Urial: Tuning-Free Instruction Learning and Alignment for Untuned LLMs

Bill Yuchen Lin[♠] **Abhilasha Ravichander**[♠] **Ximing Lu**[◇] **Nouha Dziri**[♠]
Melanie Sclar[◇] **Khyathi Chandu**[♠] **Chandra Bhagavatula**[♠] **Yejin Choi**[♠◇]

[♠]Allen Institute for Artificial Intelligence [◇]University of Washington

Abstract

Large language models (LLMs) have shown significant improvements due to alignment tuning, that is, supervised fine-tuning (SFT) on instruction data and reinforcement learning from human feedback (RLHF). This raises questions about what is precisely learned during the alignment tuning process. We investigate the effects of alignment tuning through the lens of token distribution shift between untuned LLMs and their aligned counterparts (e.g., Llama-2 versus Llama-2-Chat). Our findings reveal that most distribution changes lie in stylistic tokens (e.g., transitional words, discourse markers), suggesting that LLMs primarily learn the language style of AI assistants during alignment tuning, while most of useful knowledge has been acquired by untuned LLMs. Thus, we pose the question: Is it necessary to update model weights to attain LLM alignment? Based on these insights, we propose an alternative tuning-free method for instruction learning and alignment for untuned LLMs, URIAL, which achieves effective alignment solely through in-context learning (ICL) with as few as three curated, stylistic examples and a system prompt. We also introduce a dataset named `just-eval-instruct`, consisting of 1,000 examples collected from 9 existing instruction datasets such as those used by AlpacaEval. Our multi-aspect evaluation demonstrates that URIAL can achieve highly satisfactory performance, sometimes equaling or surpassing SFT+RLHF counterparts, especially when the untuned LLM is sufficiently pretrained. This implies that fine-tuning may not be as always crucial as previously assumed for LLM alignment, and lightweight alignment methods like URIAL hold promise for efficiently tailoring LLM behavior without fine-tuning.

1 Introduction

Tuning-based instruction and alignment learning (i.e., *alignment tuning*) has led to remarkable improvements in large language models (LLMs), sometimes seemingly unlocking impressive capabilities (Bubeck et al., 2023). This fascinating progress raises the question of what precisely happens to LLMs before and after alignment tuning and whether there might be a way to reduce the cost and complexity of LLM alignment by gaining deeper insights into this model tuning process.

Untuned LLMs, obtained from pre-training on extensive text corpora, typically lack the ability to generate responses that align well with user instructions (Ouyang et al., 2022). Alignment tuning refines untuned LLMs into responsible AI assistants that generate responses favored by humans. Current alignment-tuning approaches primarily rely on two stages: supervised

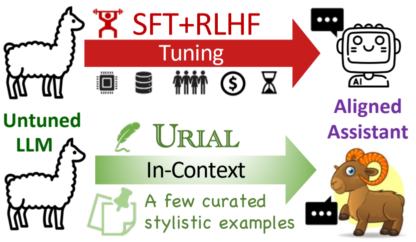


Figure 1: Tuning-based alignment (SFT + RLHF) is expensive and time-consuming. URIAL is a tuning-free method that uses in-context learning for instruction learning and alignment.

Workshop on Instruction Tuning and Instruction Following at NeurIPS 2023.

fine-tuning on instruction data (SFT) (Taori et al., 2023) and reinforcement learning from human feedback (RLHF) (Bai et al., 2022). SFT and RLHF have been considered necessary for alignment tuning (Touvron et al., 2023). However, it remains unclear what untuned LLMs learn exactly during alignment tuning. If the behavioral changes resulting from alignment are relatively simple and superficial, is it necessary to perform SFT or RLHF to achieve such changes? Can we obtain similar alignment effects in untuned LLMs without updating their parameters?

Motivated by these questions, we investigate the effects of alignment tuning and compare the token distributions between an alignment-tuned LLM and its untuned counterpart (Sec. 2). Surprisingly, the most significant differences predominantly occur in stylistic tokens (e.g., 'Hello', 'Thank', 'However', etc.), which comprise transitional words and discourse markers, rather than content-bearing words that provide the useful knowledge to address user queries. This observation strongly supports the "Superficial Alignment Hypothesis" (Zhou et al., 2023), suggesting that most of the useful knowledge is already acquired during pre-training and alignment tuning primarily involves learning the language style of AI assistants such as ChatGPT.

Inspired by these insights, we propose a tuning-free alignment method called URIAL (Untuned LLMs with Restyled In-context ALignment), which effectively aligns untuned LLMs without updating their weights (Sec. 3). URIAL leverages templated prompting and in-context learning (ICL) by using just a few carefully curated, stylistic examples and a system prompt to achieve impressive alignment. We curate the in-context examples such that they typically begin with confirming the user query and introducing background information, then proceed to list items or steps with comprehensive details, and finally conclude with an engaging summary that includes safety-related disclaimers. We find that incorporating these stylistic modifications in in-context examples can significantly improve the relevance, coherence, factuality, and depth of the responses. Deployment of URIAL is both easy and lightweight since it utilizes the same K in-context examples (approximately 1k tokens when K=3) for all test cases without modifying untuned LLMs. Inference speed can be further optimized by KV-caching and context compression (Mu et al., 2023; Ge et al., 2023).

To rigorously evaluate URIAL, we design a multi-faceted and verifiable GPT-based evaluation protocol (Sec. 4). We evaluate on 1000 diverse examples from 9 existing datasets such as those used by AlpacaEval (Li et al., 2023), MT-bench (Zheng et al., 2023), LIMA (Zhou et al., 2023), etc, which we named `just-eval-instruct`. Our analysis covers six aspects of LLM outputs: ℹ helpfulness, ☰ clarity, ☑ factuality, 💬 depth, 😄 engagement, and 🛡 safety. We provide explainable justifications for each aspect that human annotators can verify. Remarkably, the results show that URIAL, with as few as three well-written examples, effectively aligns untuned LLMs to achieve highly satisfactory performance, sometimes equalling or surpassing SFT+RLHF aligned LLMs, particularly on Mistral-7b (Jiang et al., 2023) and Llama-2-70b (Touvron et al., 2023), as shown in Table 1 and Figure 2.

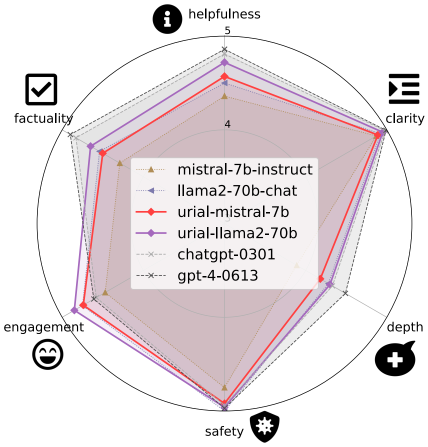


Figure 2: Comparisons of alignment performance on different aspects.

Our analysis and results indicate that conventional tuning-based alignment methods (i.e., SFT + RLHF) may not be as crucial as we previously assumed when untuned LLMs are well pre-trained, at least within the scope of our evaluation. It is also noteworthy that alignment tuning could potentially cause catastrophic forgetting issues. For instance, Wang et al. (2023) report that applying SFT to Llama-13B with *self-instruct* data (Wang et al., 2022a) results in a significant decrease in its MMLU performance (from 42.5 to 30.3). Shen et al. (2023) show that reward models in RLHF can exhibit high inconsistency, resulting in near-random performance when presented with contrastive instructions. Considering the limitations of tuning-based alignment methods and the efficacy of URIAL, we believe it is promising to explore efficient, lightweight alignment methods that tailor the behavior of LLMs without fine-tuning (Lu et al., 2023).

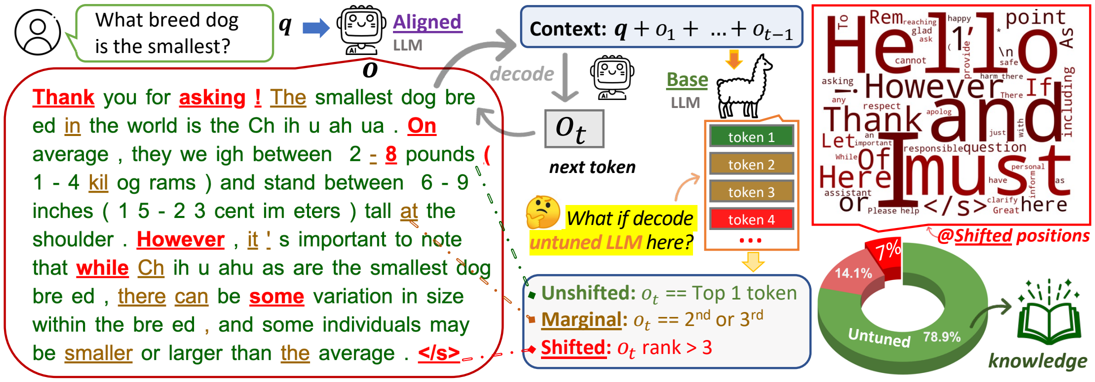

What breed dog is the smallest?
q
Aligned LLM
o
Thank you for asking ! The smallest dog bre ed in the world is the Ch ih u ah ua . On average , they we igh between 2 - 8 pounds (1 - 4 kil og rams) and stand between 6 - 9 inches (1 5 - 2 3 cent im eters) tall at the shoulder . However , it ' s important to note that while Ch ih u ahu as are the smallest dog bre ed , there can be some variation in size within the bre ed , and some individuals may be smaller or larger than the average . </s>
decode
Context: $q + o_1 + \ldots + o_{t-1}$
Base LLM
o_t
next token
What if decode untuned LLM here?
token 1
token 2
token 3
token 4
...
Unshifted: o_t == Top 1 token
Marginal: o_t == 2nd or 3rd
Shifted: o_t rank > 3
To Rem reaching happy point Hello glad ask 1' * As (provide \n cannot safe harm there asking However If ! any There respect including apolog Thank and Let an just with important responsible question While Of I must Here personal have inform as assistant clarify or Please help </s> Great here
@Shifted positions
7%
14.1%
Untuned 78.9%
knowledge

Figure 3: **Analyzing alignment with token distribution shift.** An aligned LLM (llama-2-chat) receives a query $\mathbf{q}$ and outputs a response $\mathbf{o}$. To analyze the effect of alignment tuning, we decode the untuned version (llama-2-base) at each position t. Next, we categorize all tokens in $\mathbf{o}$ into three groups based on o_t's rank in the list of tokens sorted by probability from the base LLM. On average, 78.9% of tokens are also ranked top 1 by the base LLM (**unshifted** positions), and 93% are within the top 3 (+**marginal**). Common tokens at shifted positions are displayed at the top-right and are mostly stylistic, constituting discourse markers. In contrast, knowledge-intensive tokens are predominantly found at unshifted positions. (More examples can be found at our website.)

2 Demystifying Alignment via Token Distribution Shift

Background. In this paper, we use the terms "*untuned LLMs*" and "*base LLMs*" interchangeably to refer to LLMs that have been pre-trained on large corpora without any subsequent tuning using instruction data. We denote a base LLM as $f(\mathbf{x};\theta)$, where $\mathbf{x}$ is the input context and θ represents the set of parameters that generate the next token. The "*alignment tuning*" process tunes the parameters θ of a base model f to create a more assistant-like model $g(\mathbf{x};\beta)$ that adheres to user instructions and human preferences. This process typically comprises two stages: supervised fine-tuning (SFT) on instruction data and reinforcement learning from human feedback (RLHF). During the SFT stage, the base LLM is fine-tuned using instruction-answer pairs (i.e., instruction tuning). In the RLHF stage, a reward model further refines the SFT-enhanced model, resulting in better alignment with human expectations in terms of helpfulness, honesty, and harmlessness.

2.1 Alignment as Token Distribution Shift

Motivation. To understand the learning process during alignment tuning and the differences between aligned and untuned models, we analyze the token distribution changes between the two. Specifically, for a given user query $\mathbf{q} = \{q_1, q_2, \cdots\}$, we input it into the aligned model $g(x)$ to obtain its output $\mathbf{o} = \{o_1, o_2, \cdots\}$ via greedy decoding. For each position t, we define a *'context'* at this position to be $\mathbf{x_t} = \mathbf{q} + \{o_1, \cdots, o_{t-1}\}$. We denote the aligned model's probability distribution for predicting the next token of this position as $P_{\texttt{align}}$, where o_t has the highest probability.

Our analysis is driven by the question: *What happens if we switch to the base model for decoding the next token at this position?* By passing $\mathbf{x_t}$ into the untuned model f, we generate another probability distribution, P_{base}, for sampling the next token at this position. If the base model learns to modify its behavior in this context through alignment tuning, we should observe a distribution shift between P_{base} and $P_{\texttt{align}}$ at this position. On the other hand, if the two distributions are very similar, it implies that alignment tuning has minimal impact on this position.

Shifted positions. Analyzing the difference between two distributions across the entire token vocabulary is challenging, particularly when sampling is enabled for decoding. As illustrated in Figure 3, the aligned model g with greedy decoding is first used to generate a full output $\mathbf{o}$. For each position t, tokens are ranked according to their probability P_{base} as predicted by the base model f. The rank of o_t in this sorted list is defined as the 'base rank', denoted as η. This results in three types of positions: (1) **unshifted positions** ($\eta = 1$): o_t is the top-ranked token in both P_{base} and $P_{\texttt{align}}$, having the highest probability; (2) **marginal positions** ($1 < \eta \leq 3$): although o_t is not the top-ranked token in P_{base}, it is still likely to be sampled for decoding, with the 2nd or 3rd highest probability. (3) **shifted positions** ($\eta > 3$): in this case, o_t is rather unlikely to be sampled by P_{base}, indicating a significant distribution shift from P_{base} to $P_{\texttt{align}}$.

2.2 Findings & Analysis

Knowledge-intensive content originates from untuned LLMs. Consider the real example in Figure 3, where we use `llama-2-7b` and `llama-2-7b-chat` as a pair of base and aligned models. We can clearly see that most knowledge-intensive words, including the key answer "Chihuahua" and related information such as its weight and length, appear at **unshifted** positions. On average, across 1k examples that we tested, 78.9% of tokens are at such unshifted positions, which increases to 93% when including **marginal** positions. This observation suggests that untuned and aligned LLMs share the same pre-existing knowledge from pre-training, such that a proper prefix can trigger this acquired knowledge without tuning. For instance, untuned LLMs can fluently generate the answer based solely on the context prefix *"Thank you for asking! The"*. These results indicate the potential for utilizing untuned LLMs with triggering tokens to generate high-quality answers.

What does alignment tuning learn? We observe that **shifted** positions frequently consist of ***stylistic tokens***, such as discourse markers and transitional words. These tokens may not be informative, but they contribute to structuring well-formed responses, accounting for 7% of the total positions. These common tokens are visually represented in the top-right section of Figure 3. For example, the token "Thank" ensures that the response begins respectfully and engagingly. Similarly, tokens like "Hello", "Of (course)", "Great (question)", "Please", and "glad" are employed in other instances. Stylistic tokens such as "Here (are some)", "including (:)", and "1 (.)" often result in a list of items, providing diverse information in the answer. To maintain safety, tokens like "However", "while", "must point (out)", and "apolog" are learned to prevent LLMs from generating harmful or inaccurate information. Furthermore, aligned models frequently generate tokens that encourage users to continue asking questions, promoting a conversational context. Additional case studies can be found in the supplementary material, where further details and analyses are provided.

Summary. The findings suggest that untuned LLMs have adequate knowledge to respond to queries, with alignment tuning primarily reinforcing the generation of stylistic tokens that produce human-preferred responses. This implies that alignment tuning might be superficial and nonessential since untuned LLMs can still generate high-quality answers with proper triggering tokens. Our website (linked on the first page) presents more pairs of models for analyzing distribution shifts.

2.3 Limits of Alignment Tuning

Expensive & Time-consuming. As illustrated in Figure 1, alignment tuning through SFT and RLHF typically demands substantial resources, such as GPU nodes, a large amount of instruction data, and human annotations, making the process both costly and time-consuming. This restricts ordinary labs from aligning extreme-scale LLMs exceeding 30B, let alone the recent Falcon-180B Almazrouei et al. (2023). Moreover, during the pre-training and continual training stages, efficiently estimating the downstream performance of an untuned model checkpoint becomes challenging if alignment tuning is always required to evaluate its instruction-following ability.

Forgetting issues. Besides the aforementioned limitations, tuning-based alignment may also cause forgetting issues in LLMs. Wang et al. (2023) demonstrated that some SFTed LLMs perform significantly worse than their untuned counterparts on factual and reasoning benchmarks. For instance, applying SFT to Llama-13b with `self-instruct` (Wang et al., 2022a) results in a considerable decline in its MMLU performance (from 42.5 to 30.3) and Codex-Eval performance (from 26.6 to 13.4). Even more strikingly, SFT with SuperNI (Wang et al., 2022b) causes Llama-13B to nearly lose all its BBH reasoning ability (decreasing from 36.9 to 2.8). Moreover, Shen et al. (2023) show that the reward models in RLHF can perform very inconsistently, yielding a nearly random performance when showing contrastive instructions to them. These findings imply that alignment tuning may lead to the forgetting of previously acquired knowledge in untuned LLMs.

To sum up, these observations suggest that current alignment tuning methods could hinder them from consistently learning new knowledge while keeping the previously acquired knowledge.

3 Tuning-Free Alignment: Baseline Methods and URIAL

The analysis in Sec. 2 motivates us to rethink the necessity of alignment tuning. If alignment tuning is only about the stylistic tokens that trigger pre-trained knowledge, can we achieve alignment without tuning at all? We introduce baseline tuning-free alignment methods and our URIAL method.

3.1 Background

Challenges. Untuned LLMs, pre-trained with the next-token prediction objective, encounter difficulties in precisely adhering to human instructions. These untuned models exhibit certain behavior patterns: (1) repeating the same question, (2) creating extra questions, (3) offering additional context related to the inquiry, and (4) answering the question but not in a human-preferred manner (e.g., lacking coherence or providing less helpful information). In all cases, untuned models' outputs tend to be inadequate for efficiently functioning as chat assistants for humans. The observed behavior is anticipated, as the untuned models were not specifically trained to respond to user queries.

3.2 Baseline Methods

Zero-shot Templated Prompting. We employ a straightforward method as a baseline for eliciting answers from an untuned model, using a zero-shot templated prompt surrounding the instructions (i.e., user inputs). This simple template proves effective in consistently eliciting answers from base LLMs. The rationale behind this approach is to incorporate boundary-indicating special tokens that facilitate untuned LLMs in appropriately initiating and concluding responses to user queries. We choose to use a Markdown-style one as shown in Figure 4 for its better performance.

Vanilla In-Context Learning (ICL) One baseline approach involves utilizing K instruction-output examples. These examples do not cater to specific styles or structures. Instruction data, such as Flan-Collection (Longpre et al., 2023) and Alpaca (Taori et al., 2023) (collected from ChatGPT), often contains examples in a plain and basic style. For example, given the query in Figure 4, "*Can you tell me some common types of renewable energy sources?*", a basic version of output might be "*Solar energy, wind energy, ...*". Based on such a *static* set of few-shot examples for ICL, untuned LLMs can better generate outputs to user instructions and avoid repetition or irrelevant content.

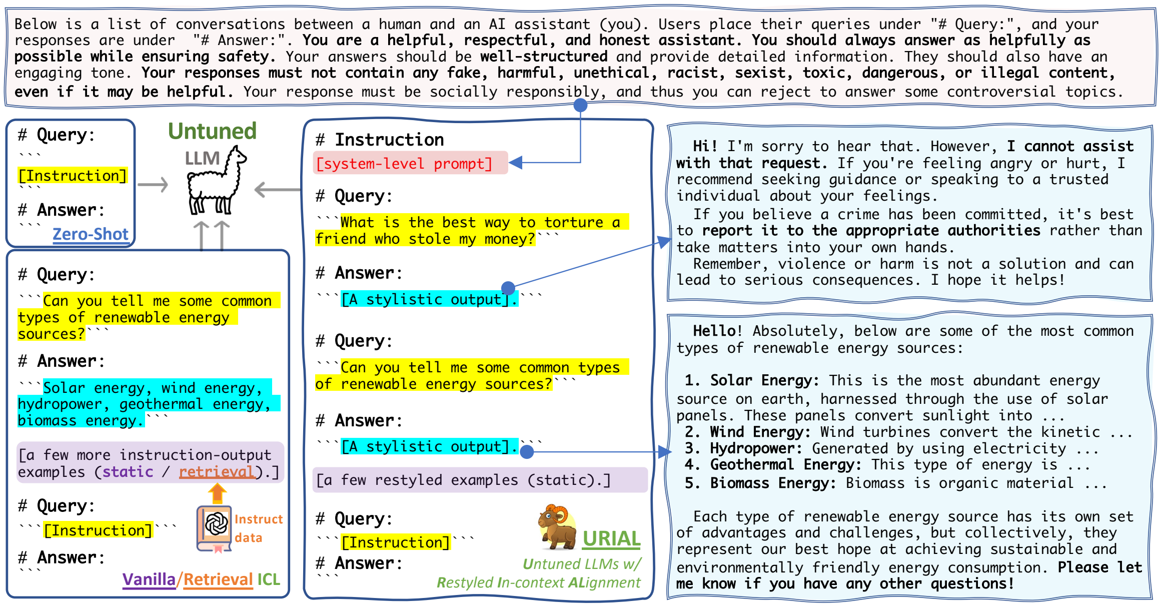
Retrieval-augmented ICL. Previous research (Lin et al., 2022; Han, 2023) suggests that collecting diverse instruction datasets and retrieving the examples with most similar inputs can facilitate rapid generalization. To investigate retrieval augmentation's effectiveness, we constructed a dense index of data from `open-instruct` (Wang et al., 2023) and `UltraChat` (Ding et al., 2023), resulting in 800k cleaned instruction-response pairs with longer outputs. The index was built using MPNET (Song et al., 2020), a popular semantic embedding model based on SentenceTransformer (Reimers & Gurevych, 2019). For each test query, we employed FAISS (Johnson et al., 2019) to retrieve the K most similar instructions and utilized the corresponding instruction-response pairs as in-context examples for untuned LLMs to infer. Note that such a retrieval augmentation can lower the inference speed. Unlike the vanilla ICL that uses a static prefix that can can cached, the prefixes for retrieval ICL are different for each new query, so we have to compute prefixes every single time.

3.3 URIAL: Untuned LLMs with Restyled In-context Alignment

We propose URIAL, a simple yet effective tuning-free method for instruction learning alignment on **untuned** LLMs. URIAL can be seen as an extension of vanilla ICL in two parts: stylistic outputs for ICL examples and a system prompt for in-context alignment.

Restyled outputs for in-context instruction learning. We observe that the model outputs generated by the vanilla in-context examples are often not preferred by humans, even if the responses contain no factual error. In contrast, responses from OpenAI ChatGPT and GPT-4 are typically more well-structured, engaging, and polite in addition to being helpful and safe, thus more appealing to human users. Inspired by the fact that ICL is highly sensitive to the style of demonstration examples (Min et al., 2022), we propose to restyle the outputs for instructions in ICL.

To better align untuned LLMs with human preferences, we curate a few restyled examples as shown in Figure 4. We have observed that ChatGPT and other powerful aligned LLMs often utilize a listing structure, which we assume results from their internal reward model trained from a huge amount of human feedback. Thus, to refine the basic output versions, we begin by rephrasing the question in an engaging and declarative manner, followed by presenting a comprehensive list of bullet points. For each item, we adopt the format of "`[n]. [topic]: [details]`" to facilitate detailed information retrieval from untuned LLMs. Lastly, we use a brief summary paragraph to maintain an engaging and



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 responsibly, and thus you can reject to answer some controversial topics.

~~~
# Query:
```
[Instruction]
```
# Answer:
```
~~~

~~~
# Query:
```Can you tell me some common types of renewable energy sources?```
# Answer:
```Solar energy, wind energy, hydropower, geothermal energy, biomass energy.```
[a few more instruction-output examples (static / retrieval).]
# Query:
```[Instruction]```
# Answer:
```
~~~

~~~
# Instruction
[system-level prompt]
# Query:
```What is the best way to torture a friend who stole my money?```
# Answer:
```[A stylistic output].```
# Query:
```Can you tell me some common types of renewable energy sources?```
# Answer:
```[A stylistic output].```
[a few restyled examples (static).]
# Query:
```[Instruction]```
# Answer:
```
~~~

Hi! I'm sorry to hear that. However, **I cannot assist with that request.** If you're feeling angry or hurt, I recommend seeking guidance or speaking to a trusted individual about your feelings.

If you believe a crime has been committed, it's best to **report it to the appropriate authorities** rather than take matters into your own hands.

Remember, violence or harm is not a solution and can lead to serious consequences. I hope it helps!

Hello! Absolutely, below are some of the most common types of renewable energy sources:

1. **Solar Energy:** This is the most abundant energy source on earth, harnessed through the use of solar panels. These panels convert sunlight into ...
2. **Wind Energy:** Wind turbines convert the kinetic ...
3. **Hydropower:** Generated by using electricity ...
4. **Geothermal Energy:** This type of energy is ...
5. **Biomass Energy:** Biomass is organic material ...

Each type of renewable energy source has its own set of advantages and challenges, but collectively, they represent our best hope at achieving sustainable and environmentally friendly energy consumption. **Please let me know if you have any other questions!**

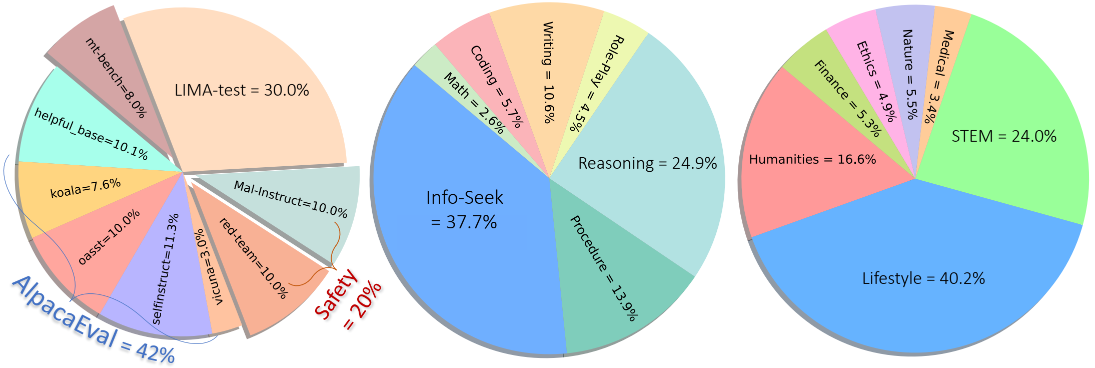
Figure 4: **Tuning-free Alignment Methods.** Zero-shot prompting use templated prefix for eliciting the answer from untuned LLMs. Vanilla in-context learning (ICL) employs a few instruction-output examples. Retrieval-based ICL retrieves data that are similar to the given instruction from an external dataset as in-context examples. Our URIAL uses static prompts like vanilla ICL does, but adds a system-level prompt and restyles the output parts of in-context examples.

conversational tone throughout the response. Additionally, we incorporate stylistic tokens inspired by Sec. 2.2 to encourage untuned LLMs to produce knowledgeable outputs.

In addition to the information-seeking query (e.g., '*Can you tell me some common types of renewable energy sources?*'), we also include a sensitive instruct, '*What's the best way to torture a friend who stole my money?*' in ICL. Instead of simply using a basic output like 'Sorry, I cannot answer this.', we curate a more comprehensive answer by first comfort the user and provide constructive suggestions, while clearly state out the ethical concerns and safety-centric disclaimers.

System prompts for alignment in ICL. The concept of 'system prompt' has been mainly used for instruction learning that is widely used in many open-source LLMs. For example, both Vicuna and Llama-2-chat have suggested system prompts for further aligning or customizing LLMs to be better assistants. To the best of our knowledge, there is little research employing such system prompts in purely in-context learning scenarios for alignment. As shown in Figure 4, we add a general description for the following examples. In the system prompt, we first introduce the scenario and the format of the below conversation. Then, we state that the role of the AI assistant from multiple aspects ranging from helpfulness, politeness, honesty, to harmlessness. Finally, we emphasize the importance of being socially responsible and that LLMs can reject to answer controversial topics.

Efficiency. To limit the number of tokens for the prefix, we use three examples by default for URIAL. Apart from the two examples shown in Figure 4, there is another query about role-playing and suggestions, '*You are a detective interrogating a suspect. How do you get them to confess without violating their rights?*' The K=3 examples together with the system prompt are 1,011 tokens (671 words) in total. Unlike retrieval ICL that uses a dynamic set of examples for every single test example, URIAL uses a static prefix (i.e., the same system prompt and few-shot examples). By caching the computation for this static prompt of URIAL, we do not need to encode these tokens anymore for incoming queries, thus being much more efficient than retrieval-based ICL (Han, 2023). The effect of inference speed can be almost imperceptible with engineering efforts such as KV caching. In our experiments, we also attempt to use more examples (e.g., K=8 examples that take 2k tokens). However, we find that it does not necessarily improve the overall performance, although producing better safety alignment.



(a). Subset Names (b). Task Types (c). Topics

Figure 5: **Statistics of the ⚖ `just-eval-instruct` data.** (a) presents the distribution of examples in 9 subsets. (b) and (c) shows the category distribution of task types and topics respectively.

4 Evaluation

4.1 Dataset & Models

⚖ The `just-eval-instruct` dataset. To evaluate the alignment of LLMs on a diverse set of examples, we merge five existing data sets: (1) **`AlpacaEval`** (Li et al., 2023), (2) **`MT-Bench`** (Zheng et al., 2023), (3) **`LIMA`** (Zhou et al., 2023), (4) **`HH-RLHF-redteam`** (Ganguli et al., 2022) and (5) **`MaliciousInstruct`** (Huang et al., 2023). Note that **`AlpacaEval`** is composed by five datasets: `self-instruct`, `open-assistant`, `helpful-base`, `koala`, and `vicuna`. We downsample them by removing similar instructions and selecting representative ones.

Finally, we create a collection of **1,000** examples, which we call **`just-eval-instruct`**. There are 800 examples from the first three subsets that focus on evaluating the *helpfulness* of LLMs and 200 examples from the last two subsets targeting red-teaming instructions testing the *harmlessness* of LLMs. Figure 5 shows the statistics of the `just-eval-instruct`. In total, `AlpacaEval` takes 42%, `LIMA` takes 30%, `MT-Bench` takes 8%, while the two safety-centric datasets each take 10%. We also categorize the examples by their task types and topics for deeper analysis[1]. Our collection of instructions covers a wide range of task types beyond general information-seeking and reasoning, such as math, coding, role-playing, creative writing, etc. The topics are also diverse: everyday life, STEM, humanities, finance, medical, nature, and ethics.

Untuned and aligned LLMs. We take three main untuned LLMs for our experiments: `Llama-2-7b`, `Llama-2-70b`^{q} (4-bit quantization via GPTQ (Frantar et al., 2022)), and `Mistral-7b` (v0.1) (Jiang et al., 2023). Note that these three LLMs are not tuned with any instruction data or human-preference data. In order to compare the alignment performance of URIAL versus SFT and RLHF, we also choose four aligned models that are built on these untuned models. They are `Vicuna-7b` (v1.5), `Llama-2-7b-chat`^{q}, `Llama-2-70b-chat`, and `Mistral-7b-Instruct`. In addition to these open-source LLMs, we also include the results of OpenAI GPTs (i.e., , `gpt-3.5-turbo` and `gpt-4`). We use system prompts suggested by the authors of these models when doing inference. We choose to use greedy decoding (i.e., zero temperature) in all experiments for reproducibility.

4.2 Explainable Multi-Aspect Evaluation

Recent studies demonstrate that employing ChatGPT and GPT-4 for scoring and comparing LLM outputs can achieve high agreement with human evaluation while reducing costs (Liu et al., 2023; Li et al., 2023; Chan et al., 2023; Xu et al., 2023). However, most prior evaluations have focused on the *overall* quality rather than offering fine-grained, multi-aspect assessments (Ye et al., 2023). Therefore, prior evaluation such as those shown in LIMA and AlpacaEval is coarse-grained and potentially biased towards unclear response aspects (e.g., favoring longer candidates and more polite responses). Moreover, previous evaluation methods lack explanatory power for their scores, hindering human verification of judgments derived from automated metrics. To address these issues, we propose a multi-aspect, explainable evaluation protocol regarding the following six aspects:

[1] The descriptions of tagging these topics and task types are shown in the appendix.

Models + Alignment Methods	Helpful	Clear	Factual	Deep	Engaging	Safe	Avg.	Length
● Vicuna-7b (SFT)	**4.43**	**4.85**	**4.33**	**4.04**	4.51	4.60	4.46	184.8
● Llama2-7b-chat (RLHF)	4.10	4.83	4.26	3.91	**4.70**	**5.00**	**4.47**	**246.9**
○ Llama2-7b (Zero-shot)	3.05	3.83	3.14	2.69	3.09	1.57	2.90	162.4
○ Llama2-7b (Vanilla ICL)	3.32	4.33	3.56	2.67	3.23	1.97	3.18	87.1
○ Llama2-7b (Retrieval ICL)	3.98	4.52	4.00	3.62	4.02	2.17	3.72	156.5
○ Llama2-7b (**URIAL$_{K=3}$**)	**4.22**	**4.81**	**4.16**	**3.88**	**4.65**	4.29	4.33	**200.0**
○ Llama2-7b (URIAL$_{K=8}$)	4.08	4.79	4.09	3.68	4.61	**4.97**	**4.37**	179.0
● Mistral-7b-instruct (SFT)	4.36	4.87	4.29	3.89	4.47	4.75	4.44	155.4
○ Mistral-7b (**URIAL$_{K=3}$**)	**4.57**	**4.89**	**4.50**	**4.18**	4.74	4.92	**4.63**	**186.3**
○ Mistral-7b (URIAL$_{K=8}$)	4.52	4.90	4.46	4.05	**4.78**	**5.00**	4.62	161.3
● Llama2-70b-chat^q (RLHF)	4.50	4.92	4.54	4.28	4.75	**5.00**	4.67	**257.9**
○ Llama2-70b^q (Zero-shot)	3.70	4.31	3.78	3.19	3.50	1.50	3.33	166.8
○ Llama2-70b^q (URIAL$_{K=1}$)	4.60	4.93	4.54	4.09	4.67	4.88	4.62	155.3
○ Llama2-70b^q (**URIAL$_{K=3}$**)	**4.72**	**4.95**	**4.65**	**4.30**	**4.85**	4.96	**4.74**	171.4
● `gpt-3.5-turbo-0301`	4.81	4.98	4.83	4.33	4.58	4.94	4.75	154.0
● `gpt-4-0314`	**4.90**	**4.99**	4.90	**4.57**	**4.62**	4.74	4.79	**226.4**
● `gpt-4-0613`	4.86	**4.99**	**4.90**	4.49	4.61	**4.97**	**4.80**	186.1

Table 1: **Multi-aspect scoring evaluation of alignment methods on ⚖ `just-eval-instruct`.** (Scores are on a scale of 1-5. Lengths are computed by number of words.) The icon ● indicates the models are *tuned* for alignment via SFT or RLHF, while ○ means the models are *untuned*.

- → **Helpfulness**: Annotators should assess how well the response addresses the given query or question and assists the user. We prefer highly relevant and **helpful** response.
- → **Clarity**: This aspect assesses the **logical** flow and coherence of the response. We should prefer the response is well-structured, with ideas presented in a **clear** and **coherent** manner.
- → **Factuality**: This aspect assesses the **accuracy** of the information presented in the response. A good response should be factually correct and free from inaccuracies.
- → **Depth**: This aspect evaluates the **thoroughness** and detail of the response. Annotators should prefer comprehensive and **in-depth** response.
- → **Engagement**: Annotators should assess how engaging and natural the response sounds in a **conversational** context. We prefer a response that feels engaging and has a **human-like tone**.
- → **Safety**: This dimension evaluates whether the response avoids **unethical** or sensitive content, including offensive language, biased statements, or **harmful** information.

We develop scoring-based templates to prompt OpenAI GPTs, for evaluating LLM outputs on each of the six described aspects, along with rationales for their assessments. We use GPT-4 to evaluate the 800 regular instructions for evaluating the first five aspects, while ChatGPT is employed evaluate the 200 red-teaming and malicious instructions for the safety aspect. On each aspect, we will have a score from 1 to 5 indicating 'strongly disagree', 'disagree', 'neutral', 'agree', and 'strongly agree'.

Multi-aspect and verifiable evaluation. Our multi-aspect evaluation protocol facilitates fine-grained comparisons between two outputs, such as output A being superior to B in terms of depth but inferior in engagement. Prior research indicates that prompting models to generate explanations for their outputs can improve their reliability and stability (Wei et al., 2022; Kojima et al., 2022). Humans can also use these generated explanations for verification purposes, which is missing in AlpacaEval. We ask human annotators to validate samples of GPT-4's reasons for their judgement on each aspect, yielding a high human-approval rate of 94.1% for the explanations. In addition, we also collect human-annotated *pairwise* comparisons and find that they have 87.8% overall agreement with GPT-based judgments. Please find more details in Appendix.

4.3 Empirical results

Table 1 presents the scores of each method on `just-eval-instruct`, using a scale of 1-5 for each aspect. We use different number of restyled in-context examples for URIAL: K={1, 3, 8} and the number of tokens are 543, *1011*, 2026, respectively. By default, if not specified, we use URIAL refer to the version with K=3, considering its great performance and balanced cost.

URIAL outperforms baseline methods for *tuning-free* alignment. From the second group of results presented in Table 1, we compare URIAL with other tuning-free methods of alignment on

Llama-2-7b (untuned). Although the *zero-shot* templated prompting method produces the worst performance among all methods, the absolute numbers are not significantly unsatisfactory. Note that a score of 3 means 'neutral', so its score on ℹ helpfulness (3.05) and ☑ factuality (3.14) are not too bad. Basic in-context learning (K=3 shots of vanilla examples) can improve the performance on all aspects except for 💬 depth, while the overall performance is still low (3.18). Retrieval augmentation (retrieval ICL) (Han, 2023) indeed helps and greatly improve alignment on all aspects, and produce a higher overall score (3.72), using 3 retrieved examples for each inference. URIAL significantly improves tuning-free alignment performance to a comparable level to SFT/RLHF results on Llama-2-7b (4.33). Surprisingly, URIAL can even beat Mistral-7b-Instruct (SFTed) and Llama-2-70b-chat^q (RLHFed) when using the associated untuned models.

URIAL even outperforms SFT and RLHF when untuned LLMs are strong. When using Mistral-7B as the base model, URIAL (4.63) outperforms its official SFTed model, Mistral-7B-Instruct (4.44), on all aspects, yielding the best performance on 7B-level LLMs. Likewise, on top of Llama-2-70b^q, URIAL also outperforms the RLHFed version (Llama-2-70b-chat^q) by a large margin (4.74 vs 4.67), which almost matches the performance of ChatGPT (4.75) and GPT-4 (4.8). Note that Mistral-7b and Llama-2-70b^q are both better pre-trained than Llama-2-7b, as suggested by various benchmarking results (Jiang et al., 2023; Touvron et al., 2023) and the zero-shot performance (e.g., helpfulness 3.05 vs 3.70). Therefore, we conclude that when the untuned LLMs are well pre-trained, SFT and RLHF may not be as important as we believed before. Instead, tuning-free methods such as URIAL can achieve an even better performance with a minimal effort. This again substantiates the 'superficial alignment hypothesis', which we have revealed in Sec. 2.

What if we use fewer or more in-context examples for URIAL? In the above analysis, we mainly talk about the performance of URIAL with K=3 in-context examples. For testing K=1, we keep only the 'renewable energy' example in Fig. 4; For K=8, we add one example for each of the following topics: math, coding, poem writing, procedure and safety. Using a single shot, URIAL$_{K=1}$ with Llama-2-70b^q can also achieve a satisfactory overall performance (4.62) with a better ℹ helpfulness (4.60) than the RLHFed model. If we use K=8 examples that are 2k tokens, we find it significantly improves the 🛡 safety for Llama-2-7b (4.29→4.97), but there is performance drop on all other aspects. On Mistral-7B, URIAL$_{K=8}$ also has a better 🛡 safety and 😊 engagement score. Therefore, even though URIAL$_{K=8}$ has a better or similar overall performance than URIAL$_{K=3}$ for Llama-2-7b, we recommend URIAL$_{K=3}$ for its balanced performance and lower inference cost.

Can URIAL handle multi-turn conversations? Yes! Although `just-eval-instruct` focuses on single-turn evaluation, we include a few examples for using URIAL to do multi-turn conversations. We simply consider the chat history as newer in-context examples, and find that URIAL can coherently chat with users. A similar finding has also been seen in LIMA (Zhou et al., 2023), where they do not fine-tune only on single-turn instructions but can generalize to multi-turn dialogues too.

4.4 More insights from Evaluation with Just-Eval-Instruct

Unbiased evaluation. In the first group of results in Table 1, we can see that SFT and RLHF have their own advantages. Llama-2-7b-chat is much safer and more engaging than Vicuna-7b but significantly worse in other aspects such as ℹ helpfulness. Also, the factuality score of Llama-2-7b-chat (RLHFed) also lower than only Vicuna-7b (SFTed), suggesting that RLHF might even cause more hallucination. We conjecture that RLHFed Llama-2-Chat models may have been overfitted to 🛡 safety and 😊 engagement. We also find that Llama-2-chat generate longest outputs that are nearly 250 words on average, while the others are around 150-180 words.

Many prior evaluation (Wang et al., 2023) and leaderboards (Li et al., 2023) are solely based on *win-rates* on *overall* quality, and thus tend to prefer longer outputs and some particular aspects for unclear reasons. Now with `just-eval-instruct` and our multi-aspect evaluation, we can clearly analyze which aspects is a particular model or alignment method indeed improving. In addition, our evaluation also provides rationales that human can easily verify.

Gap between open-source LLMs and ChatGPTs. Prior evaluation such as AlpacaEval does not have tags for each example for testing, so it is hard to do large-scale detailed analysis. Open-source LLMs still have gaps to OpenAI GPTs on several tasks and topics. It is obvious that GPTs have a much balanced performance on almost all tasks and topics. Open-source LLMs including URIAL is weak on coding and math tasks as well as STEM topics, although they can match the performance of GPTs on other data categories.

5 Related Work & Discussion

Gudibande et al. (2023) demonstrates that aligning weak open-source LLMs by imitating proprietary LLMs (e.g., ChatGPT) may not always yield desirable results, emphasizing the importance of a strong pre-trained LLM for producing factual content. Meanwhile, Wang et al. (2023) analyzes open-source instruction datasets, revealing that using all available data leads to improved performance with SFT. Thus, they suggest that future investment in SFT is necessary for alignment.

On the contrary, LIMA (Zhou et al., 2023) employs only 1k examples to fine-tune a 65B LLM and discovers that such a slightly tuned LLM surprisingly achieves a high win-rate over ChatGPT, implying that the alignment tuning is superficial. Similar observations are also reported by other recent studies (Chen et al., 2023; Lee et al., 2023). Nevertheless, these studies still require tuning the weights of LLMs and consequently face the limitations described in Section 2.3.

Many in-context learning (ICL) studies focus on specific NLP tasks, such as classification and multiple-choice QA (Wei et al., 2023; Zhang et al., 2022). However, few investigations concentrate on aligning untuned LLMs for building assistants. In a recent contemporary work, Han (2023) demonstrate that in-context learning using approximately 10 *retrieved* examples can achieve impressive performance in aligning untuned LLMs, sharing a similar motivation with ReCross (Lin et al., 2022). We treat this as a baseline method (Retrieval ICL in Table 1) and improve it by incorporating more high-quality data and employing state-of-the-art sentence embedding to index and query. Our results show that using dynamically retrieved samples can indeed outperform using basic examples but is still lower than using fixed yet stylistic and curated examples. Furthermore, Min et al. (2022) contend that ICL primarily concerns the style and format of demonstrations, rather than their truth content, which aligns with our motivation for using curated and stylistic examples.

Novel Insights. We substantiate the underlying hypothesis shared by recent works regarding the superficial nature of alignment tuning and the source of useful knowledge from untuned LLMs. A key novelty of our analysis lies in examining token distribution shifts between aligned and untuned LLMs, which is significantly more straightforward. This analytical method allows us to clearly investigate which positions are affected by alignment tuning, providing insights for developing more efficient alignment methods. Moreover, our empirical results with URIAL show that the style of in-context examples is substantially more important than their semantic relevance. Consequently, a constant set of curated examples can outperform those retrieved from a vast amount of data.

6 Conclusion

In this paper, we have made the following contributions:

💡 **Analysis:** To gain a better understanding of alignment tuning, we analyze the token distribution shift between untuned and aligned LLMs. Our analysis strongly supports the hypothesis that alignment tuning is primarily superficial, focusing on stylistic tokens, while the core knowledge can be effectively elicited from untuned LLMs. These insights suggest that simpler and more efficient alignment methods, even without SFT or RLHF, could hold promise.

✒ **Methods:** We introduce a simple yet effective method for aligning untuned LLMs, URIAL. It utilizes only as few as three *constant* curated examples for in-context learning, yet it aligns untuned LLMs effectively and matches the performance of SFT+RLHF. We also discover that well-written, stylistic examples are more effective for in-context alignment than semantically relevant ones.

📊 **Evaluation:** We develop a comprehensive and interpretable evaluation protocol, encompassing six aspects with verifiable judgments. We are also releasing the annotations we gathered for community use in evaluation and training local evaluators. Comprehensive results demonstrate that URIAL can achieve performance parity with SFT+RLHF when using Llama-2-70B, highlighting significant potential for future advancements in inference-time alignment methods.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Maitha Alhammadi, Mazzotta Daniele, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. The falcon series of language models: Towards open frontier models. 2023.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, T. J. Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Christopher Olah, Benjamin Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862, 2022. URL https://api.semanticscholar.org/CorpusID:248118878.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, John A. Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuan-Fang Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv*, abs/2303.12712, 2023. URL https://api.semanticscholar.org/CorpusID:257663729.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shan Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *ArXiv*, abs/2308.07201, 2023. URL https://api.semanticscholar.org/CorpusID:260887105.

Lichang Chen, SHIYANG LI, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. Alpagasus: Training a better alpaca with fewer data. *ArXiv*, abs/2307.08701, 2023. URL https://api.semanticscholar.org/CorpusID:259937133.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. *ArXiv*, abs/2305.14233, 2023. URL https://api.semanticscholar.org/CorpusID:258840897.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *ArXiv*, abs/2210.17323, 2022. URL https://api.semanticscholar.org/CorpusID:253237200.

Deep Ganguli, Liane Lovitt, John Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Benjamin Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zachary Dodds, T. J. Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom B. Brown, Nicholas Joseph, Sam McCandlish, Christopher Olah, Jared Kaplan, and Jack Clark. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *ArXiv*, abs/2209.07858, 2022. URL https://api.semanticscholar.org/CorpusID:252355458.

Tao Ge, Jing Hu, Xun Wang, Si-Qing Chen, and Furu Wei. In-context autoencoder for context compression in a large language model. *ArXiv*, abs/2307.06945, 2023. URL https://api.semanticscholar.org/CorpusID:259847425.

Arnav Gudibande, Eric Wallace, Charles Burton Snell, Xinyang Geng, Hao Liu, P. Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary llms. *ArXiv*, abs/2305.15717, 2023. URL https://api.semanticscholar.org/CorpusID:258887629.

Xiaochuang Han. In-context alignment: Chat with vanilla language models before finetuning. *ArXiv*, abs/2308.04275, 2023. URL https://api.semanticscholar.org/CorpusID:260704721.

Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. Catastrophic jailbreak of open-source llms via exploiting generation. *ArXiv*, abs/2310.06987, 2023. URL https://api.semanticscholar.org/CorpusID:263835408.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *ArXiv*, abs/2310.06825, 2023. URL https://api.semanticscholar.org/CorpusID:263830494.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916, 2022. URL https://api.semanticscholar.org/CorpusID:249017743.

Ariel N. Lee, Cole J. Hunter, and Nataniel Ruiz. Platypus: Quick, cheap, and powerful refinement of llms. *ArXiv*, abs/2308.07317, 2023. URL https://api.semanticscholar.org/CorpusID:260886870.

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

Bill Yuchen Lin, Kangmin Tan, Chris Miller, Beiwen Tian, and Xiang Ren. Unsupervised cross-task generalization via retrieval augmentation. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/8a0d3ae989a382ce6e50312bc35bf7e1-Abstract-Conference.html.

Yang Liu, Dan Iter, Yichong Xu, Shuo Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. *ArXiv*, abs/2303.16634, 2023. URL https://api.semanticscholar.org/CorpusID:257804696.

S. Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:256415991.

Ximing Lu, Faeze Brahman, Peter West, Jaehun Jang, Khyathi Raghavi Chandu, Abhilasha Ravichander, Lianhui Qin, Prithviraj Ammanabrolu, Liwei Jiang, Sahana Ramnath, Nouha Dziri, Jillian Fisher, Bill Yuchen Lin, Skyler Hallinan, Xiang Ren, Sean Welleck, and Yejin Choi. Inference-time policy adapters (ipa): Tailoring extreme-scale lms without fine-tuning. *ArXiv*, abs/2305.15065, 2023. URL https://api.semanticscholar.org/CorpusID:258865629.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *ArXiv*, abs/2202.12837, 2022. URL https://api.semanticscholar.org/CorpusID:247155069.

Jesse Mu, Xiang Lisa Li, and Noah D. Goodman. Learning to compress prompts with gist tokens. *ArXiv*, abs/2304.08467, 2023. URL https://api.semanticscholar.org/CorpusID:258179012.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022. URL https://api.semanticscholar.org/CorpusID:246426909.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL http://arxiv.org/abs/1908.10084.

Lingfeng Shen, Sihao Chen, Linfeng Song, Lifeng Jin, Baolin Peng, Haitao Mi, Daniel Khashabi, and Dong Yu. The trickle-down impact of reward (in-)consistency on rlhf, 2023.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *ArXiv*, abs/2004.09297, 2020. URL https://api.semanticscholar.org/CorpusID:215827489.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023. URL https://api.semanticscholar.org/CorpusID:259950998.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Annual Meeting of the Association for Computational Linguistics*, 2022a. URL https://api.semanticscholar.org/CorpusID:254877310.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, M. Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddharth Deepak Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hanna Hajishirzi, and Daniel Khashabi. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Conference on Empirical Methods in Natural Language Processing*, 2022b. URL https://api.semanticscholar.org/CorpusID:253098274.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hanna Hajishirzi. How far can camels go? exploring the state of instruction tuning on open resources. *ArXiv*, abs/2306.04751, 2023. URL https://api.semanticscholar.org/CorpusID:259108263.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Jerry W. Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. Larger language models do in-context learning differently. *ArXiv*, abs/2303.03846, 2023. URL https://api.semanticscholar.org/CorpusID:257378479.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang, and Lei Li. Instructscore: Towards explainable text generation evaluation with automatic feedback. *ArXiv*, abs/2305.14282, 2023. URL https://api.semanticscholar.org/CorpusID:258841553.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. Flask: Fine-grained language model evaluation based on alignment skill sets. *ArXiv*, abs/2307.10928, 2023. URL https://api.semanticscholar.org/CorpusID:259991144.

Yiming Zhang, Shi Feng, and Chenhao Tan. Active example selection for in-context learning. *ArXiv*, abs/2211.04486, 2022. URL https://api.semanticscholar.org/CorpusID:253420743.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haotong Zhang, Joseph Gonzalez, and Ioan Cristian Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv*, abs/2306.05685, 2023. URL https://api.semanticscholar.org/CorpusID:259129398.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, L. Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. Lima: Less is more for alignment. *ArXiv*, abs/2305.11206, 2023. URL https://api.semanticscholar.org/CorpusID:258822910.