Learning or Self-aligning? Rethinking Instruction Fine-tuning

Anonymous ACL submission

Abstract

Instruction Fine-tuning (IFT) is a critical phase in building Large Language Models (LLMs). Previous works mainly focus on the IFT's role in the transfer of behavioral norms and the learning of additional world knowledge. However, the understanding of the underlying mechanisms of IFT remains significantly limited. In this paper, we design a knowledge intervention framework to decouple the potential underlying factors of IFT, thereby enabling individual analysis of different factors. Surprisingly, our experiments reveal that attempting to learn additional world knowledge through IFT often struggles to yield positive impacts and can even lead to markedly negative effects. Further, we discover that maintaining internal knowledge consistency before and after IFT is a critical factor for achieving successful IFT. Our findings reveal the underlying mechanisms of IFT and provide robust support for some very recent and potential future works.

1 Introduction

The advent and evolution of large language models (LLMs) have marked a significant milestone in natural language processing (NLP) (Zhao et al., 2023a). As one of the core steps in the construction of LLMs, Instruction Fine-tuning (IFT) employs supervised instruction-response pairs to fine-tune LLMs, thereby facilitating the transformation of LLMs from a continuous writing model to a question-answering agent (Chung et al., 2022, Iyer et al., 2023, Jang et al., 2023).

Despite the critical role of IFT in the construction of LLMs, there is a significant lack of in-depth research on the mechanisms by which IFT operates. In traditional machine learning, supervised learning aims to fit models to specific tasks and data distributions, whereas the impact of IFT on LLMs is markedly different. As shown in Figure 1, on one hand, one of the most apparent effects of IFT is its ability to align the output of LLMs more closely with the latent behavioral norms contained within the IFT data, thereby enabling more effective parameter-entailed knowledge expression (Zhou et al., 2023). On the other hand, many existing studies aim to facilitate domain-specific adaptation of LLMs through IFT, by injecting the world knowledge contained in IFT data into LLMs (Li et al., 2023, Cui et al., 2023).

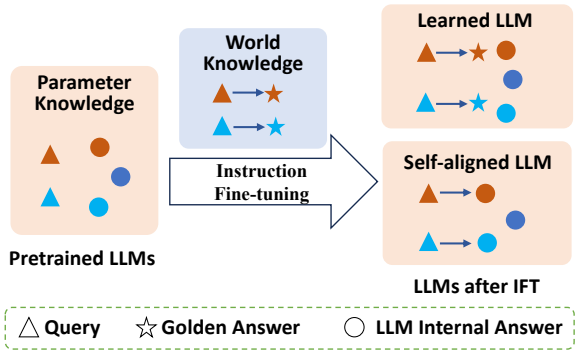


Figure 1: Two potential mechanisms for instruction fine-tuning. 1) Learning, which injects world knowledge in IFT data into the LLMs; 2) Self-aligning, which aligns queries with knowledge already in LLMs with similar behavioral norms. Elements with the same color are related.

Unfortunately, both the transfer of behavioral norms and the enhancement of domain knowledge are closely coupled with the corpus applied in IFT, rendering the analysis of IFT's true effects exceedingly challenging. Due to the interconnection of these two effects, it is challenging to discern whether the benefits derived are due to the promotion of more effective expression of parameter knowledge or the injection of additional world knowledge. The coupling between the above two factors, along with a lack of in-depth analysis of the IFT mechanism, hampers our comprehension of the effectiveness of IFT. This limitation hinders our ability to develop robust strategies for IFT data construction, model training, and model evaluation due to insufficient theoretical support. Therefore, a thorough and comprehensive analysis of the un-

derlying core factors that drive the effectiveness of IFT is crucial for achieving more effective IFT.

To this end, this paper designs a knowledge intervention framework for analyzing the underlying mechanisms of IFT. The main idea of our framework is to control the consistency between the knowledge in IFT data and the existing parameter knowledge of LLM, in order to decouple the injection of domain knowledge from the transfer of behavior norms during IFT. This allows for a separate analysis of the roles of these two crucial IFT factors. Specifically, we first employ in-context learning (ICL) (Dong et al., 2023, Brown et al., 2020) to probe the internal parameter knowledge of LLMs. Building on this, we intervene in the composition of existing parameter knowledge and the newly introduced world knowledge within IFT data and then observe the differences in model behavior after IFT under different intervention groups. Based on the framework, we conduct an in-depth analysis to answer the following two critical research questions (RQ):

- **RQ1:** *How does the world knowledge in IFT data affect LLMs?*
- **RQ2:** *What is the underlying cause of the above impact?*

For RQ1, we initially discover that significant discrepancies between the world knowledge contained in IFT data and the existing parameter knowledge within LLMs can substantially undermine the models' capabilities. Performance derived from a set of IFT data that contains incorrect world knowledge but aligns with the model's parameter knowledge is significantly better than that from a set containing correct world knowledge but inconsistent with the model's internal parameter knowledge. To dive into this phenomenon, we explicitly provide LLMs with the necessary world knowledge in the context, enabling the model to only transform the output behavioral norms rather than jointly learn the inconsistent world knowledge. We discover that the detrimental effects caused by the inconsistency between parameter knowledge and world knowledge can be significantly mitigated by explicitly providing such self-contained IFT datapoints. These two findings indicate that attempts to introduce world knowledge during the IFT phase that is inconsistent with the model's internal parameter knowledge can severely undermine the model, which suggests that the injection of world knowledge does not lie in the center of a successful IFT.

For RQ2, we further analyze the model performance under varying degrees of consistency between the knowledge contained in IFT data and the parameter knowledge in the initial LLM. We find that while the consistency between the two has a significant impact on model performance, a higher degree of consistency does not necessarily correlate with better model outcomes. However, our further research reveals a strong correlation between the model's ultimate capability after IFT and the consistency of the model's internal knowledge before and after IFT. That is to say, for the model after IFT, if its responses are more consistent with the outcomes produced by the original model from in-context learning probing, then the performance of the model after IFT is also better. The validity of the finding is independent of whether the test data belongs to the same domain as the training data, and is also unrelated to the original performance of the base LLM. This implies that the phenomenon is solely influenced by the consistency of the knowledge before and after the IFT process. Furthermore, we discover that using IFT data that is either too consistent or too inconsistent with the initial parameter knowledge can lead to a divergence in the model's internal knowledge before and after IFT, thereby resulting in a decline in performance.

Our experiments reveal the fundamental role of IFT in the construction of LLMs. Essentially, IFT is not a supervised, domain-specific **learning** process, but a process of **self-aligning** instruction with the already existing internal knowledge (e.g., that can be obtained through few-shot in-context probing) of LLMs. Our findings not only provide robust theoretical support for the very recently emerged researches on self-play (Chen et al., 2024) and super-alignment (Burns et al., 2023), etc., but also shed light on the future direction of data construction, model learning, and model evaluation for IFT.

2 Related Work

Despite the significance and wide applications of instruction fine-tuning, there is no rigorous study looking deeply into the underlying mechanisms of IFT. Current studies focus on instruction data construction (Song et al., 2023;Yuan et al., 2023;Liu et al., 2023;Xu et al., 2023)). Recently, starting from the surface alignment hypothesis (Zhou et al., 2023), an increasing number of works on self-

play (Chen et al., 2024) and self-align (Sun et al., 2023) have been proposed, increasing, attracting widespread attention. Although self-align has been verified for its effectiveness several times, current works remain at the stage of hypothesis and application (Yuan et al., 2024;Sun et al., 2023;Lu et al., 2024;Zhang et al., 2023;Wan et al., 2024) without understanding of its underlying mechanisms.

3 Knowledge Intervention Framework

During the process of IFT, the potential transfer of behavioral norms and the injection of world knowledge are coupled together. Consequently, prior research on IFT has struggled to distinguish the relative effects of these two. To further investigate the underlying mechanisms of IFT, this paper designs a knowledge intervention framework to decouple these two factors. The main idea behind our framework is to control the association between knowledge in IFT data and the existing parameter knowledge in LLMs, thereby managing the degree of potential world knowledge that would be injected during IFT. Through this, we can decouple the injection of world knowledge and the transfer of behavioral norms, by observing the effects of IFT at varying degrees of world knowledge injection.

Specifically, we select four multi-choice datasets from different domains for our experimental investigations. Based on these four datasets, we employ in-context learning to probe the existing internal parameter knowledge of LLMs. Then, based on adjusting the consistency degree of world knowledge and model parameter knowledge, we construct multiple datasets of IFT. Finally, we analyze the underlying impact of different degrees of world knowledge injection on IFT by fine-tuning LLMs on these datasets and examining their performance on the homogeneous, in-domain, and out-of-domain test sets. In the following, we will introduce our knowledge intervention framework in detail.

3.1 Domain IFT Corpus Setup

In order to facilitate more efficient knowledge consistency identification and IFT evaluation, we select all IFT corpus in the form of multiple choice from four domains: medicine, history, engineering, and jurisprudence. In the medical domain, we construct a dataset comprising 20,000 training instances, 2,206 testing instances, and 10 instances for development. This dataset is curated by filtering medmcqa (Pal et al., 2022) to extract entries with explanations and only one correct answer. For the other three domains, we procure the relevant items from the Xiezhi Benchmark (Gu et al., 2023) and held 10 for development and 200 for test [1]. Meanwhile, to prevent the potential collapse during model training due to the **exclusive** use of multiple-choice questions, we generate an explanation for the answer and incorporate an equal proportion of general instruction data sampled from alpaca-gpt4-en (Peng et al., 2023), thereby ensuring a more stable IFT.

To make a comprehensive evaluation, for each dataset, we construct three types of test sets: **1) homogeneous test set (HOMO)**, which is held out from the same multiple-choice dataset; **2) in-domain test set (ID)**, which includes data from MMLU (Hendrycks et al., 2021) that belong to the same domain of the dataset; **3) an out-of-domain test set (OOD)**, which comprises instances in MMLU that are from distinct domains with the dataset, as well as ARC (Clark et al., 2018). By observing the performance differences across these three types of test sets, we aim to examine the impact of IFT on various testing scenarios.

3.2 Parameter Knowledge Probing via Few-shot In-context Learning

Our knowledge intervention framework relies on effectively detecting the parameter knowledge of pre-trained LLMs. To this end, this paper leverages few-shot in-context learning, which is a widely-used approach for probing the capabilities and internal knowledge of the base model, to effectively identify the internal knowledge of pre-trained LLMs.

Specifically, we utilize in-context learning to probe the base model's response to each data item in domain multi-choice dataset and regard the response as the model's internal parameter knowledge for this question.

3.3 Construction of Instruction Data

Upon probing the intrinsic knowledge of pre-trained LLMs, we build instances based on the consistency between world knowledge contained within the domain data and parameter knowledge, thereby constructing IFT datasets for different groups. Specifically, the three different groups of data compared in our main experiment including:

[1]The availability of datasets tailored to multiple-choice questions within specialized domains is profoundly limited. Hence, from this benchmark, we select the three domains with the most data.

- **Harmonious setting**, which consists of data where the embedded world knowledge is consistent with model parameter knowledge, which means the pre-trained LLM can answer correctly under in-context learning. In the learning process with this dataset, the model only needs to transfer behavioral norms, without the need to acquire additional world knowledge due to the above consistency.

- **Incompatible setting**, which comprises instances where the pre-trained LLM cannot correctly answer under in-context learning. Due to the complete inconsistency, the model needs to learn both the behavioral norms and the world knowledge during its training phase.

- **Self-aligning setting**, which consists of data in which the queries are exactly the same as those in the incompatible set, but we modify the answers corresponding to each query to match the pre-trained LLM's internal knowledge. Therefore, in this group of data, all responses are incorrect, and the model cannot learn any additional world knowledge.

To ensure a fair comparison, we maintain the same size across the three groups of data.

3.4 Experiment Setup

Base Model We use LLaMA-2-7B, LLaMA-2-13B, LLaMA-2-70B (Touvron et al., 2023) and Mistral-7B (Jiang et al., 2023) as the foundation base model of our experiments.

Training Details We only calculate loss on outputs, setting epoch to 3, learning rate to $2e^{-5}$ and batch size to 256. For Mistral-7B, we set learning rate to $1e^{-5}$ [2]. We use deepspeed zero3 (Rasley et al., 2020) for LLaMA-2-70B and FSDP (Zhao et al., 2023b) for the other three. All experiments are implemented on Nvidia A100-80GB GPUs.

4 Exp-I: Does Learning Domain-specific World Knowledge Matter?

In our first group of experiments, we would like to examine how the additional world knowledge in the IFT corpus can affect LLMs. To this end, we conduct experiments on three types of IFT data including: harmonious, incompatible, and self-aligning data. By observing the performance discrepancies among these groups, we analyze the effects of injecting world knowledge into the IFT.

Finding 1. *When encompassing correct world knowledge, IFT data congruent with model parameter knowledge can lead to superior IFT outcomes.*

To show this, we compare the experimental outcomes under two settings: Harmonious and Incompatible. Note that both datasets contain correct world knowledge, meaning each query- response pair aligns with correct knowledge. Therefore, the core distinction between these two datasets lies in whether the entailed world knowledge is consistent with the internal knowledge of the LLM. Specifically, training on harmonious data only requires learning behavioral norms without the need for learning any additional knowledge, while training on incompatible data requires learning both.

Table 1 reveals that models fine-tuned on the harmonious dataset outperform those fine-tuned on the incompatible dataset across homogeneous, in-domain, and out-of-domain test sets. Specifically, the harmonious setting yields mean performance gains of 11.27%, 14.58%, and 14.57% over the incompatible setting for homogeneous, in-domain, and out-of-domain tests, respectively. The results indicate that utilizing IFT data, which is consistent with model parameter knowledge and does not inject any additional domain-specific world knowledge, yields superior fine-tuned models. This conclusion holds true across homogeneous, in-domain, and out-of-domain evaluations.

Finding 2. *Using IFT data that aligns with parameter knowledge yet is erroneous yields better performance, compared to employing those that are correct but incongruent with parameter knowledge.*

To further investigate the impact of learning domain-specific world knowledge on IFT, we conduct a more direct comparative experiment between the self-aligning group and the incompatible group. Both groups use identical queries which the pre-trained LLM cannot answer correctly under in-context learning but have different responses: the incompatible group's responses are correct, reflecting world knowledge, while the self-aligning group's responses represent model parameter knowledge, are incorrect.

Table 1 compares results between the self-aligning group and the incompatible group. Surprisingly, despite all queries of the self-aligning

[2] Training loss of Mistral-7B using learning rate $2e^{-5}$ does not converge and even spikes.

Eval	Medicine			History			Engineering			Jurisprudence		
	HAR	INC	SELF	HAR	INC	SELF	HAR	INC	SELF	HAR	INC	SELF
LLaMA-2-7B												
HOMO	**40.22**$_{11.77\uparrow}$	28.45	37.00 $_{8.55\uparrow}$	**38.80** $_{9.20\uparrow}$	29.60	33.60 $_{4.00\uparrow}$	**48.40**$_{16.00\uparrow}$	32.40	32.80 $_{0.40\uparrow}$	**37.60** $_{3.60\uparrow}$	34.00	33.20 $_{0.80\downarrow}$
ID	39.82 $_{2.56\uparrow}$	37.26	**41.46** $_{4.20\uparrow}$	**54.30**$_{23.22\uparrow}$	31.08	46.02$_{14.94\uparrow}$	**42.07**$_{11.04\uparrow}$	31.03	26.21 $_{4.82\downarrow}$	**38.86** $_{3.16\uparrow}$	35.70	36.34 $_{0.64\uparrow}$
OOD	53.24 $_{7.30\uparrow}$	45.94	**54.25** $_{8.31\uparrow}$	**52.10**$_{14.79\uparrow}$	37.32	47.87$_{10.55\uparrow}$	**53.64**$_{25.88\uparrow}$	27.75	34.82 $_{7.06\uparrow}$	**51.65**$_{10.16\uparrow}$	41.50	45.41 $_{3.91\uparrow}$
LLaMA-2-13B												
HOMO	**40.83** $_{4.78\uparrow}$	36.05	34.41 $_{1.64\downarrow}$	**48.40**$_{16.00\uparrow}$	32.40	43.60$_{11.20\uparrow}$	**58.00**$_{20.80\uparrow}$	37.20	55.20$_{18.00\uparrow}$	**44.00**$_{11.60\uparrow}$	32.40	37.60 $_{5.20\uparrow}$
ID	**55.43**$_{20.37\uparrow}$	35.06	52.13$_{17.07\uparrow}$	**68.28**$_{22.15\uparrow}$	46.13	64.09$_{17.96\uparrow}$	**45.52**$_{15.86\uparrow}$	29.66	40.00$_{10.34\uparrow}$	**54.77**$_{16.22\uparrow}$	38.55	52.77$_{14.22\uparrow}$
OOD	**69.40**$_{25.47\uparrow}$	43.94	63.57$_{19.63\uparrow}$	**66.49**$_{16.44\uparrow}$	50.05	62.61$_{12.56\uparrow}$	**65.98**$_{23.09\uparrow}$	42.89	64.25$_{21.36\uparrow}$	**65.55**$_{15.03\uparrow}$	50.52	59.71 $_{9.18\uparrow}$
LLaMA-2-70B												
HOMO	**47.95** $_{5.41\uparrow}$	42.54	46.03 $_{3.49\uparrow}$	**59.20**$_{17.20\uparrow}$	42.00	51.60 $_{9.60\uparrow}$	**62.40** $_{7.20\uparrow}$	55.20	57.60 $_{2.40\uparrow}$	**55.20** $_{7.60\uparrow}$	47.60	51.60 $_{4.00\uparrow}$
ID	**65.37** $_{3.97\uparrow}$	61.40	63.11 $_{1.71\uparrow}$	**82.37**$_{11.08\uparrow}$	71.29	81.29$_{10.00\uparrow}$	**55.17**$_{15.86\uparrow}$	39.31	54.48$_{15.17\uparrow}$	**67.69** $_{5.48\uparrow}$	62.21	67.52 $_{5.31\uparrow}$
OOD	**81.62** $_{6.38\uparrow}$	75.24	79.68 $_{4.44\uparrow}$	**80.64** $_{5.16\uparrow}$	75.48	80.41 $_{4.93\uparrow}$	**80.51** $_{3.46\uparrow}$	77.05	79.99 $_{2.94\uparrow}$	**80.33** $_{6.54\uparrow}$	73.79	79.64 $_{5.85\uparrow}$
Mistral-7B												
HOMO	**49.80**$_{15.12\uparrow}$	34.68	35.02 $_{0.34\uparrow}$	**46.80**$_{13.60\uparrow}$	33.20	40.80 $_{7.60\uparrow}$	**59.60**$_{11.20\uparrow}$	48.40	55.20 $_{6.80\uparrow}$	**48.00** $_{9.20\uparrow}$	38.80	43.60 $_{4.80\uparrow}$
ID	**58.17**$_{16.40\uparrow}$	41.77	51.83$_{10.06\uparrow}$	**67.74**$_{38.39\uparrow}$	29.35	50.11$_{20.76\uparrow}$	**44.83**$_{13.80\uparrow}$	31.03	42.07$_{11.04\uparrow}$	**55.21**$_{13.78\uparrow}$	41.43	49.38 $_{7.95\uparrow}$
OOD	**70.98**$_{26.41\uparrow}$	44.57	58.54$_{13.97\uparrow}$	**68.82**$_{24.51\uparrow}$	44.31	62.20$_{17.88\uparrow}$	**66.48**$_{10.50\uparrow}$	55.98	58.62 $_{2.65\uparrow}$	**68.81**$_{11.98\uparrow}$	56.83	66.36 $_{9.53\uparrow}$

Table 1: The performance of the four base LLMs after fine-tuning under incompatible (INC), self-aligning (SELF), and harmonious (HAR) settings. Models fine-tuned with the harmonious dataset and with the self-aligning dataset achieve superior performance compared to those fine-tuned with the incompatible dataset, across all test scenarios including homogeneous (HOMO), in-domain (ID), and out-of-domain (OOD) evaluations. The best/second-best performance in each scenario is highlighted in bold/underline.

group having incorrect answers, models fine-tuned with this data significantly outperform those needing inconsistent world knowledge injection from the incompatible group. The self-aligning group shows a considerable advantage across evaluations: an average performance increase of 5.25%, 9.78%, and 9.67% on homogeneous, in-domain, and out-of-domain evaluations, respectively.

This finding emphatically indicates that injecting additional domain knowledge through IFT also fails to bring effective improvements to LLMs even in homogeneous and in-domain evaluations. Conversely, maintaining consistency with model parameter knowledge, that is, without injecting any additional world knowledge through the self-aligning setting, can yield superior results. Moreover, this advantage holds true across all evaluations, as well as different model sizes and architectures.

The above results demonstrate a significant decline of performance in models fine-tuned using data that contain correct world knowledge but conflict with model parameter knowledge, compared to using consistent IFT data aligned with parameter knowledge. This suggests that injecting additional world knowledge through IFT, in cases where it is inconsistent with the parameter knowledge, is unlikely to yield the benefits we anticipate. Therefore, the core role of IFT may lie in facilitating the transfer of behavioral norms, rather than injecting additional domain-specific world knowledge. To further validate this conclusion, in the next section, we design a novel method to decouple the conflict knowledge contained in IFT data and present further analysis of this issue.

5 Exp-II: IFT with Contextualized Knowledge

In this section, we introduce a new analysis method called contextualized knowledge decoupling to further investigate the impact of inconsistent knowledge during IFT. This approach involves explicitly providing relevant world knowledge needed to answer a query within the context of the query itself. Under this paradigm, the model no longer needs to learn knowledge during IFT, but only needs to use the knowledge in the context to give the answer in the expected behavioral norm. This method helps prevent the model from learning additional world knowledge during IFT, which is inconsistent with parameter knowledge, thus separating knowledge injection from behavioral norm transfer.

To this end, we start with the data from the incompatible group to construct a dataset with contextualized knowledge. Specifically, given an instruction-answer pair in the incompatible group, we employ GPT-3.5[3] to generate the entailed world knowledge that is required to answer this question. The knowledge is then concatenated with the original instruction, as well as the answer, to construct an augmented instance. Finally, we use the con-

[3]https://openai.com/

Model	IFT	HOMO	ID	OOD	Overall
LLaMA-2-7B	Vanilla	31.11	33.77	38.13	34.34
	Contextualized	**37.62**	**43.70**	**50.12**	**43.82**
LLaMA-2-13B	Vanilla	34.51	37.35	46.85	39.57
	Contextualized	**41.47**	**49.48**	**61.37**	**50.77**
Mistral-7B	Vanilla	**38.77**	35.90	50.42	41.70
	Contextualized	37.22	**43.47**	**59.16**	**46.62**
Average	Vanilla	34.80	35.67	45.13	38.53
	Contextualized	**38.77**	**45.55**	**56.89**	**47.07**

Table 2: The performance comparison between models fine-tuned with vanilla IFT and models fine-tuned with contextualized IFT respectively.

structed data to fine-tune LLMs and compare them with the models fine-tuned with vanilla IFT.

Finding 3. *Ensuring that the model does not learn world knowledge conflicting with parameter knowledge during IFT enhances the effectiveness of IFT.*

Table 2 shows the results of contextualized knowledge decoupling in three base models. From the table, it can be observed that fine-tuning the model with data using explicit contextualized knowledge significantly mitigates the adverse effects caused by inconsistencies between parameter knowledge and world knowledge in IFT data. Compared to IFT using incompatible group data, our method achieves an average improvement of 9.48% on LLaMA-2-7B, 11.20% on LLaMA-2-13B, and 4.92% on Mistral-7B. Except in rare cases, IFT using data with contextualized knowledge can significantly improve the effect of IFT across homogeneous, in-domain, and out-of-domain evaluations. we should not force the model to learn additional inconsistent knowledge, when the world knowledge in IFT data is inconsistent with model parameter knowledge. Instead, by decoupling knowledge learning and the transfer of behavioral norms, the problems caused by the above knowledge conflicts can be effectively alleviated. This further verifies our observation in Section 4 that additional world knowledge injection may be ineffective or even harmful for IFT. During IFT, the model should focus on transferring the behavioral norm relying on the existing parameter knowledge and the regularity in the IFT data, rather than learning additional world knowledge. Therefore, for the RQ1, we conclude:

Conclusion 1. *For IFT, there is little, if not even causing additional damage, benefits from the learning of world knowledge incongruent with parameter knowledge.*

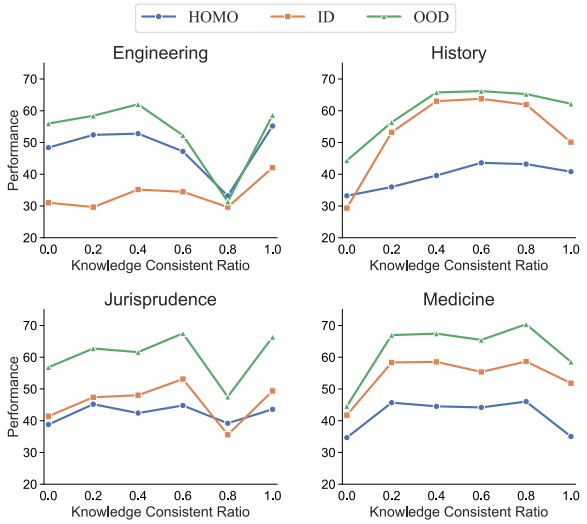


Figure 2: The performance of Mistral-7B fine-tuned with instruction datasets of varying consistency ratios. Each dataset is composed of a mixture of incompatible and self-aligning data, and the consistency ratio represents the proportion of self-aligning samples. The results of other base models are presented in the Appendix B due to page limitations.

6 Exp-III: Is Consistency All You Need?

The experimental results above seem to suggest a conclusion: For better transfer of behavioral norms, we should use IFT data that is completely consistent with model parameter knowledge without additional world knowledge. To further verify this hypothesis, we design a new set of experiments.

Specifically, by adjusting the proportion of samples derived from the incompatible group and self-aligning groups, we aim to adjust the ratios of consistency between the world knowledge in the IFT data and model parameter knowledge, thereby observing the variations in IFT outcomes under different consistency ratios.

Finding 4. *Employing IFT data that is fully consistent with model parameter knowledge does not necessarily result in optimal performance.*

Figure 4 presents the results on three base models under different consistency ratios across four domains. From the figure, we can see that: 1) The performance of the incompatible group is indeed poor, which aligns with our previous conclusion; 2) However, it is also observed that exclusively using IFT data that is consistent with model parameter knowledge does not guarantee optimal performance in the majority of scenarios; 3) Instead, an intermediate setting that incorporates a mix of incompatible and self-aligning data achieves the best performance under most conditions. Further-

more, the optimal ratio of this mixture varies across different base models and domains.

The observations indicate that although the consistency between world knowledge within the IFT data and model parameter knowledge significantly affects the performance of the fine-tuned model, it is not the fundamental factor influencing IFT performance. Consequently, elucidating the true determinants of IFT's effect necessitates further investigation into the underlying mechanisms of knowledge consistency.

7 Exp-IV: Rethinking Consistency: What Really Matters for IFT?

To further explore the underlying mechanisms of IFT, we analyze the knowledge discrepancies of different base LLMs before and after IFT on various IFT data, in order to observe the extent of internal knowledge alternation triggered by IFT. Specifically, for each sample within the test data, we initially compute the Pearson correlation coefficient between the original model's predictions, based on in-context learning, and those provided by the fine-tuned model. Building upon this, we calculate the average Pearson correlation coefficient across each test set and subsequently compare it with the performance on the same test set of the fine-tuned model. Through this experimental analysis, we aim to observe how the knowledge changes induced by IFT affect the ultimate effect of IFT.

To this end, we employ a partial correlation assessment (Spearman, 1961) to analyze the aforementioned Pearson correlation and model performance. Specifically, we utilize a total of 96 models from Exp-III, conducting tests on homogeneous, in-domain, and out-of-domain data. To eliminate the potential influence of the base model's performance on the correlation analysis, ensuring that the results only reflect the differences brought about by varying IFT data, we treat the base model's performance as a control variable. By employing partial correlation assessment under the constraint of removing this control variable's influence, we analyze the correlation between the Pearson correlation coefficients of model predictions before and after IFT and their ultimate performance.

Finding 5. *The consistency of internal knowledge within a model before and after IFT is the key factor affecting the performance of the instruction fine-tuned model.*

Model	HOMO		ID		OOD	
	r	p-value	r	p-value	r	p-value
Mistral-7B	0.78	0.00	0.81	0.00	0.84	0.00
LLaMA-2-7B	0.27	0.14	0.19	0.30	0.29	0.11
LLaMA-2-13B	0.56	0.00	0.78	0.00	0.83	0.00
All	0.43	0.00	0.57	0.00	0.47	0.00

Table 3: The Spearman partial correlation analysis between the model performance after fine-tuning, and the knowledge consistency between base model and fine-tuned model. The analysis is controlled with the base model's performance on each test set. r and $p\text{-}value$ denote partial correlation coefficient and significance respectively. For LLaMA-2-13B and Mistral-7B, p-values significantly lower than 0.05 indicates a high level of confidence.

Model	Best	Self-aligning	Incompatible
Mistral-7B	0.23	0.32	0.35
LLaMA-2-7B	0.17	0.94	0.50
LLaMA-2-13B	0.20	0.68	0.35

Table 4: KL divergence between the prediction distribution of fine-tuned model under zero-shot settings and the prediction distribution of base model using in-context learning probing.

Figure 3 presents the results of grouped regression analysis conducted on the Mistral-7b models. Table 3 displays the correlation coefficients and significance for the partial correlation analysis across different base models and different evaluations. From the aforementioned figure and table, a significant trend can be observed, namely that the correlation of predictions made by models before and after IFT on a given dataset has a substantial impact on the final performance of the fine-tuned models on that dataset. It is noteworthy that this phenomenon holds true across homogeneous, in-domain, and out-of-domain evaluations. This implies that even for homogeneous and in-domain test sets, which are within the same field as the IFT data, the injection of additional world knowledge and the alteration of internal parameter knowledge through IFT does not contribute to enhancing the final performance of the models. Instead, maintaining consistency in the knowledge of models before and after IFT significantly positively influences the performance of the fine-tuned models.

To further investigate whether this finding is the underlying mechanism responsible for Finding 4, we explore the impact of different consistency data on the correlation of model knowledge before and after IFT in Exp-III. We also analyze the probability distribution of model outputs under different IFT data conditions. Table 4 presents the related

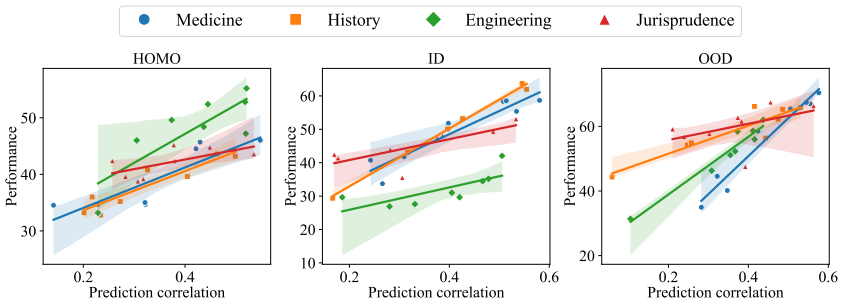


Figure 3: The regression analysis between the model performance after fine-tuning, and the knowledge consistency between base model and fine-tuned model. We show the results of Mistral-7B in three evaluations.

experimental results, which reveal that: IFT using data containing world knowledge completely inconsistent with the parameter knowledge evidently leads to a divergence in the internal knowledge of the fine-tuned model from that of the original model, thereby impairing the performance of the fine-tuned model. Furthermore, IFT using data containing world knowledge completely consistent with the parameter knowledge also may result in a divergence in the knowledge distribution between the original and fine-tuned models. Specifically, fine-tuning with only consistent IFT data may steer the model towards an ad-hoc knowledge distribution, whereas the original model's parameter knowledge exists as a relatively smooth distribution. Conversely, using a middle setting that mixes incompatible and self-aligning data allows the optimization process to maintain model parameter knowledge unchanged while preserving the distribution's smoothness, thereby enabling the fine-tuned model outputs to more closely resemble those of the original model, ultimately yielding better performance.

The experimental results mentioned above demonstrate that the key to the performance of fine-tuned models lies in the consistency of model parameter knowledge before and after IFT. In fact, the core mechanism underlying the superior performance observed in both Finding 1 and Finding 2 for the IFT data is attributed to the maintenance of consistency in model knowledge before and after IFT. Therefore, for RQ2 raised above, we conclude:

Conclusion 2. *The essence of an effective IFT lies in maintaining the consistency of model parameter knowledge before and after IFT.*

8 Conclusion and Discussion

Our experiments and conclusions indicate that the core function of IFT is not to learn domain-specific world knowledge. Instead, learning world knowledge that is inconsistent with model parameter knowledge actually undermines the performance of LLMs from all homogeneous, in-domain to out-of-domain settings. Furthermore, we discover that the consistency of model parameter knowledge before and after IFT (i.e., the knowledge probed through in-context learning before IFT and the knowledge exhibited during question-answering tasks in zero-shot mode after IFT) plays a crucial role in determining the ultimate performance of the fine-tuned model. These two findings unveil a fundamental mechanism of IFT, that is, instruction fine-tuning is not a supervised, domain-specific knowledge *learning* process, but a process of *self-aligning* instruction with the already existing parameter knowledge of LLMs. Therefore, the ultimate determinant of IFT effect is not the extent of domain knowledge injection, but rather whether the IFT process can facilitate more effective self-aligning, thereby enhancing the expression of the model's parameter knowledge under the zero-shot question-answering paradigm after IFT.

Our discovery not only provides guidance for future IFT data construction, model training, and performance evaluation of instruction fine-tuning but also offers solid evidence for a series of very recent works. For example, super-alignment (Burns et al., 2023) aims to use a weaker, smaller model to guide a stronger, larger model towards alignment. Our conclusion proves that it is entirely possible to use a weak model with less knowledge to guide a larger, more knowledgeable model for IFT. Furthermore, recent works on self-play (Chen et al., 2024) have achieved model alignment through the self-training of LLMs, without the introduction of external knowledge. Our conclusions also elucidate the viable underlying factors for such endeavors and furnish a foundation for the future advancement of these works (Yuan et al., 2024;Sun et al., 2023;Lu et al., 2024).

Limitations

In order to facilitate probing model parameter knowledge, we currently focus on multiple-choice questions. In the future, we plan to extend our framework to free-style generation. Besides, due to hardware limitations, the vast majority of experiments are conducted on models with about 10B parameters, we only explore 70B models in several experiments. Repeating our study on larger models in more domains will contribute to a deeper understanding of the IFT for larger models.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. 2023. Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision. ArXiv:2312.09390 [cs].

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models. ArXiv:2401.01335 [cs, stat].

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, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. ArXiv:2210.11416 [cs].

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. ArXiv:1803.05457 [cs].

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases. ArXiv:2306.16092 [cs].

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A Survey on In-context Learning. ArXiv:2301.00234 [cs].

Zhouhong Gu, Xiaoxuan Zhu, Haoning Ye, Lin Zhang, Jianchen Wang, Sihang Jiang, Zhuozhi Xiong, Zihan Li, Qianyu He, Rui Xu, Wenhao Huang, Zili Wang, Shusen Wang, Weiguo Zheng, Hongwei Feng, and Yanghua Xiao. 2023. Xiezhi: An Ever-Updating Benchmark for Holistic Domain Knowledge Evaluation. ArXiv:2306.05783 [cs].

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. ArXiv:2009.03300 [cs].

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2023. OPT-IML: Scaling Language Model Instruction Meta Learning through the Lens of Generalization. ArXiv:2212.12017 [cs].

Joel Jang, Seungone Kim, Seonghyeon Ye, Doyoung Kim, Lajanugen Logeswaran, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2023. Exploring the Benefits of Training Expert Language Models over Instruction Tuning. ArXiv:2302.03202 [cs].

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. ArXiv:2310.06825 [cs].

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. ArXiv:2303.14070 [cs].

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2023. What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning. ArXiv:2312.15685 [cs].

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large Language Models are Superpositions of All Characters: Attaining Arbitrary Role-play via Self-Alignment. ArXiv:2401.12474 [cs].

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In *Conference on Health,*

Inference, and Learning, pages 248–260. PMLR. ISSN: 2640-3498.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction Tuning with GPT-4. ArXiv:2304.03277 [cs].

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pages 3505–3506, New York, NY, USA. Association for Computing Machinery.

Chiyu Song, Zhanchao Zhou, Jianhao Yan, Yuejiao Fei, Zhenzhong Lan, and Yue Zhang. 2023. Dynamics of Instruction Tuning: Each Ability of Large Language Models Has Its Own Growth Pace. ArXiv:2310.19651 [cs].

C. Spearman. 1961. *The Proof and Measurement of Association Between Two Things*. Studies in individual differences: The search for intelligence. Appleton-Century-Crofts, East Norwalk, CT, US. Pages: 58.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-Driven Self-Alignment of Language Models from Scratch with Minimal Human Supervision. ArXiv:2305.03047 [cs].

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. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

Fanqi Wan, Xinting Huang, Leyang Cui, Xiaojun Quan, Wei Bi, and Shuming Shi. 2024. Mitigating Hallucinations of Large Language Models via Knowledge Consistent Alignment. ArXiv:2401.10768 [cs].

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. WizardLM: Empowering Large Language Models to Follow Complex Instructions. ArXiv:2304.12244 [cs].

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-Rewarding Language Models. ArXiv:2401.10020 [cs].

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. Scaling Relationship on Learning Mathematical Reasoning with Large Language Models. ArXiv:2308.01825 [cs].

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2023. R-Tuning: Teaching Large Language Models to Refuse Unknown Questions. ArXiv:2311.09677 [cs].

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023a. A Survey of Large Language Models. ArXiv:2303.18223 [cs].

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. 2023b. Pytorch fsdp: Experiences on scaling fully sharded data parallel.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less Is More for Alignment. ArXiv:2305.11206 [cs].

A Train and Test format

```
# Input
The following are multiple choice questions
about {domain}. Please choose the correct
answer.

{question}
{choices}

# Output
{answer}
Explanation: {explanation}
```

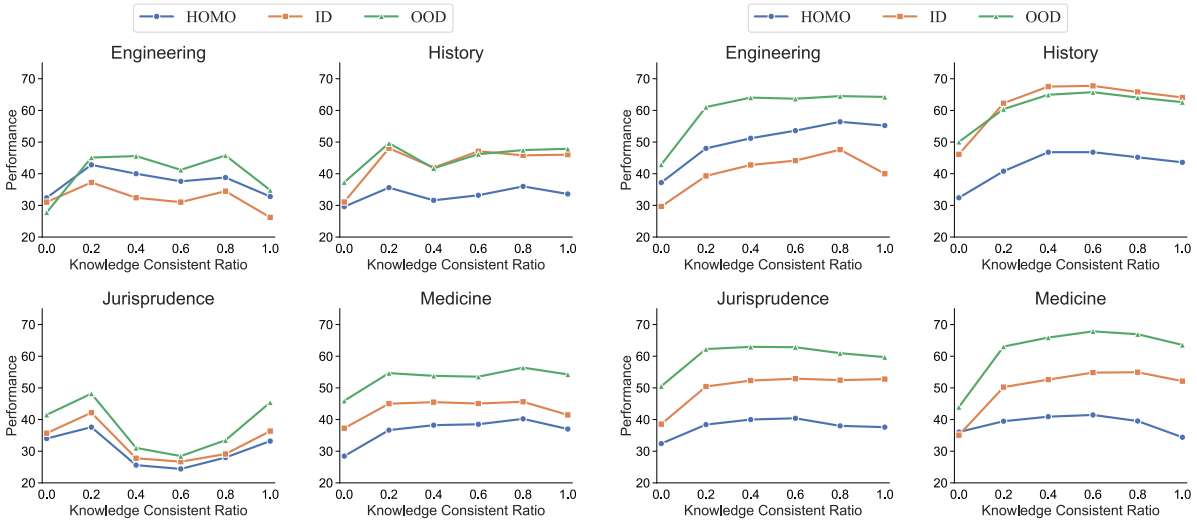
Text 1: The default format of instruction data pairs for input and output.

```
# Input
The following are multiple choice questions
about {domain}. Given the context. Please
choose the correct answer.

{context}
{question}
{choices}

# Output
{answer}
```

Text 2: The format of the instruction data pair input and output of our contextualized IFT.

B The performance of LLaMA-2-7B and LLaMA-2-13B fine-tuned with instruction datasets of varying consistency ratios

In Exp-III, due to page limitations, we only report the performance of Mistral-7B fine-tuned with instruction datasets of varying consistency ratios. In Figure 4, we supplement the results of the other two base LLMs: LLaMA-2-7B and LLaMA-2-13B.



(a) LLaMA-2-7B

(b) LLaMA-2-13B

Figure 4: The performance of LLaMA-2-13B and LLaMA-2-7B fine-tuned with instruction datasets of varying consistency ratios. Each dataset is composed of a mixture of incompatible and self-aligning data, and the consistency ratio represents the proportion of self-aligning samples. Best performance is highlighted with red color. The results of other base models are presented in the Appendix due to page limitations.