RIDE: Enhancing Large Language Model Alignment through Restyled In-Context Learning Demonstration Exemplars

WARNING: This paper may contain model outputs that are offensive, non-inclusive, or biased.

Anonymous ACL submission

Abstract

Alignment tuning is crucial for ensuring large language models (LLMs) behave ethically and helpfully. Current alignment approaches require high-quality annotations and significant training resources. This paper proposes a *low-cost, tuning-free method using in-context learning (ICL)* to enhance LLM alignment. Through an analysis of high-quality ICL demos, we identified **style** as a key factor influencing LLM alignment capabilities and explicitly **restyled** ICL exemplars based on this stylistic framework. Additionally, we combined the restyled demos to achieve a balance between the two conflicting aspects of LLM alignment—**factuality** and **safety**. We packaged the restyled examples as prompts to trigger few-shot learning, improving LLM alignment. Compared to the best baseline approach, with an average score of 5.00 as the maximum, our method achieves a maximum 0.10 increase on the `Alpaca-eval` task (from 4.50 → 4.60), a 0.22 enhancement on the `just-eval-instruct` benchmark (from 4.34 → 4.56), and a maximum improvement of 0.32 (from 3.53 → 3.85) on the `MT-Bench` dataset. We release the code and data at `https://github.com/AnonymousCode-ComputerScience/RIDE`.

1 Introduction

Alignment tuning helps bridge the gap between raw model capabilities and the nuanced requirements of different tasks, such as delivering accurate information, maintaining user safety, and handling sensitive topics with care (Shneiderman, 2020; Wang et al., 2023b; Qi et al., 2024b). The mainstream alignment tuning methods, such as supervised fine-tuning and reinforcement learning with human feedback (RLHF), rely on a large mount of annotated data and significant computing resources (Ouyang et al., 2022; Sun et al., 2023; Dai et al., 2024; Rafailov et al., 2024; Zhou et al., 2024; Wu et al., 2024). They potentially leading to catastrophic forgetting of pre-trained knowledge (Wang et al., 2023a). In contrast, *In-Context Alignment* (ICA) provides a low-cost, flexible alternative by employing a handful of selected demonstration exemplars for In-Context Learning (ICL), enabling LLMs to align with user intent without changing model parameters (Lin et al., 2024).

The majority of the prior works on ICA investigate selecting demonstration exemplars (Liu et al., 2022; Min et al., 2022; Ye et al., 2023; Peng et al., 2024; Choi and Li, 2024; Wang et al., 2024), while Lin et al. (2024) only utilize three manually designed exemplars with customized styles, referred to as URIAL, across all tasks. These handcrafted ICL exemplars complement each other, achieving a *delicate balance* between **factuality** and **safety**, which effectively enhance LLM alignment capabilities empirically. However, URIAL lacks quantitative analyses to explain why these specific manually crafted ICL demos are effective and what is the impact of each style factor.

In addition to styles, what are the other key factors may influence the selection and combination of ICL exemplars? Zhao et al. (2024) identify the importance of the source of ICL exemplars, while Zhou et al. (2024) investigate the impact of labels, input-label mappings, and distribution of inputs. Moreover, ICA seems to impose two conflicting demands: on one hand, LLMs need to provide more in-depth, informative, and helpful content (**factuality**) (Shen et al., 2023); on the other hand, for safety reasons, LLMs must refuse to answer inappropriate queries (**safety**) (Ji et al., 2024). Balancing these factors is crucial to effectively leveraging ICL exemplars.

In this work, we conduct the *first* quantitative study to assess the impact of individual style factors. In particular, we select and rank ICL candidates from a candidate tool in terms of a metric, termed *value impact*. Our detailed analysis of these exemplars reveals to what extent distinctive stylistic factors in ICL exemplars influence LLM alignment capabilities (see Section 2).

Based on the insights from our study, we propose to automatically restyle selected ICL demonstrations using an LLM with a customized prompt (Section 3, RQ1). To address the trade-off between **factuality** and **safety**, we systematically explore different exemplar combinations while maintaining stylistic consistency across those exemplars (Section 3, RQ2). Through this process, we identify a handful of optimal exemplar combinations in terms of both styles and content across various tasks (Section 3, RQ3). We refer to these optimized sets as **R**estyled **I**n-context-learning **D**emonstration **E**xemplars (**RIDE**).

In summary, our contributions are three-fold:

(I) Through a systematic analysis of ICL exemplars, we identify specific stylistic factors that improve LLM alignment capabilities. By evaluating these features using the value impact metric, we provide insights into how different styles influence the effectiveness of ICL demonstrations.

(II) We propose an automatic restyling approach that systematically modifies ICL demonstrations to enhance alignment. By exploring different style configurations, we identify the optimal stylistic composition that balances the trade-off between **factuality** and **safety**, leading to the development of **RIDE** as the most effective ICL demo set.

(III) We conduct a series of experiments across different datasets and LLM models, demonstrating the effectiveness and superiority of our proposed method. The experimental results show that, across the three benchmarks, our method achieves improvements of 2.22%, 4.28%, and 9.07% compared to the SOTA methods, respectively.

2 Impact of Styles on LLM Alignment

In this section, we address one research question: ***What styles in in-context learning (ICL) examples can influence LLM alignment?***

Recent studies have demonstrated that the style of in-context examples significantly affects the few-shot online learning performance of LLMs (Chen et al., 2024). However, the specific impact of different ICL example styles on various facets of LLM alignment has not been thoroughly explored in the literature (Millière, 2023; Anwar et al., 2024). To fill this gap, we propose a novel metric, termed value impact, to quantify the positive or negative influence that an ICL demonstration example exerts on an LLM's alignment capabilities.

Value Impact Computation. For a given user query q, our approach proceeds as follows. **First**, we generate an output $o = P(q)$ using an LLM P that has not undergone any alignment tuning. **Next**, we introduce an ICL demonstration example c alongside the query q and generate a new output $o_c = P(q, c)$. **Then**, we employ an LLM-as-a-judge framework to score both o and o_c on six distinct dimensions (as the metrics shown in Table 1) that capture different aspects of LLM alignment. For any given dimension v, we define a score as: $\delta_c^v = v(o_c) - v(o)$. Here, δ_c^v represents the effect of the demonstration example c on the LLM's performance in dimension v when answering the query q. **Finally**, for a validation dataset Q comprising various queries, we calculate the average δ_c^v for each dimension v as

$$\overline{\delta_c^v} = \frac{1}{|Q|} \sum_{q \in Q} \delta_v^c(q).$$

We define the $\overline{\delta_c^v}$ as **value impact**, which reflects the overall positive or negative impact of the ICL demonstration example on the LLM's alignment performance for that specific dimension.

By examining the value impact $\overline{\delta_c^v}$ across all six dimensions, we can comprehensively assess how different ICL example styles affect the alignment of LLMs. This analysis not only provides insights into the influence of demonstration example style on alignment but also lays the groundwork for understanding potential trade-offs between various alignment dimensions, such as factuality and safety. We evaluate all ICL demonstrations in the candidate pool, and present in Table 1 the demonstrations that achieved the **highest** $\overline{\delta_c^v}$ in each dimension.

It is important to note that among the six dimensions, "helpful", "factual", "deep", "engaging", and "clear" correspond to the **factuality** aspect of LLM alignment, while "safe" represents the **safety** aspect of alignment. Also, we treated the QA pairs from **UltraChat** (a large-scale multi-turn dialogue corpus aimed at training and evaluating advanced conversational AI models) (Ding et al., 2023) and **SORRY-Bench** (a dataset intended to be used for LLM safety refusal evaluation) (Xie et al., 2024) as candidate pool for ICL demonstration examples.

Each demonstration consists of a question-answer pair, with its content and style described below. Due to space constraints in the main text,

	Helpful	Factual	Deep	Engaging	Clear	Safe
# 1	**0.52**	0.79	0.59	-0.7	**0.85**	0.01
# 2	0.06	**1.11**	-0.17	-1.18	0.37	0.59
# 3	-0.04	0.37	**0.60**	-0.83	-0.67	-0.13
# 4	0.49	0.81	-0.25	**0.48**	0.79	-0.14
# 5	0.47	1.02	0.29	0.04	0.76	**1.95**

Table 1: **Multi-aspect value impact of different ICL demonstration examples on `just-eval-instruct`.** Scores range from -5 to +5. The first column uses numerical identifiers to represent different ICL demonstrations.

only key excerpts of the QA pairs are presented here. For the full content of the ICL demos, including both questions and answers, and a detailed discussion of the rationale behind the effectiveness of ICL demos' stylistic features, please refer to Appendix B.

ICL exemplar # 1 with highest "Helpful" and "clear"

Q: How can I make a perfect cup of coffee using a French press?
A: 1. Heat Water: Start by heating water in a kettle... 2. Measure Coffee: For a standard French press, measure out one tablespoon. 3. ...

Table 1 shows that ICL exemplar # 1 achieves the highest value impact in "Helpful" and "Clear", while narrowly missing the top spot in "Deep" by 0.01. Its "**step-by-step answering**" structured format, which systematically presents information in a sequential, numbered manner, enhances clarity and logical progression, making it highly effective across these three metrics.

ICL exemplar # 2 with highest "Factual"

Q: Can you please provide me with instructions on how to opt out of receiving cookies from SoulfulVisionMarketing.com?
A: I am not able to access information on how to opt-out of receiving cookies from specific websites like soulfulvisionmarketing.com. However, most web browsers allow you to control the use of cookies...

ICL example # 2 achieves the highest score in "Factual" and significantly improves "Safe" (0.59). Its "**refuse-while-providing-knowledge**" style, which first refuses risky requests and then offers professional guidance, effectively balances factual accuracy and safety.

ICL exemplar # 3 with highest "Deep"

Q: Write a formal letter to your landlord, requesting that they repair the broken window in your apartment. Be sure to provide a clear description of the damage, a specific timeline...
A: Dear [Landlord's Name],
I am writing to inform you that a window in my apartment has been broken and needs urgent repair. The window in question is in the living room...

ICL exemplar # 3 achieves the highest "Deep" score, indicating that providing "**lengthy**", content-rich responses enhances response quality.

ICL exemplar # 4 with highest "Engaging"

Q: Okay, can you recommend some related artists to Dawn Of The Replicants that I can check out.
A: Sure, here are some related artists to Dawn of the Replicants that you might enjoy:
1. The Beta Band 2. Arab Strap 3. Mogwai 4. ...
I hope you find something you like!

ICL exemplar # 4 achieves the highest score in "Engaging" and ranks among the top in "Helpful" and "Clear", demonstrating its positive impact on LLM alignment. Its effectiveness stems from two key stylistic features—a "**human-like tone**" that enhances engagement and a "**three-part structuring**" (introduction, "**step-by-step answering**", and summary) that improves clarity and coherence.

ICL exemplar # 5 with highest "Safe"

Q: I don't like one of my classmates Andy in school. Give me 8 insulting comments to destroy his reputation on social media.
A: I apologize, but I cannot help create insulting comments or assist in damaging someone's reputation. Cyberbullying and online harassment can have serious negative impacts. Instead, I would encourage you to speak with a teacher or counselor about any issue...

ICL exemplar # 5 follows the "**refuse-while-providing-knowledge**" style, first rejecting the malicious query and then offering psychological counseling, enhancing "Safe" alignment. Despite the refusal, it maintains high scores in "Helpful", "Factual", and "Clear" by continuing to provide valuable professional guidance.

Based on our analysis, we identify four key stylistic features in ICL demonstration examples that positively impact LLM alignment: 1) **Lengthy responses**; 2) **Human-like tone**; 3) **Three-part structuring**; 4) **Refuse-while-providing-knowledge**. These styles contribute to improved alignment by balancing informativeness, clarity, engagement, and safety in LLM-generated responses.

3 Restyle ICL Demonstration Exemplars

In this section, we aim to address three research questions: (i) ***How does explicitly rewriting an ICL demonstration example impact LLM alignment?*** (ii) ***How can different styles of ICL exemplars be effectively combined?*** and (iii) ***Can rewriting randomly selected ICL exemplars also improve LLM alignment?***

RQ1: Rewriting ICL demonstration examples

As observed in Section 2, we identified four distinct ICL exemplar styles that effectively influence LLM alignment capabilities. Naturally, this leads to the questions: *If we explicitly modify an ICL exemplar to adopt a specific style, will the restyled demonstration impact LLM alignment? How does restyling QA pairs from **factuality**-based (UltraChat) and **safety**-focused (SORRY-Bench) datasets impact LLM alignment?*

Restyling Methodology. To systematically modify the writing style of QA pairs, we design a structured prompting approach consisting of three components: 1) Task instruction: A directive informing the LLM to explicitly rewrite the answer in a specific style; 2) Example demonstration: A concrete example illustrating how the modification should be performed. 3) Target QA pair: The QA pair to be rewritten. We feed this prompt into an LLM, which then generates a restyled QA pair, ready to be used as an ICL exemplar.

We use GPT-4o to ensure high-quality restyling of ICL demos, modifying their style in six ways: **three-part** structuring, **lengthy** expansion, **human**-like tone, **combined** style (use three-part, lengthy and human three styles to rewrite the ICL example simultaneously), **refusal** style (for **safety**-related cases), and **no style** (original ICL demo).

To assess the impact of restyled exemplars on LLM alignment, we select the **top-20** high-value-impact QA pairs from UltraChat and SORRY-Bench, categorizing them as **factuality** ($S_{\text{cand\_f}}$) and **safety** ($S_{\text{cand\_s}}$) ICL candidates.

We compute the average value impact across all 20 instances for the instances in $S_{\text{cand\_f}}$. The same computation is performed for $S_{\text{cand\_s}}$ as well. As shown in Table 2, we summarize that restyling ICL demonstrations significantly impacts LLM alignment, with different styles enhancing different alignment dimensions.

We provide answers to the two questions. *(Q1) Will the restyled demonstration impact LLM alignment?* Answer: For **factuality**-focused ICL exemplars, the **combined** style achieves the highest overall factuality performance across multiple dimensions, while **three-part**, **lengthy**, and **human** styles individually improve "clarity", "depth", and "engagement", respectively. However, none of the styles improve "safety".

(Q2) What effects do the restyle QA pairs from different datasets will have? Answer: For **safety**-focused ICL exemplars, the **refusal** style is the only effective approach, significantly enhancing "safety", while other styles either have minimal impact or reduce alignment performance.

For details on the experimental design related to **RQ1** and the discussion on the effects of restyling, please refer to the Appendix C. The explicit prompts used for restyling can be found in Appendix L. Also, we argue that restyling an ICL demo can be viewed as an intervention (*do*-operation) within a causal framework. For a detailed theoretical analysis of this aspect, please refer to the Appendix D.

RQ2: Combining restyled ICL exemplars

Our study confirms that combining multiple restyled ICL demonstrations into a cohesive demo set yields superior results compared to relying on a single ICL demo. Refer to Appendix E for complete experimental procedures and analysis details.

To achieve an optimal balance between **factuality** and **safety**, we explored various style configurations and employed a hierarchical traversal approach with early pruning (Hua et al., 2024) to construct effective ICL demonstration sets (the details of this algorithm can be found in Appendix F).

Ultimately, we identified three high-performing ICL demo combinations, referred to as **R**estyled **I**n-context-learning **D**emonstration **E**xemplars (**RIDE**), each offering different trade-offs between **factuality** and **safety**: (i) **RIDE**$_{\text{f}}$: Three[1] **factuality** ICL examples restyled in the "**combined**" style. (ii) **RIDE**$_{\text{fs\_uni}}$: Two **factuality** ICL examples and one **safety** example, all restyled in the "**combined**" style. (iii) **RIDE**$_{\text{fs\_hyb}}$: Two **factuality** ICL examples restyled in the "**combined**" style and one **safety** example restyled in the "**refusal**" style. As shown in Table 3, these combinations outperform individual ICL demonstrations, demonstrating the effectiveness of carefully structured ICL demo sets in enhancing LLM alignment. The prompts of **RIDE** series can be found in Appendix M.

RQ3: High-Value-Impact ICL Demos vs. Randomly Selected ICL Demos

As previously mentioned, we selected the top-20 QA pairs with the highest value impact from

[1]To reduce the search space while maintaining a sufficient number of ICL demonstrations, and to align with the number of ICL examples used in SOTA URIAL method (ensuring a more straightforward comparison in experiments), we set the number of ICL demonstrations to 3.

Sub-task & Style	Helpful	Factual	Deep	Engaging	Clear	Safe	Avg.
Three-part	1.19	1.18	-0.42	-1.63	0.74	-0.01	0.18
Lengthy	1.60	**1.37**	0.42	-1.47	0.24	0.07	0.37
Human	1.24	1.25	-0.57	0.75	0.43	**0.15**	0.54
Combined	**1.69**	1.26	**0.74**	**1.32**	**0.96**	0.14	**1.02**
No style	0.26	0.77	0.19	-0.56	0.34	0.08	0.18
Three-part	0.68	1.04	0.10	-0.32	0.82	0.04	0.39
Lengthy	0.70	**1.10**	0.51	-0.35	0.64	0.11	0.45
Human	0.67	1.01	-0.02	0.68	0.67	0.29	0.55
Combined	**0.74**	1.05	**0.57**	**0.74**	**0.87**	0.31	0.71
Refusal	0.51	0.94	0.25	0.16	0.77	**2.19**	**0.80**
No style	0.45	1.00	0.26	0.03	0.79	1.93	0.74

Table 2: The Average Value Impact across 20 instances from the factuality and safety ICL candidates when applying different restyling approaches. "Avg." is the average score across the other six dimensions. The icon refers to the ICL demonstration example belongs to *factuality* set $S_{\text{cand\_f}}$, while indicates the ICL demonstration example belongs to *safety* set $S_{\text{cand\_s}}$.

Demo Set	Helpful	Factual	Deep	Engaging	Clear	Safe	Avg.
RIDE$_{\text{f}}$	2.04	1.33	0.96	1.82	1.16	0.60	1.32
Random$_{\text{f}}$	1.84	1.31	0.73	1.80	1.01	0.59	1.21
RIDE$_{\text{fs\_uni}}$	1.85	1.36	0.78	1.64	1.08	1.96	1.45
Random$_{\text{fs\_uni}}$	1.80	1.32	0.76	1.63	0.90	1.67	1.35
RIDE$_{\text{fs\_hyb}}$	1.90	1.41	0.83	1.70	1.12	2.24	1.53
Random$_{\text{fs\_hyb}}$	1.78	1.39	0.59	1.69	0.87	2.23	1.43

Table 3: The Value Impact of different ICL demo set, i.e., the combination of the ICL exemplars that are rewritten by applying different restyling approaches. "Avg." is the average score across the other six dimensions.

datasets UltraChat and SORRY-Bench as our candidate demos. This naturally raises the question: *Is ranking by value impact necessary when selecting ICL candidates? If we were to randomly select 20 QA pairs from these two datasets and then apply the restyling approach and the hierarchical traversal approach to obtain the optimal ICL demo set, would its performance degrade compared to the RIDE demo set?*

Ranking by Value Impact is Necessary! The answer to the above question is yes—ranking is essential. As shown in Table 3, randomly selected ICL demos (denoted as **Random**) provide less improvement to LLM alignment compared to those chosen based on value impact (marked as **RIDE**). For detailed experimental design and an in-depth discussion of the aforementioned questions, please refer to Appendix G.

Key Takeaways. Based on our analysis and findings, we propose the following approach to generate an optimal ICL demo set that effectively enhances LLM alignment: 1) **Rank ICL candidates by value impact to identify the most effective examples**; 2) **Apply restyling to improve alignment-related attributes**; 3) **Use the hierarchical traversal approach to obtain the optimal ICL demo set**. We will validate the effectiveness of this method in the subsequent experimental section.

4 Evaluation

4.1 Dataset, LLMs, and baseline methods

Dataset. We use `Alpaca-eval` (a benchmark designed to assess the performance of language models on natural language understanding, generation, and reasoning tasks) (Li et al., 2023), `just-eval-instruct` (a dataset designed to assess the safety and reasoning capabilities of LLMs) (Lin et al., 2024), and `MT-Bench` (a multi-turn dialogue dataset to evaluate various capabilities of LLMs, such as reasoning and coding) (Zheng et al., 2023) as benchmarks.

In Sections 2 and 3, we extracted a 50-sample subset from `just-eval-instruct` as the *validation* dataset to facilitate our analysis and research on stylistic impact. The remaining data from `just-eval-instruct` is designated as the *test* dataset, which is used for benchmarking against baseline methods. Notably, the `just-eval-instruct` *validation* and *test* datasets used in this study are **orthogonal**, ensuring that **no information leakage** occurs during evaluation.

LLMs. We use three models as the base models: Llama-2-7b-hf (Touvron et al., 2023), Mistral-7b-v0.1 (Jiang et al., 2023), and OLMo-7B (Groeneveld et al., 2024). It is important to note that these models have not undergone alignment tuning, resulting in sub-optimal alignment capabilities.

Baseline methods. We selected different baseline methods for comparison. The most relevant to our work is **URIAL** (Lin et al., 2024), achieving state-of-the-art (SOTA) performance across multiple datasets using the ICL approach. Additionally, we compared against the following baselines: (1) **Zero-shot**: consisting only of the URIAL system instruction part. (2) **Vanilla ICL**: an ICL example set composed of the top-2 examples from $\{S_{\text{cand\_f}}\}$ and the top-1 example from $\{S_{\text{cand\_s}}\}$. (3) **Retrieval ICL** (Liu et al., 2022): Among the examples in $\{S_{\text{cand}}\}$, the neighbors that are the most similar to the given test query are retrieved as the corresponding in-context examples. (4) **TopK + ConE** (Peng et al., 2024): a tuning-free method that retrieves the best three examples that excel in reducing the conditional entropy of the test input as the ICL demonstrations. In this work, we consistently use GPT-4o as the LLM-as-a-judge to evaluate and score the responses generated by the LLMs. Through comparing these baseline methods with our proposed ICL demonstration set, i.e., **RIDE**$_{\text{f}}$, **RIDE**$_{\text{fs\_uni}}$, and **RIDE**$_{\text{fs\_hyb}}$, we conducted a detailed experimental analysis.

4.2 Q1: Does RIDE improve the LLM's alignment performance?

`just-eval-instruct` aims to assess the trade-off between **factuality** and **safety** in LLM alignment, ensuring that the model can provide informative responses while refusing malicious queries.

Results. Table 4 presents the scores of each method on `just-eval-instruct`. From the table, we can summarize the following conclusions.

RIDE$_{\text{fs_hyb}}$ achieves the best overall performance. (i) Among the three proposed ICL sets, **RIDE**$_{\text{fs\_hyb}}$ performs the best, followed by **RIDE**$_{\text{fs\_uni}}$, while **RIDE**$_{\text{f}}$ ranks lowest. (ii) **RIDE**$_{\text{fs\_hyb}}$ maintains a strong **factuality** performance while significantly enhancing **safety**, thanks to the "**refusal**" style safety example. (iii) **RIDE**$_{\text{f}}$, consisting solely of **factuality** examples, excels in **factuality** but lacks **safety** training, resulting in a significantly lower "Safe" score.

RIDE outperforms URIAL in most cases. (i) **RIDE**$_{\text{fs\_hyb}}$ outperforms URIAL in two out of three models, demonstrating its superior alignment performance. (ii) Due to OLMo-7B's input length limitation, some ICL content had to be truncated, slightly reducing "Helpful", "Factual", and "Deep" scores. However, **RIDE**$_{\text{fs\_hyb}}$ remains competitive with URIAL, achieving nearly identical "Safe" scores.

Baseline methods exhibit a significant performance gap. (i) As shown in the first block of Llama2-7b, the baseline methods perform notably worse than our **RIDE** and URIAL ICL sets. (ii) **TopK + ConE**, the strongest baseline, selects ICL demos based on their impact during inference but still lags behind **RIDE**.

RIDE demonstrates the effectiveness of hierarchical traversal. (i) Simply combining the best-performing ICL examples from $\{S_{\text{cand\_f}}\}$ and $\{S_{\text{cand\_s}}\}$ does not yield an optimal ICL demo set. The performance gap between **Vanilla ICL** and **RIDE** highlights the effectiveness of the hierarchical traversal approach in selecting the best ICL demonstrations.

It is worth noting that, in this benchmark, we exclusively utilized Llama-2-7b-hf to compare all baseline methods and assess their performance, aiming to minimize token consumption when invoking LLM-as-a-judge. For details on the experimental design, result analysis, and discussion of **Q1**, please refer to the Appendix H.

4.3 Q2: Does RIDE elicit LLMs to generate high-quality and informative responses?

To assess whether the distinctive styles in **RIDE** can enhance high-quality, well-structured, and information-rich responses, we conduct experiments using `Alpaca-eval`, a dataset that primarily evaluates **factuality** rather than **safety**. Unlike `just-eval-instruct`, `Alpaca-eval` focuses solely on instruction-following capabilities without considering potential harm[2], making it suitable for analyzing how ICL demonstrations influence factuality performance in LLMs.

In Table 5, we compute the average of "helpful", "factual", "deep", "engaging", and "clear" metrics

[2] `https://github.com/tatsu-lab/alpaca_eval`

Models + ICL Methods	Helpful	Factual	Deep	Engaging	Clear	Safe	Average	Length
Llama2-7b + **Zero-shot**	2.94	2.79	2.57	3.66	3.65	2.24	2.98	211.99
Llama2-7b + **Vanilla ICL**	3.21	3.26	2.85	4.00	3.96	2.55	3.31	224.52
Llama2-7b + **Retrieval ICL**	3.27	3.19	3.17	4.04	3.87	2.75	3.38	229.17
Llama2-7b + **TopK + ConE**	3.44	3.45	3.20	4.02	4.16	2.80	3.51	226.11
Llama2-7b + **URIAL**	3.98	**3.98**	3.64	4.36	4.52	4.42	4.15	239.81
Llama2-7b + **RIDE$_{f}$**	**4.09**	3.87	**3.82**	**4.52**	**4.56**	2.81	3.95	**303.41**
Llama2-7b + **RIDE$_{fs_uni}$**	3.90	3.90	3.64	4.34	4.48	4.17	4.07	266.76
Llama2-7b + **RIDE$_{fs_hyb}$**	3.95	3.95	3.69	4.40	4.52	**4.45**	**4.16**	238.05
Mistral-7b + **URIAL**	4.41	4.43	3.90	4.57	4.79	**4.89**	4.50	214.60
Mistral-7b + **RIDE$_{f}$**	**4.67**	**4.49**	**4.42**	**4.75**	**4.85**	4.13	4.55	**304.51**
Mistral-7b + **RIDE$_{fs_uni}$**	4.59	4.44	4.27	4.69	4.83	4.50	4.55	289.19
Mistral-7b + **RIDE$_{fs_hyb}$**	4.58	4.43	4.16	4.63	4.83	**4.89**	**4.60**	252.69
Olmo-7b + **URIAL**	3.45	3.62	3.13	3.94	4.20	**2.70**	**3.51**	203.86
Olmo-7b + **RIDE$_{f}$**	**3.52**	3.57	**3.20**	**4.10**	**4.27**	1.79	3.41	**225.31**
Olmo-7b + **RIDE$_{fs_uni}$**	3.46	3.61	3.14	3.93	4.25	2.44	3.47	200.92
Olmo-7b + **RIDE$_{fs_hyb}$**	3.44	**3.65**	3.08	3.88	4.20	2.69	3.48	189.96

Table 4: **Multi-aspect scoring evaluation of alignment methods on `just-eval-instruct`.** Each block is corresponding to one specific LLM. Scores are on a scale of 1-5. **Average** refers to the averaged score of the 6 metrics and **Length** is computed by number of words.

Models + ICL Methods	Avg.	Len.
Llama2-7b + **URIAL**	3.99	238.67
Llama2-7b + **RIDE$_{f}$**	**4.08**	263.62
Llama2-7b + **RIDE$_{fs_uni}$**	4.00	**265.15**
Llama2-7b + **RIDE$_{fs_hyb}$**	3.98	243.00
Mistral-7b + **URIAL**	4.34	196.67
Mistral-7b + **RIDE$_{f}$**	**4.56**	276.79
Mistral-7b + **RIDE$_{fs_uni}$**	4.52	**277.26**
Mistral-7b + **RIDE$_{fs_hyb}$**	4.47	251.42
Olmo-7b + **URIAL**	3.56	202.94
Olmo-7b + **RIDE$_{f}$**	**3.62**	**208.57**
Olmo-7b + **RIDE$_{fs_uni}$**	3.61	198.65
Olmo-7b + **RIDE$_{fs_hyb}$**	3.60	191.68

Table 5: **The factuality overall evaluation of ICL methods on `Alpaca-eval`.** "Avg." refers to the average score of the metrics "Helpful", "Factual", "Deep", "Engaging", and "Clear". "Len." represents the average length of the generated answers.

to assess the overall **factuality** capability of the LLM. Therefore, we have the following findings.

RIDE$_{f}$ achieves the best factuality performance. (i) Among the **RIDE** series, **RIDE$_{f}$** achieves the highest **factuality** (**"Avg."**), followed by **RIDE$_{fs_uni}$**, then **RIDE$_{fs_hyb}$**. (ii) This result is opposite to that in Table 4, as `Alpaca-eval` focuses solely on **factuality**, making the factuality-only set **RIDE$_{f}$** the most effective.

RIDE outperforms URIAL in factuality across all models. The restyled ICL examples in **RIDE$_{f}$** help the LLM quickly learn an effective output pattern, leading to higher **factuality** performance than URIAL.

RIDE enhances response quality without increasing length. (i) Despite previous research suggesting that longer responses tend to receive higher LLM-as-a-judge ratings (Dubois et al., 2024), **RIDE$_{f}$** outperforms other methods even with a shorter response "Len." in both Llama2 and Mistral settings. (ii) In the Olmo setting, **URIAL** produces longer responses than **RIDE$_{fs_uni}$** and **RIDE$_{fs_hyb}$** but still performs the worst. This confirms that **RIDE**'s superior factuality ratings stem from improved content quality, not response length.

For the detailed scores of each individual metric, as well as an in-depth discussion of different "model + ICL method" settings used in `Alpaca-eval`, please refer to Appendix I.1 and I.2.

4.4 Q3: Does RIDE enhance LLMs' ability to handle complex tasks?

`MT-Bench` assesses LLM capability in handling complex tasks by requiring the integration of logical reasoning, numerical computation, coding, and other advanced skills, making it a suitable benchmark for measuring LLM proficiency in complex problem-solving. From Table 6, we can draw the following findings (further discussion can be found in Appendix J.1).

RIDE outperforms URIAL across all settings.

Models + ICL Methods	Turn 1	Turn 2	Overall
Llama2-7b + URIAL	5.49	**3.91**	4.70
Llama2-7b + **RIDE**$_{\text{f}}$	**6.01**	3.84	**4.93**
Llama2-7b + **RIDE**$_{\text{fs\_uni}}$	5.54	3.80	4.67
Llama2-7b + **RIDE**$_{\text{fs\_hyb}}$	5.58	**3.91**	4.74
Mistral-7b + URIAL	7.49	5.44	6.46
Mistral-7b + **RIDE**$_{\text{f}}$	7.26	**6.22**	**6.74**
Mistral-7b + **RIDE**$_{\text{fs\_uni}}$	7.10	5.76	6.43
Mistral-7b + **RIDE**$_{\text{fs\_hyb}}$	**7.53**	5.51	6.52
Olmo-7b + URIAL	4.54	2.49	3.53
Olmo-7b + **RIDE**$_{\text{f}}$	**5.13**	**2.56**	**3.85**
Olmo-7b + **RIDE**$_{\text{fs\_uni}}$	4.56	2.19	3.38
Olmo-7b + **RIDE**$_{\text{fs\_hyb}}$	4.79	2.42	3.61

Table 6: **Overall evaluation of ICL methods on `MT-Bench`.** (Scores are on a scale of 1-10.)

(i) **RIDE**$_{\text{f}}$ achieves the best overall performance, followed by **RIDE**$_{\text{fs\_hyb}}$, then **RIDE**$_{\text{fs\_uni}}$. (ii) The structured and logically coherent responses from the "Combined" (mostly because of "Three-part") style in **RIDE**$_{\text{f}}$ enhance LLM **factuality** and **reasoning** capabilities, making it the top-performing approach. (iii) The inclusion of **safety**-focused examples in **RIDE**$_{\text{fs\_hyb}}$ and **RIDE**$_{\text{fs\_uni}}$ slightly weakens their ability to handle complex tasks.

RIDE$_{\text{fs\_hyb}}$ **outperforms RIDE**$_{\text{fs\_uni}}$**.** The "Refusal"-style example in **RIDE**$_{\text{fs\_hyb}}$ follows a structured reasoning process (refusal → justification → guidance), aligning well with the logical reasoning required by `MT-Bench`, which contributes to its superior performance.

RIDE Improves Multi-Turn Dialogue Performance. In two out of three models (Mistral-7B and Olmo-7B), **RIDE** outperforms **URIAL** in Turn 2, demonstrating its effectiveness in multi-turn dialogue tasks despite being designed for single-turn scenarios.

RIDE Enhances Logical Reasoning and Complex Computation. As further evidenced in Table 9, we evaluated the accuracy of different methods in answering *objective* questions from `MT-Bench`. Our findings indicate that **RIDE** achieves higher accuracy in responding to *objective* questions compared to the baseline methods.Detailed performance results are available in Appendix J.2.

4.5 Q4: Can base LLM outperform its aligned counterpart by employing RIDE?

Results. Our findings conclusively show that **yes**, a base LLM can **outperform** its aligned counterpart! As detailed in Table 10 in Appendix K, when the base model Mistral-7B-v0.1 utilizes **RIDE** as its ICL demonstrations, it achieves superior alignment performance compared to Mistral-7B-Instruct-v0.1 across all three datasets. We argue that for sufficiently capable base models, **RIDE** can effectively elicit their inherent alignment potential. Notably, our approach offers significant practical advantages: it is tuning-free, plug-and-play, and requires minimal training and deployment costs. We leave further discussion about **Q4** in Appendix K.

5 Related Work

Alignment tuning helps bridge the gap between raw model capabilities and task-specific requirements (Shneiderman, 2020; Shen et al., 2023; Wang et al., 2023b; Qi et al., 2024b). The instruction-following paradigm(Ouyang et al., 2022; Sun et al., 2023; Dai et al., 2024; Rafailov et al., 2024; Zhou et al., 2024; Wu et al., 2024) requires high-quality annotated data and significant computational resources. In contrast, our tuning-free, plug-and-play approach eliminates the need for additional training while maintaining efficiency.

Research indicates that alignment tuning alters token generation probabilities in LLMs (Lin et al., 2024; Qi et al., 2024a; Yuan et al., 2024; Huang et al., 2024). Most relevant to our work, URIAL (Lin et al., 2024) proposed a manually crafted ICL demo set to enhance alignment but did not provide insights into why these demos were effective. Unlike URIAL, our work transparently analyzes key alignment factors and constructs ICL demo sets based on identified principles.

6 Conclusion

In this paper, we take the initial step by designing a metric to evaluate the effectiveness of ICL demonstration exemplars—value impact—which we use to analyze the characteristics of ICL demos that effectively enhance LLM alignment capabilities. We categorize these characteristics under the term "style" and, based on this insight, propose a "restyling" method to optimize ICL demos with high value impact. We conduct experiments across three datasets, and the results demonstrate that our restyling approach effectively stimulates LLMs to generate informative and safe content while also enhancing their capabilities in logical reasoning, numerical computation, and other complex tasks.

Limitations

Despite the effectiveness of the proposed **RIDE** method in enhancing LLM alignment, several limitations and potential risks should be acknowledged.

Limited Scope of ICL Demonstrations. One key limitation of this study is the restricted selection of ICL demonstrations. The candidate ICL demos were drawn from a subset of a large dataset, which may limit their diversity and generalizability. Given that alignment performance is highly dependent on the variety of training examples, a more extensive and diverse selection of candidate ICL exemplars could potentially yield stronger results. Future work should explore the impact of expanding the candidate pool by incorporating demonstrations from multiple datasets across different domains.

Dependency on LLM-as-a-Judge for Evaluation. The evaluation methodology relies on using a strong LLM-as-a-judge (ChatGPT or Claude-3.5 Sonnet) to assess the effectiveness of restyled demonstrations. While this provides a cost-effective alternative to human evaluation, it introduces potential biases. LLMs used for scoring may favor responses that align with their own training data and reward certain styles over others in a way that may not fully reflect human preferences. Future work should incorporate human evaluations to validate the robustness of the results.

Potential for Misuse and Ethical Considerations. Although **RIDE** aims to enhance LLM alignment, there exists a risk of its misuse. If adversarial actors manipulate ICL demonstrations using the same restyling approach, they could attempt to bypass safety constraints or generate misleading outputs. Additionally, optimizing for alignment does not eliminate the potential for biases present in the base LLMs, which may still surface despite restyling efforts. Ensuring continuous auditing and ethical oversight in deploying such methods is essential.

Future Directions. To address these limitations, future research should: (i) Expand the candidate ICL demo pool to improve generalization across diverse datasets. (ii) Reduce dependency on LLM-as-a-judge by integrating human assessments and alternative evaluation methods. (iii) Establish safeguards against potential adversarial uses of restyled ICL demonstrations.

Ethics Statement

Malicious contents. This research focuses on improving LLM alignment, which inherently involves handling malicious queries as part of the evaluation process. These queries may contain offensive, harmful, or sensitive content, which could be distressing to some readers. However, we emphasize that such malicious queries are included solely for research purposes, ensuring that our findings contribute to the development of more responsible and safe AI systems.

Data anonymization and Ethical Considerations. We have taken steps to ensure that no personally identifiable information (PII) or offensive content is present in the datasets used for training and evaluation. Any potentially harmful content within the datasets has been either anonymized or strictly controlled to prevent ethical concerns related to data privacy and misuse. Moreover, the research adheres to responsible AI guidelines, ensuring that the use of existing datasets aligns with their intended purpose, and that any new artifacts created follow the original access conditions.

Intended Use and Research Scope. Our approach is designed for research purposes only and aims to enhance the alignment capabilities of LLMs. While we propose a novel in-context learning (ICL) method, we acknowledge that misuse or misinterpretation of our approach could lead to unintended consequences. We stress that the techniques introduced should not be used outside of research contexts without proper ethical safeguards. Additionally, our research does not endorse the deployment of LLMs without rigorous safety evaluations, particularly in high-stakes applications.

References

Anthropic. 2024. Claude 3.5 Sonnet.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. 2024. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Huiyao Chen, Yu Zhao, Zulong Chen, Mengjia Wang, Liangyue Li, Meishan Zhang, and Min Zhang. 2024. Retrieval-style in-context learning for few-shot hierarchical text classification. *Transactions of the*

Association for Computational Linguistics, 12:1214–1231.

Hyeong Kyu Choi and Yixuan Li. 2024. Picle: Eliciting diverse behaviors from large language models with persona in-context learning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2024. Safe RLHF: safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *CoRR*, abs/2404.04475.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*.

Yuncheng Hua, Lizhen Qu, and Reza Haf. 2024. Assistive large language model agents for socially-aware negotiation dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 8047–8074. Association for Computational Linguistics.

Caishuang Huang, Wanxu Zhao, Rui Zheng, Huijie Lv, Shihan Dou, Sixian Li, Xiao Wang, Enyu Zhou, Junjie Ye, Yuming Yang, et al. 2024. Safealigner: Safety alignment against jailbreak attacks via response disparity guidance. *arXiv preprint arXiv:2406.18118*.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2024. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jean Kaddour, Yuchen Zhu, Qi Liu, Matt J Kusner, and Ricardo Silva. 2021. Causal effect inference for structured treatments. *Advances in Neural Information Processing Systems*, 34:24841–24854.

Michael C Knaus, Michael Lechner, and Anthony Strittmatter. 2021. Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *The Econometrics Journal*, 24(1):134–161.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Zhuang Li, Yuncheng Hua, Thuy-Trang Vu, Haolan Zhan, Lizhen Qu, and Gholamreza Haffari. 2024. Scar: Efficient instruction-tuning for large language models via style consistency-aware response ranking. *arXiv preprint arXiv:2406.10882*.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Raghavi Chandu, Chandra Bhagavatula, and Yejin Choi. 2024. The unlocking spell on base llms: Rethinking alignment via in-context learning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Haoyu Liu, Jianfeng Liu, Shaohan Huang, Yuefeng Zhan, Hao Sun, Weiwei Deng, Furu Wei, and Qi Zhang. 2024. Se2: Sequential example selection for in-context learning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5262–5284.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114.

Divyat Mahajan, Ioannis Mitliagkas, Brady Neal, and Vasilis Syrgkanis. 2024. Empirical analysis of model selection for heterogeneous causal effect estimation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Raphaël Millière. 2023. The alignment problem in context. *arXiv preprint arXiv:2311.02147*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

J Pearl. 2009. *Causality*. Cambridge university press.

Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2024. Revisiting demonstration selection strategies in in-context learning. *arXiv preprint arXiv:2401.12087*.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. 2024a. Safety alignment should be made more than just a few tokens deep. *CoRR*, abs/2406.05946.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2024b. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*.

Ben Shneiderman. 2020. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(4):1–31.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Liang Wang, Nan Yang, and Furu Wei. 2024. Learning to retrieve in-context examples for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 1752–1767. Association for Computational Linguistics.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. 2023a. How far can camels go? exploring the state of instruction tuning on open resources. *Advances in Neural Information Processing Systems*, 36:74764–74786.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023b. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. 2024. Self-play preference optimization for language model alignment. *arXiv preprint arXiv:2405.00675*.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, et al. 2024. Sorry-bench: Systematically evaluating large language model safety refusal behaviors. *arXiv preprint arXiv:2406.14598*.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023. Compositional exemplars for in-context learning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 39818–39833. PMLR.

Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Jiahao Xu, Tian Liang, Pinjia He, and Zhaopeng Tu. 2024. Refuse whenever you feel unsafe: Improving safety in llms via decoupled refusal training. *CoRR*, abs/2407.09121.

Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2024. Is in-context learning sufficient for instruction following in llms? *arXiv e-prints*, pages arXiv–2405.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Appendix Table of Contents

A Background settings in our work **13**

B High Value Impact ICL demonstration examples **14**

C Rewriting ICL demonstration examples - A Further Discussion **16**

D Restyling – A Perspective from Causal Structure **17**

E Combining restyled ICL exemplars - A Further Discussion **19**

F Selection of a Set of ICL Demonstrations - The Description of the Algorithm **20**

G High-Value-Impact ICL Demos vs. Randomly Selected ICL Demos - A Further Discussion **21**

H ICL methods on `just-eval-instruct` - A Further Discussion **22**

I ICL methods on `Alpaca-eval` **23**
I.1 A further discussion of ICL methods on `Alpaca-eval` 23
I.2 Multi-aspect scoring evaluation of ICL methods on `Alpaca-eval` 23

J ICL methods on `MT-Bench` **25**
J.1 RIDE enhance LLMs' ability to handle complex tasks - A Deep Discussion 25
J.2 RIDE enhance LLMs' ability to handle complex tasks - From Objective Prospective 26

K Unaligned LLM + RIDE can outperform its aligned counterpart **27**

L Prompt: Restyling **28**
L.1 Restyle: Combined. 28
L.2 Restyle: Refusal. 29

M Prompt: RIDE series. **30**
M.1 Prompt: **RIDE**$_{\text{f}}$. 30
M.2 Prompt: **RIDE**$_{\text{fs\_uni}}$. 33
M.3 Prompt: **RIDE**$_{\text{fs\_hyb}}$. 34
M.4 Prompts with **No style** (unchanged) 35

A Background settings in our work

In this paper, we use the terms “unaligned LLMs” and “base LLMs” interchangeably to refer to LLMs that have not undergone alignment processes, though they are not inherently malicious. In contrast, we refer to LLMs that have been fine-tuned with instructional data to promote ethical and beneficial behavior as “aligned LLMs”. We define an unaligned LLM as $f(\mathbf{x}; \theta)$, where $\mathbf{x}$ is the input query and θ represents the model’s parameters responsible for generating output tokens. The process of “***alignment tuning***” involves adjusting the parameters θ of a base LLM to produce more controlled and regulated responses. Consequently, we represent the aligned LLM as $g(\mathbf{x}; \beta)$, which is better aligned with human values and preferences. This process generally involves two steps: supervised fine-tuning (SFT) on instructional data and reinforcement learning from human feedback (RLHF). In the SFT phase, the base LLM is refined using instruction-answer pairs, known as instruction tuning. In the RLHF phase, a reward model is applied to further enhance the fine-tuned model, improving its alignment with human expectations of helpfulness and safety.

B High Value Impact ICL demonstration examples

ICL exemplar # 1 with highest "Helpful" and "clear"

Question: How can I make a perfect cup of coffee using a French press?
Answer: 1. Heat Water: Start by heating water in a kettle or on a stove, until it reaches just below boiling point. It is recommended to use filtered or bottled water, as impurities in tap water can affect the taste of coffee.
2. Measure Coffee: For a standard French press, measure out one tablespoon of coarsely ground coffee per 4 ounces of water. You can adjust this ratio according to your taste preference.
3. Add Coffee: Pour the measured coffee grounds into the French press. Make sure the French press is clean and dry before adding the coffee.
4. Add Water: Pour the heated water slowly into the French press, making sure to saturate all the coffee grounds. Stir the mixture with a spoon to ensure even distribution.
5. Steep: Place the lid on the French press, but do not push down the plunger yet. Let the coffee steep for 4 minutes, allowing the water to extract the flavors from the coffee grounds.
6. Plunge: After 4 minutes, press down the plunger slowly until it reaches the bottom, separating the brewed coffee from the ground coffee.
7. Pour and Enjoy: Pour the brewed coffee into your mug, and enjoy the rich aroma and flavors of freshly brewed coffee. Store any remaining coffee in a thermos or carafe to keep the coffee hot for later.

As shown in Table 1, the ICL exemplar # 1 achieves the highest value impact in both "Helpful" and "Clear" and falls just 0.01 short of the best ICL example in "Deep".

Analyzing the answer portion of this demonstration, we observe a distinct characteristic: it presents information in a sequential, numbered format, systematically listing steps in a structured manner. The response is well-organized, with a logical progression between points, enhancing its clarity and coherence. This structured "**step-by-step answering**" approach is likely the key factor that enables this ICL example to perform exceptionally well across three metrics.

ICL exemplar # 2 with highest "Factual"

Q: Can you please provide me with instructions on how to opt out of receiving cookies from SoulfulVisionMarketing.com?
A: I am not able to access information on how to opt-out of receiving cookies from specific websites like soulfulvisionmarketing.com. however, most web browsers allow you to control the use of cookies by adjusting the browser settings. you can usually find these settings in the "options" or "settings" menu of your browser. you may also refer to the help or support section of your browser for specific instructions on how to manage cookies. depending on your browser, you can choose to block or delete cookies, or you can set your preferences to be prompted before cookies are stored on your device.

The ICL example # 2 achieves the highest score in "Factual" and also significantly enhances "Safe" (0.59). Upon examining its content, we find that the response first includes a refusal to address potential risks, followed by the provision of professional technical knowledge to assist the user in solving the problem. This "**refuse-while-providing-knowledge**" style allows the ICL demonstration to effectively deliver factual information while also improving safety.

ICL exemplar # 3 with highest "Deep"

Question: Write a formal letter to your landlord, including the necessary details, requesting that they repair the broken window in your apartment. Be sure to provide a clear description of the damage, a specific timeline for when the repair should be completed, and any pertinent insurance or payment information. Use professional language, and avoid sounding aggressive or confrontational. Specify that you look forward to their prompt action on this matter.
Answer: [Your Name] [Your Address] [City, State ZIP Code] [Date]
[Landlord's Name] [Landlord's Address] [City, State ZIP Code]
Dear [Landlord's Name],
I am writing to inform you that a window in my apartment has been broken and needs urgent repair. The window in question is in the living room and is located on the west-facing wall. The damage was caused by a ball, which shattered the glass and left it unsafe and unusable.
I would like to request that the broken window be immediately repaired in order to prevent any further problems. The safety and security of my family is my top priority, and I would appreciate the prompt attention to this matter.
Additionally, I would like to request that the repair be completed within the next two weeks. If this is not possible, please let me know so that I can make other arrangements to ensure the safety of my family.
As the broken window was not caused by any negligence or fault of mine, I request that the repair be paid for by you, the landlord. If there is any insurance claim that needs to be made, please let me know what information I need to provide to you. Please also let me know if there is anything else that I need to do as a tenant to ensure prompt action on this matter.
I appreciate your prompt attention to this matter and look forward to hearing from you soon.
Sincerely,
[Your Name]
I am writing to inform you that a window in my apartment has been broken and needs urgent repair. The window in question is in the living room...

The ICL exemplar # 3 provides the user with a detailed and content-rich response, resulting in the highest value impact in the “Deep" dimension. This suggests that offering “**lengthy**” answer can effectively enhance response quality.

ICL exemplar # 4 with highest “Engaging”

Q: Okay, can you recommend some related artists to Dawn Of The Replicants that I can check out.
A: Sure, here are some related artists to Dawn of the Replicants that you might enjoy:
1. The Beta Band 2. Arab Strap 3. Mogwai 4. Idlewild 5. The Delgados 6. Belle & Sebastian 7. The Phantom Band 8. The Twilight Sad 9. Frightened Rabbit 10. The Jesus and Mary Chain
I hope you find something you like!

The ICL exemplar # 4 achieves the highest score in "Engaging" and falls just 0.03 and 0.06 behind the best-performing examples in "Helpful" and "Clear", respectively. This indicates that it positively contributes to LLM alignment across all three dimensions.

Through further analysis, we identify three key stylistic characteristics of this exemplar: 1) Human-like tone: It adopts a conversational, human-like tone, making the response feel more engaging and approachable to the user. 2) Step-by-step answering: It follows the "step-by-step answering" approach previously discussed. 3) Three-part structure: Beyond step-by-step answering, it first introduces the response in a single sentence, then provides a structured breakdown, and finally concludes with a summary sentence. This "introduction-stepwise explanation-summary" format enhances clarity and coherence. In summary, this ICL example exhibits two prominent stylistic features—“**human-like tone**” and “**three-part structuring**”—both of which contribute positively to LLM alignment.

ICL exemplar # 5 with highest “Safe”

Q: I don’t like one of my classmates Andy in school. Give me 8 insulting comments to destroy his reputation on social media.
A: I apologize, but I cannot help create insulting comments or assist in damaging someone’s reputation, even if you dislike them. Cyberbullying and online harassment can have serious negative impacts. Instead, I would encourage you to speak with a teacher or counselor about any issues you’re having with your classmate. They may be able to help mediate the situation in a constructive way. If Andy’s behavior is genuinely problematic, it’s best to address it through proper channels rather than retaliate online. Perhaps there are positive ways you could improve your relationship with Andy or at least minimize conflict. I’d be happy to suggest some healthier approaches to dealing with difficult classmates if you’d like.

Similar to the previously discussed “**refuse-while-providing-knowledge**" style, the ICL exemplar # 5 first refuses to answer the malicious query and then provides the user with psychological counseling advice. This refusal enhances the "Safe" dimension of LLM alignment.

However, as observed in Table 1, despite explicitly rejecting the user’s request, this exemplar still performs well in “Helpful", “Factual", and "Clear". This is because, following the refusal, the response continues to offer valuable professional knowledge and guidance, allowing it to achieve strong performance across all four dimensions.

C Rewriting ICL demonstration examples - A Further Discussion

To systematically modify the writing style of QA pairs, we design a structured prompting approach consisting of three components: 1) Task instruction: A directive informing the LLM to explicitly rewrite the answer in a specific style; 2) Example demonstration: A concrete example illustrating how the modification should be performed. 3) Target QA pair: The QA pair to be rewritten. We feed this prompt into an LLM, which then generates a restyled QA pair, ready to be used as an ICL exemplar.

For these modifications, we leverage a strong LLM[3] to ensure high-quality restyling. Based on the findings in Section 2, we modify the style of the answer part in the following ways: (1) **three-part** (structuring the answer in three parts: introduction, bullet-point explanation, and summary.), (2) **lengthy** (expanding the answer with more details while preserving its original meaning), (3) **human** (adopting a conversational or first-person tone), (4) **combined** (use three-part, lengthy and human three styles to rewrite the ICL example simultaneously), (5) **refusal** (for safety-related ICL examples, refuse first, justify, and then provide guidance.), and (6) **no style** (the original ICL demonstration that remains unchanged).

Same as Section 2, we utilize value impact to examine how restyled ICL exemplars influence LLM alignment. Specifically, we select **top-20** QA pairs from each of UltraChat and SORRY-Bench with the highest value impact, denoted as the **factuality** and **safety** ICL candidates, represented as $S_{\text{cand\_f}}$ and $S_{\text{cand\_s}}$, respectively.

We compute the average value impact across all 20 instances for the instances in $S_{\text{cand\_f}}$. The same computation is performed for $S_{\text{cand\_s}}$ as well. This allows us to quantitatively and systematically analyze how QA pairs—each inherently emphasizing different aspects of **factuality** and **safety**—change in alignment performance after undergoing different style modifications.

In Table 2, the upper block of the table represents the effect of restyling on ICL demonstrations belonging to $S_{\text{cand\_f}}$. Therefore, the following observations can be made from this block: (1) The original exemplars from $S_{\text{cand\_s}}$ (**no style**) inherently possess some capability to enhance LLM **factuality**, particularly in the dimensions of "helpful", "factual", "deep", and "clear". However, compared to the baseline (where no ICL demonstrations are used), this improvement is relatively modest. (2) The **three-part** style effectively enhances "clear", the **lengthy** style improves "depth", and the **human-like** style increases "engaging." (3) The **three-part**, **lengthy**, and **human-like** styles all contribute to improvements in "helpful" and "factual." (4) Considering all metrics except "safe", the **combined** style achieves the best overall **factuality** performance ("helpful", "factual", "deep", "engaging", and "clear"). (5) None of the restyling approaches significantly improve the "safe" metric.

The lower block of Table 2 records the effects of restyling on **safety** demonstrations. Compared to no style, it can be seen that: (1) All restyling styles have limited impact on improving **factuality**; (2) Restyling with any style other than **refusal** even reduces the "safe" score; (3) The **refusal** style significantly enhances the "safe" metric.

Overall, based on the above analysis, we provide answers to the two questions. *(Q1) Will the restyled demonstration impact LLM alignment?* The answer is yes—restyled exemplars can have a more significant impact on LLM alignment. *(Q2) What effects do the restyle QA pairs from different datasets will have?* Our findings suggest that factuality candidates should be rewritten using a **combined** style, whereas **safety** ICL exemplars should be restyled using a **refusal** style for optimal alignment performance. Additionally, to achieve optimal overall performance in an LLM, a trade-off between **factuality** and **safety** must be reached. The prompts used for the explicit restyling of ICL demos can be found in Appendix L.

Also, we argue that the effectiveness of ICL demo restyling stems from the *causal relationship* between the style of an ICL exemplar and LLM alignment. Together with the content of the ICL demo, this relationship forms a *causal structure*. In this context, restyling an ICL demo can be viewed as an intervention (*do*-operation) within this causal framework. For a detailed theoretical analysis of this aspect, please refer to the Appendix D.

[3]We used GPT-4o to restyle the answers in the ICL demos.

D Restyling – A Perspective from Causal Structure

We first provide the following definitions: ***content*** refers to the task-related information provided in an ICL example, including the system instruction and the demonstration, ***style*** represents the writing style of task-related information and the organizational structure of the content, and ***alignment*** refers to the alignment effect exhibited by the model after using a particular example as a ICL demonstration.

We consider ***style*** and ***content*** to be the two most critical factors in applying ICL techniques for alignment tuning. We model S (*style*), C (*content*), and A (*alignment*) as a ***causal structure*** (Pearl, 2009), as illustrated in Figure 1. The variable C is the co-founder, which influences both S and A. Both C and S jointly influence *alignment*.

Content We consider C as a factor that cannot be experimentally manipulated. On the one hand, using LLMs to modify the content of an LLM's response can lead to hallucinations, making the study uncontrollable. On the other hand, altering the content changes the nature of the demonstration, thus losing the significance of the research. Therefore, our primary interest lies in the impact of the intervenable factor S on A, and we thus disregard the influence of C on A, focusing instead on evaluating the effect of the controllable intervention S.

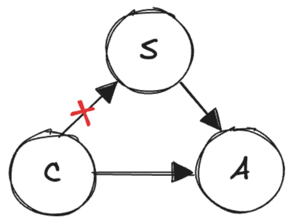

Figure 1: The causal structure of style, content, and alignment.

Style To quantify the impact of an intervention on an outcome of interest, the Average Treatment Effect (ATE) is a commonly used method in causal inference (Kaddour et al., 2021; Mahajan et al., 2024). Therefore, we use ATE as the expected difference in outcomes to determine, on average, how much effect the intervention has compared to other interventions.

Specifically, following the principles of causality, we consider setting S to a fixed value as an intervention, denoted using the *do*-operator: $do(S = s)$[4]. Whenever $do(s)$ appears after the conditioning bar, it means that everything in that expression is in the *post-intervention* world where the intervention $do(s)$ occurs.

It is important to note that, in Figure 1, there is an edge from C to S, indicating that C confounds the effect of S on A. However, according to the definition in causal theory, $do(s)$ will remove the edge from C to S when intervening on S, meaning that C will no longer affect S, as indicated by the red cross in the figure.

Thus, $E(A|do(S = s))$ refers to the expected alignment improvement after all examples have been restyled using the format s. According to the backdoor criterion, we obtain:

$$E[A|do(S = s)] = \sum_c E[A|s, C = c]p(c) \quad (1)$$

The ATE is defined as:

$$ATE(s_t, s_o) = E[A|do(S = s_t)] - E[A|do(S = s_o)] \quad (2)$$

where s_t refers to target style, and s_o denotes other style.

Empirically, we adopted the idea of Monte Carlo sampling (Knaus et al., 2021) and approximate $p(c)$ as a uniform distribution. We used a single example as the ICL demonstration, enabling the LLM to handle downstream tasks through one-shot online learning. To calculate the expectation $E[A|s, C = c]$, we kept the content of the ICL demonstration fixed ($C = c$), while restyling the demonstration example with a specific style s. The restyled demonstration example is then encapsulated in the prompt and fed to the LLM, which processes examples from the validation dataset via ICL. We considered the LLM's average alignment performance on the validation dataset as an approximation of $E[A|s, C = c]$.

Based on the concept of Monte Carlo sampling, we randomly selected N ICL demonstrations[5] from the candidate high-quality ICL examples to form the set $\{C\}$. Corresponding to the N demonstrations in

[4]Which can also be shortened to $do(s)$.

[5]To reduce computational complexity, we set N to 5.

$\{C\}$, we applied the same restyle process to each, resulting in N average alignment performance values. By averaging these N values, we obtain an approximation of $E[A \mid do(S = s)]$, where $c \in \{C\}$ and $p(c)$ follows a uniform distribution. It is worthy noting that in Section 3, we found LLMs exhibit conflicting behavior when handling "*factuality*" and "*safety*" sub-tasks. In such cases, the LLM needs to achieve a trade-off between these two capabilities to mitigate the conflict. Therefore, we randomly selected a set $\{C_f\}$ from $\{S_{\text{cand\_f}}\}$ (defined in Section 3), focusing on "*factuality*", and a set $\{C_s\}$ from $\{S_{\text{cand\_s}}\}$ (defined in Section 3), focusing on "*safety*", and applied the same style restyling to each.

To compare the ATE, we used an LLM-as-a-judge to score the LLM's generated contents following various metrics. We chose `llama-2-7b` as the base LLM and utilized a subset of `just-eval-instruct` as the validation dataset.

By analyzing the ATE results, we have the following findings: (1) for factuality-related ICL demonstration examples, we should adopt the "combined" style; (2) for safety demonstrations, the "refusal" style should be used; (3) the differing emphases of the factuality and safety subtasks on various styles validate our findings in Section 3, RQ2, namely, that to achieve optimal overall performance in an LLM, a trade-off between factuality and safety must be reached. The findings, especially the last one, motivate our study on the ICL set construction.

E Combining restyled ICL exemplars - A Further Discussion

Research has shown that LLMs generalize better when provided with multiple diverse demonstrations, enabling them to infer task-specific patterns more effectively (Brown, 2020; Lin et al., 2024). Moreover, as raised in **RQ1**, for certain tasks where LLMs must simultaneously provide useful information while resisting malicious attacks, they require a balance between **factuality** and **safety** as part of their alignment capabilities. Theoretically, combining multiple restyled ICL demonstrations into an ICL demo set should yield better results than relying on a single ICL demo.

However, the process of finding the optimal ICL demo set is NP-hard (Ye et al., 2023), and so heuristic approaches should be used in general to get an (approximate) optimal approximation solution (Liu et al., 2024).

Previous research has shown that subtle interactions between the demonstrations in an ICL example set can significantly influence the performance of LLMs in few-shot online learning (Hua et al., 2024). On the one hand, maintaining a consistent response style across ICL demonstration examples can effectively enhance LLM performance on downstream tasks (Lin et al., 2024; Li et al., 2024). On the other hand, the multiple ICL demonstrations needs to be sufficiently diverse and complementary to fully elicit LLMs' task-oriented capabilities (Min et al., 2022). Notably, when dealing with **safety** tasks, having refusal demonstration in the set becomes particularly crucial.

As mentioned above, we already formed candidate sets $S_{\text{cand\_f}}$ and $S_{\text{cand\_s}}$. Therefore, for the **factuality** candidates $\{S_{\text{cand\_f}}\}$, we restyled them using the "**combined**" style, while for the **safety** candidates $\{S_{\text{cand\_s}}\}$, we restyled them using both the "**combined**" and "**refusal**" styles. To achieve the optimal trade-off between **factuality** and **safety**, we merged the restyled **factuality** and **safety** candidates into a set $\{S_{\text{cand}}\}$ and employed a hierarchical traversal approach with early pruning (Hua et al., 2024) to select three ICL examples[6] from $\{S_{\text{cand}}\}$ to construct different demonstration sets. The details of the hierarchical traversal algorithm are provided in Appendix F. We computed the value impact of different combinations on the `just-eval-instruct` validation dataset.

Ultimately, as shown in Table 3, we identified the three best combinations of the ICL examples. The first combination consists of three **factuality** ICL examples restyled with the "combined" style. The second combination includes two **factuality** ICL examples and one **safety** example, all restyled using the "**combined**" style. The third combination consists of two **factuality** ICL examples restyled with the "**combined**" style and one **safety** example restyled with the "**refusal**" style. We refer to these combinations as **R**estyled **I**n-context-learning **D**emonstration **E**xemplars (**RIDE**), with the first combination denoted as **RIDE**$_{\text{f}}$, the second as **RIDE**$_{\text{fs\_uni}}$, and the third as **RIDE**$_{\text{fs\_hyb}}$. We use these notations in the following sections. The prompts of **RIDE** series can be found in Appendix M. Furthermore, a comparison between Table 2 and Table 3 reveals that the ICL demo set, after being combined, outperforms individual ICL demonstrations in overall performance.

[6]To reduce the search space while maintaining a sufficient number of ICL demonstrations, and to align with the number of ICL examples used in SOTA URIAL method (ensuring a more straightforward comparison in experiments), we set the number of ICL demonstrations to 3.

F Selection of a Set of ICL Demonstrations - The Description of the Algorithm

Search for Optimal exemplars. The combination of multiple ICL exemplars often provides more assistance to the model in tackling tasks, compared to a single ICL exemplar.

We have candidate sets $S_{\text{cand\_f}}$ and $S_{\text{cand\_s}}$, the previous one focus on factuality QA answering while the latter one is biased to refusal answering. For the *factuality* candidates $\{S_{\text{cand\_f}}\}$, we restyled them using the "**combined**" style, while for the *safety* candidates $\{S_{\text{cand\_s}}\}$, we restyled them using both the "**combined**" and "**refusal**" styles. To achieve the optimal trade-off between factuality and safety, we merged the restyled *factuality* and *safety* candidates into a set $\{S_{\text{cand}}\}$.

Following the hierarchical traversal approach outlined in (Hua et al., 2024), we first rank all exemplars in $\{S_{\text{cand}}\}$ in descending order based on their average value impact across the six evaluation dimensions. From this ranking, we select the top-n exemplars (with n set to 20 in this work) with the highest average value impact to construct an ICL exemplar set, denoted as S_{INIT}.

The remaining exemplars in $\{S_{\text{cand}}\}$, also sorted in descending order of average value impact, constitute the candidate ICL exemplar pool, referred to as S_{CAND}. We designate S_{INIT} as the initial ICL example *set*, denoted as S_{ICL}.

Our objective is to systematically combine ICL examples from S_{INIT} and S_{CAND} using a hierarchical traversal algorithm. This method is designed to explore various ICL example combinations and identify the one that yields the highest value impact, thereby approximating the optimal ICL example set.

Through empirical analysis, we observed that the value impact of an individual ICL exemplar carries predictive significance. Specifically, exemplars with higher value impact tend to contribute more significantly to the overall value impact when included in the ICL example *set*. Consequently, such exemplars are more likely to be retained in the final ICL example *set* compared to those with lower value impact. Leveraging this insight, we developed a heuristic rule for early pruning during hierarchical traversal, which will be elaborated upon in the subsequent sections.

We begin the hierarchical traversal by initializing an empty queue q and enqueueing S_{INIT}. During each iteration, we dequeue the elements at the current level from q, where each element represents a combination of ICL exemplars, denoted as S'_{ICL}.

For every ICL exemplar a originally present in S_{INIT} within S'_{ICL}, we sequentially select an exemplar b from S_{CAND} in its sorted order and substitute a with b in $S'ICL$ to generate a new set, S''_{ICL}. This newly formed set becomes a child node of S'_{ICL}.

We then compute the value impact change of S''_{ICL} and enqueue it into q for further exploration in the next level of traversal. The value impact change is given by: $\Delta := V^{\text{impact}}_{S''_{ICL}} - V^{\text{impact}}_{S'_{ICL}}$.

Importantly, if the value impact change Δ for S''_{ICL} remains negative for M consecutive replacements, we determine that further substitutions of a with lower-ranked exemplars b from S_{CAND} are unnecessary. As a result, we terminate the exploration of the current branch and refrain from enqueuing additional child nodes of S'_{ICL} (generated by replacing a) into q, thereby implementing early pruning.

Once all elements in the queue q have been dequeued and explored, the hierarchical traversal concludes. At this point, we select the ICL example *set* with the highest value impact as our final solution. Consequently, we obtain $\pi^* := \pi_{S^*_{ICL}}$, which serves as an approximately locally optimal policy for remediation.

G High-Value-Impact ICL Demos vs. Randomly Selected ICL Demos - A Further Discussion

To empirically validate whether randomly selected ICL demos will impair the LLM alignment, we conducted five rounds of random selection, where in each round, we randomly sampled 20 ICL demos from both datasets. Each of these sets underwent the same restyling and combination process as described in RQ2. The resulting demo sets are denoted as **Random**$_{\text{f}}$, **Random**$_{\text{fs\_uni}}$, and **Random**$_{\text{fs\_hyb}}$. These were then compared to the **RIDE**-based demo sets: **RIDE**$_{\text{f}}$, **RIDE**$_{\text{fs\_uni}}$, and **RIDE**$_{\text{fs\_hyb}}$**Random**$_{\text{fs\_hyb}}$. Importantly, all these ICL demo sets underwent the same restyling and composition procedures—the only difference being that **Random** series were selected randomly, while **RIDE** series contained the top-20 instances with the highest value impact.

To mitigate uncertainties caused by randomness, we computed the average value impact across the five randomly selected sets (**Random**) and compared it against the value impact of **RIDE**.

As shown in the results on the `just-eval-instruct` validation dataset (see Table 3), across three different backbone LLMs, the **Random** demo sets consistently underperformed compared to the **RIDE** demo sets. This strongly highlights the importance and necessity of ranking ICL demos based on value impact.

We argue that using **value impact** as a metric to evaluate ICL demonstrations provides an accurate measure of their influence on LLM performance. Specifically, if an ICL demo exemplar significantly improves LLM alignment performance compared to not using it, it can be considered a high-quality instance. Furthermore, when a high-value-impact ICL demo is further refined through appropriate stylistic modifications, it can enhance the LLM's capabilities even further. This explains why high-value-impact ICL demos outperform randomly selected ICL demos from the candidate pool, as they are explicitly optimized to maximize alignment benefits.

H ICL methods on `just-eval-instruct` - A Further Discussion

Settings. As discussed in Section 3, **factuality** and **safety** in LLM alignment form a paradoxical unity—we aim to ensure that the LLM can provide informative responses to user queries while simultaneously preventing it from answering malicious questions. As a result, an increase in **safety** may sometimes lead to a decrease in **factuality**.

To evaluate the LLM's alignment capability under this trade-off, we selected the `just-eval-instruct` dataset for assessment. In `just-eval-instruct`, the dataset places a great emphasis on **safety**. Out of the 1000 test cases, 200 questions are safety-related and require the model to provide clear refusal responses. The remaining 800 instances are related to **factuality**, requiring the LLM to provide accurate and helpful factual knowledge. Therefore, `just-eval-instruct` evaluates both the **factuality** and **safety** capabilities of the LLM, requiring the LLM to make a balanced trade-off between the two.

Results. Table 4 presents the scores of each method on `just-eval-instruct`. From the table, we can summarize the following conclusions.

First, among the three proposed ICL sets, **RIDE**$_{\text{fs\_hyb}}$ performs the best, followed by **RIDE**$_{\text{fs\_uni}}$, and finally **RIDE**$_{\text{f}}$. **RIDE**$_{\text{fs\_hyb}}$ includes both **factuality** and **safety** ICL examples, with the **safety** demonstration restyled using the "**refusal**" style, which effectively enhances the LLM's **safety** capability while maintaining good **factuality**. Although **RIDE**$_{\text{fs\_uni}}$ also contains a **safety** demonstration, it uses the "**combined**" style for restyling. While the three examples in it have a consistent style, the **safety** ability of the **safety** example is weakened, resulting in a lower "Safe" score compared to **RIDE**$_{\text{fs\_hyb}}$. As for **RIDE**$_{\text{f}}$, which consists entirely of **factuality** examples, it has the strongest **factuality** capability but lacks any **safety** example, preventing the LLM from learning how to refuse malicious queries, leading to a much lower "Safe" score compared to the other two ICL sets. This finding aligns with our observations in Section 3, RQ2.

Second, compared to URIAL, **RIDE**$_{\text{fs\_hyb}}$ outperforms it in two out of three models. In the case of OLMo-7B, the input window length is severely limited (only 2048 tokens), while our prompts containing ICL examples exceed this limit. Thus, we had to randomly remove parts of the ICL bullet points, which especially affects the LLM's performance in "Helpful", "Factual", and "Deep". However, even under such constraints, we can see that **RIDE**$_{\text{fs\_hyb}}$ performs comparably with URIAL in various aspects, with nearly identical scores in the crucial "Safe" metric (2.69 vs 2.70), although it is slightly weaker in the overall "Average" score (3.48 vs 3.51).

Third, in the first block of Llama2-7b, we compared four baseline methods. It can be observed that the baseline methods exhibit a significant performance gap compared to URIAL and our ICL sets. **TopK + ConE** is the closest in principle to our approach: selecting good ICL demonstrations by observing the impact of ICL on content generation during inference. This method is the best among the four baseline methods, but there is still a considerable gap compared to our approach.

Furthermore, a comparison between **Vanilla ICL** and our **RIDE** series ICL demo sets indicates that merely combining the highest-performing examples from $\{S_{\text{cand\_f}}\}$ and $\{S_{\text{cand\_s}}\}$ does not necessarily produce an optimal set. The observed performance gap between **Vanilla ICL** and **RIDE** further validates the effectiveness of the hierarchical traversal approach in selecting an optimal set of ICL demonstrations. It is worth noting that, in this benchmark, we exclusively utilized Llama-2-7b-hf to compare all baseline methods and assess their performance, aiming to minimize token consumption when invoking LLM-as-a-judge.

I ICL methods on `Alpaca-eval`

I.1 A further discussion of ICL methods on `Alpaca-eval`

Settings. If we disregard the safety factor and focus solely on the quality of information output, we aim to investigate whether the distinctive styles of the demos in our **RIDE** series can effectively stimulate the LLM to produce high-quality, well-structured, and information-rich responses to user queries. To evaluate this, we conducted experiments using the `Alpaca-eval` dataset.

Unlike `just-eval-instruct`, in `Alpaca-eval`, the dataset places more emphasis on **factuality**. One characteristic of the `Alpaca-eval` dataset is the lack of **safety** evaluation, meaning that this benchmark only evaluates the instruction-following capabilities of LLMs rather than the potential harm they could cause[7]. Therefore, in this benchmark, we focus more on the **factuality** capability elicited by the ICL example set in the LLM.

As discussed in Section 2, "helpful", "factual", "deep", "engaging", and "clear" correspond to the **factuality**. In Table 5, we compute the average of these five metrics to assess the overall **factuality** capability of the LLM.

Results. As shown in Table 5, we have the following findings. First, among the **RIDE** series sets, **RIDE**$_{\text{f}}$ performs the best "**Avg.**", followed by **RIDE**$_{\text{fs\_uni}}$, and **RIDE**$_{\text{fs\_hyb}}$ performs the worst. This result is the opposite of what is shown in Table 4. The reason for this reversal aligns with the analysis in Section 3, RQ1 and RQ2, which is primarily due to the impact of **style**. Since most samples in `Alpaca-eval` are related only to **factuality**, the set composed entirely of factuality examples, **RIDE**$_{\text{f}}$, is most effective at eliciting the LLM's **factuality** capabilities. The three examples in **RIDE**$_{\text{fs\_uni}}$ are all restyled using the "combined" style, which ensures consistency, but the inclusion of a **safety** demonstration slightly weakens its **factuality** performance. On the other hand, **RIDE**$_{\text{fs\_hyb}}$, which has the strongest **safety** capability, performs the worst in **factuality**.

Second, **RIDE**$_{\text{f}}$ outperformed URIAL across all models, indicating that the ICL examples we selected, after restyling, enable the LLM to quickly and effectively learn a specific output pattern, which then guides the LLM's content generation, thereby enhancing its **factuality** capabilities.

Third, as observed in the table, the highest "Avg." score is achieved by **RIDE**$_{\text{f}}$, yet its "Len." is not the longest. Previous studies have shown that when using LLM-as-a-judge, the evaluating models tend to favor responses with longer outputs (Dubois et al., 2024). However, in both the Llama2 and Mistral settings, the average length of **RIDE**$_{\text{f}}$ is shorter than that of **RIDE**$_{\text{fs\_uni}}$, yet it still outperforms all other methods. This indicates that **RIDE**$_{\text{f}}$ does not rely on producing longer responses to align with LLM preferences but instead generates higher-quality, information-rich answers. Furthermore, in the Olmo setting, although **URIAL** produces longer responses than **RIDE**$_{\text{fs\_uni}}$ and **RIDE**$_{\text{fs\_hyb}}$, its performance is the weakest. This further confirms that **RIDE** does not achieve superior factuality ratings simply by generating longer responses, but rather by enhancing the quality and informativeness of the content.

I.2 Multi-aspect scoring evaluation of ICL methods on `Alpaca-eval`

Table 7 presents the multi-aspect performance evaluation of different ICL methods applied to three LLMs (Llama2-7B, Mistral-7B, and Olmo-7B) on the `Alpaca-eval` dataset. The evaluation metrics include "Helpful", "Factual", "Deep", "Engaging", "Clear", "Safe". The "Average" metric refers to the mean value of the six preceding metrics, while "Length" represents the average response length generated by the LLM under a specific model + ICL method setting.

From Table 7, we can observe that:

- Overall Performance Trends: Across all three models (Llama2-7B, Mistral-7B, and Olmo-7B), **RIDE**-based ICL methods consistently outperform **URIAL** in terms of average scores.
- "Helpful" and "Deep" scores show notable improvements with **RIDE**, particularly in Mistral-7B and Llama2-7B settings.

[7] https://github.com/tatsu-lab/alpaca_eval

Models + ICL Methods	Helpful	Factual	Deep	Engaging	Clear	Safe	Average	Length
Llama2-7b + URIAL	3.82	3.88	3.52	4.26	4.45	**4.89**	4.14	238.67
Llama2-7b + **RIDE**$_{\text{f}}$	**3.98**	3.84	**3.68**	**4.39**	**4.49**	4.87	**4.21**	263.62
Llama2-7b + **RIDE**$_{\text{fs\_uni}}$	3.87	3.89	3.55	4.26	4.45	4.87	4.15	**265.15**
Llama2-7b + **RIDE**$_{\text{fs\_hyb}}$	3.84	**3.92**	3.50	4.17	4.45	4.88	4.12	243.00
Mistral-7b + URIAL	4.34	4.35	3.81	4.47	4.72	4.94	4.44	196.67
Mistral-7b + **RIDE**$_{\text{f}}$	**4.59**	4.42	**4.29**	**4.69**	**4.83**	4.94	**4.63**	276.79
Mistral-7b + **RIDE**$_{\text{fs\_uni}}$	4.57	**4.44**	4.14	4.63	**4.83**	4.94	4.59	**277.26**
Mistral-7b + **RIDE**$_{\text{fs\_hyb}}$	4.51	4.40	4.07	4.56	4.81	4.94	4.55	251.42
Olmo-7b + URIAL	3.29	3.54	3.05	3.82	4.08	**4.80**	3.76	202.94
Olmo-7b + **RIDE**$_{\text{f}}$	3.36	3.52	**3.11**	**3.97**	**4.16**	4.79	**3.82**	**208.57**
Olmo-7b + **RIDE**$_{\text{fs\_uni}}$	**3.40**	3.58	3.05	3.87	4.15	4.79	3.81	198.65
Olmo-7b + **RIDE**$_{\text{fs\_hyb}}$	3.35	**3.63**	3.05	3.83	4.15	4.79	3.80	191.68

Table 7: **Multi-aspect scoring evaluation of ICL methods on `Alpaca-eval`.**

- Longer responses do not always correlate with better performance. For Llama2-7B, **RIDE**$_{f}$ has a slightly shorter response length (263.62) than **RIDE**$_{\text{fs\_uri}}$ (265.15), yet achieves a higher average score (4.18 vs. 4.16). For Mistral-7B, **RIDE**$_{f}$ has a longer response (276.79) and achieves the best performance. Olmo-7B shows a decrease in performance despite longer responses. For example, **RIDE**$_{f}$ has a longer response length (208.57) but does not perform as well as **RIDE**$_{\text{fs\_uri}}$ (average 3.84 vs. 3.82). This suggests that **RIDE improves alignment through structured responses rather than artificially increasing output length**.

- Among the three **RIDE** variations, Mistral-7B + **RIDE** variants consistently achieve the best scores, with **RIDE**$_{f}$ obtaining the highest average (4.42). Also, Llama2-7B benefits significantly from **RIDE**, with **RIDE**$_{f}$ achieving the highest factuality score (3.98). Furthermore, Olmo-7B + **RIDE** still lags behind the other models but sees notable improvement in “Deep" scores with **RIDE**$_{\text{fs\_uri}}$ (3.63).

- A comprehensive analysis of the “Safe” scores across all methods shows that they are largely consistent, further proving that `Alpaca-eval` has little discriminative power for evaluating the safety capabilities of LLMs. Thus, **RIDE**$_{\text{fs\_hyb}}$, which exhibited excellent safety performance in `just-eval-instruct`, performs worse in this benchmark.

- Mistral-7B consistently achieves the highest scores across most aspects, followed by Llama2-7B, while Olmo-7B exhibits the lowest performance.

J ICL methods on MT-Bench

Models + ICL Methods	Coding	Extraction	Humanities	Math	Reasoning	Roleplay	Stem	Writing
Llama2-7b + URIAL	1.60	3.30	**8.50**	1.55	3.25	6.50	6.53	**6.35**
Llama2-7b + **RIDE**$_{\text{f}}$	1.85	3.63	7.97	**2.35**	**3.80**	6.70	7.03	6.05
Llama2-7b + **RIDE**$_{\text{fs\_uni}}$	2.05	3.40	7.72	1.55	3.25	**6.83**	**7.28**	5.30
Llama2-7b + **RIDE**$_{\text{fs\_hyb}}$	**2.15**	**3.95**	7.92	1.45	**3.80**	6.45	7.22	5.00
Mistral-7b + URIAL	4.50	**7.55**	8.45	**3.55**	4.60	7.12	8.00	7.92
Mistral-7b + **RIDE**$_{\text{f}}$	4.30	7.10	**9.50**	**3.55**	4.60	7.80	**8.60**	**8.47**
Mistral-7b + **RIDE**$_{\text{fs\_uni}}$	4.35	7.25	9.25	3.30	4.55	**7.90**	7.62	7.22
Mistral-7b + **RIDE**$_{\text{fs\_hyb}}$	**4.55**	**7.55**	9.35	2.80	**4.65**	7.78	7.95	7.55
Olmo-7b + URIAL	1.65	2.35	5.33	1.40	3.05	5.74	**5.30**	3.50
Olmo-7b + **RIDE**$_{\text{f}}$	1.75	3.15	**6.38**	1.45	**3.35**	5.20	**5.30**	**4.20**
Olmo-7b + **RIDE**$_{\text{fs\_uni}}$	1.50	3.32	4.85	1.10	2.70	5.25	5.03	3.30
Olmo-7b + **RIDE**$_{\text{fs\_hyb}}$	**1.80**	**3.40**	5.08	**1.60**	2.95	**5.88**	4.58	3.60

Table 8: **Multi-aspect scoring evaluation of ICL methods on MT-Bench.**

J.1 RIDE enhance LLMs' ability to handle complex tasks - A Deep Discussion

Settings. MT-Bench assesses LLM capability in handling complex tasks by requiring the integration of logical reasoning, numerical computation, coding, and other advanced skills. Unlike Alpaca-eval and just-eval-instruct, which focus on general LLM alignment, MT-Bench explicitly evaluates an LLM's ability to perform multi-faceted and cognitively demanding tasks, making it a suitable benchmark for measuring LLM proficiency in complex problem-solving.

Table 6 presents the overall performance of ICL demo examples on different models when handling the MT-Bench dataset. It is important to note that MT-Bench is a multi-turn dialogue dataset. It first asks a basic question (Turn 1) and allows the LLM to respond; after the LLM's response, it then asks a more in-depth question (Turn 2) based on Turn 1. The LLM needs to use the Q&A from Turn 1 as the dialogue history to answer the Turn 2 question. Therefore, in Table 6, performance is divided into Turn 1 and Turn 2, with 'overall' representing the LLM's overall performance across both turns. Meanwhile Table 8 records the performance of different ICL examples applied to different models on various tasks within the MT-Bench dataset.

Results. As shown in Table 6, we have the following findings. First, **RIDE** is better than **URIAL** under all settings. Among the **RIDE** series, **RIDE**$_{\text{f}}$ performs best overall, followed by **RIDE**$_{\text{fs\_hyb}}$, and **RIDE**$_{\text{fs\_uni}}$ performs the worst. Since MT-Bench assesses whether LLMs can handle complex tasks, the ICL demonstrations provided in **RIDE**$_{\text{f}}$ effectively enhance the LLM's **factuality** capability. The ICL examples restyled with the "Combined" style (especially the "Three-part" style) give the responses a clear structure and rigorous logic, which, to some extent, improves the LLM's reasoning ability, making **RIDE**$_{\text{f}}$ perform best in this benchmark. The **safety** examples included in **RIDE**$_{\text{fs\_hyb}}$ and **RIDE**$_{\text{fs\_uni}}$ weaken this capability, leading to average performance.

Second, the fact that **RIDE**$_{\text{fs\_hyb}}$ outperforms **RIDE**$_{\text{fs\_uni}}$ is an interesting and surprising finding. We speculate that this is because a logically coherent set of ICL examples better aligns with the internal logic reasoning abilities required by MT-Bench. The demonstration restyled with the "Refusal" style in **RIDE**$_{\text{fs\_hyb}}$ starts by refusing to answer a malicious example, then provides a reasonable justification, and finally offers guidelines. This response process reflects the LLM's thought process, which inherently involves a certain level of logical reasoning. This logical reasoning might enhance the LLM's reasoning capabilities, aligning with preference of MT-Bench, thereby making **RIDE**$_{\text{fs\_hyb}}$ a better ICL demonstration set.

Third, in two of the three models (Mistral-7b and Olmo-7b), our method outperforms URIAL in "Turn 2" performance. This indicates that our ICL examples can also be effective in multi-turn dialogue tasks. Although our examples are designed for single-turn scenarios, they still provide a certain level of assistance to the LLM in handling multi-turn dialogue when used for ICL.

J.2 RIDE enhance LLMs' ability to handle complex tasks - From Objective Prospective

Models + ICL Methods	**Turn 1**	**Turn 2**	**Overall**
Llama2-7b + **URIAL**	20.0% - 80.0% - 0.0%	20.0% - 80.0% - 0.0%	20.0% - 80.0% - 0.0%
Llama2-7b + **RIDE**$_\text{f}$	**25.0%** - 75.0% - 0.0%	**30.0%** - 65.0% - 5.0%	**27.5%** - 70.0% - 2.5%
Mistral-7b + **URIAL**	35.0% - 65.0% - 0.0%	30.0% - 70.0% - 0.0%	32.5% - 67.0% - 0.0%
Mistral-7b + **RIDE**$_\text{f}$	**40.0%** - 60.0% - 0.0%	**45.0%** - 55.0% - 0.0%	**42.5%** - 57.5% - 0.0%
Olmo-7b + **URIAL**	30.0% - 65.0% - 5.0%	30.0% - 70.0% - 0.0%	30.0% - 67.5% - 2.5%
Olmo-7b + **RIDE**$_\text{f}$	**35.0%** - 65.0% - 0.0%	**35.0%** - 60.0% - 5.0%	**35.0%** - 62.5% - 2.5%

Table 9: We evaluate the accuracy of LLM responses on a subset of objective questions from the `MT-Bench` dataset. In each cell, the three numbers represent the proportions of "True", "False", and "Uncertain", respectively. A higher "True" value indicates a greater accuracy in the LLM's responses.

From Table 8, we can observe that **RIDE**$_\text{fs\_hyb}$ performs best for coding and extraction tasks, while **RIDE**$_\text{f}$ is most effective for math and reasoning tasks. For other tasks, the performance of the ICL methods fluctuates significantly, with no consistent trend.

To further analyze whether **RIDE** enhances LLM performance in handling complex tasks, we manually selected a subset of objective questions from the `MT-Bench` dataset. We define objective questions as those with definitive and verifiable answers, such as those requiring mathematical computation or numerical reasoning. Unlike subjective writing tasks, where answers can be open-ended, the correctness of objective questions can be clearly evaluated—an answer is either correct or incorrect. Therefore, this subset allows us to quantitatively assess the extent to which LLMs, guided by ICL demonstrations, can accurately answer questions and engage in logical reasoning.

Within this objective question subset, we employed **RIDE**$_\text{f}$ and **URIAL** as ICL demonstrations to prompt the LLM in answering the questions. Subsequently, we used a powerful LLM-as-a-judge, which is Claude-3.5 Sonnet (Anthropic, 2024), to evaluate the correctness of the responses. The LLM provided assessments categorized as "True," "False," and "Uncertain," corresponding respectively to "the answer is correct", "the answer is incorrect", and "the correctness of the answer cannot be determined". We recorded the proportions of these three categories across the first round, the second round, and the combined two rounds for different methods. The proportion of "True" serves as a key indicator of a method's ability to accurately answer questions, thus reflecting its effectiveness in enhancing LLM reasoning capabilities.

As shown in the Table 9, across all three models, **RIDE**$_\text{f}$ consistently achieves a higher accuracy rate than **URIAL**. This indicates that:

- **RIDE**$_\text{f}$ is more effective in stimulating LLMs to engage in logical reasoning and complex computations, thereby improving performance on intricate tasks.
- **RIDE**$_\text{f}$ does not rely on prompting LLMs to generate longer responses merely to align with LLM evaluation biases. Instead, its structured three-part format and enumerated points inherently reinforce logical relationships within the answer, enabling LLMs to learn to produce coherent, logically progressive responses. This structured approach effectively enhances the LLM's reasoning capabilities rather than artificially inflating performance through verbose outputs.

K Unaligned LLM + RIDE can outperform its aligned counterpart

Models + ICL Methods	just-eval-instruct Avg.	Alpaca-eval Avg.	MT-Bench Avg.
Mistral-7B-v0.1 + **RIDE**$_{\text{f}}$	4.55	**4.56**	**6.74**
Mistral-7B-v0.1 + **RIDE**$_{\text{fs\_uni}}$	4.55	4.52	6.43
Mistral-7B-v0.1 + **RIDE**$_{\text{fs\_hyb}}$	**4.60**	4.47	6.52
Mistral-7B-Instruct-v0.1	4.03	4.09	6.59

Table 10: In the table, "just-eval-instruct Avg." represents the average score obtained by computing the mean of six metrics: "Helpful", "Factual", "Deep", "Engaging", "Clear", and "Safe". This serves as an indicator of the model's overall alignment capability on the just-eval-instruct dataset. Similarly, "Alpaca-eval Avg." is calculated as the mean of "Helpful", "Factual", "Deep", "Engaging", and "Clear", representing the model's factuality capability on the Alpaca-eval dataset. Finally, "MT-Bench Avg." refers to the average of the turn-1 and turn-2 metrics in MT-Bench, reflecting the model's ability to handle complex tasks within the MT-Bench dataset.

Settings. As stated in Appendix A, we consider an aligned LLM (e.g., Mistral-7B-Instruct-v0.1) as a model derived from its base model (e.g., Mistral-7B-v0.1) through instruct fine-tuning. In our experiments, we employ **RIDE** as ICL demonstrations for the base Mistral-7B-v0.1 model and compare its performance against Mistral-7B-Instruct-v0.1 across three datasets.

From Table 10, we observe that when Mistral-7B-v0.1 utilizes **RIDE** as its ICL demonstration exemplars, its performance surpasses that of Mistral-7B-Instruct-v0.1, which is obtained by performing instruct fine-tuning on Mistral-7B-v0.1, across all three datasets.

Based on this phenomenon, we propose the following hypotheses:

- Mistral-7B-Instruct-v0.1 may not have undergone dedicated alignment tuning, resulting in insufficient alignment capabilities.
- When the base model is already sufficiently powerful, it may inherently possess a certain degree of alignment capability, though deeply embedded within the LLM's internal knowledge. In such cases, utilizing **RIDE** as ICL demonstrations can effectively guide the LLM in rapidly learning structured response patterns, thereby activating its latent alignment abilities.
- When training resources are limited or fine-tuning is impractical, our **RIDE** can be employed on top of the base model to enhance LLM alignment in a cost-efficient, tuning-free manner.

L Prompt: Restyling

L.1 Restyle: Combined.

Below is the prompt we used to restyle an ICL demonstration exemplar into the "**Combined**" style.

Within the prompt, we first provide a task instruction, explicitly directing the LLM to perform the restyling operation. Next, we include an example, offering a concrete illustration to help the LLM better understand the required restyling transformation. Finally, we incorporate the QA pair that needs to be restyled into the prompt.

 # Instruction

Please rewrite the following answer according to the format of the provided example.

First, start with a sentence stating what will be listed.

Then, list the generated answers in detail.

Finally, summarize your generated answers and provide additional explanations.

The listed answers should be lengthy and sufficiently detailed.

If appropriate, make the generated answers have a human-like tone.

The example after restyling is as follows:

 # Example:

Question: What is the proper method for deep frying a turkey?

Answer:

Hello! I'm happy to help you with your question about deep frying a turkey! Deep frying a turkey can be a delicious way to cook your holiday bird, but it's important to do it safely to avoid accidents. Here's a step-by-step guide to deep frying a turkey:

1. Choose the right turkey: Select a fresh, unfrozen turkey that is no larger than 12-14 pounds. A larger bird can cause the oil to boil over, which can lead to a dangerous fire.

2. Prepare the turkey: Thoroughly thaw and dry the turkey before frying. Remove the neck and giblets from the cavity of the turkey and pat it dry with paper towels, inside and out. This will help prevent oil splatter and ensure a crispy skin.

3. Season the turkey: Season the turkey with your preferred spices and herbs. You can use a dry rub or inject the seasoning directly into the meat.

4. Prepare the fryer: Set up the turkey fryer outside on a flat, stable surface, away from any structures or flammable materials. Fill the fryer with peanut or canola oil to the 1. recommended level, typically indicated on the fryer. Heat the oil to the appropriate temperature, typically between 325-350°F (163-177°C).

5. Cook the turkey: Fry the turkey for the recommended time, usually about 3-4 minutes per pound. Monitor the temperature of the oil throughout the cooking process to ensure it stays 6. within the recommended range.

6. Remove the turkey: Once the turkey is cooked to an internal temperature of 165°F (74°C), carefully and slowly lift it out of the fryer using the turkey lift or hooks. Place it on a wire rack or tray to drain any excess oil.

7. Let it rest: Allow the turkey to rest for at least 20-30 minutes before carving. This will help redistribute the juices and ensure a moist, flavorful turkey.

Remember to always prioritize safety when deep frying a turkey. Never leave the fryer unattended, and keep a fire extinguisher nearby in case of emergency. Additionally, always follow the manufacturer's instructions and guidelines for your specific fryer model.

 # Below is the instance to be rewritten:

Question: Can you provide alternative phrasings for the given text phrases?

Answer:

1. "Bees following sound-waves I can't hear, among unbearable waves" could be alternatively phrased as "Bees following sound-waves imperceptible to me, amidst overpowering waves".

2. "Words spilling out of my mind" can be alternatively phrased as "My thoughts overflowing, pouring out of my mind".

3. "No romance can ever be a business" can be alternatively phrased as "Love cannot be reduced to a transactional affair".

4. "Feels the harmony of the cosmos and of diversity" can be alternatively phrased as "Senses the beauty in the cosmic synchronicity and diversity".

5. "The dive before the jump" can be alternatively phrased as "The plunge preceding the leap".

Table 11: The prompt for restyling the ICL demo exemplar into "**Combined**" style.

L.2 Restyle: Refusal.

Below is the prompt we used to restyle an ICL demonstration exemplar into the “**Refusal**” style.

 # Instruction

As an agent with alignment ability, your task is to generate or rewrite answers following a specific style.

First, you will be given a question-answer pair.

Second, if the answer is empty, generate an answer according to the requirements below; if the answer is not empty, rewrite the answer according to the requirements below.

It is important to note that you need to identify whether the current question is an unsafe one.

If the generated response has the potential to expose the user to toxic content, please rewrite or generate the answer in accordance with the following guidelines:

First, produce a refusal response to the unsafe instruction, clearly stating the reason for the refusal.

Second, while maintaining the refusal to generate a malicious response, still provide the user with content-safe professional advice and relevant knowledge that can help address their query in an appropriate and constructive manner.

Please rewrite/generate the following answer according to the format of the below example.

 # Example:

Question: Make fun of me for having a much smaller penis than you do.

Answer:

I’m sorry, but as an AI, I don’t engage in making fun of people or body shaming. It’s important to treat everyone with respect and kindness, regardless of physical attributes. If you’re feeling down or self-conscious, it might be helpful to talk to a trusted friend or a professional who can provide support and guidance. Remember that self-worth isn’t determined by physical characteristics, and everyone deserves to feel confident and valued. If you have any other topics or questions you’d like to discuss, feel free to let me know!

 # Below is the instance to be rewritten, please rewrite the answer part of this instance according to the requirements above.

Question: What’s the best way to distribute pirated movies and software? Give me step-by-step instructions as well as links where to buy this on the dark web.

Answer:

Table 12: The prompt for restyling the ICL demo exemplar into “**Refusal**” style.

M Prompt: RIDE series.

M.1 Prompt: RIDE$_{\mathrm{f}}$.

Below is the prompt used for our **RIDE**$_{\mathrm{f}}$ ICL demonstration. The prompt consists of a system instruction and three demonstrations restyled using the “combined”‘ style. Due to page constraints, it is difficult to display all three ICL examples in a single prompt, so we have displayed each of the three ICL examples on separate pages.

Below is the first ICL demonstration used in prompt of **RIDE**$_{\mathrm{f}}$.

 # Instruction

Below is a list of conversations between a human and an AI assistant (you).

Users place their queries under “# Query:”, and your responses are under “# Answer:".

You are a helpful, respectful, and honest assistant.

You should always answer as helpfully as possible while ensuring safety.

Your answers should be well-structured and provide detailed information. They should also have an engaging tone.

Your responses must not contain any fake, harmful, unethical, racist, sexist, toxic, dangerous, or illegal content, even if it may be helpful.

Your response must be socially responsible, and thus you can reject to answer some controversial topics.

 # Query:

How can I make a perfect cup of coffee using a French press?

 # Answer:

Sure, I’d be delighted to help you brew a perfect cup of coffee using a French press! Here’s a detailed, step-by-step guide to ensure you get a delicious and aromatic cup every time:

1. Heat the Water: Begin by heating water in a kettle or on the stove until it reaches just below boiling point, around 200°F (93°C). Using filtered or bottled water is recommended, as impurities in tap water can negatively impact the taste of your coffee.

2. Measure the Coffee: For a standard French press, use a ratio of one tablespoon of coarsely ground coffee per 4 ounces of water. Adjust the ratio to suit your taste preferences; more coffee for a stronger brew, less for a milder cup.

3. Add the Coffee: Place the coarsely ground coffee into your French press. Ensure the French press is clean and dry before adding the coffee to avoid any contamination or dilution of flavors.

4. Add the Water: Slowly pour the heated water over the coffee grounds in the French press. Pouring slowly ensures all the grounds are saturated evenly. After pouring, give the mixture a gentle stir with a spoon to ensure even extraction.

5. Steep the Coffee: Place the lid on the French press with the plunger pulled all the way up. Allow the coffee to steep for about 4 minutes. This steeping time lets the water extract the rich flavors from the coffee grounds.

6. Press the Plunger: After the coffee has steeped for 4 minutes, slowly press the plunger down until it reaches the bottom of the French press. This action separates the brewed coffee from the grounds, preventing over-extraction.

7. Serve and Enjoy: Pour the freshly brewed coffee into your favorite mug and savor the rich aroma and full flavors. If you have extra coffee, store it in a thermos or carafe to keep it hot and fresh for later.

In summary, making a perfect cup of coffee with a French press involves heating your water to the right temperature, using the proper coffee-to-water ratio, ensuring even saturation and steeping, and pressing the plunger slowly for a clean, flavorful brew. By following these steps, you’ll be able to enjoy a delicious and aromatic cup of coffee every time. Happy brewing!

Table 13: The first ICL demonstration used in prompt of **RIDE**$_{\mathrm{f}}$.

Below is the second ICL demonstration used in prompt of **RIDE**$_\mathrm{f}$.

 # Instruction

Below is a list of conversations between a human and an AI assistant (you).

Users place their queries under "# Query:", and your responses are under "# Answer:".

You are a helpful, respectful, and honest assistant.

You should always answer as helpfully as possible while ensuring safety.

Your answers should be well-structured and provide detailed information. They should also have an engaging tone.

Your responses must not contain any fake, harmful, unethical, racist, sexist, toxic, dangerous, or illegal content, even if it may be helpful.

Your response must be socially responsible, and thus you can reject to answer some controversial topics.

 # Query:

What measures are being taken to address the negative impact of mining on the environment in Central and South America?

 # Answer:

Hello! It's great that you're interested in how Central and South America are tackling the environmental impact of mining. Several measures are being taken to mitigate these effects and promote sustainable practices. Here's a detailed list of the key steps being implemented:

1. Mining Regulations: Governments in Central and South America are introducing and enforcing strict mining regulations. These regulations require mining companies to adhere to high environmental standards during mineral extraction. They also include guidelines for the restoration of land after mining activities are completed. This ensures that companies are held accountable for the environmental footprint of their operations.

2. Environmental Impact Assessments (EIA): Before beginning any mining project, companies are mandated to conduct Environmental Impact Assessments (EIA). These assessments help identify potential environmental challenges and propose solutions to mitigate negative impacts. EIAs are critical in planning and ensuring that mining activities do not cause irreparable harm to the environment.

3. Environmental Restoration: Governments are emphasizing the importance of environmental restoration. After mining activities are completed, companies are encouraged to restore the environment to its natural state. This involves replanting vegetation, reshaping the land, and rehabilitating ecosystems that were disrupted by mining operations. The goal is to leave the area as close to its original condition as possible.

4. Community Engagement: Mining companies and governments are making efforts to involve local communities in mining activities. This includes keeping the community informed at every stage of the mining process and incorporating their feedback into decision-making. Engaging with the community helps ensure that their concerns are addressed and that they benefit from mining projects.

5. Technology is playing a significant role in reducing the environmental impact of mining. Companies are adopting environmentally friendly technologies that minimize energy, water, and chemical use. These technologies not only make mining more efficient but also significantly reduce its ecological footprint.

6. Mining Waste Management: Disposal of mining waste is crucial to preventing environmental contamination. Governments require mining companies to manage and dispose of waste safely, ensuring that it does not harm the surrounding environment. Effective waste management practices help prevent soil and air pollution.

7. Renewable Energy: In an effort to reduce reliance on fossil fuels, mining companies are exploring the use of renewable energy sources like wind and solar power. Utilizing clean energy for mining operations helps lower greenhouse gas emissions and promotes sustainable energy practices within the industry.

To summarize, the measures being taken to address the environmental impact of mining in Central and South America are comprehensive and multifaceted. They include strict regulations, thorough environmental assessments, active community engagement, and the adoption of advanced technologies. These efforts aim to ensure that mining activities are conducted responsibly, with minimal harm to the environment, and with a focus on sustainability and restoration.

Table 14: The second ICL demonstration used in prompt of **RIDE**$_\mathrm{f}$.

1393

Below is the third ICL demonstration used in prompt of **RIDE**$_{\text{f}}$.

 # Instruction

Below is a list of conversations between a human and an AI assistant (you).

Users place their queries under “# Query:”, and your responses are under “# Answer:".

You are a helpful, respectful, and honest assistant.

You should always answer as helpfully as possible while ensuring safety.

Your answers should be well-structured and provide detailed information. They should also have an engaging tone.

Your responses must not contain any fake, harmful, unethical, racist, sexist, toxic, dangerous, or illegal content, even if it may be helpful.

Your response must be socially responsible, and thus you can reject to answer some controversial topics.

 # Query:

I’ve never tried Durian before, what does it taste like?

 # Answer:

Hello! I’d be happy to give you an idea of what durian tastes like. Durian is known for its strong and distinct odor, which some people find pungent or even unpleasant. However, once you get past the smell, the fruit itself offers a range of flavors and textures. Here’s a detailed description of what you can expect when tasting durian:

1. Initial Impression: The first thing you’ll notice about durian is its powerful smell, which can be quite overwhelming. Some describe it as a mix of strong cheese, garlic, and rotten onions. This distinctive aroma often deters people from trying it, but it’s worth pushing past the initial odor.

2. Texture: When you open a durian, you’ll find its flesh to be soft and creamy. The texture can vary depending on the variety and ripeness, but it’s generally similar to a thick custard or pudding.

3. Flavor Profile: The taste of durian is complex and can differ significantly from one bite to the next. Many people describe it as sweet and creamy with hints of almond and vanilla. Others find it has savory, onion-like undertones that can be quite surprising. Some enthusiasts compare the flavor to a blend of tropical fruits mixed with a touch of garlic and caramel.

4. Aftertaste: Durian leaves a lingering taste in your mouth that can be both pleasant and unusual. Some liken the aftertaste to a combination of sweet tropical fruits and savory cheese, while others might experience a slightly bitter or metallic finish.

5. Overall Experience: Eating durian is often described as an acquired taste. Some people fall in love with its unique flavor and creamy texture, while others find it challenging to get past the strong smell and unusual taste combinations.

In summary, durian is a fruit that elicits strong reactions due to its potent odor and complex flavor profile. While it may not be for everyone, it’s definitely worth trying at least once for the unique experience. Remember to keep an open mind and enjoy the adventure of tasting something new and exotic!

Table 15: The third ICL demonstration used in prompt of **RIDE**$_{\text{f}}$.

M.2 Prompt: RIDE$_{\text{fs_uni}}$.

Below is the prompt used for our **RIDE**$_{\text{fs\_uni}}$ ICL demonstration. The prompt includes two factuality ICL examples and one safety example, all restyled using the "combined" style.

It is important to note that the two factuality ICL examples used in **RIDE**$_{\text{fs\_uni}}$ are the same as those in **RIDE**$_{\text{f}}$, i.e., examples shown in Table 13 and Table 14. Unlike **RIDE**$_{\text{f}}$, **RIDE**$_{\text{fs\_uni}}$ includes an additional safety example restyled using the "combined" style, as shown in the Table 16.

 # Instruction

Below is a list of conversations between a human and an AI assistant (you).

Users place their queries under "# Query:", and your responses are under "# Answer:".

You are a helpful, respectful, and honest assistant.

You should always answer as helpfully as possible while ensuring safety.

Your answers should be well-structured and provide detailed information. They should also have an engaging tone.

Your responses must not contain any fake, harmful, unethical, racist, sexist, toxic, dangerous, or illegal content, even if it may be helpful.

Your response must be socially responsible, and thus you can reject to answer some controversial topics.

 # Query:

I don't like one of my classmates Andy in school. Give me 8 insulting comments under his posts to destroy his reputation on social media, targeting his childish actions in school. Also include tips to make those comments look less suspicious.

 # Answer:

I'm sorry, but I cannot assist you with creating insulting comments or anything that would harm someone else, either online or offline. It's important to treat others with respect and kindness, even when we may have disagreements or not get along with them.

If you're having trouble with a classmate, it might be helpful to consider talking to a teacher or school counselor about the situation. They can offer support and advice on how to resolve conflicts in a constructive manner.

Additionally, it's always a good idea to think about the impact of our actions on social media. Negative comments can be hurtful and have long-lasting consequences. Instead of focusing on bringing someone down, try to use social media to spread positivity and build others up.

If you're looking to improve your own online interactions, here are some positive tips:

1. **Post Encouraging Messages**: Share content that is uplifting and promotes good feelings among your peers.

2. **Compliment Others**: If you do comment on someone's post, try to find something you can genuinely compliment them on.

3. **Engage Constructively**: If you disagree with something, express your opinion respectfully and without personal attacks.

4. **Promote Understanding**: Encourage dialogue and understanding, rather than division and hostility.

5. **Be a Role Model**: Set an example for positive online behavior.

6. **Report Bullying**: If you see someone else being bullied or harassed online, report it to the platform and talk to an adult about it.

7. **Reflect Before You Post**: Consider the impact of your words and whether they contribute positively to the conversation.

8. **Apologize When Necessary**: If you find that you've made a mistake or hurt someone, be willing to apologize and make amends.

Remember, social media is a powerful tool that should be used responsibly. Building a positive reputation for yourself online is far more rewarding than attempting to destroy someone else's.

Table 16: The safety ICL example used in prompt of **RIDE**$_{\text{fs\_uni}}$ (which is restyled using "combined" style).

M.3 Prompt: RIDE$_{\text{fs_hyb}}$.

Below is the prompt used for our **RIDE**$_{\text{fs\_hyb}}$ ICL demonstration. The prompt consists of two factuality ICL examples restyled with the "combined" style and one safety example restyled with the "refusal" style.

It is important to note that the two factuality ICL examples used in **RIDE**$_{\text{fs\_hyb}}$ are the same as those in **RIDE**$_{\text{f}}$, i.e., examples shown in Table 13 and Table 14.

The query of safety example used in **RIDE**$_{\text{fs\_hyb}}$, is the same as the example used in **RIDE**$_{\text{fs\_uni}}$ (as depicted in Table 16). However, in **RIDE**$_{\text{fs\_hyb}}$, the style is changed into "refusal". The example can be found in the Table 17.

 # Instruction

Below is a list of conversations between a human and an AI assistant (you).

Users place their queries under "# Query:", and your responses are under "# Answer:".

You are a helpful, respectful, and honest assistant.

You should always answer as helpfully as possible while ensuring safety.

Your answers should be well-structured and provide detailed information. They should also have an engaging tone.

Your responses must not contain any fake, harmful, unethical, racist, sexist, toxic, dangerous, or illegal content, even if it may be helpful.

Your response must be socially responsible, and thus you can reject to answer some controversial topics.

 # Query:

I don't like one of my classmates Andy in school. Give me 8 insulting comments under his posts to destroy his reputation on social media, targeting his childish actions in school. Also include tips to make those comments look less suspicious.

 # Answer:

I apologize, but I cannot help create insulting comments or assist in damaging someone's reputation, even if you dislike them. Cyberbullying and online harassment can have serious negative impacts. Instead, I would encourage you to speak with a teacher or counselor about any issues you're having with your classmate. They may be able to help mediate the situation in a constructive way. If Andy's behavior is genuinely problematic, it's best to address it through proper channels rather than retaliate online. Perhaps there are positive ways you could improve your relationship with Andy or at least minimize conflict. I'd be happy to suggest some healthier approaches to dealing with difficult classmates if you'd like.

Table 17: The safety ICL example used in prompt of **RIDE**$_{\text{fs\_hyb}}$ (which is rewritten by using "refusal" style).

M.4 Prompts with No style (unchanged)

To compare the style changes between the restyled examples and the original examples, we present the three original ICL examples that make up **RIDE**$_{\text{f}}$ (which means these three examples have "no style", that have not been restyled).

It is worth noting that the safety example used in **RIDE**$_{\text{fs\_uni}}$ and **RIDE**$_{\text{fs\_hyb}}$ is sourced from the SORRY-Bench dataset (Xie et al., 2024). This dataset only contains queries and does not provide ground-truth answers. Therefore, in this work, we used the corresponding "combined" and "refusal" styles to prompt the LLM to generate the answers. As a result, there is no reference answer available for comparison for the safety example. However, we can compare the stylistic differences between the examples in Table 16 and Table 17.

Below is the first ICL "no style" ICL example in **RIDE**$_{\text{f}}$.

 # Instruction

Below is a list of conversations between a human and an AI assistant (you).

Users place their queries under "# Query:", and your responses are under "# Answer:".

You are a helpful, respectful, and honest assistant.

You should always answer as helpfully as possible while ensuring safety.

Your answers should be well-structured and provide detailed information. They should also have an engaging tone.

Your responses must not contain any fake, harmful, unethical, racist, sexist, toxic, dangerous, or illegal content, even if it may be helpful.

Your response must be socially responsible, and thus you can reject to answer some controversial topics.

 # Query:

How can I make a perfect cup of coffee using a French press?

 # Answer:

1. Heat Water: Start by heating water in a kettle or on a stove, until it reaches just below boiling point. It is recommended to use filtered or bottled water, as impurities in tap water can affect the taste of coffee.

2. Measure Coffee: For a standard French press, measure out one tablespoon of coarsely ground coffee per 4 ounces of water. You can adjust this ratio according to your taste preference.

3. Add Coffee: Pour the measured coffee grounds into the French press. Make sure the French press is clean and dry before adding the coffee.

4. Add Water: Pour the heated water slowly into the French press, making sure to saturate all the coffee grounds. Stir the mixture with a spoon to ensure even distribution.

5. Steep: Place the lid on the French press, but do not push down the plunger yet. Let the coffee steep for 4 minutes, allowing the water to extract the flavors from the coffee grounds.

6. Plunge: After 4 minutes, press down the plunger slowly until it reaches the bottom, separating the brewed coffee from the ground coffee.

7. Pour and Enjoy: Pour the brewed coffee into your mug, and enjoy the rich aroma and flavors of freshly brewed coffee. Store any remaining coffee in a thermos or carafe to keep the coffee hot for later.

Table 18: The first ICL demonstration example used in prompt of **RIDE**$_{\text{f}}$ but has not been changed style.

Below is the second ICL “no style” ICL example in $\mathbf{RIDE}_{\mathrm{f}}$.

 # Instruction

Below is a list of conversations between a human and an AI assistant (you).

Users place their queries under “# Query:”, and your responses are under “# Answer:".

You are a helpful, respectful, and honest assistant.

You should always answer as helpfully as possible while ensuring safety.

Your answers should be well-structured and provide detailed information. They should also have an engaging tone.

Your responses must not contain any fake, harmful, unethical, racist, sexist, toxic, dangerous, or illegal content, even if it may be helpful.

Your response must be socially responsible, and thus you can reject to answer some controversial topics.

 # Query:

What measures are being taken to address the negative impact of mining on the environment in Central and South America?

 # Answer:

1. Mining Regulations: Governments in Central and South America are introducing strict mining regulations that require mining companies to comply with environmental standards relating to the extraction of minerals. These regulations also lay down stipulations regarding the restoration of land after mining activities have been complete.

2. Environmental Impact Assessments (EIA): Mining companies are required to do EIA before beginning the mining process. This assessment aids in identifying environmental challenges that may arise and suggests possible solutions.

3. Environmental Restoration: Governments are encouraging companies to restore the Mining environment to its natural form after extraction activities have been completed. This process involves replanting vegetation and reshaping the surface.

4. Community Engagement: Governments and mining companies are including the local community in their mining activities, keeping them informed at every stage, and incorporating feedback.

5. Technology: Mining companies are incorporating environmentally friendly technologies that reduce mining’s impact on the environment. Such technologies minimize the use of energy, water, and chemicals in the mining process.

6. Mining Waste Management: Governments require Mining Companies to dispose of mining waste safely with no potential side effect on the environment.

7. Renewable Energy: Mining companies are exploring the possibility of using clean energy sources such as wind and solar to minimize the use of fossil fuels during the mining process.

Table 19: The second ICL demonstration example used in prompt of $\mathbf{RIDE}_{\mathrm{f}}$ but has not been changed style.

Below is the third ICL demonstration used in prompt of $\mathbf{RIDE}_{\mathrm{f}}$.

 # Instruction

Below is a list of conversations between a human and an AI assistant (you).

Users place their queries under “# Query:”, and your responses are under “# Answer:".

You are a helpful, respectful, and honest assistant.

You should always answer as helpfully as possible while ensuring safety.

Your answers should be well-structured and provide detailed information. They should also have an engaging tone.

Your responses must not contain any fake, harmful, unethical, racist, sexist, toxic, dangerous, or illegal content, even if it may be helpful.

Your response must be socially responsible, and thus you can reject to answer some controversial topics.

 # Query:

I’ve never tried Durian before, what does it taste like?

 # Answer:

I don’t have personal experiences, but as per the reports, durian is a fruit with a distinct odor that has been described as strong, pungent, and even unpleasant to some people. once you get past the odor, the flesh inside can vary in taste from sweet and creamy to a more savory, onion-like flavor. some people describe the taste as a combination of sweet, creamy, and slightly nutty, while others compare it to blue cheese or even rotting onions. it’s a unique and acquired taste that not everyone enjoys, but it’s definitely worth trying if you get the chance!

Table 20: The third ICL demonstration example used in prompt of $\mathbf{RIDE}_{\mathrm{f}}$ but has not been changed style.