约鲁巴语的**GlobalMMLU**数据集可靠性图。

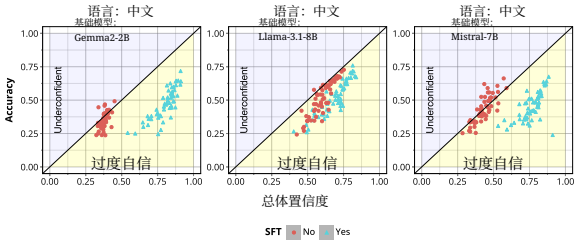


图126：中文的**GlobalMMLU**数据集可靠性图。

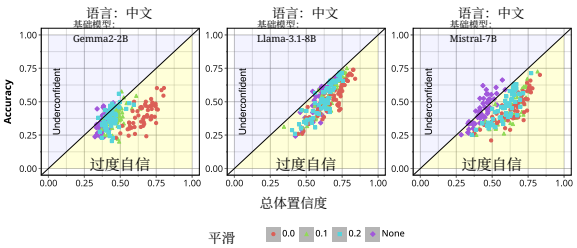


图127：在**Tulu3Mixture** 数据集上进行指令微调后，中文的**GlobalMMLU**数据集可靠性图。

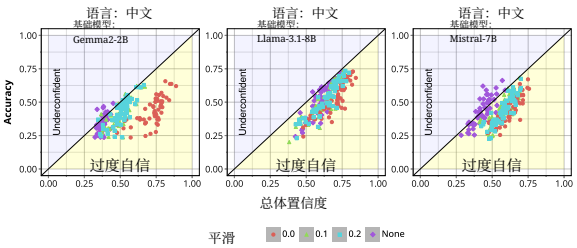


图128：在**OpenHermes** 数据集上进行指令微调后，zh 语言在**GlobalMMLU**数据集上的可靠性图。

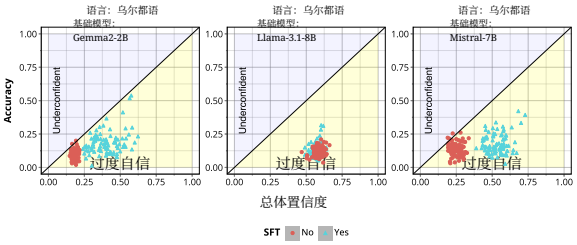


图129：ur 语言在**MMLU-ProX**数据集上的可靠性图。

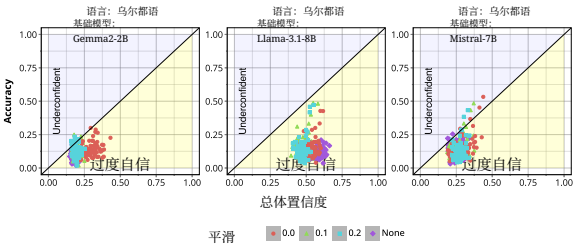


图130：在**Tulu3Mixture** 数据集上进行指令微调后，ur 语言在**MMLU-ProX**数据集上的可靠性图。

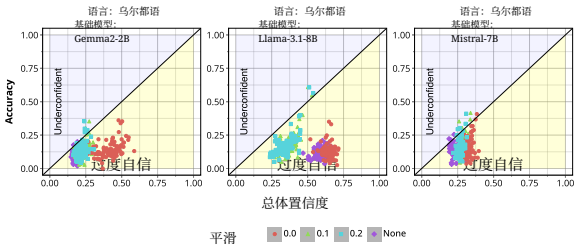


图131：在**MMLU-ProX**数据集上，乌尔都语的可靠性图；在**OpenHermes** 数据集上进行指令微调后。

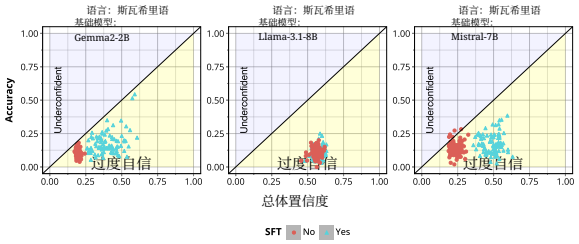


图132：**MMLU-ProX**数据集上斯瓦希里语的可靠性图。

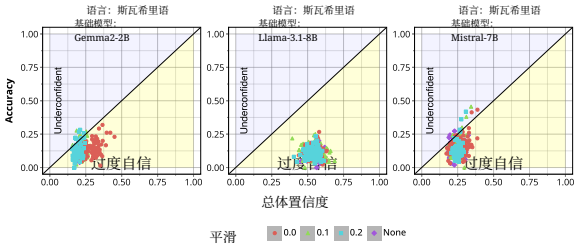


图133：在**MMLU-ProX**数据集上，斯瓦希里语的可靠性图；在**Tulu3Mixture** 数据集上进行指令微调后。

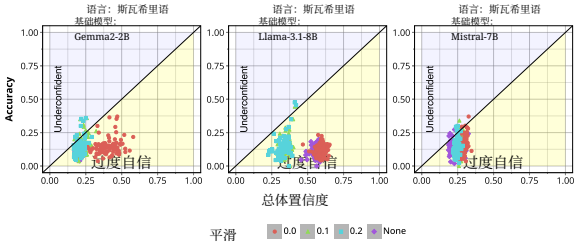


图134：在**MMLU-ProX**数据集上，斯瓦希里语的可靠性图；在**OpenHermes** 数据集上进行指令微调后。

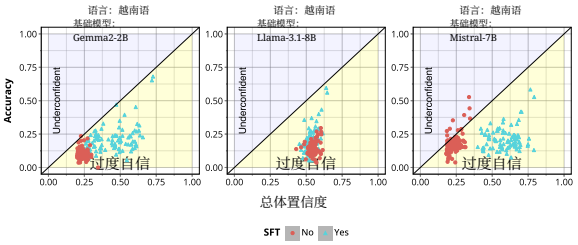


图135：**MMLU-ProX**数据集上越南语的可靠性图。

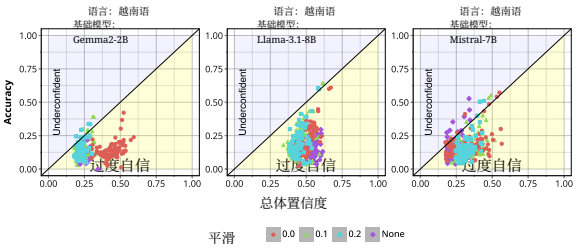


图136：在**Tulu3Mixture** 数据集上进行指令微调后，越南语在**MMLU-ProX**数据集上的可靠性图。

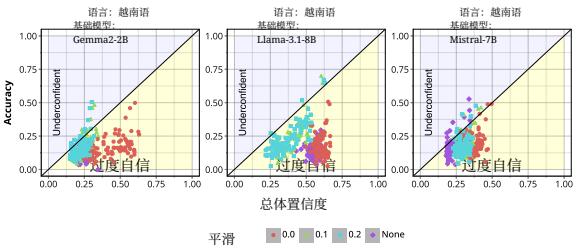


图137：在**OpenHermes** 数据集上进行指令微调后，越南语在**MMLU-ProX**数据集上的可靠性图。

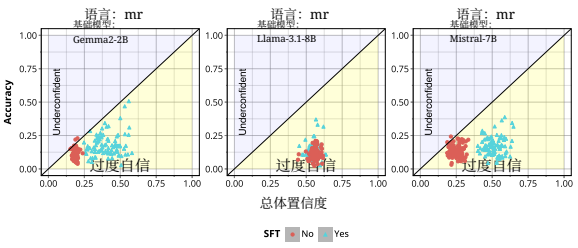


图138：马拉地语在**MMLU-ProX**数据集上的可靠性图。

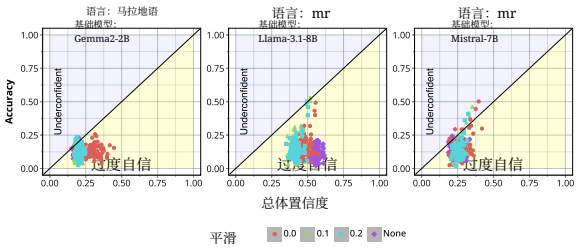


图139：在**MMLU-ProX**数据集上，针对马拉地语的可靠性图；该模型在**Tulu3Mixture** 数据集上进行指令微调后。

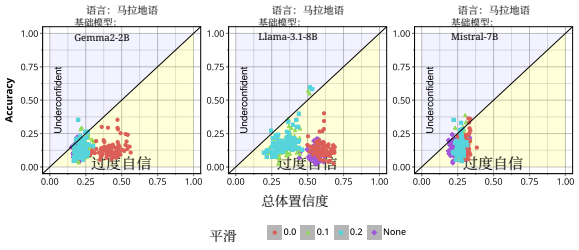


图140：在**MMLU-ProX**数据集上，针对马拉地语的可靠性图；该模型在**OpenHermes** 数据集上进行指令微调后。

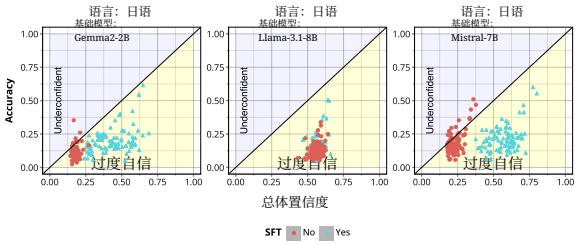


图141：针对日语的**MMLU-ProX**数据集可靠性图。

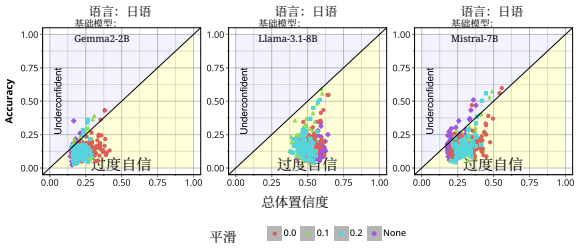


图142：在**MMLU-ProX**数据集上，针对日语的可靠性图；该模型在**Tulu3Mixture** 数据集上进行指令微调后。

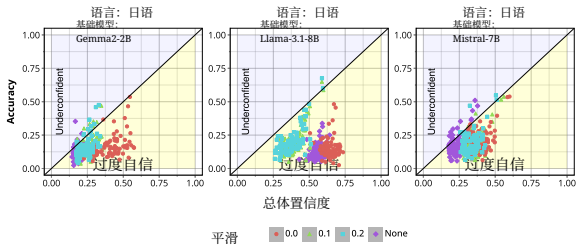


图143：在**OpenHermes** 数据集上进行指令微调后，**MMLU-ProX**数据集的日语可靠性图。

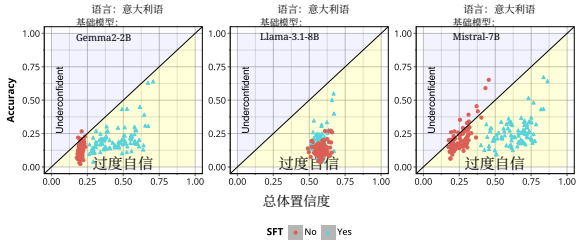


图144：**MMLU-ProX**数据集的意大利语可靠性图。

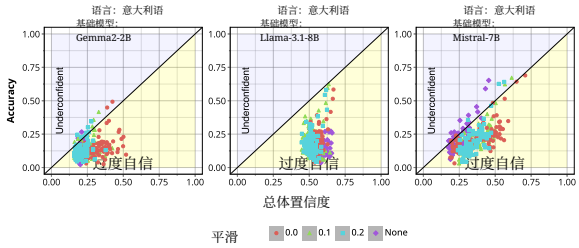


图145：在**Tulu3Mixture** 数据集上进行指令微调后，**MMLU-ProX**数据集的意大利语可靠性图。

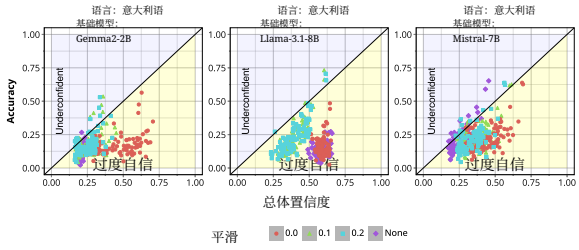


图146：在**OpenHermes** 数据集上进行指令微调后，**MMLU-ProX**数据集的意大利语可靠性图。

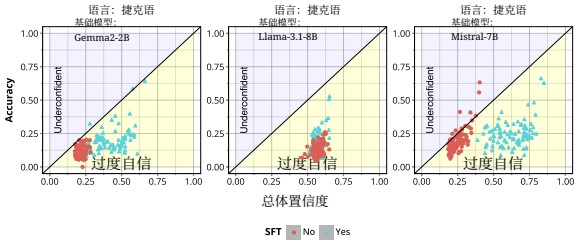


图147：捷克语**MMLU-ProX**数据集的可靠性图。

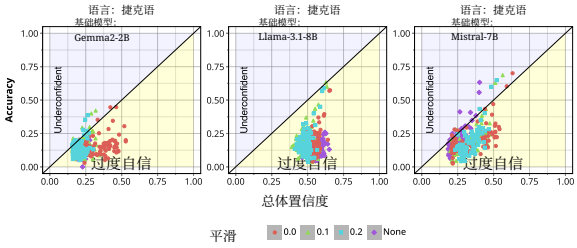


图148：在**Tulu3Mixture** 数据集上进行指令微调后，捷克语**MMLU-ProX**数据集的可靠性图。

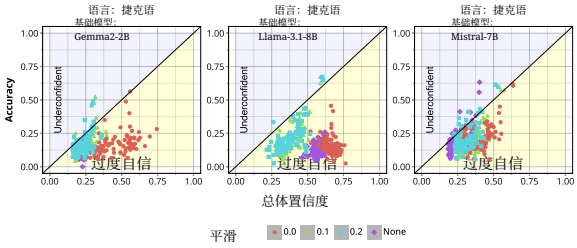


图149：在**OpenHermes** 数据集上进行指令微调后，捷克语**MMLU-ProX**数据集的可靠性图。

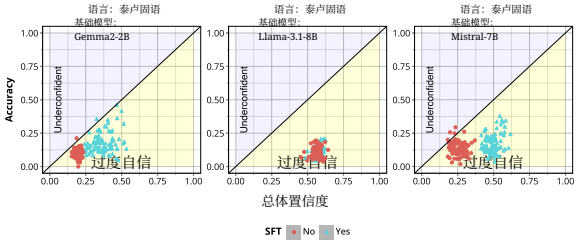


图150：泰卢固语**MMLU-ProX**数据集的可靠性图。

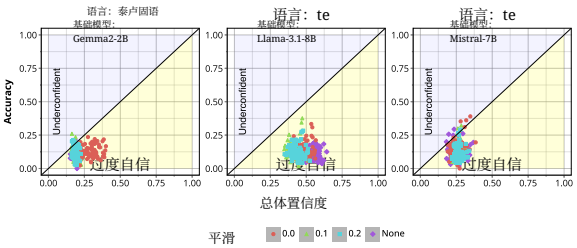


图151：在**Tulu3Mixture** 数据集上进行指令微调后，泰卢固语**MMLU-ProX**数据集的可靠性图。

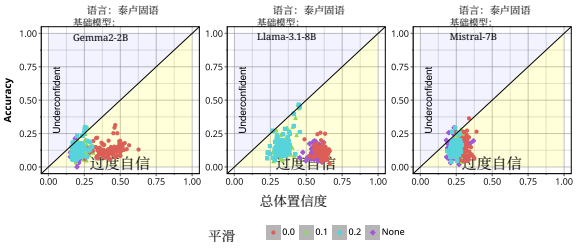


图152：在 **OpenHermes** 数据集上进行指令微调后，泰卢固语在**MMLU-ProX**数据集上的可靠性图。

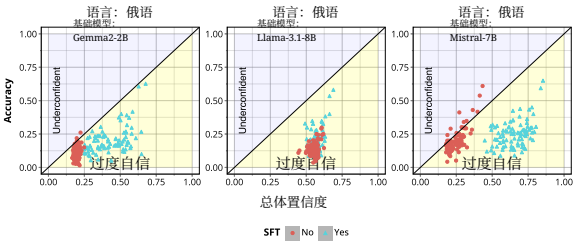


图153：俄语在**MMLU-ProX**数据集上的可靠性图。

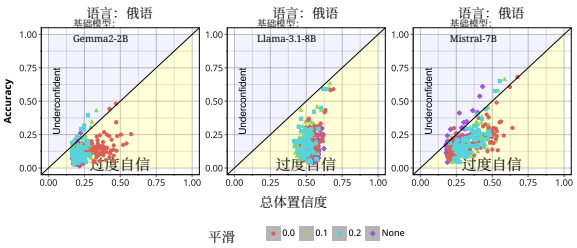


图154：在 **Tulu3Mixture** 数据集上进行指令微调后，俄语在**MMLU-ProX**数据集上的可靠性图。

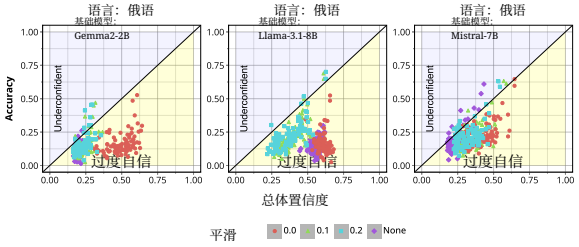


图155：在**OpenHermes** 数据集上进行指令微调后，俄语的**MMLU-ProX**数据集可靠性图。

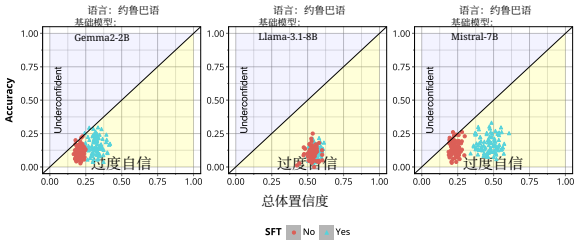


图156：约鲁巴语的**MMLU-ProX**数据集可靠性图。

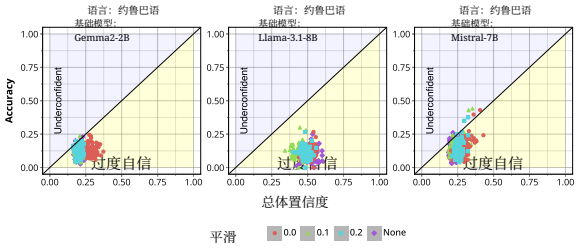


图157：在**Tulu3Mixture** 数据集上进行指令微调后，约鲁巴语的**MMLU-ProX**数据集可靠性图。

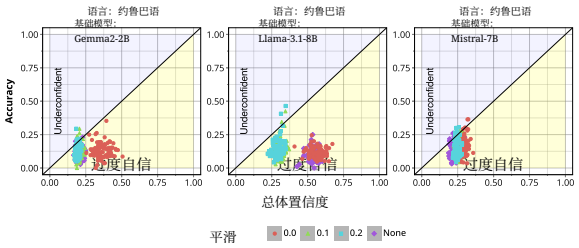


图158：在**OpenHermes** 数据集上进行指令微调后，约鲁巴语的**MMLU-ProX**数据集可靠性图。

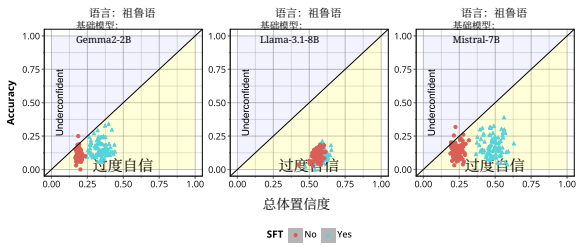


图159：祖鲁语的**MMLU-ProX**数据集可靠性图。

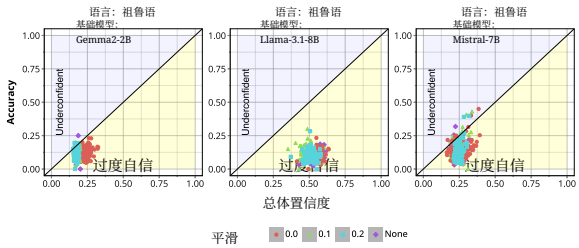


图 160：在**Tulu3Mixture** 数据集上进行指令微调后，祖鲁语的**MMLU-ProX**数据集的可靠性图。

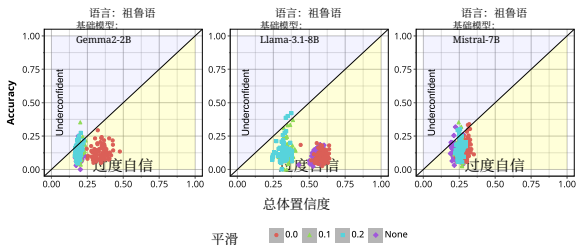


图 161：在**OpenHermes** 数据集上进行指令微调后，祖鲁语的**MMLU-ProX**数据集的可靠性图。

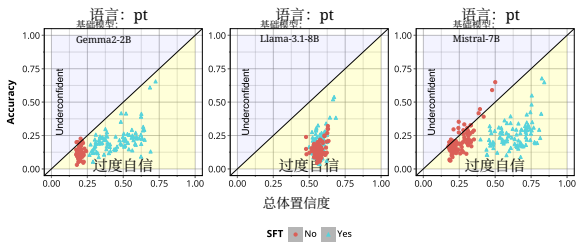


图 162：葡萄牙语的**MMLU-ProX**数据集的可靠性图。

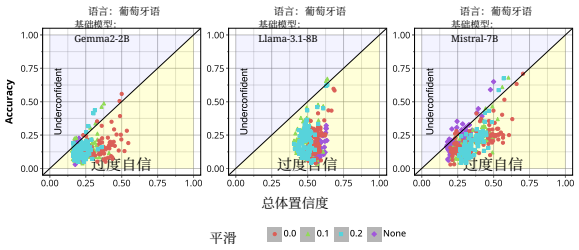


图 163：在**MMLU-ProX**数据集上，针对葡萄牙语的可靠性图；模型在**Tulu3Mixture** 数据集上进行指令微调后。

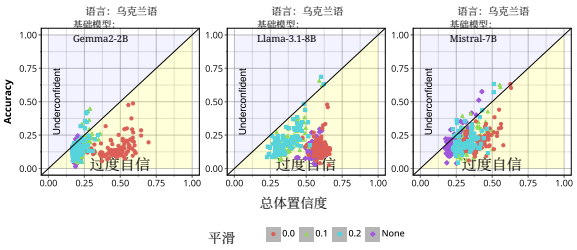


图167：在**OpenHermes** 数据集上进行指令微调后，乌克兰语的**MMLU-ProX**数据集可靠性图。

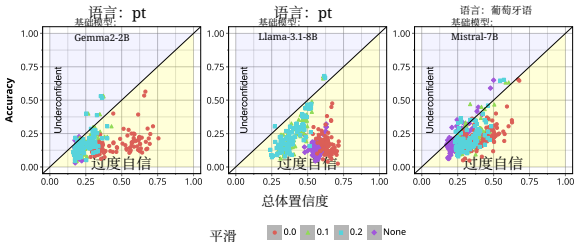


图 164：在**MMLU-ProX**数据集上，针对葡萄牙语的可靠性图；模型在**OpenHermes** 数据集上进行指令微调后。

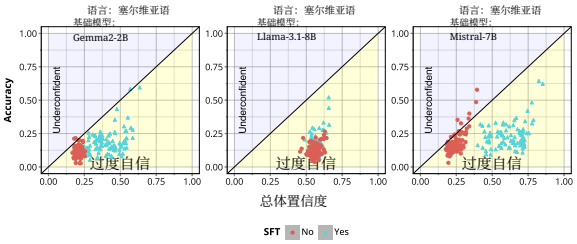


图168：塞尔维亚语的**MMLU-ProX**数据集可靠性图。

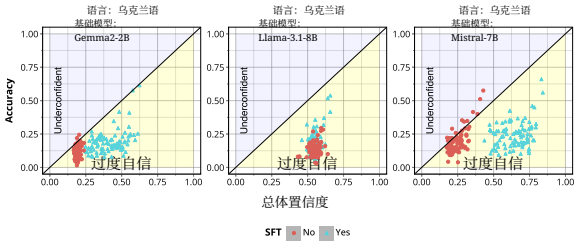


图 165：针对乌克兰语的**MMLU-ProX**数据集可靠性图。

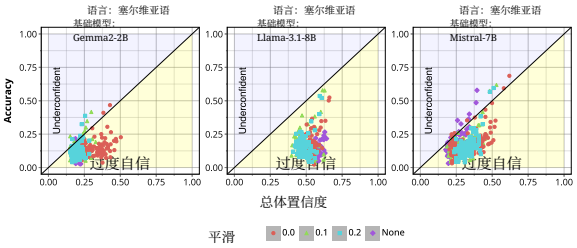


图169：在**Tulu3Mixture** 数据集上进行指令微调后，塞尔维亚语的**MMLU-ProX**数据集可靠性图。

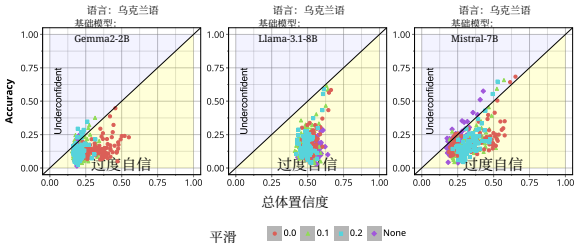


图 166：针对乌克兰语的**MMLU-ProX**数据集可靠性图；模型在**Tulu3Mixture** 数据集上进行指令微调后。

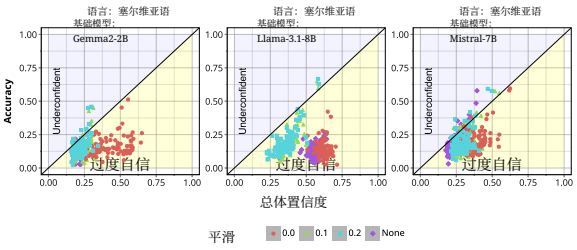


图170：在**OpenHermes** 数据集上进行指令微调后，塞尔维亚语的**MMLU-ProX**数据集可靠性图。

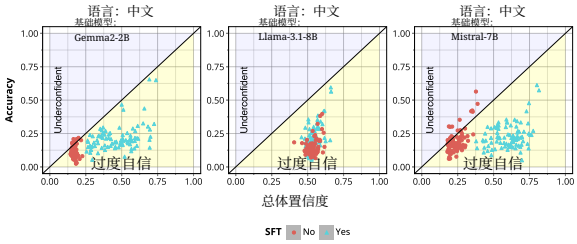
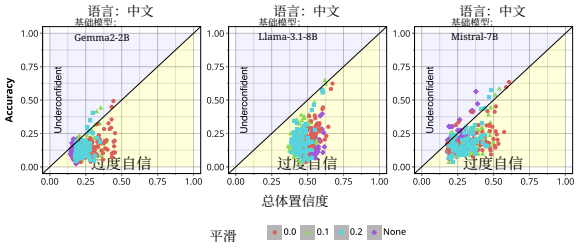
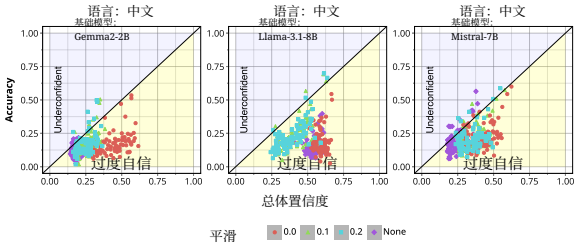


图 171：中文语言的**MMLU-ProX**数据集可靠性图。



在**Tulu3Mixture** 数据集上进行指令微调后，中文语言的**MMLU-ProX**数据集可靠性图。



在**OpenHermes** 数据集上进行指令微调后，中文语言的**MMLU-ProX**数据集可靠性图。

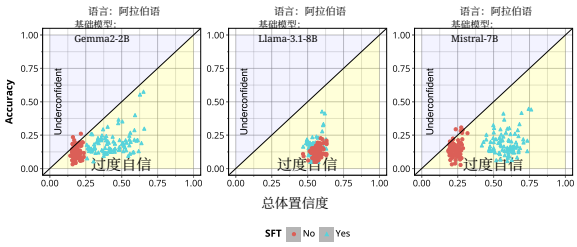


图 174：阿拉伯语的**MMLU-ProX**数据集可靠性图。

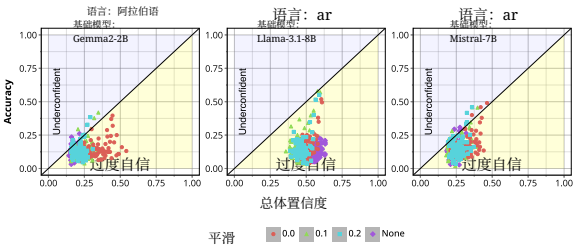


图 175：在**Tulu3Mixture** 数据集上进行指令微调后，阿拉伯语的**MMLU-ProX**数据集可靠性图。

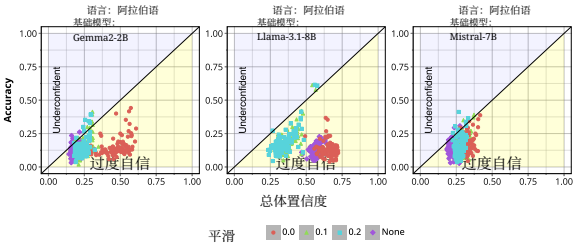


图 176：在**OpenHermes** 数据集上进行指令微调后，**MMLU-ProX**数据集中阿拉伯语的可靠性图。

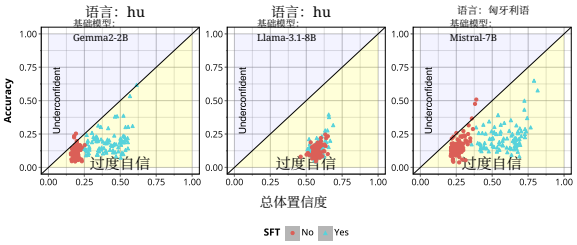


图 177：**MMLU-ProX**数据集中匈牙利语的可靠性图。

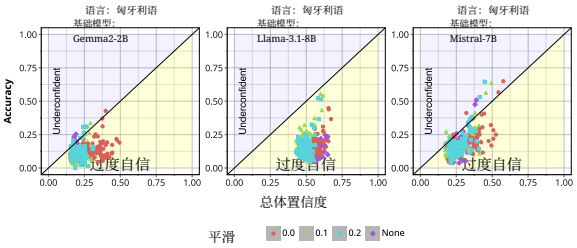


图 178：在**Tulu3Mixture** 数据集上进行指令微调后，**MMLU-ProX**数据集中匈牙利语的可靠性图。

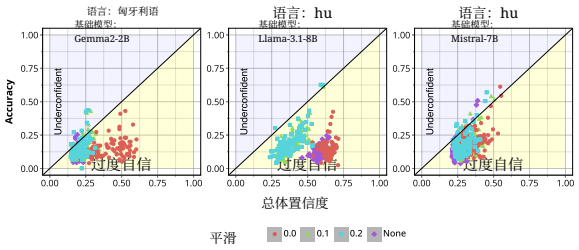


图 179：在**MMLU-ProX**数据集上，匈牙利语经过**OpenHermes**数据集指令微调后的可靠性图。

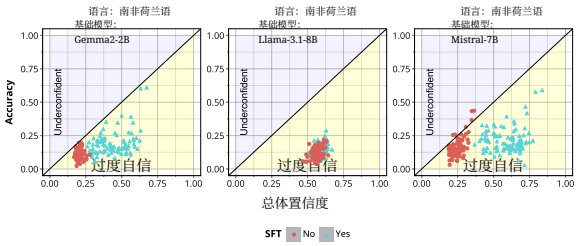


图 180：**MMLU-ProX**数据集上南非荷兰语的可靠性图。

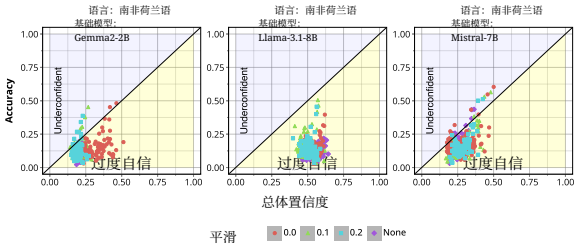


图 181：在**MMLU-ProX**数据集上，南非荷兰语经过**Tulu3Mixture**数据集指令微调后的可靠性图。

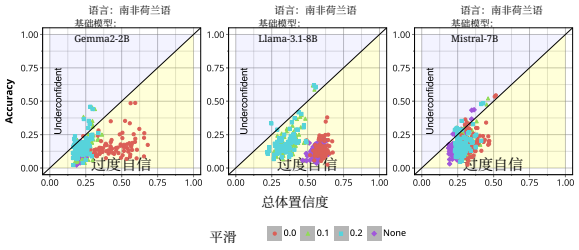


图 182：在**MMLU-ProX**数据集上，南非荷兰语经过**OpenHermes**数据集指令微调后的可靠性图。

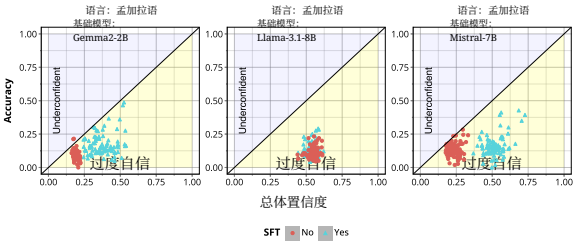


图 183：**MMLU-ProX**数据集上孟加拉语的可靠性图。

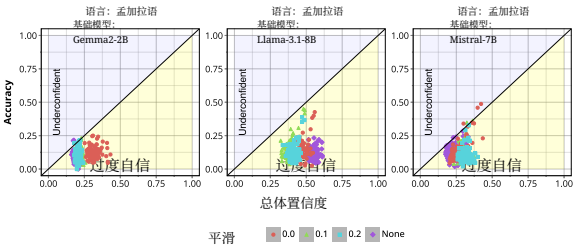


图 184：在**Tulu3Mixture**数据集上进行指令微调后，孟加拉语在**MMLU-ProX**数据集上的可靠性图。

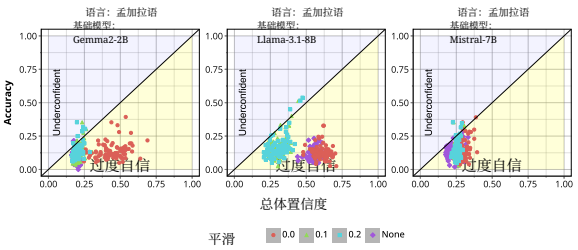


图 185：在**OpenHermes**数据集上进行指令微调后，孟加拉语在**MMLU-ProX**数据集上的可靠性图。

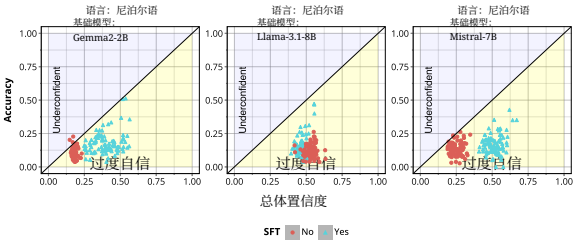


图 186：尼泊尔语在**MMLU-ProX**数据集上的可靠性图。

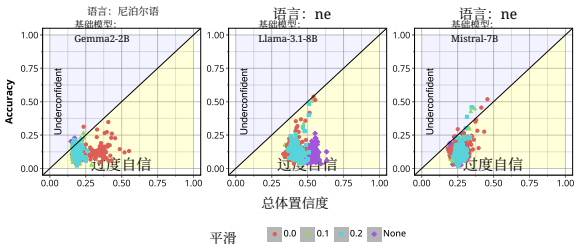


图187：在**MMLU-ProX**数据集上，尼泊尔语在**Tulu3Mixture**数据集上进行指令微调后的可靠性图。

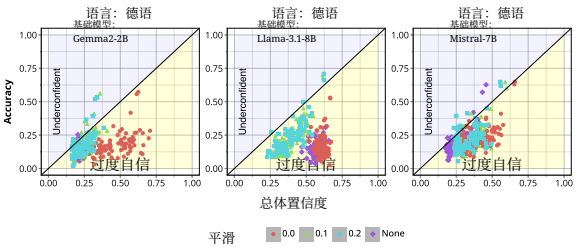


图191：对指令微调后德语的**MMLU-ProX**数据集在**OpenHermes**数据集上进行评估所得的可靠性图。

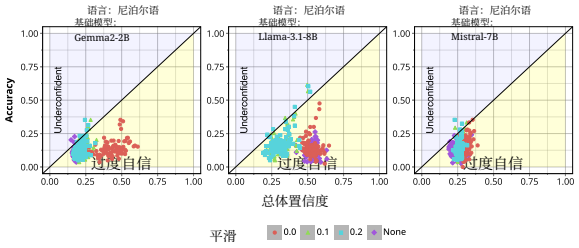


图188：在**MMLU-ProX**数据集上，尼泊尔语在**OpenHermes**数据集上进行指令微调后的可靠性图。

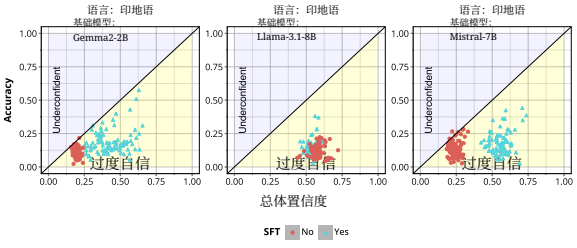


图192：印地语**MMLU-ProX**数据集的可靠性图。

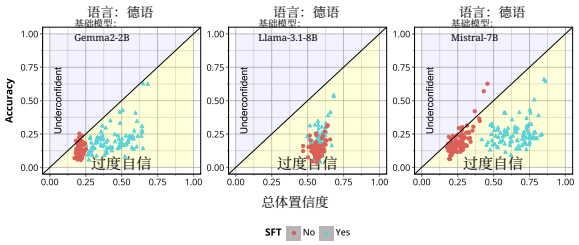


图189：**MMLU-ProX**数据集上德语的可靠性图。

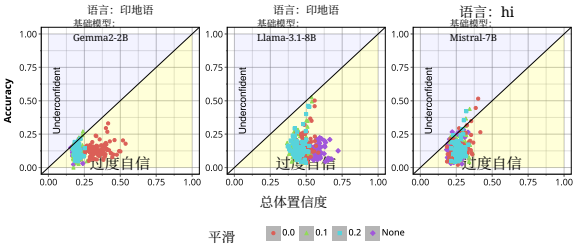


图193：对指令微调后印地语的**MMLU-ProX**数据集在**Tulu3Mixture**数据集上进行评估所得的可靠性图。

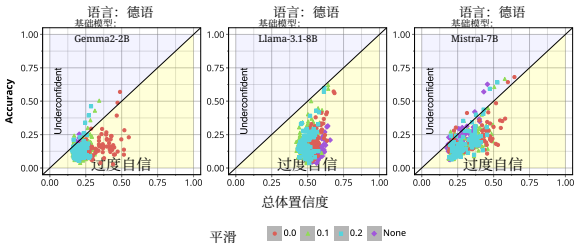


图190：在**MMLU-ProX**数据集上，德语在**Tulu3Mixture**数据集上进行指令微调后的可靠性图。

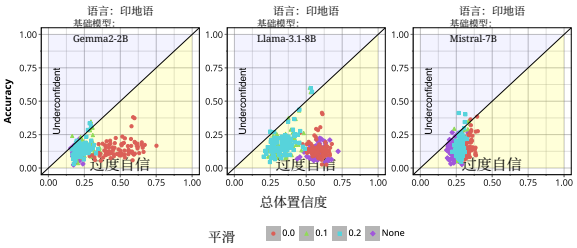


图194：对指令微调后印地语的**MMLU-ProX**数据集在**OpenHermes**数据集上进行评估所得的可靠性图。

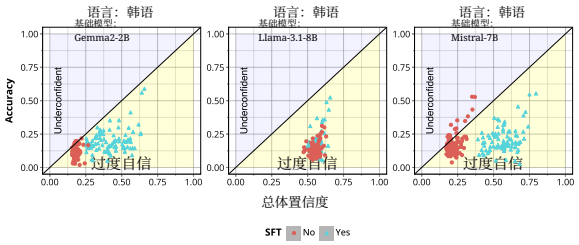


图195：韩语 MMLU-ProX 数据集的可靠性图。

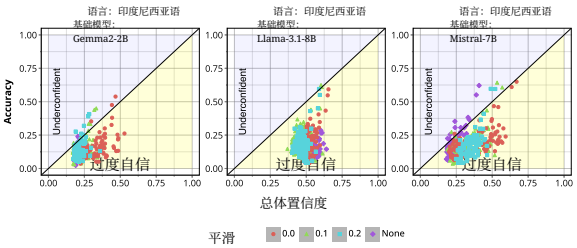


图199：在 Tulu3Mixture 数据集上进行指令微调后，印尼语 MMLU-ProX 数据集的可靠性图。

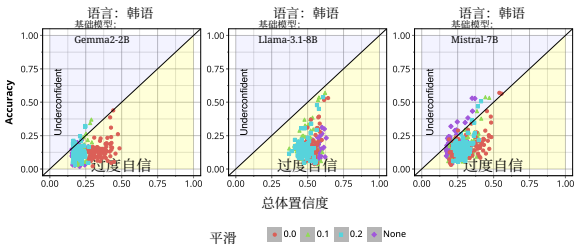


图196：在 Tulu3Mixture 数据集上进行指令微调后，韩语 MMLU-ProX 数据集的可靠性图。

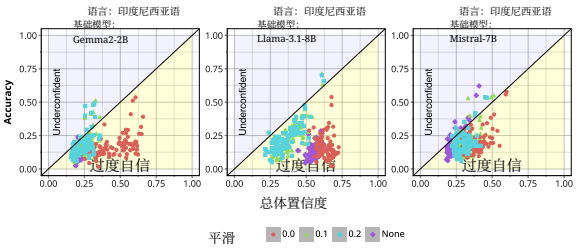


图200：在**OpenHermes** 数据集上进行指令微调后，印度尼西亚语在**MMLU-ProX**数据集上的可靠性图。

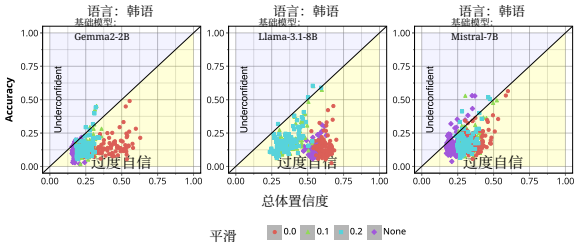


图197：在 OpenHermes 数据集上进行指令微调后，韩语 MMLU-ProX 数据集的可靠性图。

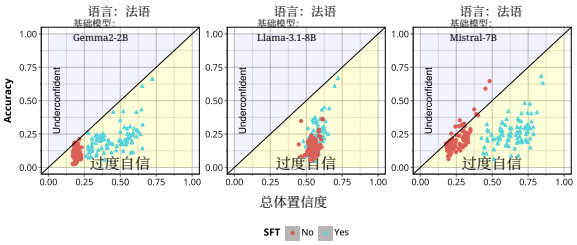


图201：法语在**MMLU-ProX**数据集上的可靠性图。

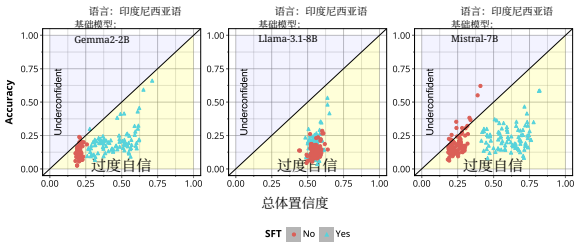


图198：印尼语 MMLU-ProX 数据集的可靠性图。

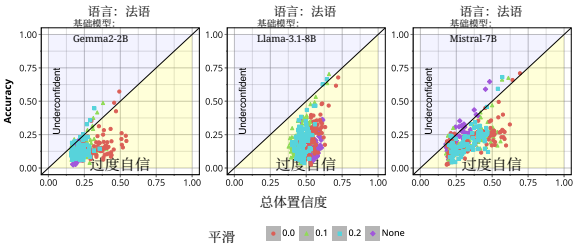


图202：在**Tulu3Mixture** 数据集上进行指令微调后，法语在**MMLU-ProX**数据集上的可靠性图。

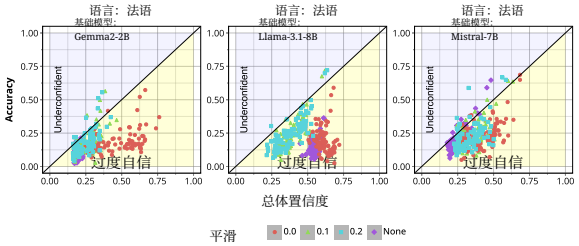


图 203：在**MMLU-ProX**数据集上，法语在使用**OpenHermes**数据集进行指令微调后的可靠性图。

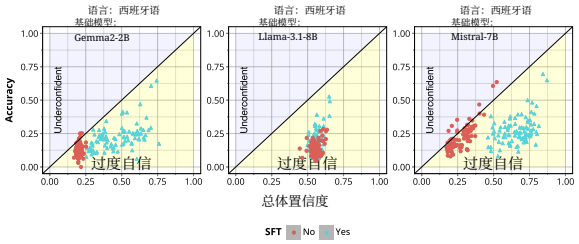


图 204：**MMLU-ProX**数据集的西班牙语可靠性图。

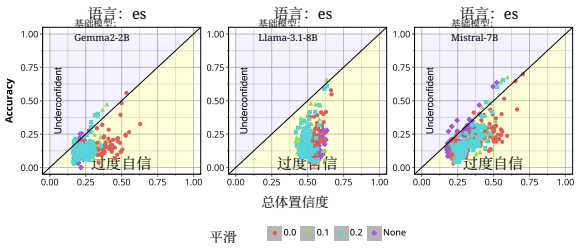


图 205：**MMLU-ProX**数据集的西班牙语在使用**Tulu3Mixture**数据集进行指令微调后的可靠性图。

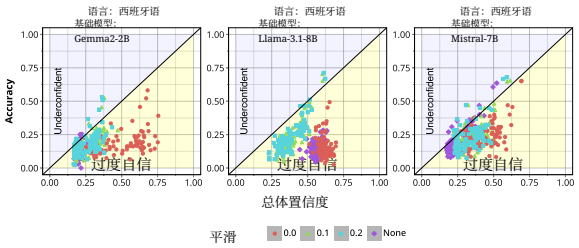


图 206：**MMLU-ProX**数据集的西班牙语在使用**OpenHermes**数据集进行指令微调后的可靠性图。

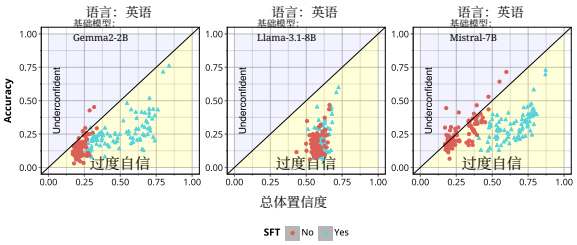


图 207：**MMLU-ProX**数据集的英语可靠性图。

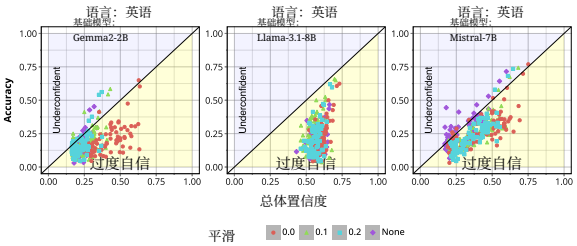


图208：在**Tulu3Mixture**数据集上进行指令微调后，英语语言的**MMLU-ProX**数据集的可靠性图。

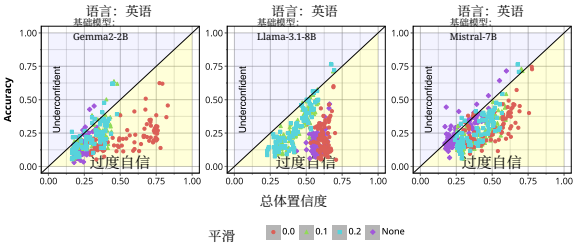


图209：在**OpenHermes**数据集上进行指令微调后，英语语言的**MMLU-ProX**数据集的可靠性图。

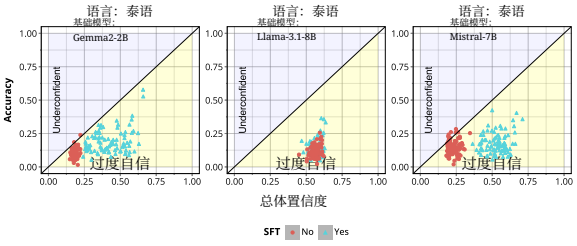


图210：泰语的**MMLU-ProX**数据集的可靠性图。

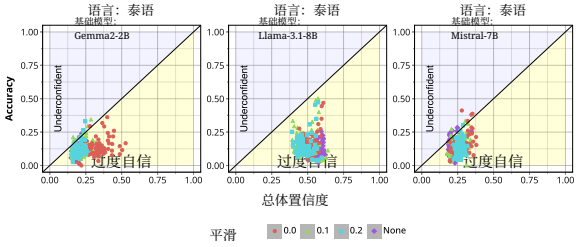


图 211：**MMLU-ProX**数据集在泰语上的可靠性图，该语言经过**Tulu3Mixture** 数据集的指令微调。

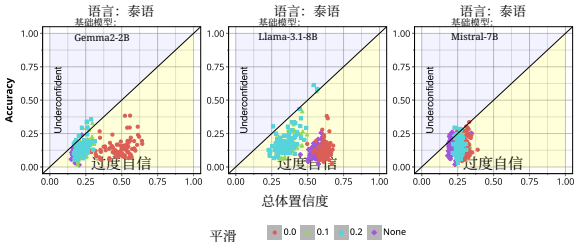


图 212：**MMLU-ProX**数据集在泰语上的可靠性图，该语言经过**OpenHermes** 数据集的指令微调。

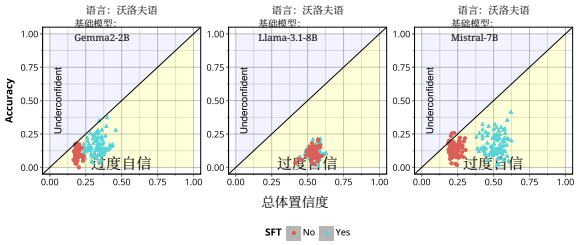


图 213：**MMLU-ProX**数据集在沃洛夫语上的可靠性图。

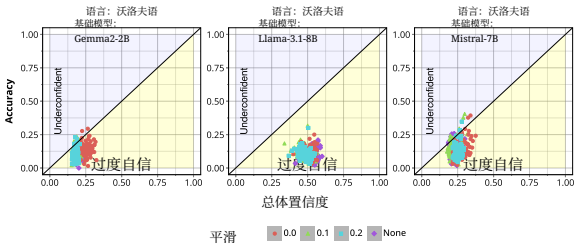


图 214：**MMLU-ProX**数据集在沃洛夫语上的可靠性图，该语言经过**Tulu3Mixture** 数据集的指令微调。

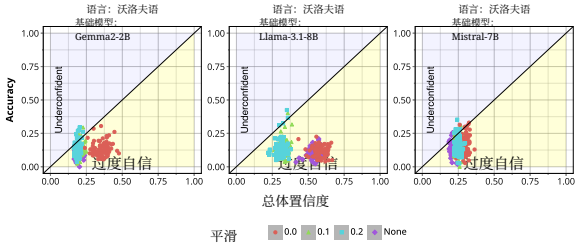


图 215：**MMLU-ProX**数据集在沃洛夫语上的可靠性图，该语言经过**OpenHermes** 数据集的指令微调。