INVESTIGATING UNCERTAINTY CALIBRATION OF ALIGNED LANGUAGE MODELS UNDER THE MULTIPLE-CHOICE SETTING

Guande He[1], **Peng Cui**[1], **Jianfei Chen**[1*], **Wenbo Hu**[2] **& Jun Zhu**[1*]
[1]Dept. of Comp. Sci. & Tech., Institute for AI, Tsinghua-Bosch Joint Center for ML,
BNRist Center, State Key Lab for Intell. Tech. & Sys., Tsinghua University, Beijing, China
[2] Hefei University of Technology
guande.he17@outlook.com, xpeng.cui@gmail.com, wenbohu@hfut.edu.cn
{jianfeic, dcszj}@tsinghua.edu.cn

ABSTRACT

Despite the significant progress made in practical applications of aligned language models (LMs), they tend to be overconfident in output answers compared to the corresponding pre-trained LMs. In this work, we systematically evaluate the impact of the alignment process on logit-based uncertainty calibration of LMs under the multiple-choice setting. We first conduct a thoughtful empirical study on how aligned LMs differ in calibration from their pre-trained counterparts. Experimental results reveal that there are two distinct uncertainties in LMs under the multiple-choice setting, which are responsible for the answer decision and the format preference of the LMs, respectively. Then, we investigate the role of these two types of uncertainty on aligned LM's calibration through fine-tuning in synthetic alignment schemes and conclude that one reason for aligned LMs' overconfidence is the alteration of their answer uncertainty. We hope our findings could provide insights into the design of more reliable alignment processes for LMs.

1 INTRODUCTION

Aligning pre-trained language models (LMs) with human feedback, e.g., ChatGPT (Ouyang et al., 2022), LLaMA (Touvron et al., 2023b), and Vicuna (Chiang et al., 2023), has achieved remarkable success in a broad spectrum of real-world application scenarios. However, recent works show that the aligned LMs tend to be more overconfident in their answers compared to the pre-trained LMs and result in poor calibration (Kadavath et al., 2022; OpenAI, 2023; Tian et al., 2023; Zhao et al., 2023), which makes it challenging to distinguish truthful and hallucinated answers of the models. As a result, this issue hinders the deployment of aligned LMs in safety-critical domains.

Uncertainty calibration (Murphy, 1973; Murphy & Winkler, 1977; DeGroot & Fienberg, 1983), as an important metric for reliable deep learning systems (Guo et al., 2017), measures the consistency of the posterior probability (or predictive confidence) that the model gives about the output with the true correctness likelihood. For example, when a well-calibrated model gives predictions each with 0.8 confidence, then 80% of predictions should be correct, i.e., *the model knows what it knows*. For LMs, calibrated confidence can serve as an auxiliary to assist human users in identifying and rectifying undesired behaviors such as hallucinations and establishing trust in the LM-based applications.

One plausible way to evaluate LMs' calibration is quantifying their confidence through the logit-based likelihood over the output tokens. With the established background that advanced large pre-trained LMs are well-calibrated while aligned LMs are poorly calibrated due to overconfidence in logit-based evaluation (Kadavath et al., 2022; OpenAI, 2023), previous works mainly focus on alternative approaches to elicit LMs' confidence (Kuhn et al., 2023; Tian et al., 2023; Lin et al., 2023) or the correlations between calibrations and other metrics (Liang et al., 2023). Nevertheless, how the alignment process deteriorates LMs' calibration remains unexplored.

*Corresponding authors.

In light of this, we first conduct a thoughtful empirical study to examine how pre-trained and aligned LMs differ in calibration under the multiple-choice setting, and our main findings are: 1). in-context learning plays an important role in pre-trained LMs' calibration by demonstrating the response's format while keeping LM's confidence in their answer; 2). aligned LMs are inherently overconfident with altered predictive distribution on both answer decision and response format under the multiple-choice setting.

To further investigate how aligned LMs' calibration distorts, we formulate two types of uncertainty related to LMs' calibration, namely answer uncertainty and format uncertainty, which correspond to making decisions among candidates and formatting LMs' responses, respectively. By analyzing the impact of alignment on the two types of uncertainty with synthetic fine-tuning schemes, we conclude that one reason for the miscalibration of aligned LMs is that current alignment processes cannot distinguish these two types of uncertainty, and the shifted answer uncertainty leads to overconfidence.

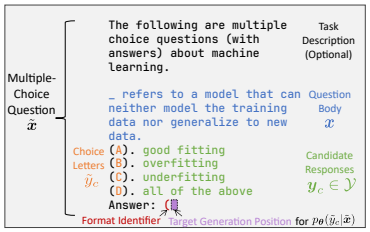


Figure 1: An example MCQ prompt for MMLU (0-shot).

2 How Pre-trained and Aligned LMs Differ in Calibration

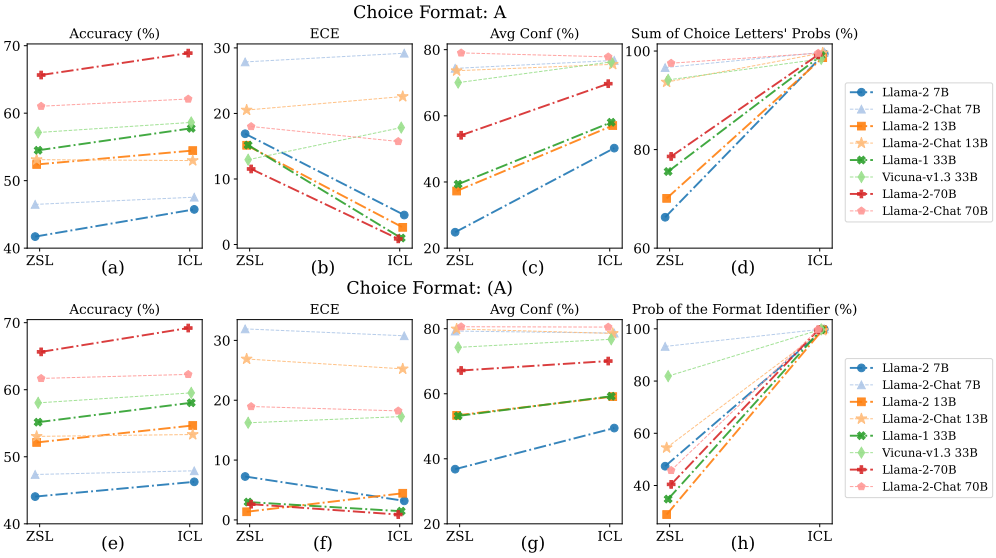


Figure 2: The accuracy, ECE, and average predictive confidence of ZSL and ICL with choice format "A' and "(A)" on MMLU. We also report the average confidence of the predicted choice, sum of the choice letter's probabilities for choice format "A" and the probability of the format identifier for choice format "(A)".

Task Definition Let us consider a class of tasks whose sample consists of an instruction (or question) $\boldsymbol{x}$ and a set of candidate responses $\boldsymbol{y}_c \in \mathcal{Y}$. To estimate $p(\boldsymbol{y}_c|\boldsymbol{x})$ through an auto-regressive LM $p_{\boldsymbol{\theta}}(x_t|\boldsymbol{x}_{<t})$, we can format the sample $(\boldsymbol{x}, \mathcal{Y})$ into a multiple-choice question (MCQ). More precisely, we create the MCQ $\tilde{\boldsymbol{x}}$ that concatenates the task description, question body, and all candidate responses using a mapping $\tilde{\boldsymbol{x}} = f(\boldsymbol{x}, \mathcal{Y})$ and assign a choice letter $\tilde{y}_c$ for each candidate $\boldsymbol{y}_c$, as illustrated in Fig. 1. To perform uncertainty calibration, we can estimate $p_{\boldsymbol{\theta}}(\boldsymbol{y}_c|\boldsymbol{x})$ using the probability of the choice letter $p_{\boldsymbol{\theta}}(\tilde{y}_c|\tilde{\boldsymbol{x}})$ calculated from the token logits given by the LM at the target generation position. For aligned LMs, we consider two common paradigms – supervised fine-tuning (SFT) and learning from pairwise feedback (LPF).

Evaluation Setup In this work, we adopt expected calibration error (ECE) with 10 equal-sized bins to measure LM's calibration. We evaluate the LM with both zero-shot learning (**ZSL**) and five-shot in-context learning (**ICL**). All data are adapted to MCQs, as shown in Fig. 1. We employ two choice formats: "A" and "(A)", where we refer to "(" as a format identifier that serves as a hint

for the LM, guiding it to answer MCQs directly with one of the choice letters. The detailed setup is available in Appendix B.1.

First, we examine how pre-trained and aligned LMs differ in their ECE, confidence, and other relevant probabilities in ZSL and ICL settings with a typical MCQ task MMLU (Hendrycks et al., 2021). (Comprehensive results are available in Appendix C)

For pre-trained LMs, as shown in Fig. 2b and Fig. 2c we observe that they are underconfident in the ZSL setting with the choice format "`A`", yielding high ECE. Meanwhile, Fig. 2d shows that the sum of probabilities for all choice letters is widely low, revealing that pre-trained LMs tend not to start responses with choice letters for MCQs. We refer to this phenomenon as the *format preference*, leading to low predictive confidence of pre-trained LMs in the ZSL setting. Upon examining the ICL setting, as shown in Fig. 2d, pre-trained LMs can alter their format preference based on in-context examples. Notably, the result predictive confidence produced by ICL is well-calibrated out-of-the-box, as depicted by the clear decrease of ECE between ZSL and ICL in Fig. 2b.

For the choice format "`(A)`", the effect of ICL in altering the LMs' format preference is reflected in the probability of the format identifier "`(`". As shown in Fig. 2d & 2h, the probability of the format identifier under "`(A)`" and the sum of choice letters' probabilities under "`A`" exhibit a similar trend from ZSL to ICL. Such phenomenon suggests that the format identifier "separates" a part of LMs' format preference from the choice letters. Consequentially, the probabilities of the LMs over the choice letters could more accurately reflect their confidence in candidate answers. As compared in Fig. 2b & 2f and Fig. 2c & 2g, the difference in the ECE and confidence between ZSL and ICL is evidently narrowed under the choice format "`(A)`". Along with the results for choice format "`A`" in the previous paragraph, we conclude that *in-context learning calibrates pre-trained LMs by both changing LMs' format preference and maintaining LMs' confidence in their answer prediction.*

In contrast to the pre-trained LMs, the **aligned LMs** are overconfident in both ZSL and ICL settings across all choice formats. Additionally, they prefer to directly output the choice letter to answer MCQs, i.e., aligned LMs have different format preferences compared to pre-trained LMs. However, unlike pre-trained LMs, ICL can only change aligned LMs' format preference while having a marginal effect on their calibration in this setting which suggests that the alignment process destroys the well-calibrated predictive distributions of pre-trained LMs and cannot be restored by ICL. Based on these findings, we also propose a sample-efficient post-hoc calibration technique in Appendix E.

3 How Alignment Process Impacts LMs' Uncertainty Calibration

3.1 The Two Uncertainties of LMs in Multiple-Choice Questions

Inspired by the empirical observations in §2, we propose to decompose LMs' uncertainty in MCQs into two parts: 1). Answer uncertainty: the uncertainty about choosing an answer among all candidates; 2). Format uncertainty: the uncertainty about the response format for answering a question in general, e.g., start the response with a choice decision or other texts like "Let's think step by step" for answering MCQs. Formally, let F be a discrete random variable representing the LM's preference for structuring its response $\boldsymbol{y}$ among all possible formats $\mathcal{F}$ given an instruction $\boldsymbol{x}$. Suppose each instruction-response pair $(\boldsymbol{x}, \boldsymbol{y})$ corresponds to an *unique* format F, i.e., either $p(F|\boldsymbol{x}, \boldsymbol{y}) = 1$ or $p(F|\boldsymbol{x}, \boldsymbol{y}) = 0$. Then, similar to the decomposition of the model uncertainty and data uncertainty (Gal, 2016; Kendall & Gal, 2017; Malinin & Gales, 2018), we could decompose the LM's predictive probability $p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$ as (detail in Appendix D):

$$p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}) = \underbrace{p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}, F)}_{\text{Answer}} \underbrace{p_{\boldsymbol{\theta}}(F|\boldsymbol{x})}_{\text{Format}} \tag{1}$$

Here the format uncertainty $p_{\boldsymbol{\theta}}(F|\boldsymbol{x})$ for a question $\boldsymbol{x}$ is induced by: $p_{\boldsymbol{\theta}}(F|\boldsymbol{x}) = \sum_{y \in \mathcal{Y}_F} p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$, where $\mathcal{Y}_F = \{\boldsymbol{y} | p_{\boldsymbol{\theta}}(F|\boldsymbol{x}, \boldsymbol{y}) = 1\}$, i.e., all responses $\boldsymbol{y}$ that correspond to the same format F. We have the following two hypotheses under this formulation based on empirical observations:

Hypothesis 3.1. *For MCQs, the answer uncertainty $p^{\mathrm{PT}}_{\boldsymbol{\theta}}(\tilde{y}_c|\tilde{\boldsymbol{x}}, F_{\mathrm{MC}})$ of pre-trained LMs under the format F_{MC} is well-calibrated, where F_{MC} denotes the format that directly makes a choice decision.*

Hypothesis 3.2. *The altered answer uncertainty during alignment leads to the miscalibration of aligned LMs.*

We first discuss empirical evidence supporting Hypothesis 3.1. In practice, for answering general MCQs such as MMLU, once the in-context examples provide enough signals for F_{MC} via the format of S_K, the LM's prediction under ICL would be largely independent of the content of S_K (Min et al., 2022), which suggests that $p_{\boldsymbol{\theta}}^{\mathrm{PT}}(\tilde{y}_c|\tilde{\boldsymbol{x}}, S_K)$ would be a good approximator of the LM's answer uncertainty $p_{\boldsymbol{\theta}}^{\mathrm{PT}}(\tilde{y}_c|\tilde{\boldsymbol{x}}, F_{\mathrm{MC}})$ when S_K conforms to the format F_{MC}. We also include an extended empirical study to demonstrate this point in Appendix C.5. Meanwhile, as shown in §2 and Fig. 2, pre-trained LMs' predictive confidence $p_{\boldsymbol{\theta}}^{\mathrm{PT}}(\tilde{y}_c|\tilde{\boldsymbol{x}}, S_K)$ under ICL or given format identifier is generally calibrated, indicating pre-trained LMs' answer uncertainty $p_{\boldsymbol{\theta}}^{\mathrm{PT}}(\tilde{y}_c|\tilde{\boldsymbol{x}}, F_{\mathrm{MC}})$ is calibrated.

3.2 Common Alignment Processes Conflate the Two Uncertainties of LMs

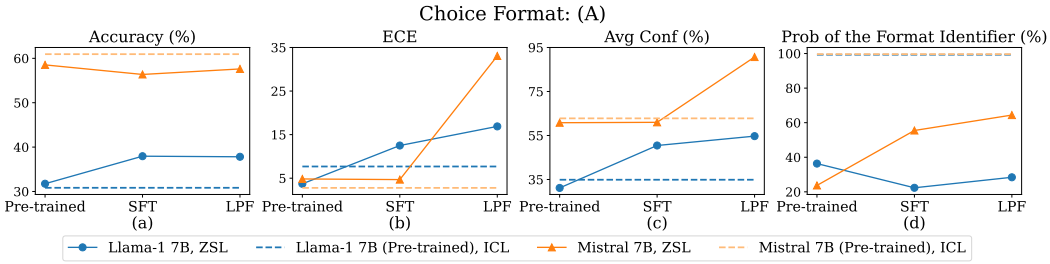


Figure 3: ZSL results of different alignment stages on MMLU validation set.

In the subsequent analysis, our objective is to demonstrate evidence that supports Hypothesis 3.2. In this section, we examine the effect of common alignment stages on LMs' two uncertainties in MCQs. Specifically, we choose two sets of full-process (i.e., both SFT and LPF) alignment LMs, Alpaca-Farm (Dubois et al., 2023), which aligns pre-trained Llama-1 7B with SFT and PPO (Schulman et al., 2017), and Zephyr (Tunstall et al., 2023), which aligns pre-trained Mistral 7B (Jiang et al., 2023) with SFT and DPO (Rafailov et al., 2023). We track the same four metrics in §2 at different stages of alignment under choice format "`(A)`". We perform the evaluation using the officially released checkpoints of aligned LMs in the HuggingFace[1]. For Alpaca-Farm (Dubois et al., 2023), we adopt the `alpaca-farm-sft10k` and `alpaca-farm-ppo-human` for SFT and PPO version of Llama-1 7B. We use `Mistral-7B-v0.1`, `mistral-7b-sft-beta`, and `zephyr-7b-beta` for the pre-trained, SFT, and DPO version of Mistral 7B (Jiang et al., 2023; Tunstall et al., 2023). All LMs are aligned with free-form QA datasets labeled with preference, where the Alpaca-Farm generates the data using self-instruct (Wang et al., 2023), and the Zephyr pipeline adopts the UltraChat (Ding et al., 2023) dataset.

As shown in Fig. 3, aligning pre-trained LMs on human-preference dialog data with SFT and LPF impacts both LM's answer uncertainty (indicated by confidence) and format uncertainty (est. with the probability of the format identifier) in MCQs. For aligning Llama-1 with the Alpaca-Farm pipeline, the LM's confidence keeps increasing during the whole alignment process, whereas for aligning from Mistral to Zephyr, the LM's confidence remains unchanged during SFT and sharply increases when performing DPO. Meanwhile, the format uncertainty of the LMs has varying degrees of change at all stages of alignment. Interestingly, the SFT version of Mistral is the only aligned LM that preserves the calibration of the pre-trained LMs on MCQs. These results highlight that the two uncertainties of LMs in MCQs undergo uncontrolled changes during alignment, i.e., *common alignment processes conflate the two uncertainties of LMs*. Crucially, the observations in Fig. 2 and Fig. 3 together show that, although the alignment processes do not involve much domain knowledge required for MMLU, aligning pre-trained LMs with preference data results in a monotonic increasing pattern in LM's confidence in this task regardless of the changes in accuracy or format uncertainty, which suggests current alignment processes will likely lead to overall overconfidence in MCQs.

3.3 How Uncertainty Conflation During Alignment Impacts LMs' Calibration

Intuitively, the mixed effect of alignment on LMs' uncertainties in MCQs may stem from the choice of examples in SFT and the way of reward modeling in LPF, where the LMs are optimized towards both correct answers and human-preferred formats simultaneously. To better under-

[1] `https://huggingface.co/`

stand how aligned LMs become miscalibrated, we design a series of synthetic alignment schemes where the pre-trained LM performs controlled optimization of answer uncertainty and format uncertainty on a synthetic MCQ task. In specific, we experiment with three variants for SFT and DPO (Rafailov et al., 2023), respectively: 1). **SFT-Format**: calculate loss on the format identifier, i.e., optimize $p_{\boldsymbol{\theta}}(F_{\text{MC}}|\tilde{\boldsymbol{x}})$ only; 2). **SFT-Choice**: calculate loss on the choice letters, i.e., optimize $p_{\boldsymbol{\theta}}(\tilde{y}_c|\tilde{\boldsymbol{x}}, F_{\text{MC}})$ only; 3). **SFT-Mixed**: calculate loss on both kinds of tokens, i.e. optimize $p_{\boldsymbol{\theta}}(F_{\text{MC}}|\tilde{\boldsymbol{x}})$ & $p_{\boldsymbol{\theta}}(\tilde{y}_c|\tilde{\boldsymbol{x}}, F_{\text{MC}})$; 4). **DPO-Format**: the preference pair $(\boldsymbol{y}_w, \boldsymbol{y}_l)$ has same choice but different format; 5). **DPO-Choice**: the preference pair $(\boldsymbol{y}_w, \boldsymbol{y}_l)$ has same format but different choice; 6). **DPO-Mixed**: the preference pair $(\boldsymbol{y}_w, \boldsymbol{y}_l)$ has different choices and formats.

For the synthetic MCQ task, we adopt the one used in Lieberum et al. (2023), where the LM must pick one choice corresponding to a particular English word specified in the question from four candidates, as shown in Fig. 4. We set the preferred format in DPO to directly output the choice letter, e.g., "`(A)`", and set the undesired format to "`It's (A)`". We choose Llama-1 7B as our base pre-trained LM since we observe that it only achieves 70% zero-shot accuracy on this synthetic task, ensuring the alignment process is non-degenerate. Since the synthetic MCQ data has limited scale and diversity, we opt to use LoRA (Hu et al., 2022) to avoid overfitting. The detailed training setup is presented in Appendix B.2.

```
Which of the following options
corresponds to "Foo"?
(A). "Foo"
(B). "Bar"
(C). "Baz"
(D). "Qux"

Answer:
```

Figure 4: An example of the synthetic MCQ.

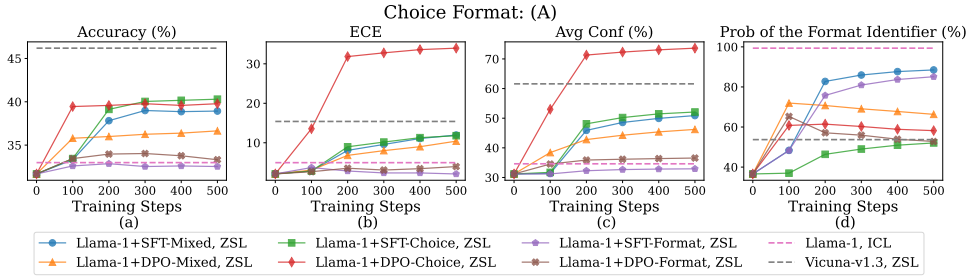


Figure 5: Results of all synthetic alignment schemes for Llama-1 7B on MMLU validation set.

Fig. 5 illustrates the result on MMLU after aligning the pre-trained LM with the synthetic schemes. Both SFT-Format and DPO-Format have close accuracy, ECE, and confidence to the pre-trained LM in the ICL setting while increasing the model's likelihood on the format identifier. In comparison, the Choice and Mixed schemes exhibit overconfidence in the evaluation task, among which the DPO-Choice scheme causes the most severe overconfident tendency on MMLU. We also present the accuracy on the synthetic task of these schemes in Appendix C.6.

These results provide empirical evidence for validating Hypothesis 3.2, i.e., *the typically observed overconfidence and miscalibration of aligned LMs in MCQs primarily arise from alterations in answer uncertainty during the alignment process.* The synthetic alignment schemes show that even updating the answer uncertainty of LMs on such a simple task can lead to overconfidence in other multiple-choice tasks. Arguably, updating the answer uncertainty of LMs during alignment could also boost the accuracy of the overall performance for MCQs, as shown in Fig. 5a. However, such benefit becomes marginal for strong pre-trained LMs, such as Mistral (in Fig. 3) and Llama-2 70B (in Fig. 2). Thus, in order to preserve the calibration of the tasks not covered during the alignment, we need to design the alignment process elaborately, and our analysis suggests that focusing on the format uncertainty might be a potential direction.

4 Conclusions

In this work, we complement the line of study on logit-based uncertainty calibration of LMs by delving into the differences in the calibration of pre-trained and aligned LMs in ZSL and ICL settings with MCQs. Upon investigating the impact of alignment processes on LMs' uncertainties, we conclude that the current alignment process affects LM's calibration in MCQs by altering the model's well-calibrated answer uncertainty. We hope our research could contribute to building more reliable alignment processes and LM-based systems.

REFERENCES

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2022.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, 2019.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality, March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations. *arXiv preprint arXiv:2305.14233*, 2023.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback, 2023.

Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations*, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022.

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

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language Models (Mostly) Know What They Know, 2022.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 2017.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *The Eleventh International Conference on Learning Representations*, 2023.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*, 2023.

Tom Lieberum, Matthew Rahtz, János Kramár, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. Does Circuit Analysis Interpretability Scale? Evidence from Multiple Choice Capabilities in Chinchilla. *arXiv preprint arXiv:2307.09458*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models, 2023.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. LogiQA: A Challenge Dataset for Machine Reading Comprehension with Logical Reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 2020.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 2018.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *EMNLP*, 2018.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

Allan H Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600, 1973.

Allan H Murphy and Robert L Winkler. Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 26(1): 41–47, 1977.

OpenAI. GPT-4 Technical Report, 2023.

Long Ouyang, Jeff 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 Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Peter Salamon, David Salamon, V. Adrian Cantu, Michelle An, Tyler Perry, Robert A. Edwards, and Anca M. Segall. Estimating Classification Confidence Using Kernel Densities, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms, 2017.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023b.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct Distillation of LM Alignment, 2023.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An Explanation of In-context Learning as Implicit Bayesian Inference. In *International Conference on Learning Representations*, 2022.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Jize Zhang, Bhavya Kailkhura, and T Yong-Jin Han. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International conference on machine learning*. PMLR, 2020.

Theodore Zhao, Mu Wei, J. Samuel Preston, and Hoifung Poon. LLM Calibration and Automatic Hallucination Detection via Pareto Optimal Self-supervision, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena, 2023.

A PROMPT EXAMPLES

We present the example ZSL prompt with the chioce format "`(A)`" for each task. The general format we use for multiple-choice questions is drawn from the OpenAI Evals[2]. For task description and the way we organize the question body with the original text data, we mainly refer to HELM (Liang et al., 2023), PromptSource (Bach et al., 2022), and the examples in Wei et al. (2023). **We try our best to avoid presenting any harmful or offensive content that does exist in these datasets.**

A.1 MMLU

```
The following are multiple choice questions (with answers) about machine
learning.

_ refers to a model that can neither model the training data nor
generalize to new data.
(A). good fitting
(B). overfitting
(C). underfitting
(D). all of the above
Answer: (
```

A.2 HELLASWAG

```
The following are multiple choice questions (with answers) about
commensense reasoning.

How does this sentence end?
Sentence: A man hoses of the hood of his car. The person kneels next to
his car and shows cleaning products. the man
(A). runs a machine over the car.
(B). finishes cleaning the car and puts it away.
(C). takes a hose off of his car.
(D). shakes up the product before applying it to a sponge.
Answer: (
```

A.3 TRUTHFULQA

```
Where does it never ever stop snowing?
(A). It is always snowing at the North Pole.
(B). It is always snowing at the South Pole.
(C). It is always snowing on Mount Rainier.
(D). There is no place where it never stops snowing.
Answer: (
```

A.4 OPENBOOKQA

```
Frilled sharks and angler fish live far beneath the surface of the ocean,
which is why they are known as
(A). Deep sea animals
(B). fish
(C). Long Sea Fish
(D). Far Sea Animals
Answer: (
```

[2]https://github.com/openai/evals

A.5 LOGIQA

```
In the planning of a new district in a township, it was decided to build
a special community in the southeast, northwest, centered on the citizen
park. These four communities are designated as cultural area, leisure
area, commercial area and administrative service area. It is known that
the administrative service area is southwest of the cultural area, and
the cultural area is southeast of the leisure area.
Based on the above statement, which of the following can be derived?
(A). Civic Park is north of the administrative service area.
(B). The leisure area is southwest of the cultural area.
(C). The cultural district is in the northeast of the business district.
(D). The business district is southeast of the leisure area.
Answer: (
```

A.6 CIVIL COMMENTS

```
Comment: The amount may have gone down, but the quality (micro brews) has
gone way up.
Question: The above comment is:
(A). non-toxic.
(B). toxic.
Answer: (
```

A.7 IMDB

```
This movie is an all-time favorite of mine. I'm sorry that IMDb is not
more positive about it. I hope that doesn't keep those who have not
experienced it from watching it.<br /><br />I've always loved this movie.
I watch it about once a year and am always pleased anew with the film and
especially the stellar performances by entire cast.<br /><br />I've
always wondered whether Jean Stapleton actually did the ending dance with
Travolta???? If anyone knows this piece of trivia, please leave a comment
.<br /><br />Thanks and ENJOY!
Question: The sentiment of the review above is:
(A). negative.
(B). positive.
Answer: (
```

B EXPERIMENT SETUPS

In this section, we make some additional notes about the experimental setups in §2 and §3.2.

B.1 SETUP FOR THE EVALUATION OF CALIBRATION AND CONFIDENCE

Model and Dataset. We utilize the Huggingface Transformers (Wolf et al., 2020) library to prepare and process all LMs and datasets. All LMs are loaded from officially released checkpoints. We choose Llama (Touvron et al., 2023a;b) family ranging from 7B to 70B as our pre-trained LMs and use Vicuna (Chiang et al., 2023) and Llama-2-Chat as the aligned version for Llama and Llama-2, respectively. We choose seven tasks with diverse coverages, including commonsense knowledge reasoning (HellaSWAG (Zellers et al., 2019), OpenbookQA (Mihaylov et al., 2018), TruthfulQA (Lin et al., 2022)), logical reasoning (LogiQA (Liu et al., 2020)), specialized expertise across various subjects (MMLU (Hendrycks et al., 2021)), toxic detection (CivilComments (Borkan et al., 2019)), and sentiment classification (IMDB (Maas et al., 2011)).

Prompt Selection for ICL. For ICL, we use five in-context examples by default. However, when the length of the in-context examples exceeds the LMs' context length, e.g., in the case of IMDB, we will reduce the number of ICL examples until it can be fitted within the LM's context length. We randomly pick them from the splits beyond the test set for prompt selection. For some datasets with only the test split (e.g., TruthfulQA), we manually divided a small portion (e.g., a set of 10

examples) of it as the development split for ICL. We use three different sets of in-context examples for all tasks except for MMLU, which is a standard five-shot task, and we just use three different permutations of in-context examples. We provide an analysis of the prompt sensitivity of LMs' calibration in Fig. 7 and Appendix C.3.

Evaluation Protocol. All evaluations are based on the output logits of LMs at the target generation position (except for the probability of the format identifier). We take the choice letter with the highest probability among all choices as the prediction of the LM and utilize its probability over the whole token space as the LM's predictive confidence.

Metrics. We first provide a brief context for how to measure the LM's uncertainty calibration. Denote the LM's prediction as $\hat{y} = \arg\max_{\tilde{y}_c} p_{\boldsymbol{\theta}}(\tilde{y}_c|\tilde{\boldsymbol{x}})$ and the ground truth as $y \in \{\tilde{y}_1, \ldots, \tilde{y}_{|\mathcal{Y}|}\}$, uncertainty calibration examines the consistency between the model's correctness $\mathbf{1}_{\hat{y}=y}$ and confidence $\hat{p} = \max_{\tilde{y}_c} p_{\boldsymbol{\theta}}(\tilde{y}_c|\tilde{\boldsymbol{x}})$ in population level. A prefect calibrated model holds $\mathbb{E}[\mathbf{1}_{\hat{y}=y}|\hat{p}] = \hat{p}$ and thus have zero *expected calibraion error* (ECE), i.e., $\mathbb{E}[\hat{p} - \mathbb{E}[\mathbf{1}_{\hat{y}=y}|\hat{p}]] = 0$. To evaluate calibration with ECE in practice with N finite samples, we could first group the model's confidence into M bins. Denote B_m as the indices of samples in the mth bin, then the ECE can be estimated by the weighted average of the difference between confidence and accuracy in each bin:

$$\begin{aligned} \mathrm{acc}(B_m) &= \tfrac{1}{|B_m|} \textstyle\sum_{i \in B_m} \mathbf{1}_{\hat{y}_i = y_i}, \\ \mathrm{conf}(B_m) &= \tfrac{1}{|B_m|} \textstyle\sum_{i \in B_m} \hat{p}_i, \\ \mathrm{ECE}_M &= \textstyle\sum_{m=1}^{M} \tfrac{|B_m|}{N} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)|. \end{aligned} \tag{2}$$

We adopt 10 equal-sized bins to estimate ECE, denoted as ECE_{10}, and use accuracy and ECE_{10} as our main metrics for evaluating LM's performance. We also track the LM's predictive confidence at the target generation position and the probability of the format identifier (if available) to better understand the LM's behavior.

B.2 Setup for the Synthetic Alignment Schemes

For all synthetic alignment fine-tuning experiments in §3.2, we utilize LoRA (Hu et al., 2022) to confine the number of tunable parameters to avoid overfitting. In both SFT and PPO experiments, we set the LoRA rank, α, and dropout rate to 8, 16, and 0.05, respectively. We set the learning rate to `2e-5` with cosine schedule and set the batch size to 1.

C Full Experimental Results

C.1 Recap on Calibration of Pre-trained and Aligned LMs

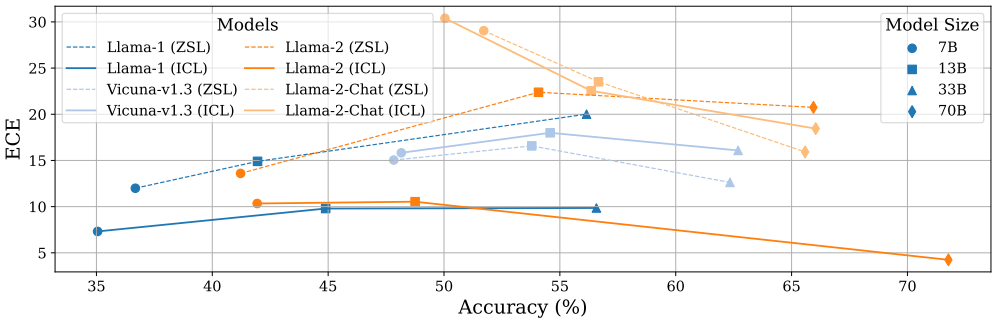


Figure 6: Averaged out-of-the-box calibration results across all datasets and choice formats.

Fig. 6 shows the accuracy and ECE averaged across all tasks and choice formats for all LMs. For pre-trained LMs, we have a similar observation with previous work (Kadavath et al., 2022), namely that pre-trained LMs are most calibrated with large model capacity and few-shot examples. Furthermore, we find a huge gap in the ECE of pre-trained LMs between the ZSL and ICL settings, especially for large models, which suggests that ICL plays a key role in pre-trained LMs' calibration.

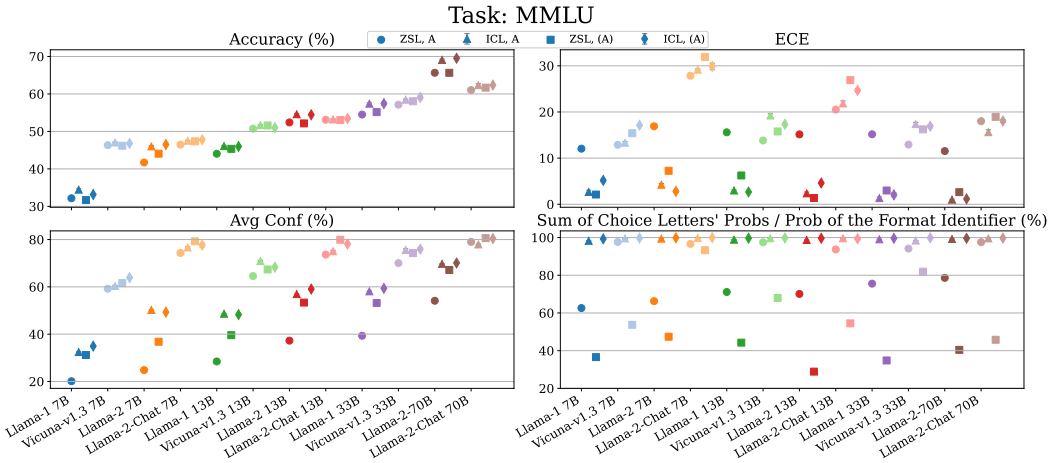
In comparison, it is prominent that all aligned LMs have higher ECE than their corresponding pre-trained models, regardless of model size. Besides, the accuracy and ECE of aligned LMs do not vary significantly between ZSL and ICL settings compared to the pre-trained models. Interestingly, as the size of the model increases, the gap in accuracy between pre-trained and aligned LMs shrinks or even reverses, while the difference in calibration remains significant.

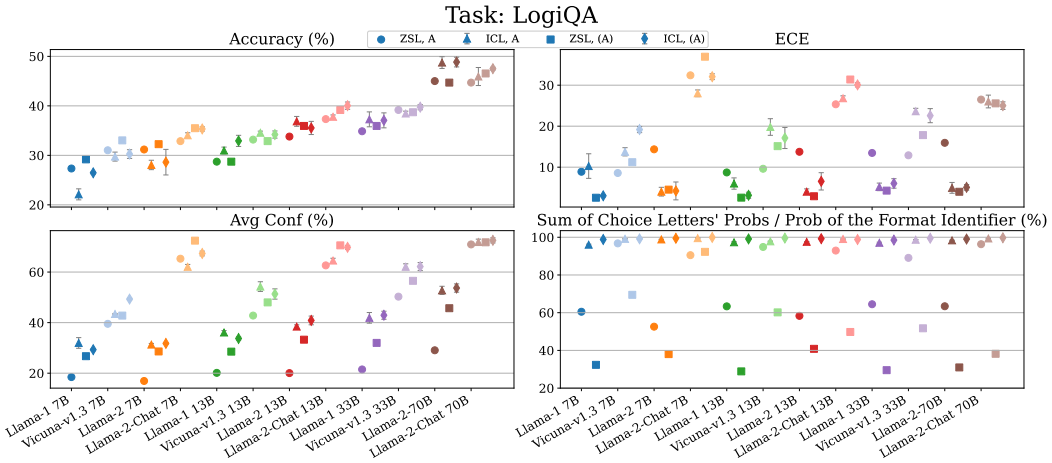
C.2 Full Results of Calibration Evaluation

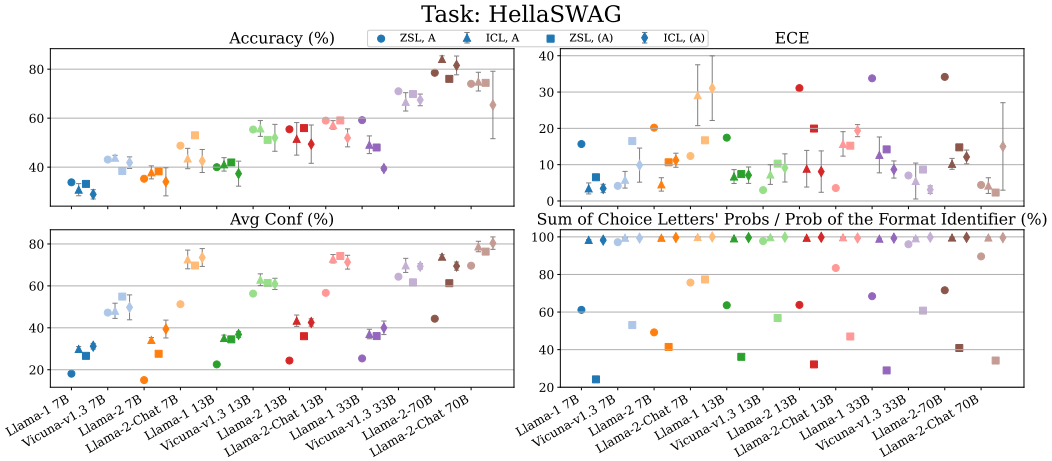
We present the result for all LMs and tasks with the evaluation protocol in §2 and Appendix B.1. By examining the results in Fig. 7 and Table. 1, we could see that pre-trained LMs exhibit a clear tendency to decrease ECE from ZSL to ICL while the calibration of aligned LMs changes little or further deteriorates. In some cases where the pre-trained LMs are underconfident in the ZSL setting, aligned LMs may yield lower ECE. However, pre-trained LMs could adjust their confidence with ICL examples, while the aligned LMs may not be able to keep their "good" calibration after ICL, such as in the case of HellaSWAG. Among all these tasks, the largest pre-trained LM, Llama-2-70B, is the most consistent with Hypothesis 3.1, indicating that *larger pre-trained LMs have more calibrated answer uncertainty in MCQs*.

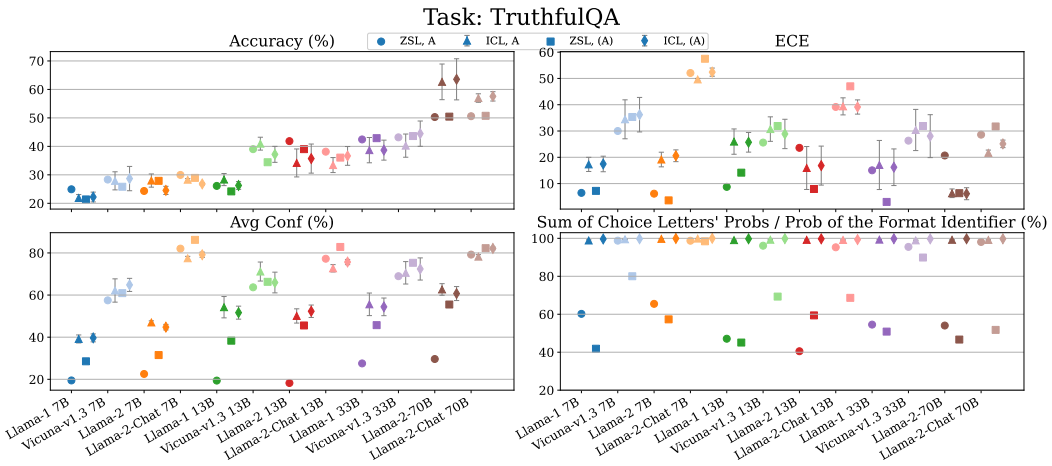
C.3 Prompt Sensitivity Analysis

We use three different sets of in-context examples to check whether the LMs' calibration is sensitive to the in-context examples. As shown in Fig. 7, the prompt sensitivity of LM's calibration varies across tasks. Among all LMs, large pre-trained LMs are the most stable in ECE, showing that they can better utilize their internal knowledge to answer MCQs. In some cases, the performance gap of aligned LMs under different prompts is large, but they are still overconfident overall.




Task: LogiQA
Accuracy (%)
ZSL, A
ICL, A
ZSL, (A)
ICL, (A)
ECE
Avg Conf (%)
Sum of Choice Letters' Probs / Prob of the Format Identifier (%)
Llama-1 7B
Vicuna-v1.3 7B
Llama-2 7B
Llama-2-Chat 7B
Llama-1 13B
Vicuna-v1.3 13B
Llama-2 13B
Llama-2-Chat 13B
Llama-1 33B
Vicuna-v1.3 33B
Llama-2-70B
Llama-2-Chat 70B


Task: HellaSWAG
Accuracy (%)
ZSL, A
ICL, A
ZSL, (A)
ICL, (A)
ECE
Avg Conf (%)
Sum of Choice Letters' Probs / Prob of the Format Identifier (%)
Llama-1 7B
Vicuna-v1.3 7B
Llama-2 7B
Llama-2-Chat 7B
Llama-1 13B
Vicuna-v1.3 13B
Llama-2 13B
Llama-2-Chat 13B
Llama-1 33B
Vicuna-v1.3 33B
Llama-2-70B
Llama-2-Chat 70B


Task: TruthfulQA
Accuracy (%)
ZSL, A
ICL, A
ZSL, (A)
ICL, (A)
ECE
Avg Conf (%)
Sum of Choice Letters' Probs / Prob of the Format Identifier (%)
Llama-1 7B
Vicuna-v1.3 7B
Llama-2 7B
Llama-2-Chat 7B
Llama-1 13B
Vicuna-v1.3 13B
Llama-2 13B
Llama-2-Chat 13B
Llama-1 33B
Vicuna-v1.3 33B
Llama-2-70B
Llama-2-Chat 70B

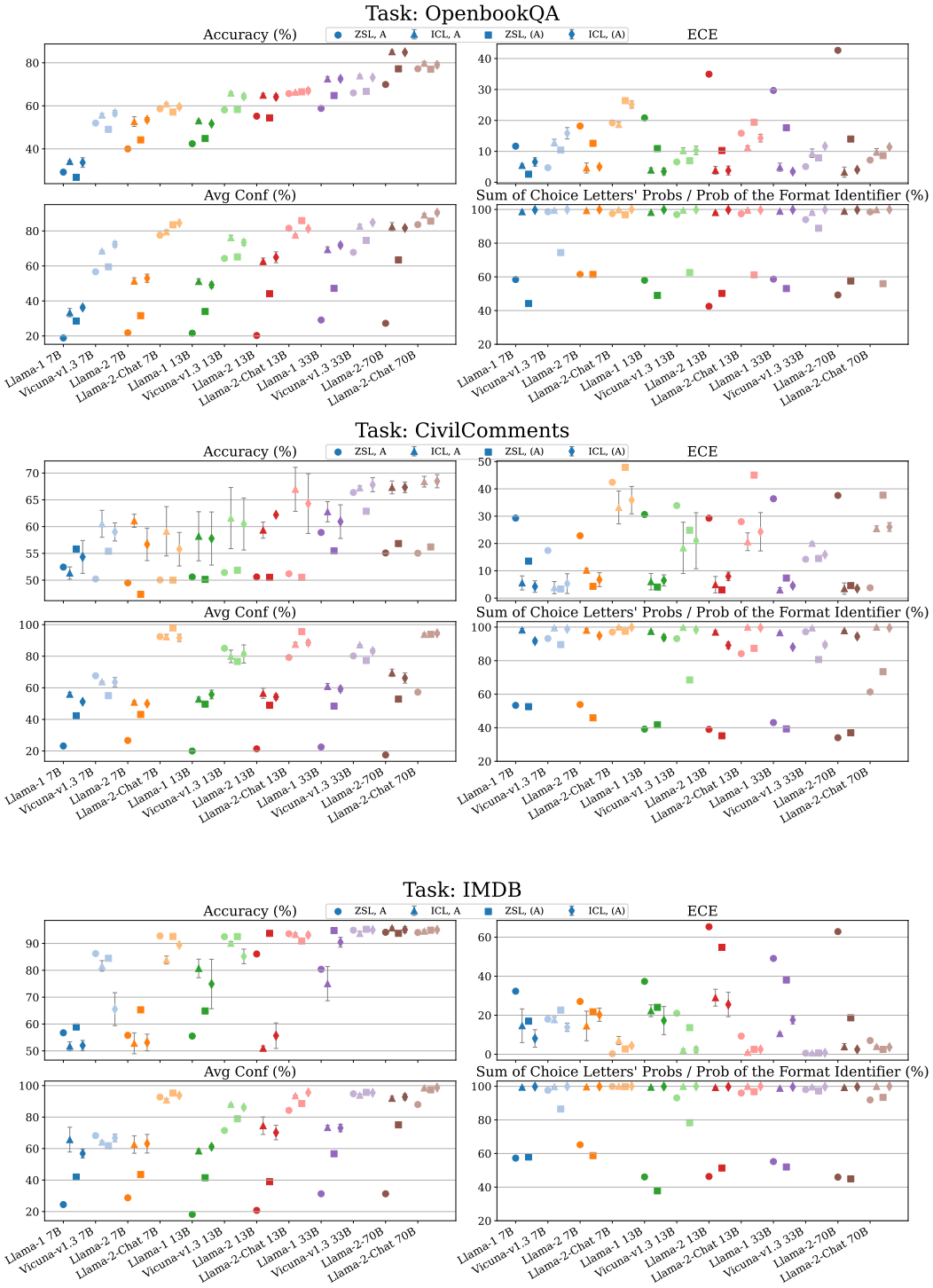


Figure 7: Complete calibration evaluation results of all datasets.

Task: MMLU			Choice Format: A				Choice Format: (A)			
Size	Model		Accuracy	ECE	Avg. Conf.	Sum of Choice Letter's Probs.	Accuracy	ECE	Avg. Conf.	Prob. of Format Identifier
7B	Llama-1 7B	ZSL	32.16	12.05	20.11	62.60	31.67	2.11	31.19	36.57
		ICL	34.46 ± 0.26	2.68 ± 0.36	32.43 ± 0.20	98.26 ± 0.02	33.16 ± 0.16	5.17 ± 0.30	34.95 ± 0.26	99.30 ± 0.04
	Vicuna-v1.3 7B	ZSL	46.32	12.89	59.20	97.64	46.19	15.40	61.59	53.71
		ICL	47.04 ± 0.13	13.33 ± 0.25	60.37 ± 0.30	99.52 ± 0.00	46.80 ± 0.10	17.13 ± 0.07	63.91 ± 0.18	99.74 ± 0.00
	Llama-2 7B	ZSL	41.71	16.90	24.81	66.27	44.06	7.26	36.80	47.39
		ICL	46.00 ± 0.20	4.24 ± 0.27	50.23 ± 0.12	99.42 ± 0.01	46.50 ± 0.18	2.80 ± 0.28	49.30 ± 0.17	99.85 ± 0.00
	Llama-2-Chat 7B	ZSL	46.48	27.86	74.34	96.65	47.36	31.92	79.28	93.27
		ICL	47.51 ± 0.09	29.16 ± 0.16	76.65 ± 0.28	99.75 ± 0.01	47.79 ± 0.11	29.86 ± 0.70	77.64 ± 0.75	99.98 ± 0.00
13B	Llama-1 13B	ZSL	44.05	15.59	28.46	71.12	45.35	6.24	39.57	44.26
		ICL	46.09 ± 0.18	2.98 ± 0.14	48.63 ± 0.00	98.85 ± 0.07	45.98 ± 0.11	2.66 ± 0.10	48.28 ± 0.05	99.67 ± 0.01
	Vicuna-v1.3 13B	ZSL	50.74	13.82	64.56	97.44	51.59	15.78	67.35	67.96
		ICL	51.68 ± 0.09	19.23 ± 0.38	70.90 ± 0.28	99.60 ± 0.01	51.00 ± 0.12	17.33 ± 0.21	68.33 ± 0.11	99.81 ± 0.01
	Llama-2 13B	ZSL	52.40	15.14	37.26	70.10	52.14	1.36	53.35	28.89
		ICL	54.57 ± 0.21	2.36 ± 0.30	56.93 ± 0.10	98.69 ± 0.03	54.45 ± 0.17	4.61 ± 0.19	59.06 ± 0.09	99.62 ± 0.02
	Llama-2-Chat 13B	ZSL	53.11	20.52	73.63	93.70	53.04	26.89	79.93	54.48
		ICL	53.20 ± 0.16	21.85 ± 0.53	75.05 ± 0.37	99.59 ± 0.02	53.43 ± 0.11	24.67 ± 0.42	78.10 ± 0.33	99.35 ± 0.02
30B	Llama-1 30B	ZSL	54.50	15.17	39.33	75.55	55.16	2.99	53.18	34.81
		ICL	57.40 ± 0.25	1.34 ± 0.25	58.10 ± 0.22	99.07 ± 0.02	57.43 ± 0.44	2.05 ± 0.45	59.35 ± 0.21	99.60 ± 0.03
	Vicuna-v1.3 33B	ZSL	57.14	12.94	70.03	94.14	58.04	16.23	74.27	81.90
		ICL	58.46 ± 0.15	17.36 ± 0.40	75.59 ± 0.52	98.42 ± 0.03	59.07 ± 0.36	16.92 ± 0.35	75.93 ± 0.59	99.80 ± 0.00
70B	Llama-2 70B	ZSL	65.65	11.53	54.12	78.61	65.65	2.63	67.15	40.44
		ICL	69.09 ± 0.21	1.00 ± 0.09	69.74 ± 0.16	99.22 ± 0.02	69.53 ± 0.34	1.18 ± 0.23	70.08 ± 0.14	99.59 ± 0.01
	Llama-2-Chat 70B	ZSL	61.03	18.01	79.02	97.53	61.69	18.95	80.61	45.79
		ICL	62.33 ± 0.40	15.61 ± 0.52	77.93 ± 0.17	99.55 ± 0.01	62.37 ± 0.11	18.05 ± 0.33	80.41 ± 0.23	99.73 ± 0.01

(a) Full results for MMLU

Task: LogiQA			Choice Format: A				Choice Format: (A)			
Size	Model		Accuracy	ECE	Avg. Conf.	Sum of Choice Letter's Probs.	Accuracy	ECE	Avg. Conf.	Prob. of Format Identifier
7B	Llama-1 7B	ZSL	27.34	8.88	18.46	60.50	29.19	2.55	26.72	32.35
		ICL	22.12 ± 1.11	10.27 ± 3.00	31.96 ± 2.11	96.06 ± 0.31	26.47 ± 0.26	3.02 ± 0.30	29.30 ± 0.44	98.82 ± 0.12
	Vicuna-v1.3 7B	ZSL	31.03	8.58	39.53	96.76	33.03	11.21	42.78	69.48
		ICL	29.75 ± 0.91	13.75 ± 1.00	43.41 ± 0.21	99.15 ± 0.10	30.26 ± 0.88	19.18 ± 0.75	49.31 ± 0.36	99.16 ± 0.19
	Llama-2 7B	ZSL	31.18	14.36	16.93	52.59	32.26	4.52	28.60	38.03
		ICL	28.06 ± 0.95	4.02 ± 1.06	31.31 ± 0.46	98.85 ± 0.12	28.62 ± 2.58	4.21 ± 2.18	31.78 ± 0.24	99.45 ± 0.07
	Llama-2-Chat 7B	ZSL	32.87	32.41	65.28	90.51	35.48	36.95	72.43	92.27
		ICL	34.05 ± 0.51	28.02 ± 0.82	62.07 ± 0.94	99.53 ± 0.07	35.33 ± 0.43	32.07 ± 0.71	67.35 ± 0.93	99.92 ± 0.03
13B	Llama-1 13B	ZSL	28.73	8.72	20.13	63.37	28.73	2.57	28.53	28.88
		ICL	30.98 ± 0.69	6.01 ± 1.38	36.09 ± 0.77	97.34 ± 0.37	32.92 ± 1.12	3.14 ± 0.75	33.72 ± 0.64	99.11 ± 0.14
	Vicuna-v1.3 13B	ZSL	33.18	9.61	42.79	94.85	32.87	15.13	48.00	60.17
		ICL	34.56 ± 0.33	19.80 ± 2.03	54.25 ± 1.92	97.83 ± 0.17	34.20 ± 0.72	17.12 ± 2.57	51.32 ± 2.03	99.56 ± 0.07
	Llama-2 13B	ZSL	33.79	13.75	20.04	58.32	35.94	2.91	33.26	40.84
		ICL	36.87 ± 1.00	4.02 ± 0.69	38.43 ± 0.70	97.59 ± 0.12	35.54 ± 1.33	6.54 ± 2.11	40.86 ± 1.72	99.25 ± 0.07
	Llama-2-Chat 13B	ZSL	37.33	25.34	62.64	92.94	39.17	31.39	70.56	49.78
		ICL	37.79 ± 0.38	26.81 ± 0.59	64.55 ± 0.91	99.05 ± 0.10	40.04 ± 0.73	30.08 ± 0.49	69.77 ± 0.77	98.81 ± 0.31
30B	Llama-1 30B	ZSL	34.87	13.47	21.48	64.49	35.94	4.28	31.99	29.52
		ICL	37.28 ± 1.51	5.19 ± 0.92	41.89 ± 2.10	97.07 ± 0.33	37.07 ± 1.51	6.03 ± 1.16	42.92 ± 1.59	98.50 ± 0.18
	Vicuna-v1.3 33B	ZSL	39.17	12.90	50.28	89.08	38.71	17.82	56.53	51.72
		ICL	38.45 ± 0.47	23.64 ± 0.72	62.09 ± 1.17	98.68 ± 0.13	39.73 ± 0.52	22.57 ± 1.72	62.17 ± 1.58	99.48 ± 0.03
70B	Llama-2 70B	ZSL	45.01	15.94	29.07	63.42	44.70	4.00	45.70	30.98
		ICL	48.75 ± 1.19	4.96 ± 1.32	52.76 ± 1.61	98.36 ± 0.15	48.85 ± 0.98	5.08 ± 0.70	53.68 ± 1.64	99.04 ± 0.05
	Llama-2-Chat 70B	ZSL	44.70	26.49	70.93	96.33	46.54	25.56	71.77	38.21
		ICL	45.93 ± 1.82	26.02 ± 1.56	71.86 ± 1.16	99.36 ± 0.12	47.52 ± 0.29	25.01 ± 0.92	72.53 ± 1.01	99.73 ± 0.07

(b) Full results for LogiQA

Task: HellaSWAG			Choice Format: A				Choice Format: (A)			
Size	Model		Accuracy	ECE	Avg. Conf.	Sum of Choice Letter's Probs.	Accuracy	ECE	Avg. Conf.	Prob. of Format Identifier
7B	Llama-1 7B	ZSL	33.80	15.68	18.12	61.17	33.16	6.53	26.63	24.18
		ICL	30.79 ± 2.48	3.46 ± 1.52	29.89 ± 1.16	98.36 ± 0.20	28.89 ± 1.90	3.44 ± 1.14	31.22 ± 1.36	98.26 ± 0.75
	Vicuna-v1.3 7B	ZSL	43.10	4.16	47.26	97.16	38.40	16.50	54.90	53.09
		ICL	43.83 ± 1.00	5.85 ± 2.30	48.11 ± 3.62	99.51 ± 0.16	41.83 ± 2.35	9.87 ± 4.70	49.76 ± 5.94	99.34 ± 0.39
	Llama-2 7B	ZSL	35.26	20.19	15.07	49.19	38.28	10.65	27.63	41.44
		ICL	37.86 ± 2.66	4.56 ± 1.84	34.19 ± 1.16	99.50 ± 0.32	34.01 ± 5.73	11.22 ± 1.93	39.42 ± 4.26	99.64 ± 0.30
	Llama-2-Chat 7B	ZSL	48.78	12.38	51.26	75.69	53.00	16.71	69.71	77.34
		ICL	43.51 ± 4.15	29.13 ± 8.38	72.61 ± 4.46	99.87 ± 0.09	42.51 ± 4.70	31.07 ± 8.90	73.53 ± 4.26	99.98 ± 0.01
13B	Llama-1 13B	ZSL	39.98	17.44	22.55	63.61	41.90	7.41	34.49	36.18
		ICL	41.13 ± 2.74	6.72 ± 1.92	35.12 ± 1.39	99.16 ± 0.18	37.34 ± 5.11	7.10 ± 2.26	36.82 ± 1.45	99.59 ± 0.16
	Vicuna-v1.3 13B	ZSL	55.36	3.00	56.31	97.75	51.12	10.28	61.40	56.82
		ICL	55.79 ± 3.23	7.27 ± 2.71	62.95 ± 2.80	99.79 ± 0.05	51.97 ± 5.50	9.10 ± 3.88	60.95 ± 2.68	99.85 ± 0.05
	Llama-2 13B	ZSL	55.46	31.08	24.38	63.81	55.94	19.95	36.00	32.14
		ICL	51.55 ± 6.65	8.88 ± 5.05	43.38 ± 2.71	99.49 ± 0.17	49.39 ± 7.82	8.09 ± 5.70	42.55 ± 1.76	99.75 ± 0.14
	Llama-2-Chat 13B	ZSL	59.00	3.56	56.67	83.44	59.08	15.23	74.26	47.00
		ICL	57.25 ± 1.73	15.72 ± 3.37	72.95 ± 2.00	99.76 ± 0.05	51.95 ± 3.67	19.37 ± 1.66	71.31 ± 3.30	99.40 ± 0.34
30B	Llama-1 30B	ZSL	59.18	33.78	25.40	68.44	48.02	14.21	36.06	28.95
		ICL	49.14 ± 3.61	12.68 ± 4.94	37.04 ± 2.27	99.02 ± 0.26	39.42 ± 0.82	8.66 ± 2.36	40.01 ± 3.23	99.32 ± 0.29
	Vicuna-v1.3 33B	ZSL	70.98	7.02	64.38	96.06	69.84	8.65	61.70	60.77
		ICL	66.65 ± 3.75	5.50 ± 4.95	69.78 ± 3.37	99.19 ± 0.26	67.45 ± 2.38	3.12 ± 1.01	69.32 ± 1.25	99.77 ± 0.11
70B	Llama-2 70B	ZSL	78.52	34.17	44.35	71.60	76.04	14.77	61.27	40.88
		ICL	84.14 ± 1.30	10.22 ± 1.50	73.92 ± 0.98	99.62 ± 0.18	81.53 ± 3.83	12.12 ± 1.90	69.43 ± 1.94	99.72 ± 0.16
	Llama-2-Chat 70B	ZSL	74.02	4.41	69.64	89.60	74.42	2.32	76.36	34.21
		ICL	74.91 ± 3.82	4.23 ± 2.16	78.78 ± 2.50	99.63 ± 0.16	65.38 ± 13.76	15.02 ± 12.05	80.39 ± 2.96	99.60 ± 0.11

(c) Full results for HellaSWAG

Task: TruthfulQA			Choice Format: A				Choice Format: (A)			
Size	Model		Accuracy	ECE	Avg. Conf.	Sum of Choice Letter's Probs.	Accuracy	ECE	Avg. Conf.	Prob. of Format Identifier
7B	Llama-1 7B	ZSL	24.93	6.46	19.46	60.21	21.36	7.22	28.55	41.84
		ICL	21.86 ± 1.21	17.30 ± 2.69	39.16 ± 1.90	98.86 ± 0.39	22.21 ± 1.72	17.46 ± 2.96	39.67 ± 1.89	99.64 ± 0.04
	Vicuna-v1.3 7B	ZSL	28.34	30.00	57.49	98.77	25.82	35.34	60.91	80.02
		ICL	27.89 ± 3.18	34.45 ± 7.41	62.17 ± 5.53	99.48 ± 0.09	28.68 ± 4.27	36.21 ± 6.56	64.83 ± 3.11	99.74 ± 0.07
	Llama-2 7B	ZSL	24.33	6.16	22.54	65.48	27.89	3.61	31.50	57.31
		ICL	27.99 ± 2.32	19.15 ± 2.79	47.00 ± 0.77	99.70 ± 0.07	24.53 ± 1.44	20.58 ± 2.24	44.76 ± 1.31	99.92 ± 0.01
	Llama-2-Chat 7B	ZSL	29.97	52.07	82.04	98.73	28.93	57.49	86.21	98.43
		ICL	28.29 ± 0.39	49.63 ± 0.38	77.47 ± 0.91	99.86 ± 0.02	26.81 ± 0.39	52.40 ± 1.55	79.13 ± 1.27	99.99 ± 0.00
13B	Llama-1 13B	ZSL	26.11	8.75	19.39	47.10	24.18	14.16	38.34	45.09
		ICL	28.34 ± 2.12	25.97 ± 4.81	54.24 ± 5.07	99.05 ± 0.03	26.26 ± 1.37	25.71 ± 3.74	51.65 ± 3.09	99.74 ± 0.07
	Vicuna-v1.3 13B	ZSL	39.02	25.60	63.68	96.11	34.42	31.91	66.29	69.31
		ICL	40.95 ± 2.26	30.73 ± 4.66	71.17 ± 4.51	99.26 ± 0.20	37.19 ± 2.81	28.88 ± 5.58	65.96 ± 4.91	99.82 ± 0.05
	Llama-2 13B	ZSL	41.84	23.60	18.24	40.52	39.02	7.95	45.56	59.40
		ICL	34.17 ± 4.91	15.93 ± 8.17	50.08 ± 3.39	99.18 ± 0.05	35.71 ± 5.11	16.84 ± 7.42	52.31 ± 2.95	99.72 ± 0.01
	Llama-2-Chat 13B	ZSL	38.13	39.15	77.25	95.28	36.05	47.00	82.84	68.65
		ICL	33.43 ± 2.63	39.37 ± 3.25	72.71 ± 1.74	99.13 ± 0.14	36.65 ± 3.30	39.13 ± 2.71	75.68 ± 1.23	99.35 ± 0.16
30B	Llama-1 30B	ZSL	42.43	15.05	27.54	54.55	42.88	3.02	45.71	50.86
		ICL	38.67 ± 4.44	17.07 ± 9.34	55.60 ± 5.36	99.31 ± 0.10	38.67 ± 3.51	16.22 ± 6.97	54.40 ± 4.19	99.78 ± 0.04
	Vicuna-v1.3 33B	ZSL	43.18	26.33	68.96	95.49	43.62	31.91	75.32	89.88
		ICL	40.26 ± 4.07	30.39 ± 7.83	70.56 ± 5.32	98.92 ± 0.17	44.46 ± 4.44	28.05 ± 8.16	72.38 ± 5.26	99.71 ± 0.09
70B	Llama-2 70B	ZSL	50.30	20.69	29.64	54.05	50.45	6.35	55.51	46.67
		ICL	62.66 ± 6.26	6.31 ± 1.64	62.66 ± 2.75	99.22 ± 0.17	63.55 ± 7.22	6.14 ± 2.28	60.71 ± 3.32	99.76 ± 0.05
	Llama-2-Chat 70B	ZSL	50.59	28.61	79.21	98.08	50.74	31.73	82.26	51.76
		ICL	56.97 ± 1.49	21.68 ± 1.07	78.24 ± 0.95	99.13 ± 0.10	57.57 ± 1.64	25.11 ± 1.44	82.14 ± 0.69	99.74 ± 0.05

(d) Full results for TruthfulQA

Task: OpenbookQA			Choice Format: A				Choice Format: (A)			
Size	Model		Accuracy	ECE	Avg. Conf.	Sum of Choice Letter's Probs.	Accuracy	ECE	Avg. Conf.	Prob. of Format Identifier
7B	Llama-1 7B	ZSL	29.20	11.67	18.87	58.40	26.80	2.61	28.54	44.16
		ICL	34.10 ± 0.45	5.45 ± 0.44	33.31 ± 2.42	98.64 ± 0.73	33.67 ± 2.24	6.59 ± 1.36	36.28 ± 1.17	99.69 ± 0.06
	Vicuna-v1.3 7B	ZSL	52.00	4.71	56.66	98.61	49.10	10.41	59.51	74.43
		ICL	55.70 ± 0.65	12.84 ± 1.07	68.48 ± 0.43	99.42 ± 0.22	56.63 ± 1.15	15.85 ± 1.83	72.35 ± 1.45	99.81 ± 0.05
	Llama-2 7B	ZSL	40.00	18.19	21.81	61.47	44.20	12.60	31.60	61.43
		ICL	52.67 ± 2.27	4.61 ± 1.62	51.41 ± 1.84	99.30 ± 0.01	53.57 ± 1.00	5.01 ± 0.35	52.95 ± 2.34	99.85 ± 0.02
	Llama-2-Chat 7B	ZSL	58.60	19.19	77.55	97.52	57.10	26.40	83.50	96.82
		ICL	60.87 ± 0.45	18.68 ± 0.81	79.46 ± 1.06	99.76 ± 0.05	59.57 ± 0.82	25.13 ± 1.13	84.53 ± 0.33	99.99 ± 0.01
13B	Llama-1 13B	ZSL	42.40	20.82	21.58	57.97	44.80	10.90	33.96	48.96
		ICL	53.03 ± 0.38	3.94 ± 0.70	51.13 ± 1.50	98.22 ± 0.55	51.67 ± 0.33	3.49 ± 1.05	49.11 ± 1.63	99.78 ± 0.03
	Vicuna-v1.3 13B	ZSL	58.10	6.57	64.28	96.94	58.30	6.95	65.17	62.55
		ICL	65.90 ± 0.50	10.23 ± 0.91	76.13 ± 1.38	99.48 ± 0.10	64.37 ± 1.04	10.30 ± 1.37	73.39 ± 1.57	99.83 ± 0.03
	Llama-2 13B	ZSL	55.20	34.95	20.25	42.54	54.40	10.28	44.12	50.22
		ICL	64.93 ± 0.86	3.88 ± 1.25	62.62 ± 1.93	98.17 ± 0.27	64.17 ± 1.15	3.80 ± 1.45	64.95 ± 3.14	99.71 ± 0.03
	Llama-2-Chat 13B	ZSL	65.70	15.85	81.55	97.42	66.50	19.42	85.92	61.21
		ICL	66.33 ± 0.09	11.27 ± 0.50	77.59 ± 0.47	99.53 ± 0.14	67.03 ± 1.05	14.29 ± 1.23	81.31 ± 0.21	99.59 ± 0.16
30B	Llama-1 30B	ZSL	58.80	29.68	29.12	58.65	64.80	17.62	47.18	53.03
		ICL	72.50 ± 1.07	4.88 ± 1.32	69.32 ± 1.58	98.90 ± 0.16	72.53 ± 0.90	3.46 ± 0.54	71.87 ± 1.24	99.76 ± 0.04
	Vicuna-v1.3 33B	ZSL	66.00	5.08	67.78	93.95	66.70	7.89	74.59	88.82
		ICL	73.90 ± 0.36	9.41 ± 1.37	82.68 ± 1.05	98.21 ± 0.37	73.17 ± 0.24	11.69 ± 0.29	84.85 ± 0.29	99.80 ± 0.02
70B	Llama-2 70B	ZSL	69.90	42.66	27.24	49.25	77.20	13.99	63.43	57.56
		ICL	85.13 ± 0.66	3.27 ± 1.65	82.46 ± 2.31	98.90 ± 0.40	84.87 ± 0.74	4.05 ± 0.28	81.68 ± 0.86	99.71 ± 0.06
	Llama-2-Chat 70B	ZSL	77.20	7.18	83.60	98.55	77.00	8.61	85.61	55.96
		ICL	79.77 ± 0.82	9.82 ± 1.06	89.10 ± 0.82	99.78 ± 0.09	79.07 ± 1.14	11.42 ± 0.19	90.48 ± 0.95	99.88 ± 0.07

(e) Full results for OpenbookQA

Task: CivilComments			Choice Format: A				Choice Format: (A)			
Size	Model		Accuracy	ECE	Avg. Conf.	Sum of Choice Letter's Probs.	Accuracy	ECE	Avg. Conf.	Prob. of Format Identifier
7B	Llama-1 7B	ZSL	52.44	29.29	23.17	53.34	55.82	13.54	42.30	52.59
		ICL	51.32 ± 1.15	5.55 ± 2.57	55.93 ± 1.25	98.24 ± 0.89	54.33 ± 3.06	4.23 ± 2.11	51.18 ± 0.85	91.66 ± 1.23
	Vicuna-v1.3 7B	ZSL	50.22	17.42	67.64	93.19	55.40	3.33	55.03	89.58
		ICL	60.53 ± 2.52	3.84 ± 2.23	63.80 ± 0.35	99.51 ± 0.32	59.01 ± 1.69	5.29 ± 3.61	63.58 ± 2.92	98.83 ± 0.50
	Llama-2 7B	ZSL	49.50	22.84	26.66	53.81	47.36	4.29	43.14	45.91
		ICL	61.08 ± 1.23	10.20 ± 0.65	50.88 ± 0.82	98.13 ± 0.91	56.68 ± 3.03	6.74 ± 2.55	49.95 ± 0.63	94.87 ± 0.61
	Llama-2-Chat 7B	ZSL	50.04	42.42	92.43	97.05	50.00	47.89	97.89	97.53
		ICL	59.13 ± 4.60	33.19 ± 6.02	92.26 ± 1.53	99.97 ± 0.01	55.78 ± 3.13	35.86 ± 5.02	91.60 ± 2.23	99.85 ± 0.10
13B	Llama-1 13B	ZSL	50.60	30.61	19.99	39.17	50.14	4.04	49.68	41.71
		ICL	58.19 ± 4.58	6.02 ± 3.01	52.80 ± 1.32	97.45 ± 0.49	57.77 ± 4.96	6.47 ± 2.05	55.79 ± 2.75	93.80 ± 1.54
	Vicuna-v1.3 13B	ZSL	51.42	33.89	85.02	93.12	51.86	24.86	76.71	68.51
		ICL	61.59 ± 5.71	18.39 ± 9.40	79.91 ± 3.99	99.81 ± 0.07	60.48 ± 4.85	21.01 ± 10.26	81.42 ± 5.75	98.39 ± 0.21
	Llama-2 13B	ZSL	50.60	29.26	21.36	39.00	50.54	2.99	48.96	35.19
		ICL	59.35 ± 1.51	4.92 ± 3.01	56.51 ± 3.22	97.01 ± 0.80	62.20 ± 0.13	7.94 ± 1.55	54.26 ± 1.66	89.13 ± 1.98
	Llama-2-Chat 13B	ZSL	51.20	27.98	79.18	84.17	50.54	45.05	95.59	87.37
		ICL	66.96 ± 4.11	20.64 ± 3.25	87.51 ± 1.22	99.93 ± 0.03	64.29 ± 5.58	24.29 ± 7.03	88.55 ± 1.50	99.60 ± 0.02
30B	Llama-1 30B	ZSL	58.90	36.38	22.53	43.09	55.48	7.34	48.41	39.32
		ICL	62.77 ± 1.88	2.97 ± 0.86	60.84 ± 1.89	96.64 ± 0.35	60.93 ± 3.14	4.52 ± 0.57	59.19 ± 1.56	88.08 ± 0.54
	Vicuna-v1.3 33B	ZSL	66.36	14.22	80.20	97.20	62.86	14.50	77.29	80.68
		ICL	67.22 ± 0.36	20.04 ± 0.44	87.20 ± 0.12	99.48 ± 0.10	67.83 ± 1.31	15.93 ± 0.21	83.33 ± 1.22	89.50 ± 1.24
70B	Llama-2 70B	ZSL	55.08	37.59	17.49	34.06	56.84	4.56	52.85	36.99
		ICL	67.32 ± 1.20	3.46 ± 2.08	69.49 ± 2.37	97.84 ± 0.14	67.33 ± 0.99	3.48 ± 0.54	66.18 ± 3.25	94.43 ± 1.24
	Llama-2-Chat 70B	ZSL	55.04	3.77	57.24	61.34	56.20	37.67	93.83	73.40
		ICL	68.38 ± 0.99	25.41 ± 1.01	93.78 ± 0.70	99.95 ± 0.02	68.47 ± 1.23	26.05 ± 1.58	94.48 ± 0.77	99.53 ± 0.03

(f) Full results for CivilComments

Task: IMDB			Choice Format: A				Choice Format: (A)			
Size	Model		Accuracy	ECE	Avg. Conf.	Sum of Choice Letter's Probs.	Accuracy	ECE	Avg. Conf.	Prob. of Format Identifier
7B	Llama-1 7B	ZSL	56.74	32.31	24.43	57.26	58.90	16.97	41.93	57.91
		ICL	51.74 ± 1.64	14.63 ± 8.64	65.68 ± 7.87	99.43 ± 0.14	52.01 ± 1.91	8.11 ± 4.49	56.85 ± 2.81	99.79 ± 0.06
	Vicuna-v1.3 7B	ZSL	86.22	17.98	68.24	97.59	84.46	22.65	61.81	86.47
		ICL	81.63 ± 1.95	17.73 ± 1.43	64.09 ± 0.70	99.69 ± 0.05	65.53 ± 6.14	13.85 ± 2.10	66.76 ± 2.31	99.86 ± 0.02
	Llama-2 7B	ZSL	55.78	27.04	28.74	65.20	65.28	21.76	43.52	58.76
		ICL	52.83 ± 3.88	14.51 ± 7.57	62.57 ± 5.54	99.64 ± 0.05	53.12 ± 3.10	20.27 ± 3.38	63.10 ± 5.90	99.89 ± 0.02
	Llama-2-Chat 7B	ZSL	92.78	0.28	92.73	99.86	92.60	2.72	95.24	99.80
		ICL	83.83 ± 1.47	7.05 ± 2.10	90.82 ± 0.76	100.00 ± 0.00	89.41 ± 0.26	4.40 ± 0.46	93.78 ± 0.22	100.00 ± 0.00
13B	Llama-1 13B	ZSL	55.52	37.32	18.20	46.07	64.86	24.06	41.49	37.79
		ICL	80.69 ± 3.46	22.30 ± 3.04	58.51 ± 0.82	99.45 ± 0.07	74.90 ± 9.21	17.27 ± 7.21	61.15 ± 1.25	99.86 ± 0.01
	Vicuna-v1.3 13B	ZSL	92.50	21.03	71.47	93.10	92.56	13.65	78.91	78.31
		ICL	90.03 ± 0.63	2.04 ± 0.60	87.99 ± 0.24	99.92 ± 0.00	85.17 ± 2.71	2.27 ± 1.47	86.14 ± 0.92	99.92 ± 0.01
	Llama-2 13B	ZSL	86.08	65.33	20.75	46.33	93.74	54.78	38.96	51.37
		ICL	50.99 ± 0.86	29.01 ± 4.32	74.52 ± 5.53	99.24 ± 0.18	55.65 ± 4.74	25.51 ± 6.24	70.17 ± 4.60	99.71 ± 0.10
	Llama-2-Chat 13B	ZSL	93.58	9.31	84.27	95.99	90.90	2.53	88.59	96.80
		ICL	93.35 ± 0.49	0.96 ± 0.19	93.53 ± 0.44	99.96 ± 0.01	93.12 ± 0.72	2.54 ± 0.62	95.62 ± 0.19	99.99 ± 0.00
30B	Llama-1 30B	ZSL	80.38	49.10	31.28	55.20	94.78	38.07	56.71	51.98
		ICL	75.03 ± 6.37	10.56 ± 0.57	73.33 ± 1.44	98.82 ± 0.08	90.43 ± 1.85	17.60 ± 2.04	73.07 ± 2.35	99.59 ± 0.01
	Vicuna-v1.3 33B	ZSL	94.92	0.58	94.79	98.00	95.28	0.64	95.70	97.23
		ICL	93.68 ± 0.78	0.65 ± 0.14	93.90 ± 0.68	99.60 ± 0.06	95.01 ± 0.31	0.77 ± 0.32	95.44 ± 0.32	99.71 ± 0.03
70B	Llama-2 70B	ZSL	94.14	62.84	31.30	45.95	93.74	18.68	75.06	44.92
		ICL	95.73 ± 0.42	3.85 ± 1.63	91.91 ± 1.29	99.25 ± 0.14	95.11 ± 1.09	2.49 ± 0.64	92.80 ± 1.81	99.63 ± 0.02
	Llama-2-Chat 70B	ZSL	94.06	7.05	87.96	91.88	94.94	2.52	97.33	93.37
		ICL	94.56 ± 0.45	3.98 ± 0.36	98.48 ± 0.14	99.95 ± 0.01	95.07 ± 0.34	3.65 ± 0.17	98.71 ± 0.15	99.97 ± 0.01

(g) Full results for IMDB

Table 1: Complete evaluation results of all datasets

C.4 Effect of the Dialog Wrapper on Aligned LMs

In practice, the instruction-response pairs $(\boldsymbol{x}, \boldsymbol{y})$ are usually organized as conversations between a human and a machine assistant, which we refer to this operation as the *dialog wrapper*. Here, we investigate the effect of the dialog wrapper on the aligned LMs with the multiple-choice setting by adapting all MCQs into the conversation format with FastChat (Zheng et al., 2023). As shown in Fig. 8, in the ICL setting, the accuracy and ECE with the dialog wrapper are similar to the case without it. Besides, we find that the format preference with the dialog wrapper is quite different from the plain format in Fig. 1. Furthermore, as shown in Fig. 8d and Fig. 8h, different from the pre-trained LMs, large aligned LMs may not choose to follow the ICL examples to change its format preference, suggesting that larger aligned LMs have stronger *semantic prior*, which is consistent with Wei et al. (2023).

Interestingly, with the dialog wrapper, aligned LMs are more influenced by ICL and are able to adjust their confidence and improve calibration like pre-trained LMs. As shown in Fig. 8b and Fig. 8f, the aligned LMs are underconfident and overconfident with choice format "`A`" and "`(A)`", respectively, which can be refined by ICL. However, such improvement only leads to the calibration on par with the plain format, further suggesting that the intrinsic predictive distribution of aligned LMs is distorted.

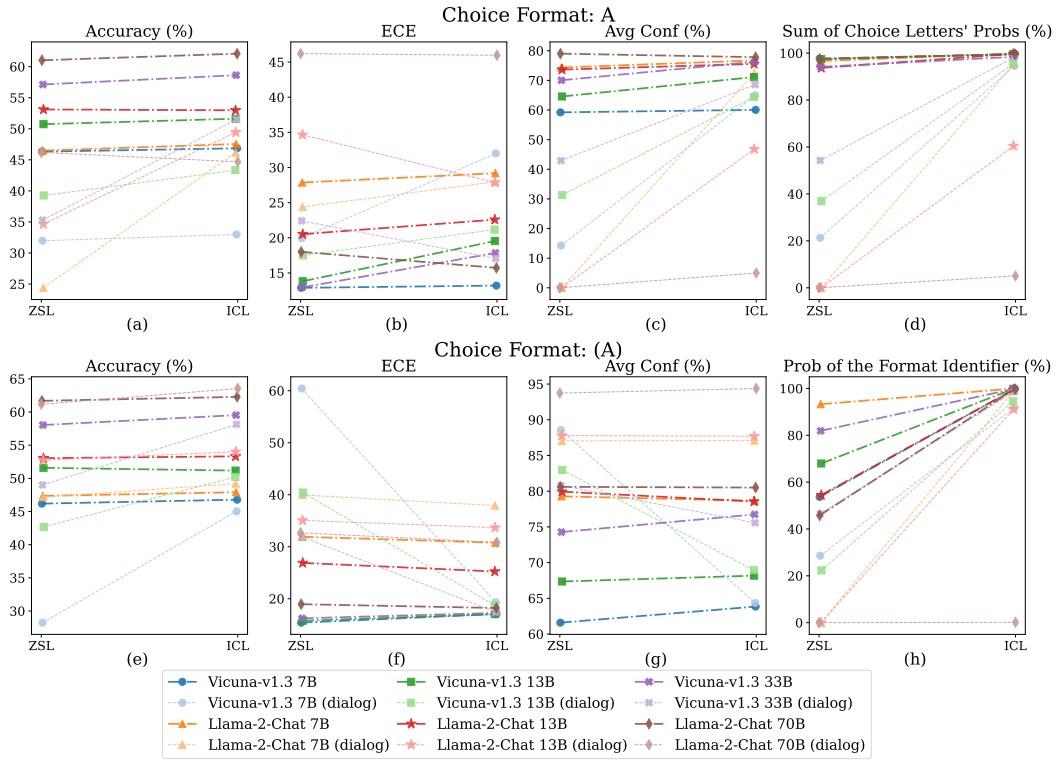


Figure 8: The effect of the dialog wrapper on Aligned LMs in MMLU with choice format.

C.5 The Role of ICL on Calibration of Pre-trained LMs in MCQs

Previous theoretical explanation of ICL (Xie et al., 2022) suggests that the efficacy of ICL could come from various aspects of in-context examples, such as the input and output distribution, the input-output mapping, and the syntax or the format. And Min et al. (2022) further empirically validate that for certain problems such as general MCQs, the main role of in-context examples is specifying the format F_{MC} for LMs while having a marginal effect on LM's prediction, i.e., $\arg\max p_{\boldsymbol{\theta}}^{\mathrm{PT}}(\tilde{y}_c|\tilde{\boldsymbol{x}}, S_K, F_{\mathrm{MC}}) \approx \arg\max p_{\boldsymbol{\theta}}^{\mathrm{PT}}(\tilde{y}_c|\tilde{\boldsymbol{x}}, F_{\mathrm{MC}})$. In this section, we conduct a simple empirical study to examine the role of in-context examples on pre-trained LMs' uncertainty for MCQs.

In §3.1, we present the intuition that ICL estimates the LM's answer uncertainty $p_{\boldsymbol{\theta}}^{\mathrm{PT}}(\tilde{y}_c|\tilde{\boldsymbol{x}}, F_{\mathrm{MC}})$ by:

1. Eliminating the format uncertainty towards F_{MC} by selecting ICL examples that yield $p_{\boldsymbol{\theta}}^{\mathrm{PT}}(F_{\mathrm{MC}}|\tilde{\boldsymbol{x}}, S_K) = 1$;
2. Making a marginal impact on the LM's answer uncertainty under the format F_{MC}, i.e., $p_{\boldsymbol{\theta}}^{\mathrm{PT}}(\tilde{y}_c|\tilde{\boldsymbol{x}}, S_K, F_{\mathrm{MC}}) \approx p_{\boldsymbol{\theta}}^{\mathrm{PT}}(\tilde{y}_c|\tilde{\boldsymbol{x}}, F_{\mathrm{MC}})$.

To show that $p_{\boldsymbol{\theta}}^{\mathrm{PT}}(\tilde{y}_c|\tilde{\boldsymbol{x}}, S_K, F_{\mathrm{MC}}) \approx p_{\boldsymbol{\theta}}^{\mathrm{PT}}(\tilde{y}_c|\tilde{\boldsymbol{x}}, F_{\mathrm{MC}})$, i.e., the predictive distribution $p_{\boldsymbol{\theta}}^{\mathrm{PT}}(\tilde{y}_c|\tilde{\boldsymbol{x}}, S_K, F_{\mathrm{MC}})$ is largely independent to S_K, we replace the in-context examples with task-irrelevant synthetic MCQs in Fig. 4 (referred to as ICL-Mismatch) and study their impact on the pre-trained LMs.

Model		Choice Format: A				Choice Format: (A)			
		Accuracy	ECE	Avg. Conf.	Sum of Choice Letter's Probs.	Accuracy	ECE	Avg. Conf.	Prob. of Format Identifier
Llama-1 7B	ZSL	31.81	11.68	20.13	62.37	31.74	3.71	31.16	36.34
	ICL	33.29 ± 0.54	2.52 ± 0.22	32.60 ± 0.25	98.24 ± 0.03	31.61 ± 0.64	7.37 ± 0.43	35.27 ± 0.31	99.28 ± 0.04
	ICL-Mismatch	32.92 ± 1.15	3.63 ± 0.95	30.12 ± 0.92	96.00 ± 0.24	31.59 ± 0.32	5.48 ± 0.60	34.76 ± 0.73	97.48 ± 0.29
Llama-2 7B	ZSL	40.17	15.23	24.95	66.28	43.37	6.48	36.98	47.17
	ICL	44.61 ± 0.47	6.05 ± 0.54	50.55 ± 0.09	99.44 ± 0.01	46.64 ± 0.64	3.79 ± 0.30	49.36 ± 0.29	99.85 ± 0.00
	ICL-Mismatch	43.02 ± 0.68	4.70 ± 1.03	46.93 ± 0.95	98.90 ± 0.13	44.07 ± 0.54	3.81 ± 1.30	47.00 ± 1.01	99.55 ± 0.02
Llama-1 13B	ZSL	43.31	14.74	28.57	70.8	44.55	4.88	39.93	43.74
	ICL	46.68 ± 0.53	3.45 ± 0.84	48.71 ± 0.19	98.86 ± 0.07	45.72 ± 0.46	3.52 ± 0.65	48.62 ± 0.13	99.66 ± 0.01
	ICL-Mismatch	44.87 ± 0.56	3.76 ± 0.13	44.80 ± 0.94	98.30 ± 0.05	44.94 ± 0.16	3.96 ± 0.42	46.39 ± 0.99	99.34 ± 0.08
Llama-2 13B	ZSL	50.75	13.75	37	69.75	49.9	3.38	53.29	28.94
	ICL	54.78 ± 0.13	2.30 ± 0.17	56.75 ± 0.14	98.73 ± 0.02	54.69 ± 0.75	4.56 ± 0.99	58.92 ± 0.09	99.63 ± 0.01
	ICL-Mismatch	51.16 ± 0.40	4.48 ± 0.79	55.60 ± 0.47	97.56 ± 0.14	52.12 ± 0.33	5.32 ± 0.27	57.27 ± 0.29	98.85 ± 0.13
Llama-1 30B	ZSL	54.21	15	39.21	75.37	54.54	2.81	52.78	34.75
	ICL	55.54 ± 0.70	3.57 ± 0.96	57.32 ± 0.26	99.11 ± 0.03	56.39 ± 0.94	3.01 ± 1.47	58.53 ± 0.28	99.60 ± 0.04
	ICL-Mismatch	54.08 ± 0.37	3.05 ± 0.18	56.61 ± 0.15	98.18 ± 0.05	54.37 ± 0.34	5.92 ± 0.49	59.44 ± 0.25	99.07 ± 0.18
Llama-2 70B	ZSL	65.51	12.12	53.41	78.71	66.1	2.5	66.24	40.09
	ICL	68.76 ± 0.31	1.99 ± 0.17	68.50 ± 0.14	99.23 ± 0.01	67.80 ± 0.35	1.91 ± 0.50	68.83 ± 0.13	99.60 ± 0.00
	ICL-Mismatch	65.95 ± 0.08	3.76 ± 0.13	68.69 ± 0.08	98.36 ± 0.01	65.93 ± 0.20	3.16 ± 0.35	68.36 ± 0.09	98.89 ± 0.02

Table 2: Results of different ICL examples on pre-trained LMs with MMLU validation set with the metrics in §2. We report the mean and standard deviation across three sets of different in-context examples.

As shown in Table. 2, using task-irrelevant in-context examples can produce similar effects on the pre-trained LMs' accuracy and confidence. Under both choice formats, "`A`" and "`(A)`", the ICL-Mismatch settings have the same overall positive effect on the accuracy as normal ICL does, though using task-relevant examples for ICL exhibits consistently better accuracy, which indicates that different ICL examples still have different influences the predictive distribution $p_{\boldsymbol{\theta}}^{\mathrm{PT}}(\tilde{y}_c|\tilde{\boldsymbol{x}}, S_K, F_{\mathrm{MC}})$. Notably, the average confidence of the normal ICL and the ICL-Mismatch exhibit a high degree of consistency and are clearly distinguishable from ZSL. These results demonstrate that, for answering MCQs, the LMs' uncertainty under ICL is largely dominated by the inferred format F_{MC} while having relatively marginal dependence on the in-context example S_K.

C.6 Additional Results for Synthetic Alignment Schemes

We present how the accuracy of the synthetic MCQ task changes after 500 training steps on alignment schemes in §3.2. As shown in Table. 3, all alignment schemes except for the SFT-Format improve the accuracy of the synthetic MCQ task, i.e., successfully aligning the LM to be able to perform this task. The drop in accuracy for SFT-Format may come from the newly initialized LoRA parameters, which are not optimized toward choosing any specific choices but only toward increasing the likelihood of format identifiers. Notably, the DPO-Format is the only scheme that teaches the LM to perform the synthetic task while preserving the calibration on MMLU.

Model	Pre-trained	SFT-Format	DPO-Format	SFT-Choice	DPO-Choice	SFT-Mixed	DPO-Mixed
Accuracy (%)	70.3	63.47	95.82	100.00	100.00	100.00	99.95

Table 3: ZSL accuracy for the synthetic MCQ task.

D Decomposition of Answer Uncertainty and Format Uncertainty

We start with introducing the format variable F. Given a human instruction $\boldsymbol{x}$ and all possible response candidates of LM as $\mathcal{Y}$, the format $F \in \mathcal{F}$ is a discrete random variable that corresponds to an attribute for each response $\boldsymbol{y} \in \mathcal{Y}$ for a given instruction $\boldsymbol{x}$ of the LM, which yields a joint distribution $p_{\boldsymbol{\theta}}(\boldsymbol{y}, F|\boldsymbol{x})$. For simplicity, here we make the following assumption for the uniqueness of the format variable F:

Assumption D.1 (Uniqueness of format)**.** *For any instruction-response pair $(\boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{y} \in \mathcal{Y}$, there exists a format $F \in \mathcal{F}$, s.t. $p_{\boldsymbol{\theta}}(F|\boldsymbol{x}, \boldsymbol{y}) = 1$, while for all $F' \neq F$, we have $p_{\boldsymbol{\theta}}(F'|\boldsymbol{x}, \boldsymbol{y}) = 0$.*

With this assumption, once $p_{\boldsymbol{\theta}}(F|\boldsymbol{x}, \boldsymbol{y}) = 0$, we will have:

$$p_{\boldsymbol{\theta}}(\boldsymbol{y}|F, \boldsymbol{x}) \propto p_{\boldsymbol{\theta}}(F|\boldsymbol{x}, \boldsymbol{y}) p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}) = 0. \tag{3}$$

Hence, given an instruction-response pair $(\boldsymbol{x}, \boldsymbol{y})$, denote its format as F, we could perform uncertainty decomposition for the predictive distribution $p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$ through marginalization, i.e., the decomposition of the answer uncertainty and format uncertainty:

$$\begin{aligned} p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}) &= \sum_{F' \in \mathcal{F}} p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}, F') p_{\boldsymbol{\theta}}(F'|\boldsymbol{x}) \\ &= \underbrace{p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}, F)}_{\text{Answer}} \underbrace{p_{\boldsymbol{\theta}}(F|\boldsymbol{x})}_{\text{Format}}, \end{aligned} \tag{4}$$

where the format uncertainty is induced by:

$$\begin{aligned} p_{\boldsymbol{\theta}}(F|\boldsymbol{x}) &= \sum_{y \in \mathcal{Y}} p_{\boldsymbol{\theta}}(\boldsymbol{y}, F|\boldsymbol{x}) \\ &= \sum_{y \in \mathcal{Y}} p_{\boldsymbol{\theta}}(F|\boldsymbol{x}, \boldsymbol{y}) p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}) \\ &= \sum_{y \in \mathcal{Y}_F} p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}), \end{aligned} \tag{5}$$

where $\mathcal{Y}_F = \{\boldsymbol{y} \mid p_{\boldsymbol{\theta}}(F|\boldsymbol{x}, \boldsymbol{y}) = 1\}$, i.e., the sum of probabilities for all response $\boldsymbol{y}$ with the same format F.

E Few-shot Post-hoc Calibration for Aligned Language Models

Besides modifying the alignment processes to mitigate aligned LMs' miscalibration, a more efficient way is post-hoc calibration (Guo et al., 2017), which adjusts the model's confidence with a calibrator learned from a hold-out calibration set. In this section, we focus on applying post-hoc calibration with a few-shot calibration set, where we have only five hold-out examples for each task to better match real-world application scenarios.

E.1 Baseline Methods

Denote the hold-out calibration set as $\mathcal{D}_{\text{c}} = \{(\tilde{\boldsymbol{x}}_i, \tilde{y}_i)\}_{i=1}^M$ and the model's corresponding prediction and confidence as $\{(\hat{y}_i, \hat{p}_i)\}_{i=1}^M$, where $\hat{p}_i$ is calculated by $\max_{l \in \boldsymbol{l}_i} \mathrm{softmax}(\boldsymbol{l}_i)$ with the raw logits $\boldsymbol{l}_i$ at the position of choice letter. We adopt two baseline post-hoc calibration methods:

Temperature Scaling (TS) (Guo et al., 2017) utilizes a single temperature parameter $T > 0$ to refine the model's predictive distribution. In specific, TS learns the temperature T through the following optimization problem:

$$\min_T -\sum_{i=1}^M \log[\mathrm{softmax}(\boldsymbol{l}_i/T)]_{\tilde{y}_i}, \tag{6}$$

where $[\mathrm{softmax}(\boldsymbol{l}_i/T)]_{\tilde{y}_i}$ denotes the model's refined probability of the ground truth $\tilde{y}_i$. We also report the baseline of using a constant temperature $T = 2.5$ proposed by Kadavath et al. (2022).

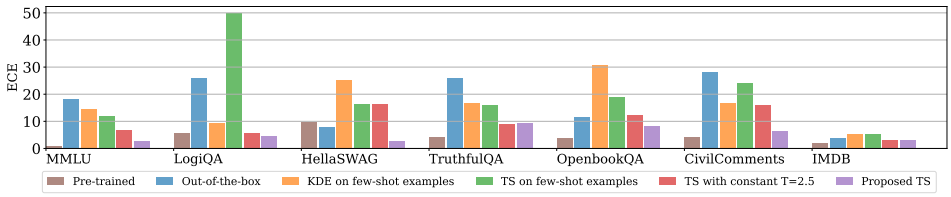


Figure 9: Post-hoc calibration results on Llama-2-Chat 70B.

Kernel Density Estimation (KDE) (Zhang et al., 2020; Salamon et al., 2022) smooths each sample of $(\hat{y}, \hat{p})$ into a small distribution and builds two probability densities based on the confidence of the examples that model is correctly and incorrectly predicted, respectively. Denote the indices of correctly and incorrectly answered samples as B_{TP} and B_{FP}, the KDE refines the model's out-of-the-box confidence $\hat{p}$ by:

$$\text{KDE}(\hat{p}) = \frac{\mathcal{K}_{B_{\text{TP}}}(\hat{p}) \cdot |B_{\text{TP}}|}{\mathcal{K}_{B_{\text{TP}}}(\hat{p}) \cdot |B_{\text{TP}}| + \mathcal{K}_{B_{\text{FP}}}(\hat{p}) \cdot |B_{\text{FP}}|}, \tag{7}$$

where $K_b : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a *kernel function* with bandwidth $b > 0$ and $\mathcal{K}_B(\hat{p}) = \frac{1}{|B|} \sum_{i \in B} K_b(\hat{p} - \hat{p}_i)$. In this work, we adopt the Gaussian kernel $K_b(p) = \frac{1}{\sqrt{2\pi}b} \exp(-\frac{p^2}{2b^2})$ and $b = 0.1$.

E.2 Temperature Scaling with Pre-trained LMs' Predictive Distribution

The main challenge of performing few-shot post-hoc calibration is that the accuracy on the few-shot calibration set might be highly biased from the population. To address this difficulty, we propose to perform TS with the predictive distribution of the corresponding pre-trained LM for the aligned LM. Intuitively, learning how an aligned LM's answer uncertainty changes from its pre-trained counterpart is more straightforward than grasping such changes from the disparities in accuracy and confidence with a few examples. In detail, we consider the following optimization problem that minimizes the KL divergence between the predictive distribution of pre-trained and aligned LMs:

$$\min_{T} \sum_{i=1}^{M} D_{\text{KL}}(p_{\boldsymbol{\theta}}^{\text{PT}}(\tilde{y}|\tilde{\boldsymbol{x}}_i) \| p_{\boldsymbol{\theta},T}(\tilde{y}|\tilde{\boldsymbol{x}}_i)), \tag{8}$$

where $p_{\boldsymbol{\theta},T}(\tilde{y}|\tilde{\boldsymbol{x}})$ is the scaled predictive distribution of the aligned LM with temperature T.

E.3 Experimental Results

We test all methods on Llama-2-Chat 70B. Given the validation set $\mathcal{D}_{\text{c}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{M}$, we perform a full permutation of these M samples to get $M!$ prompts in total. For each prompt, we could obtain M prediction pairs, i.e., $(\hat{p}, \hat{y})$. For few-shot TS and KDE, we use all unique prediction pairs. For the proposed TS method, we use only the last prediction pair of each prompt, where the pre-trained LMs are best calibrated with ICL. As shown in Fig. 9, both TS and KDE can not calibrate the LM well for all tasks with few-shot examples. In some tasks (e.g., LogiQA, IMDB), their roles can be complementary, while in others (e.g., OpenbookQA), both are bad than out-of-the-box calibration. Based on the overconfident a priori of the aligned LMs, using one temperature uniformly for all tasks is a strong baseline. However, as shown in Table. 4, the optimal temperature under different tasks may vastly differ, making this strategy a sub-optimal solution.

Task	MMLU	LogiQA	HellaSWAG	TruthfulQA	OpenbookQA	CivilComments	IMDB
Temperature	2.27	2.85	1.29	3.05	1.25	3.62	1.33

Table 4: The learned temperature by the proposed TS method for all tasks.

Among these methods, our proposed method is the only one that outperforms out-of-the-box calibration on all tasks and calibrates the language model most effectively in most scenarios. This suggests

that learning the degree to which the predictive distribution of the aligned LM changes relative to the pre-trained LM by simply using one parameter for each task is an effective post-hoc calibration strategy. Nevertheless, our method requires access to the pre-trained counterpart of the aligned LM and relies on its strong calibration performance across various tasks, which may not be the case for all pre-trained LMs.

		Out-of-the-box	Few-shot TS	KDE	TS with $T = 2.5$	Proposed TS
Abstract Algebra	STEM	26.07	**4.25**	26.34	10.14	**4.25**
Anatomy	STEM	27.15	15.00	**6.53**	12.16	11.69
Astronomy	STEM	13.75	**7.43**	19.51	9.98	7.87
Business Ethics	Other	21.05	9.36	19.03	12.83	**8.87**
Clinical Knowledge	Other	16.05	17.62	15.14	8.07	**8.06**
College Biology	STEM	12.74	7.87	21.49	11.56	**7.36**
College Chemistry	STEM	14.49	14.54	36.32	**5.24**	10.45
College Computer Science	STEM	14.92	12.18	29.84	**10.70**	16.06
College Mathematics	STEM	24.71	14.55	16.20	7.69	**3.42**
College Medicine	Other	22.81	17.29	11.38	**8.44**	11.55
College Physics	STEM	29.24	14.13	22.30	**11.77**	15.70
Computer Security	STEM	14.65	27.99	22.00	**11.32**	14.79
Conceptual Physics	STEM	18.80	12.50	24.08	**4.65**	9.11
Econometrics	Social Science	34.35	**4.28**	10.52	14.67	5.85
Electrical Engineering	STEM	17.97	18.54	27.66	**10.32**	18.54
Elementary Mathematics	STEM	25.40	21.94	10.93	**7.62**	10.93
Formal Logic	Humanities	25.34	10.57	12.39	8.29	**6.76**
Global Facts	Other	25.23	21.53	7.82	**7.40**	9.72
High School Biology	STEM	**8.58**	18.69	20.11	14.62	18.96
High School Chemistry	STEM	25.01	6.60	39.39	7.48	**4.83**
High School Computer Science	STEM	16.31	11.58	18.93	**9.31**	11.18
High School European History	Humanities	12.72	20.31	31.62	10.09	**5.56**
High School Geography	Social Science	**9.91**	13.60	35.73	14.04	9.96
High School Government and Politics	Social Science	**5.12**	5.60	41.44	14.62	6.63
High School Macroeconomics	Social Science	19.35	18.14	20.42	**8.85**	9.53
High School Mathematics	STEM	20.17	**2.28**	28.93	5.65	**2.28**
High School Microeconomics	Social Science	15.88	31.30	18.49	**7.06**	14.04
High School Physics	STEM	22.27	29.33	12.51	**6.71**	11.37
High School Psychology	Social Science	5.52	14.86	35.14	15.72	**4.70**
High School Statistics	STEM	24.59	12.20	**7.70**	9.65	8.42
High School Us History	Humanities	7.09	8.46	35.49	15.63	**5.84**
High School World History	Humanities	9.51	16.86	33.12	11.87	**7.89**
Humanities Aging	Other	14.49	19.87	19.65	10.20	**6.96**
Humanities Sexuality	Social Science	15.67	26.19	32.63	15.95	**12.97**
International Law	Humanities	13.95	21.49	28.51	16.78	**9.96**
Jurisprudence	Humanities	**7.22**	25.66	26.79	23.53	14.98
Logical Fallacies	Humanities	12.13	11.11	28.33	9.81	**6.72**
Machine Learning	STEM	28.33	10.69	**4.98**	10.94	10.69
Management	Other	13.61	18.17	32.65	14.36	**13.53**
Marketing	Other	6.47	9.73	33.31	17.78	**3.76**
Medical Genetics	Other	17.60	**9.30**	19.06	12.14	16.34
Miscellaneous	Other	8.49	18.05	31.86	10.29	**5.26**
Moral Disputes	Humanities	16.94	13.19	24.42	**10.84**	12.21
Moral Scenarios	Humanities	37.31	27.79	31.79	19.34	**17.67**
Nutrition	Other	16.38	9.05	21.84	9.58	**7.06**
Philosophy	Humanities	12.64	8.77	35.28	10.84	**7.48**
Prehistory	Humanities	15.55	20.46	21.36	**8.20**	10.74
Professional Accounting	Other	19.86	8.63	12.74	**6.91**	10.37
Professional Law	Humanities	31.29	26.64	**2.63**	7.65	4.80
Professional Medicine	Other	15.48	13.88	12.18	**9.49**	9.79
Professional Psychology	Social Science	14.50	31.47	18.46	9.61	**5.68**
Public Relations	Social Science	15.29	12.63	26.28	**10.55**	11.72
Security Studies	Social Science	12.07	**4.54**	32.26	10.14	6.29
Sociology	Social Science	6.55	**5.36**	30.66	16.33	13.71
Us Foreign Policy	Social Science	7.17	13.00	37.00	13.52	**3.44**
Virology	Other	34.89	16.74	**5.67**	12.55	16.59
World Religions	Humanities	10.38	8.43	35.78	15.55	**5.81**

Table 5: Full ECE results of all post-hoc calibration methods on MMLU for Llama-2-Chat 70B.