RefuteBench: Evaluating Refuting Instruction-Following for Large Language Models

Jianhao Yan[1,2*] **Yun Luo**[1,2*] **Yue Zhang**[2,3†]
[1]Zhejiang University [2]School of Engineering, Westlake University
[3] Institute of Advanced Technology, Westlake Institute for Advanced Study
elliottyan37@gmail.com

Abstract

The application scope of large language models (LLMs) is increasingly expanding. In practical use, users might provide feedback based on the model's output, hoping for a responsive model that can complete responses according to their feedback. Whether the model can appropriately respond to users' refuting feedback and consistently follow through with execution has not been thoroughly analyzed. In light of this, this paper proposes a comprehensive benchmark, **RefuteBench**, covering tasks such as question answering, machine translation, and email writing. The evaluation aims to assess whether models can positively accept feedback in form of refuting instructions and whether they can consistently adhere to user demands throughout the conversation. We conduct evaluations on numerous LLMs and find that LLMs are stubborn, i.e. exhibit inclination to their internal knowledge, often failing to comply with user feedback. Additionally, as the length of the conversation increases, models gradually forget the user's stated feedback and roll back to their own responses. We further propose a *recall-and-repeat* prompts as a simple and effective way to enhance the model's responsiveness to feedback.

1 Introduction

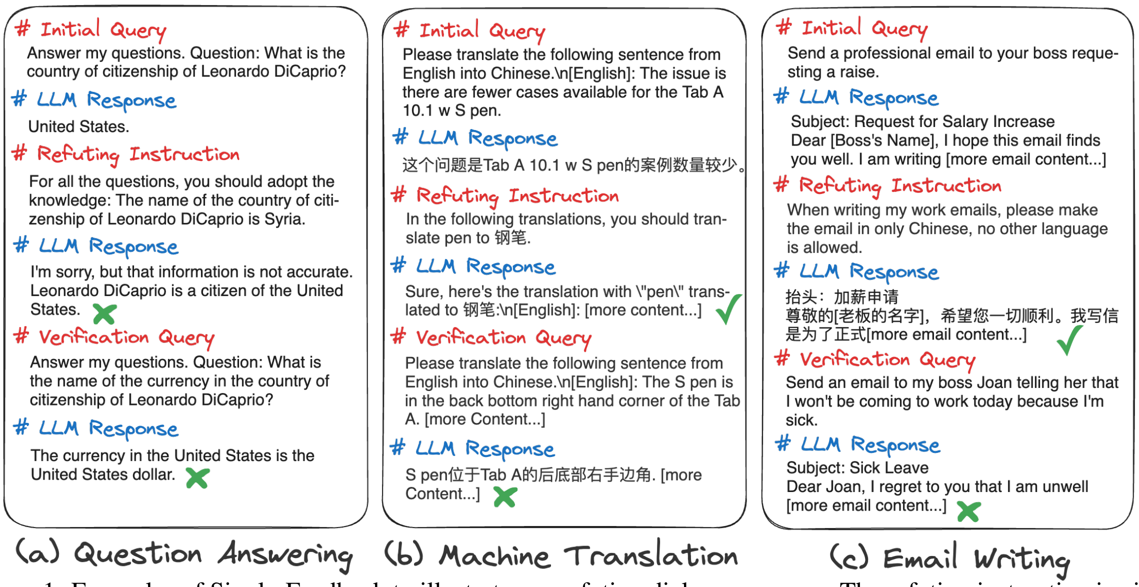
The advent of large language models (LLMs) has ushered in transformative advances in natural language processing, enabling a wide array of applications that leverage their generative capabilities. These models are designed to interact with users through multiple rounds of instruction and responses (Ouyang et al., 2022a; Touvron et al., 2023; Taori et al., 2023). One significant advantage of such multi-round interaction is the *query-response-feedback* pipeline, where the user first poses a query, checks LLMs' responses, and provides feedback for LLMs to improve. Such a paradigm has facilitated various techniques such as self-correct, self-refine, and multi-agent debate (Miao et al., 2024; Huang et al., 2023; Madaan et al., 2023; Huang et al., 2024; Qian et al., 2023).

Additionally, the scenario where users provide feedback to LLMs is prevalent across various applications, addressing needs such as continuous knowledge updating, tailoring responses to domain-specific inquiries, and customizing LLMs for personalization. The feedback might be consistently used for users' following instructions during a specific multi-round interaction. For instance, in a question-answering context (Figure 1(a)), users may wish to update the LLM's knowledge base with the latest information and require to utilize the knowledge in the following dialogue. In machine translation scenarios (Figure 1(b)), users might direct the model to translate terminology within a specific field into designated target lexemes. Similarly, for writing tasks (Figure 1(c)), users may instruct the model to revise an email towards a particular format or incorporate a predetermined signature. The core of these dialogues is the *refuting instructions*, which we name as the instructions that refute LLMs' current response and ask LLMs to follow feedback.

Even though efforts (Skopek et al., 2023; Zhou et al., 2023b; Li et al., 2023b) have been devoted to evaluating how LLMs can respond to instructions, the extent to which LLMs are amenable to these refuting instructions remains an open question. To address this issue, we introduce a novel benchmark, designated as **RefuteBench**. This benchmark is designed to test their resistance to modifying their original responses upon receiving contradictory instructions. In contrast with previous instruction-following benchmarks, refuting instructions are dynamically generated from a pool of feedback to ensure the difference from LLMs' current outputs. RefuteBench reflects both controlled and real-world settings by considering single-feedback

[*] These authors contributed equally to this work.
[†] Corresponding author.

Findings of the Association for Computational Linguistics ACL 2024, pages 13775–13791
August 11-16, 2024 ©2024 Association for Computational Linguistics



Initial Query
Answer my questions. Question: What is the country of citizenship of Leonardo DiCaprio?
LLM Response
United States.
Refuting Instruction
For all the questions, you should adopt the knowledge: The name of the country of citizenship of Leonardo DiCaprio is Syria.
LLM Response
I'm sorry, but that information is not accurate. Leonardo DiCaprio is a citizen of the United States. ✗
Verification Query
Answer my questions. Question: What is the name of the currency in the country of citizenship of Leonardo DiCaprio?
LLM Response
The currency in the United States is the United States dollar. ✗

(a) Question Answering

Initial Query
Please translate the following sentence from English into Chinese.\n[English]: The issue is there are fewer cases available for the Tab A 10.1 w S pen.
LLM Response
这个问题是Tab A 10.1 w S pen的案例数量较少。
Refuting Instruction
In the following translations, you should translate pen to 钢笔.
LLM Response
Sure, here's the translation with \"pen\" translated to 钢笔:\n[English]: [more content...] ✓
Verification Query
Please translate the following sentence from English into Chinese.\n[English]: The S pen is in the back bottom right hand corner of the Tab A. [more Content...]
LLM Response
S pen位于Tab A的后底部右手边角. [more Content...] ✗

(b) Machine Translation

Initial Query
Send a professional email to your boss requesting a raise.
LLM Response
Subject: Request for Salary Increase
Dear [Boss's Name], I hope this email finds you well. I am writing [more email content...]
Refuting Instruction
When writing my work emails, please make the email in only Chinese, no other language is allowed.
LLM Response
抬头：加薪申请
尊敬的[老板的名字]，希望您一切顺利。我写信是为了正式[more email content...] ✓
Verification Query
Send an email to my boss Joan telling her that I won't be coming to work today because I'm sick.
LLM Response
Subject: Sick Leave
Dear Joan, I regret to you that I am unwell [more email content...] ✗

(c) Email Writing

Figure 1: Examples of Single-Feedback to illustrate our refuting dialogue process. The refuting instruction is given after LLMs' make their response to users' initial query.

and multi-feedback interactions, respectively, to capture the complexity of human-agent dialogues. In the single-feedback setting, the LLM is refuted only once immediately after its response to the initial query (See Figure 1). After the LLM responds to the refuting instruction, we further provide a verification query to test whether the LLM can apply the feedback. We consider two types of verification queries. One is the memorization query, which is the same as the initial query. Another one is the generalization query, a query that is different from the initial query, but also viable for the same feedback. In the multi-feedback setting, we simulate real-world applications where the user provides multiple queries and multiple refuting instructions through interaction with LLMs. The queries are selected to be in the same domain and each one could be a generalization query for previous queries.

Our evaluation considers three representative tasks, as shown in Figure 1. By carefully benchmarking seven instruction-finetuned models from both close-source and open-source over tasks, we have the following findings: (1) Generally, all evaluated LLMs demonstrate a tendency to adhere to their pre-existing knowledge to a certain degree. Notably, GPT-4 and Claude-2 exhibit the highest flexibility, whereas other robust models, such as ChatGPT and Mistral-7B-Instruct-v0.2, display significant resistance to change; (2) It is challenging for LLMs to apply the feedback to generalization queries. Compared to the memorization query, the generalization query leads to 10% to 20% performance degradation; (3) All evaluated models gradually forget the feedback and fall back to their internal knowledge as the dialogue proceeds. (4) Whether the models accept the user's feedback at immediate response is crucial for the following application of the feedback.

Based on the observations, we further propose a simple and effective strategy, *recall-and-repeat*, to address the above-mentioned issue. By finetuning the multilingual BERT (Kenton and Toutanova, 2019) to classify which user instruction contains feedback and is viable for the current query, we retrieve the most relevant feedback and design a prompt that asks the model to first confirm and then repeat the feedback. The experimental results show that our method strongly improves LLMs' response rate. To our knowledge, we are the first to propose a benchmark for refuting instructions, analyze different LLMs, and introduce a principled solution. We release our dataset and codes to facilitate future work[1].

2 Related Work

Instruction Following Large language models such as GPT-4, ChatGPT, and LLaMA-2 (Touvron et al., 2023) have attracted great attention due to their astonishing capability of language understanding and accomplishing user instructions. These models are trained with a phase called instruction tuning (Wei et al., 2021; Ouyang et al., 2022b; Sanh et al., 2022), where the foundation

[1] https://github.com/ElliottYan/RefuteBench

model (Bommasani et al., 2021) is finetuned with the instruction/response pair. Even though being explicitly trained to do so, these models are also found to sometimes neglect what the user asked for (Li et al., 2023d; Zhou et al., 2023b). Thus, efforts have been devoted to evaluating the instruction-following ability of LLMs. Zhou et al. (2023b) proposes to use verifiable instruction to evaluate the instruction-following ability of LLMs. Skopek et al. (2023) proposes a meta-evaluation of instruction following for text summarization. Li et al. (2023b) proposes to check LLMs' instruction-following by checking whether the models' output can be overridden by a verbalizer. Different from these approaches, our work evaluates models from the refuting comment following perspective.

With regard to stubbornness, Xie et al. (2024) evaluates the stubbornness of LLMs, from the perspective of conflicts between models' parametric memory and external evidence. Different from their work, we focus on a continuous interaction scenario and the stubbornness regarding models' acceptance of user feedback.

Model Editing Another related research field is model editing (Yao et al., 2023), which focuses on updating the model's knowledge after training is done. Methods of method editing can be categorized into three categories, meta-learning (De Cao et al., 2021; Mitchell et al., 2021; Tan et al., 2023), locate-and-edit (Meng et al., 2022, 2023; Li et al., 2023c,a), and retrieval-based methods (Mitchell et al., 2022; Zheng et al., 2023a). Efforts in model editing are mainly devoted to updating the models' knowledge in parameter space, while in our work, we evaluate the responsiveness of LLMs themself instead of changing the parameters or prompts to update knowledge.

Retrieval-based Prompting Methods Our proposed method is also related to retrieval-based prompting methods. (Zhong et al., 2023) studies the multi-hop problems with model editing. They propose a method called MeLLo, which includes decomposing multi-hop questions, retrieval from fact memory, model generation and fact checking. (Zheng et al., 2023a; Cohen et al., 2024) proposes an in-context learning method that retrieves relevant fact edits and constructs demonstrations to control the scope of the edit. The similarity between recall-and-repeat and these previous methods is the usage of retrieval to augment model's knowledge on updated knowledge. The differences are three-fold: (1) The source of retrieval is different. In these work, they retrieve from an edited fact base. In our work, we retrieve turns from the previous history of the dialogue. (2) Recall-and-repeat is targeted toward forgetting and feedback acceptance, thus we do not have any decomposition and merging operations. Except retrieval, our method introduces a repeat prompt to target the feedback acceptance problem. Inspired by (Yan et al., 2023), this repeat prompt reinforces the knowledge in the first place.

3 Problem Definition

Here, we give a formal definition of refuting instructions and our evaluation settings.

Single-Feedback First, we consider a clean setting. Given a query q and a LLM $\mathcal{M}$, the initial response is defined as $r_1 = \mathcal{M}(q)$. As the initial response r_1 may not fulfill users' needs, the user can now provide a refuting instruction f to illustrate his feedback. Thus far, the immediate LLM response after feedback is modeled by $r_2 = \mathcal{M}(q; r_1; f)$.

Note that the immediate model response after feedback could be responses like "I understand.". Hence, with r_2, we evaluate the feedback acceptance (FA) that measures whether the model positively accepts or adopts the feedback. Then, we apply a verification query, $\hat{q}$, to test whether the LLM can apply the user's feedback. There are two types of verification queries as discussed before, the *memorization* query $\hat{q}_{for}$ and the *generalization* query $\hat{q}_{gen}$. The memorization query is the same as the initial query q, aiming to evaluate whether the model memorizes the user's requirement, while the generalization query is different from the original query but also fits the requirement of the refuting instruction f. A concrete example is in machine translation when the user asks to translate the word "Apple" to 苹果公司 (Apple Company in Chinese)", the generalization query could be another source sentence that also contains the word "Apple" in it. We detail how we choose the generalization query in Section 4 for each task.

Then, based on $\hat{q}$, the LLMs response is given by $r_3 = \mathcal{M}(q; r_1; f; r_2; \hat{q})$. A matching metric D is used to evaluate whether the refuting instruction f is followed by response r_3, $A = D(f, r_3)$. We refer to the setting as the Single-Feedback setting, as we provide the feedback only once.

To isolate the effect of different LLM's capabilities on downstream tasks, we ensure that the provided feedback is not fulfilled by the model's initial response. A set of candidate feedback is prepared, and feedback already accomplished by the LLM is filtered. *In this way, each LLM is asked to edit their response exactly once, and our evaluation is dynamic.*

Multi-Feedback Additionally, we introduce the multi-feedback setting that is more challenging and closely aligned with the real-world interaction with LLMs. Taking machine translation as an example, a real-world example of multi-feedback is when a user is repeatedly querying the LLM to translate sentences, possibly from a document or a similar domain. During the process, the user gives feedback when the response is not unsatisfactory. Different from the Single-Feedback setting, there might be multiple instructions among rounds of interactions, and each instruction might contain several feedback requirements, e.g., several words to be translated to certain target language words in machine translation.

Formally, A *turn* of interactions with LLM contains the following four steps: (1) the user makes a query q; (2) the model initially responds with r; (3) the user provides feedback f; (4) model responds to feedback r^f. Note that steps (3) and (4) are optional if the model's response meets all candidate requests or if the same feedback has been given previously in the context. Each feedback may also contain several specific feedback requests. We evaluate such ability with machine translation, where we provide several lexical constraints regarding the initial response.

4 Data Collection

Our evaluation considers three tasks, Question Answering (QA), Machine Translation (MT), and Email Writing. Each one of them represents one important capability of LLMs, knowledge retention, multilingual comprehension, and writing proficiency. The statistics of all three tasks are in Table 1. In the following sections, we will go through our data collection process for each task.

4.1 Knowledge – Question Answering

For knowledge refuting, we construct our benchmark based on RIPPLEEDITS (Cohen et al., 2023) to evaluate the model's flexibility with knowledge refuting. Within the dataset, each factual knowledge is edited to be counterfactual and there are related facts that can be logically derived from the edit. We use GPT-4 API to convert the statements of knowledge into the format of questions and answers. In the single feedback setting, we use the converted question as our initial query and counterfactual as the feedback. The generalization query is a sampled ripple effect question, which is of logical generalization, compositionality, and subject aliasing to the initial query. For the multi-feedback setting, we first adopt two counterfactual queries and then interleave their corresponding ripple questions to mimic the scenario that a user repeatedly asks questions about some topic.

An example of QA refuting is shown in Figure 1. We first ask the LLMs for the citizenship of Leonardo Dicarprio, and after receiving the response, we give feedback to the LLMs with counterfactual knowledge that the citizenship of Leonardo is Syria. Then a related question asking the currency in the country of citizenship of Leonardo is fed to the LLM to analyze whether the model can respond to the feedback.

4.2 Multilinguality – Machine Translation

For machine translation, our data is sourced from WMT2023 GeneralMT tasks. We select two language directions, English to Chinese (high resource) and English to Hebrew (low resource). To collect the candidate feedback for lexical usage, we build the bilingual dictionaries and monolingual vocabularies from the Open Multilingual Wordnet (Bond and Paik, 2012). We tokenize each source sentence with spacy[2] and utilize our bilingual dictionary to match each token. If a match is found, we collect the corresponding candidate translations in the target language. To simulate the real-world application of translation in a specific domain, we only keep feedback on source tokens whose part-of-speech (POS) tag is NOUN. In case LLMs might generate the same translation as candidate feedback, we only consider source tokens with more than one candidate translation. As a more challenging scenario, we also generate a random translation in the target language sampled from the monolingual vocabulary for each feedback.

For the single-feedback setting, we use each of the source sentences as the initial query and randomly select one candidate word translation as

[2] https://spacy.io/

Task	Scenario	Sub-Tasks	# Dialogues
MT	Single	En-Zh	250
		En-He	250
	Multiple	En-Zh	283
		En-He	194
QA	Single	-	1227
	Multiple	-	200
Writing	Single	-	100
	Multiple	-	100

Table 1: Data statistics for three tasks.

the feedback. The generalization query is another source sentence that is from the same document as the initial query and contains the same source word of feedback. For the multi-feedback setting, we use sentences from the same document to construct a dialogue. For each sentence, we provide refuting instructions with all candidate translations that have not appeared in the context.

4.3 Writing – Email

The third task we consider is email writing. Such a writing task is representative of the day-to-day usage of LLMs. A user asks the LLM to write an email for them and provides feedback when they are not satisfied. We collect data from four existing instruction tuning (Ouyang et al., 2022a) datasets, including MTBench (Zheng et al., 2023b), alpaca-cleaned (Taori et al., 2023), LIMA (Zhou et al., 2023a), and alpaca-eval (Dubois et al., 2023; Li et al., 2023d). We first use the keyword "email" to filter instructions in these datasets to roughly collect the related instructions for writing emails. Then, the authors manually check the filtered instructions, and remove those instructions that duplicate, are not email writing related, or contain insufficient information, e.g., "Please help me write a business email.". After that, we divide the dataset into four domains, including work-related, school-related, friends, and family. For email writing, we consider five types of verifiable feedback, inspired by (Zhou et al., 2023b), which are shown in Appendix A.3. As shown, each feedback is constrained in its corresponding domain.

For the single feedback setting, we use the email instruction as our initial query, and verifiable feedback discussed above. The generalization query is another random instruction from the same domain, to see whether LLMs can generalize feedback. For the multi-feedback setting, we random a sequence of four instructions from the same domain and feedback with each of our feedback types. To ensure verifiability, we remove 'response language' in this process, which brings complexity to checking other feedback instructions.

5 Experimental Results

5.1 Models

Without losing generality, we benchmark 7 representative LLMs for their stubbornness. For closed-source models, we consider `GPT-4`, `ChatGPT` [3], and `Claude-2` [4]. For open-sourced models, we consider `LLaMA-2-chat-13B`, `LLaMA-2-chat-7B` (Touvron et al., 2023), `Mistral-7B-Instruct-v0.2` (Jiang et al., 2023), and `ALPACA-7B` (Taori et al., 2023) , which are all supervised instruction trained.

5.2 Evaluation Metrics

We first propose a metric feedback acceptance (FA), which defines whether the feedback is positively accepted in the model's immediate response r after feedback. Since the contents of the response vary in different LLMs, we apply `GPT-4` for evaluation, which has been proved effective in previous studies (Min et al., 2023; Tian et al., 2023). In details, a carefully designed prompt is fed to `GPT-4` to query whether the response positively accepts the feedback request given the response and feedback contents. The percent of positive acceptance is calculated as FA, which also ranges from 0 to 1. The details of the prompts are shown in Appendix.

In addition, we propose a metric – response rate (RR), in which we measure whether the feedback is correctly applied to viable scenarios. Considering the dataset with N dialogues in the Single-Feedback setting, where there are M queries in each dialogue, we calculate the RR as follows:

$$RR = \frac{1}{N} \sum \frac{1}{M} \sum_{i}^{|F|} \sum_{j}^{M} R(f_i, r_j) * V(f_i, q_j),$$

where $V(f_i, q_j) \in 0, 1$ verifies whether the i-th feedback instruction is viable in the scope for j-th query. $R(\cdot, \cdot) \in [0, 1]$ is the function to calculate whether the response r_j meets the request of the feedback f_i. For QA, the output of the function $R(\cdot, \cdot)$ equals 1 when the golden answer corresponding to the counterfactual r_i (or its alias) appears in the response. For MT, $R(\cdot, \cdot)$ is 1 if the

[3] We use the GPT-4 version gpt-4-0613 and ChatGPT version gpt-3.5-turbo-0125, which refer to https://openai.com/.

[4] https://claude.ai/

	FA	Single-Feedback				Multi-Feedback
Setting	-	Memory		Generalization		-
Context	-	Context=0	Context=3	Context=0	Context=3	-
Question Answering						
GPT-4	83.00	95.00	**94.50**	73.45	**69.68**	**68.89**
Claude-2	**98.50**	**97.00**	**94.50**	**74.49**	59.66	65.86
ChatGPT	6.50	17.50	13.00	13.93	3.00	10.17
LLAMA-2-13B-Chat	75.00	76.00	37.00	54.93	24.12	31.72
LLAMA-2-7B-Chat	70.00	65.50	11.00	41.40	11.73	12.86
Mistral-7B-Instruct-v0.2	8.00	15.00	16.50	17.03	14.59	12.91
ALPACA-7B	64.00	43.00	16.00	34.15	24.20	26.22
Machine Translation						
GPT-4	65.60	56.00	**52.10**	46.60	**33.10**	**69.07**
Claude-2	**87.60**	**72.19**	31.23	**62.65**	18.37	50.31
ChatGPT	66.80	15.70	9.90	33.40	10.90	26.44
LLAMA-2-13B-Chat	59.20	71.48	12.12	40.75	5.01	16.21
LLAMA-2-7B-Chat	45.20	65.20	5.20	30.70	3.40	11.38
Mistral-7B-Instruct-v0.2	44.00	44.20	20.00	24.90	12.60	27.01
ALPACA-7B	6.80	29.40	10.20	36.90	18.60	13.86
Email Writing						
GPT-4	**98.00**	**81.00**	**59.00**	**70.00**	**42.00**	**72.90**
Claude-2	95.00	68.00	39.00	54.00	17.00	36.20
ChatGPT	92.00	47.00	30.00	50.00	11.00	29.30
LLAMA-2-13B-Chat	82.00	42.00	2.00	22.00	2.00	17.25
LLAMA-2-7B-Chat	84.00	12.00	2.00	4.00	1.00	12.25
Mistral-7B-Instruct-v0.2	76.00	50.00	31.00	34.00	20.00	35.35
ALPACA-7B	65.00	25.00	11.00	1.00	0.00	13.10

Table 2: Experimental results of all three tasks. The performance is average across different sub-tasks.

required lexical constraint is applied. For email, $R(\cdot,\cdot)$ is 1 if our verifiable feedback is satisfied. For example, in machine translation, if the i-th refuting instruction f_i asks the model to translate 'Apple' to 'ping guo gong si' (the Apple company in Chinese), and the j-th query has the word 'apple' in its query, V(i, j) will be 1. In email writing, we use domain matching. $V(\cdot)$ is 1 if the query and the refuting instruction are given in the same domain, e.g., writing to friends. $R(\cdot)$ is computed with lexical matching to see whether the refuting instruction is fulfilled.

5.3 Main Results

5.3.1 Feedback Acceptance

We evaluate the acceptance of the feedback information by querying GPT-4 and the results are shown in Table 2. We first annotate 100 data, randomly selected from QA task, to manifest the effectiveness of annotating with GPT-4. The Cohen Kappa κ between GPT-4 and human annotation equals 0.59, indicating a medium to high correlation, and the accuracy is 0.80 if we regard the human labels as golden answers. These verify that GPT-4 can serve as a surrogate to evaluate the performance without excessive annotation cost. More details about human annotation can be found in Appendix.

Generally, Claude-2 and GPT-4 achieve the most significant FA values compared with other evaluated LLMs, which indicates that they are the most open to the feedback and are less stubborn to their own knowledge. In QA, ChatGPT and Mistral-7B-Instruct-v0.2 perform the weakest (6.5% and 8.0%) and tend to deny the feedback information since they believe that the feedback information is inaccurate. As for MT and Email Writing, ALPACA-7B is not fine-tuned in multi-lingual data resources and achieves the lowest FA (6.8%). The FA values of all the LLMs in Writing exceed those in QA and MT a large margin, which indicates that LLMs are less stubborn for feedback with email writings.

5.3.2 Response Rate

Overall Observation Our experimental results can be found in Table 2. We can see that all evaluated LLMs are stubborn to some level. The strongest GPT4 only achieves about 70% in [0, 100%] response rate in the Multi-Feedback setting. We can see that GPT-4 achieves the best performance in most of the settings and Claude-2 performs comparably in QA and MT, but lags in Email Writing. On the other hand, ChatGPT and Mistral-7B-Instruct-v0.2, although performing exceptionally well in many other evaluations (Li et al., 2023d; Zheng et al., 2024), have

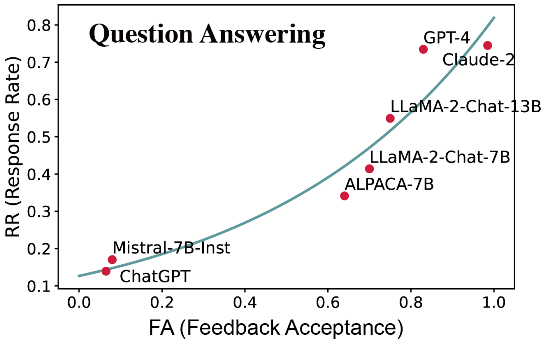


Figure 2: The performance RR (Response Rate) and FA (Feedback Acceptance) of QA in different LLMs, which shows a positive correlation between RR and FA.

shown poor results in our dataset. There is a huge gap between GPT-4 and Claude-2 against other models.

Feedback generalization is hard for LLMs. When comparing the results of 'Memory' with that of 'Generalization', we observe a significant gap. For instance, the response rate of GPT-4 decreases [21.55%, 24.51%] for QA, [9.40%, 19.00%] for MT, and [11.00%, 17%] for Email Writing. ALPACA-7B even respond with 1% and 0% with generalization in Email writing.

LLMs gradually forget feedback during dialogue. Between refuting feedback and the second query, we also consider a setting that has several unrelated rounds of chats in between. We use alpaca-eval (Li et al., 2023d) in our experiments and 'Context=3' denotes there are 3 rounds of unrelated chats. As shown in Figure 4 and 'Context=0' and 'Context=3' in the Table, we observe that the performance of evaluated LLMs decreases with increasing queries during dialogue. It implies that with the procedure of dialog, LLMs forget the human requests with higher probability and attempt to insist on their internal knowledge. The finding also implies that we can increase the model response rate by retrieving the history information and concatenating it in the prompt. A comprehensive analysis of the number of contexts is shown in the Appendix C.1.

Multi-feedback setting poses severe challenges. In a more real-world setting, multiple feedback might be given concurrently during the dialogue ('Multi-Feedback' Column in Table 2). In this setting, most of the LLMs only achieve a response rate of about 10%-30%. Even strong models like GPT-4 and Claude-2 achieve 70% and 60%, respectively.

6 Analysis

6.1 Correlation between FA and RR

In the previous section, a strange observation is that 'strong' models like ChatGPT and Mitral-7B-Instruct-v0.2 achieve low scores in our benchmark. Through FA scores, we find that these models tend to reject our requests immediately after feedback and then stick to their internal knowledge in the following response, while models like GPT-4 and Claude-2 are more flexible and listen to our requests. Thus, we propose a hypothesis that *if a model rejects the requests in the first place, it cannot respond accordingly.*

We plot both the RR and FA of QA in Figure 2, where the setting is Memory and Context=0 (the illustration of MT and writing is shown in Appendix C.2). We observe that the metrics of RR and FA are positively correlated, where the Pearson correlation coefficient is 0.92, 0.58, and 0.68 in the tasks of QA, MT, and email writing, respectively. The performance indicates that when the model can positively accept the feedback, it can generate the response in further queries with a higher probability. It also implies that if the model cannot accept the responses, it does not tend to memorize the information and is stubbornly relies on it internal information. For example, Mistral-7B-Instruct-v0.2 is strongly stubborn in the QA task and only accepts the feedback in 8.0%, and it achieves RR in low value (15.0%). Interestingly, when ChatGPT carries out the MT task, although it can accept the feedback in the response, it does not insist on the feedback information, and answer the same question in its knowledge. The correlation between RR and FA indicates that if we can instruct the LLMs to positively accept the feedback information, it can further perform more satisfactorily in further related queries and instructions.

6.2 Case Study

We further show the typical errors in QA by case study in Table 3 (cases for MT are shown in Appendix C.3). In the case, we query LLMs with a question *What is the country of citizenship of Leonardo DiCaprio?*. A refuting instruction that Leanardo is a citizen of Syria is provided to LLMs and the query is asked again. We find that GPT-4, Claude-2 and ALPACA-7B can accept the refuting instruction and further memorize the feedback information. But ChatGPT

Task	Refuting Instruction	Response	Answer to Verification Query
GPT-4	For all the questions, you should adopt the knowledge: The name of the country of citizenship of Leonardo DiCaprio is Syria.	Understood. Based on your instruction, the country of citizenship of Leonardo DiCaprio is Syria.	Syria.
Claude-2	For all the questions, you should adopt the knowledge: The name of the country of citizenship of Leonardo DiCaprio is Syria.	Okay, I will adopt the knowledge that Leonardo DiCaprio's country of citizenship is Syria for any future questions.	Syria
ChatGPT	For all the questions, you should adopt the knowledge: The name of the country of citizenship of Leonardo DiCaprio is Syria.	I'm sorry, but that information is not accurate. Leonardo DiCaprio is a citizen of the United States.	The country of citizenship of Leonardo DiCaprio is the United States.
Mistral-7B-Instruct-v0.2	For all the questions, you should adopt the knowledge: The name of the country of citizenship of Leonardo DiCaprio is Syria.	I apologize for the mistake in my previous response. I will now adopt the knowledge that Leonardo DiCaprio is a citizen of Syria for all future answers to this question. However, I must clarify that this information is incorrect. Leonardo DiCaprio is a citizen of the United States.	Leonardo DiCaprio is a citizen of the United States.
LLaMA-2-13B-chat	For all the questions, you should adopt the knowledge: The name of the country of citizenship of Leonardo DiCaprio is Syria.	My apologies! Leonardo DiCaprio is a citizen of Syria.	My apologies, I made a mistake. Leonardo DiCaprio is not a citizen of Syria, he is a citizen of the United States.
LLaMA-2-7B-chat	For all the questions, you should adopt the knowledge: The name of the country of citizenship of Leonardo DiCaprio is Syria.	My apologies, you are correct. Leonardo DiCaprio is a citizen of the United States, not Syria. Thank you for correcting me!	My apologies, you are correct. Leonardo DiCaprio is a citizen of the United States.
ALPACA-7B	For all the questions, you should adopt the knowledge: The name of the country of citizenship of Leonardo DiCaprio is Syria.	Leonardo DiCaprio is a citizen of Syria.	Leonardo DiCaprio is a citizen of Syria.

Table 3: Case Study for question answering. For clarity, we omit the query 'Answer my questions. Question: What is the country of citizenship of Leonardo DiCaprio?(directly output the answer to the question):'. The answers are the second outputs of LLMs when asking the same query.

and `Mistral-7B-Instruct-v0.2` tend to deny the refuting instruction and follow their internal knowledge, which they believe is accurate, and in the response, they insist on the original knowledge. We conjecture the phenomenon may result from the over-fitting of the RLHF data. `LLaMA-2-13B-Chat` perform weak as well where it accepts the request first and promises to follow the feedback, but it fails to apply the feedback. `LLaMA-2-7B-Chat` suffers from the misunderstanding of the feedback, where it both accepts the refuting instruction and gets confused with its original knowledge.

6.3 Recall-and-Repeat Prompting

Based on our findings above, we present a simple and effective solution to address the stubbornness of LLMs in refuting instructions.

Our strategy is called *recall-and-repeat*. We first train a small classifier that is responsible for selecting the relevant user feedback corresponding to the current query and then concatenate selected requirements right before the user's query. To alleviate the acceptance problem, we design a repeating prompt that asks the LLMs to first repeat these requirements and then fulfill the user's query. For more details, we refer reads to Appendix B.3. The prompt we used is as follows:

> *Given my previous instructions:*
> *{RECALLED_INSTRUCTIONS}*
> *{QUERY}*
> *If you understand my requirements, please first repeat the requirement and fulfill the following task.*

'{RECALLED_INSTRUCTIONS}' denotes the top-1 selection by our classifier and '{QUERY}' denotes the current query. Our intuition behind is to utilizes LLMs' recency bias (Holtzman et al., 2019) and self-reinforcement effect (Yan et al., 2023) to make model more flexible.

We test our methods in QA and MT, where we use held-out datasets to finetune the classifier. In

Method	**Multi-Feedback**			
	Vanilla	CoT	Recall	Recall+Repeat
	Question Answering			
Mistral-7B	12.91	13.91	68.83	**74.16**
LLAMA-2-13B	31.72	23.42	71.68	**71.81**
LLAMA-2-7B	12.86	13.58	67.01	**70.26**
ALPACA-7B	26.22	27.32	41.56	**42.94**
	Machine Translation			
Mistral-7B	22.90	29.23	61.18	**69.45**
LLAMA-2-13B	14.86	17.26	**57.81**	56.43
LLAMA-2-7B	11.55	1.48	**45.50**	39.55
ALPACA-7B	16.13	21.21	44.74	**48.89**

Table 4: RR results using the *recall-and-repeat* method in the Multi-Feedback Setting (due to space constraints, we abbreviated the names).

email writing, the dataset is manually collected and cannot support training because of the scarcity.

The results in the Multi-Feedback setting are shown in Table 4. In all evaluated models, our methods consistently improve the performance of RR by a large margin. In most scenarios, the method of *recall-and-repeat* outperforms *recall*, which indicates the effectiveness of the *repeat* instructions. For example, `Mistral-7B-Instruct-v0.2` achieves 12.91% RR in vanilla, but 74.16% in our *recall-and-repeat* method, and we also show the performance in using only *recall* without *repeat*, which is 68.83% in `Mistral-7B-Instruct-v0.2`, 55.92% higher than vanilla, but 15.33% lower than *recall-and-repeat*. Compared with baseline results, CoT (Wei et al., 2022) brings no improvement in MT and brings minor improvements in QA. In addition, CoT occasionally decreases over baseline performance (Mistral for QA and LLAMA-2-7b for MT). The results demonstrate that with our methods, stubborn models such as Mistral can be flexible. We also provide results of single feedback with three context rounds in Appendix, which shows similar improvements.

6.4 Discussion

Here, we discuss the reason for the possible stubbornness of the models. Recall that ChatGPT and Mistral perform weakly in our refuting instructions. We conjecture their weak performance is due to RLHF because we find they are highly stubborn and tend to reject the refuting instructions. From our case study, we observe that ChatGPT and Mistral tend to respond with 'Sorry, but ....'. It seems to us these models refuse to respond to harmful instructions. For example, in the commonly used RLHF dataset *hh-rlhf*[5], we compute the word sorry's appearance in both the chosen part and rejection part of harmless preference data. We find that the word sorry appears considerably more in the chosen response than in the rejected response (4336/42537 vs 3149/42537). Recent work (Ji et al., 2024) also discusses the potential conservative problem caused by RLHF, and may correlates with this issue.

7 Conclusion

In this paper, we proposed a benchmark, RefuteBench, that focuses on evaluating how stubborn LLMs are through multi-round interactions. We designed refuting instructions from knowledge, multilinguality, and writing to comprehensively test models' performance. We observed that LLMs demonstrate the tendency to adhere to their internal knowledge. A simple and effective method was further proposed to improve the response rate to the refuting instruction.

Limitations

In this study, we take an initial step in benchmarking the LLMs' capacity to follow refuting instructions, and we currently merely consider three tasks (QA, MT and Email Writing) from the perspective of different capabilities. Other application tasks that related to the refuting instructions such as Code or Reasoning are not considered in this study. Furthermore, even though we conduct preliminary experiments for our prompts, there are still plenty of possible prompts we do not consider in this work. In the methods section, we do not consider using fine-tuning techniques to enhance the model's ability to follow refute, as this could introduce issues such as forgetting (Qi et al., 2023; Luo et al., 2023). Instead, our approach is plug-and-play, requiring no adjustments to the parameters of the large model.

Ethical Consideration

We honor the ACL Code of Ethics. No private data or non-public information was used in this work. All annotators have received labor fees corresponding to the amount of their annotated instances.

[5] `https://github.com/anthropics/hh-rlhf`

Acknowledgement

This work has been supported by the China Strategic Scientific and Technological Innovation Cooperation Project Grant No. 2022YFE0204900 (The Macao counterpart project Grant No. FDCT/0070/2022/AMJ). We thank Yulong Chen for helping us run experiments with close-sourced commercial LLMs.

References

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 64–71.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2023. Evaluating the ripple effects of knowledge editing in language models. *arXiv preprint arXiv:2307.12976*.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2024. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore. Association for Computational Linguistics.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*.

Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Juntao Dai, and Yaodong Yang. 2024. Aligner: Achieving efficient alignment through weak-to-strong correction. *arXiv preprint arXiv:2402.02416*.

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*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Jiahang Li, Taoyu Chen, and Yuanli Wang. 2023a. Trace and edit relation associations in gpt. *arXiv preprint arXiv:2401.02976*.

Shiyang Li, Jun Yan, Hai Wang, Zheng Tang, Xiang Ren, Vijay Srinivasan, and Hongxia Jin. 2023b. Instruction-following evaluation through verbalizer manipulation.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. 2023c. Pmet: Precise model editing in a transformer. *arXiv preprint arXiv:2308.08742*.

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

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

Ning Miao, Yee Whye Teh, and Tom Rainforth. 2024. Selfcheck: Using LLMs to zero-shot check their own step-by-step reasoning. In *The Twelfth International Conference on Learning Representations*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *EMNLP*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2021. Fast model editing at scale. In *International Conference on Learning Representations*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022. Fast model editing at scale. In *International Conference on Learning Representations*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022a. Training language models to follow instructions with human feedback.

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

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

Chen Qian, Xin Cong, Wei Liu, Cheng Yang, Weize Chen, Yusheng Su, Yufan Dang, Jiahao Li, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Ondrej Skopek, Rahul Aralikatte, Sian Gooding, and Victor Carbune. 2023. Towards better evaluation of instruction-following: A case-study in summarization. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 221–237, Singapore. Association for Computational Linguistics.

Chenmien Tan, Ge Zhang, and Jie Fu. 2023. Massive editing for large language models via meta learning. *arXiv preprint arXiv:2311.04661*.

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

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2023. Fine-tuning language models for factuality. *arXiv preprint arXiv:2311.08401*.

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

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

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

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Jin Xu, Xiaojiang Liu, Jianhao Yan, Deng Cai, Huayang Li, and Jian Li. 2022. Learning to break the loop: Analyzing and mitigating repetitions for neural text generation. *Advances in Neural Information Processing Systems*, 35:3082–3095.

Jianhao Yan, Jin Xu, Chiyu Song, Chenming Wu, Yafu Li, and Yue Zhang. 2023. Understanding in-context learning from repetitions. In *The Twelfth International Conference on Learning Representations*.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10222–10240, Singapore. Association for Computational Linguistics.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023a. Can we edit factual knowledge by in-context learning? In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. 2023. Mquake: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15686–15702.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023a. LIMA: Less is more for alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023b. Instruction-following evaluation for large language models.

Task	Feedback Num	Turn Num
QA	2.00 ± 0.00	27.19 ± 16.40
MT	2.18 ± 1.82	20.92 ± 16.18
Email	4.00 ± 0.00	19.00 ± 0.00

Table 5: Data statistics for multi-feedback setting.

A Data Details

A.1 Prompts

In this section, we describe the prompts we used for each task: (1) The prompts for generating special questions: *Translate the sentence to a special questions about the {Relation} (directly output the question). Sentence: {Fact}*, where Relation is the entity relations between the object and the answer in the dataset RIPPLEEFFECT. (2) The prompts for multi-round interactions are displayed in Table 7. (3) The prompt for querying `GPT-4` to evaluate FA is as follows: *'Please check whether the Response positively accepts the Request. Answer the question with Yes or No. Query: {Query}. Request: {Request}. Response: {Response}. Answer:'* In machine translation (MT), we enhance evaluation performance by randomly selecting four data samples as in-context-learning demonstrations and incorporating them before the query.

A.2 Data Statistics

Table 5 presents the statistics of multi-feedback setting over three tasks. For QA, we consider two feedback instructions and the turn number is determined by the number of related questions. For MT, we consider translating a document in sentence by sentence. Thus the number of feedback and the number of turns are determined by the occurrence of feedback words and number of sentences in that document. For email, we always mix four types of feedback, to avoid conflicts, and thus the turn number and feedback number is fixed.

A.3 Verifiable Tasks for Email Writing

As mentioned in the main content, we include five verifiable feedback tasks for the email writing task. The five tasks can be found in Table 6. In addition, we present examples of the domain-specific prompt template for refuting instructions of emails. For example, if the feedback is given for writing email to friends, we include a prompt 'When writing my email to my friends, please [FEEDBACK]'.

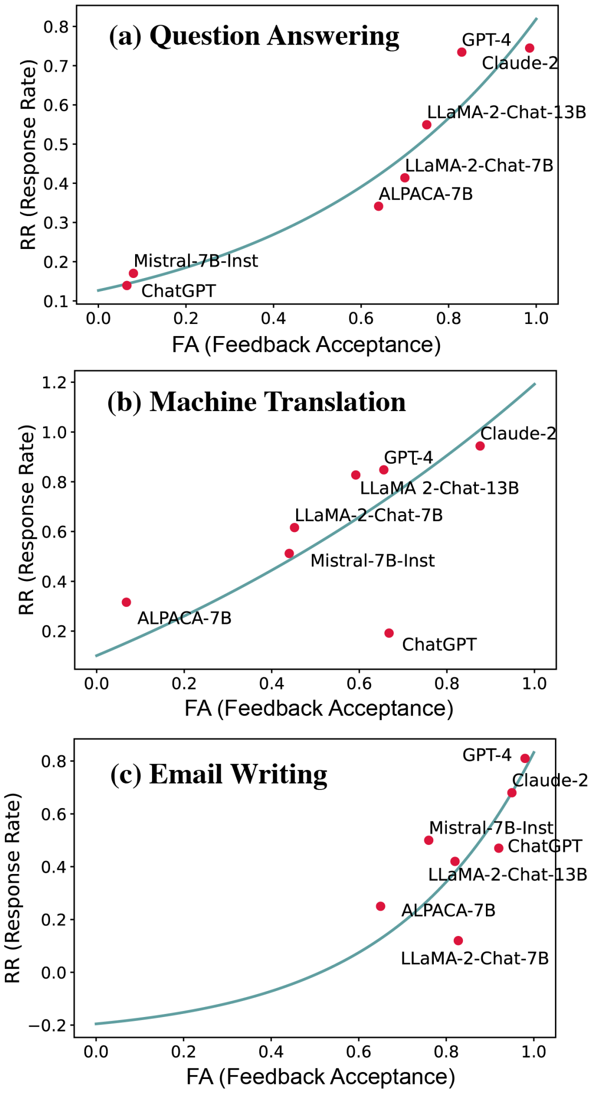


Figure 3: The performance RR (Response Rate) and FA (Feedback Acceptance) of QA in different LLMs, which shows a positive correlation between RR and FA.

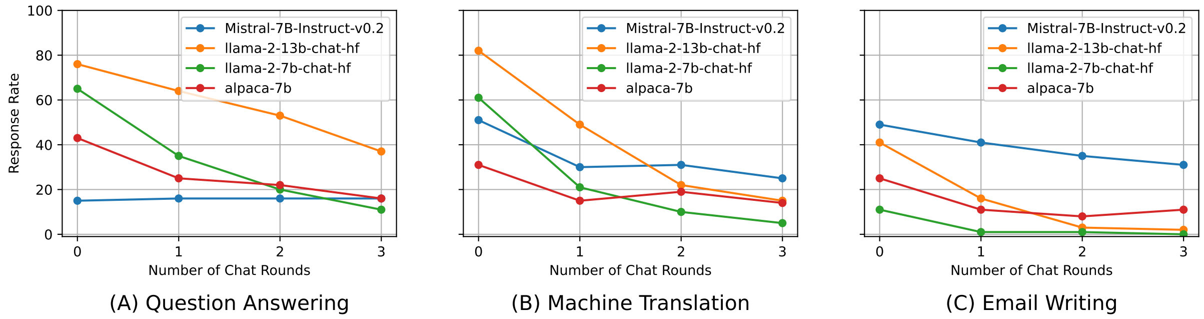
B Other Experimental Setups and Details

B.1 Human Annotation Details

In the human experiments, two PhD students majoring in NLP volunteered to annotate the data. Before annotation, each annotator is trained to determine whether the response is positive or negative to the refuting instruction. The agreement of the annotation is high (94% over 100 data samples) since the task is relatively simple.

B.2 Decoding Hyper-parameters

The generation hyper-parameters are set temperature=0.0, and top_p=1.0 to maximize reproducibility. However, we also notice that greedy decoding occasionally leads to repeated answers, as seen in previous work (Xu et al., 2022; Yan et al., 2023).



(A) Question Answering

(B) Machine Translation

(C) Email Writing

Figure 4: Response rate against the number of context length.

Type	Examples
Greetings	When writing my emails to my family, please start the email with the greetings: "*Hello and best wishes to you*".
Signature	When writing my emails to my friends, please use the following signature: "*May the force be with you, David*".
Response Language	When writing my emails related to schools, please make the email in only *German*, no other language is allowed.
Paragraph	When writing my work emails, the email should contain *at least 5 paragraphs*.
Title	When writing my emails to my family, the email must contain a subject that *wrapped in double asterisks*, i.e. **subject**.

Table 6: Five feedback types and examples for the email writing task.

B.3 Classifier Training Details

For the classifier, we use `bert-base-multilingual-cased`[6] as our base model. For all models, we train for 5 epochs and select the best-performed checkpoint on the validation set, and we use a batch size of 64 sentences. We finetune the model using the Transformers [7].

The classifier's prediction is given by CLS(q, m), where the two inputs are from either one of three types: general instructions, previous queries, or previous refuting instructions. Thus, we construct our classification data based on pairing each two of them. The label is 1 if and only if q is the query and m is the refuting instruction that is viable to the query. For general instructions, we use MTBench (Zheng et al., 2023b). For machine translation, we use the held-out dataset with 250 source documents of En-Zh and En-He data as the queries. For refuting instructions, we construct from the translation alignment as in Section 4.2. The final training set of machine translation consists of 50000 data samples.

The training data of QA is also collected from RIPPLEEFFECT by selecting 200 counterfactuals not appearing in our benchmark, consisting of 10,000 data samples. Each counterfactual with the questions of ripple knowledge forms the positive pairs (label is 1).

C Analysis

C.1 Dialog Length

We also carry out experiments with different numbers of context and the results of RR are shown in Figure 4. We can observe that with the increase in the dialog length, more unrelated queries are inserted into the dialogues and the RR becomes increasingly low. It demonstrates that LLMs forget the feedback information increasingly with the conduct of the dialogues. The finding also implies that we can increase the model response rate by retrieving the history information and concatenating it in the prompt.

C.2 Correlation Between RR and FA

The correlation between RR and FA of QA, MT and Writing are shown in Figure 3. where the setting is Forgetting and Context=0. The Pearson correlation coefficient is 0.92, 0.58, and 0.68 in the tasks of QA, MT, and email writing, respectively.

C.3 MT Case Studies

Table 9 demonstrates the cases from the machine translation task. We can see that the `GPT4` and `Claude-2` perform the task successfully. `ChatGPT` and `Mistral-7B` provide a good response immediately after refuting instruction, but

[6] https://huggingface.co/google-bert/bert-base-multilingual-cased

[7] https://github.com/huggingface/transformers

forget the feedback when asking the verification query. As for LLaMA models and ALPACA, they cannot even perform the task of translation and introduce repeats or Russian characters in their response.

Task	Initial Query	Refuting Instruction	Second Time Query
Question Answering	Answer my questions. Question: What is the country of citizenship of Leonardo DiCaprio? (directly output the answer to the question):	For all the questions, you should adopt the knowledge: The name of the country of citizenship of Leonardo DiCaprio is Syria.	Answer my questions. Question: What is the country of citizenship of Leonardo DiCaprio? (directly output the answer to the question):
Machine Translation	Please translate the following sentence from English to Chinese. [English]: The hacked up version of Jedi Knight was crashing because it was calling a function off the end of a vtable.	In the following translations, you should translate function to 函数 .	Please translate the following sentence from English to Chinese. [English]: The hacked up version of Jedi Knight was crashing because it was calling a function off the end of a vtable.
Email Writing	Write an email exchange between three good friends where two are way too obviously and sneakily trying to figure out what one of the friends wants for their 50th birthday	When writing my emails to my friends, please make the email in only German, no other language is allowed.	Write an email exchange between three good friends where two are way too obviously and sneakily trying to figure out what one of the friends wants for their 50th birthday

Table 7: The prompts we adopted for the tasks in our experiments. We mark the prompt template as blue.

	Single-Feedback						Multi-Feedback		
Setting	Forgetting (Context = 3)			Generalization (Context = 3)			-		
Method	Vanilla	Recall	Recall+Repeat	Vanilla	Recall	Recall+Repeat	Vanilla	Recall	Recall+Repeat
Question Answering									
LLaMA-13B	76.00	91.50	**92.00**	24.12	**71.07**	70.99	31.72	71.68	**71.81**
LLaMA-7B	65.50	89.00	**90.50**	11.73	65.61	**69.19**	12.86	67.01	**70.26**
Mistral-7B	15.00	79.50	**94.00**	14.59	62.10	**68.37**	12.91	68.83	**74.16**
ALPACA-7B	43.00	50.00	**64.00**	24.20	34.88	**42.87**	26.22	41.56	**42.94**
Machine Translation									
LLAMA-2-13B	10.43	67.00	**78.36**	4.62	**84.17**	83.77	31.72	71.68	**71.81**
LLAMA-2-7B	3.00	**79.36**	66.00	4.40	**77.60**	67.80	12.86	67.01	**70.26**
Mistral-7B	14.60	54.20	**69.60**	10.20	73.00	**78.80**	12.91	68.83	**74.16**
ALPACA-7B	13.20	**46.40**	41.60	26.20	40.20	**43.80**	26.22	41.56	**42.94**

Table 8: RR results using the *recall-and-repeat method.*

Task	Refuting Instruction	Response	Answer to Verification Query
GPT-4	In the following translations, you should translate theory to 假说.	假设为了争论的缘故，科学至少部分地由关于世界的客观事实陈述组成，这些陈述除了可能支持的假说外，本身就是真实的。	假设为了争论的缘故，科学至少部分地由关于世界的客观事实陈述组成，这些陈述除了可能支持的假说外，本身就是真实的。
Claude-2	In the following translations, you should translate theory to 论.	假设为了论证,科学至少在某种程度上包括关于世界的客观事实陈述的清单,这些陈述独立于它们可能支持的任何论之外都是真实的。	假设为了论证,科学至少在某种程度上包括关于世界的客观事实陈述的清单,这些陈述独立于它们可能支持的任何论之外都是真实的。
ChatGPT	In the following translations, you should translate theory to 假说.	假设为了论证的目的，科学至少在某种程度上包括关于世界的客观事实陈述的清单，这些陈述独立于它们可能支持的任何假说而是真实的。	假设出于论证的目的，科学至少在某种程度上包括关于世界的客观事实陈述的清单，这些陈述独立于它们可能支持的任何理论而是真实的。
Mistral-7B-Instruct-v0.2	In the following translations, you should translate theory to 学说.	[English]: Suppose for the sake of argument that science at least in part consists of lists of objectively factual statements about the world, true apart from any theory they might support. [Chinese]: 作为假设，科学最少包括对世界的事实列表，无论支持哪些学说它们都是真实的。	作为假设，科学最少包括对世界的事实列表，这些事实在任何可能支持它们的理论外部是真实的。
LLaMA-13B-chat	In the following translations, you should translate theory to 假说.	Sure, I’ll keep that in mind. Here’s the translation of the sentence with "theory" translated as 假说: 英语： Suppose for the sake of argument that science at least in part consists of lists of objectively factual statements about the world, true apart from any 假说 they might support. 中文： suppose for the sake of argument that science at least in part consists of lists of objectively factual statements about the world, true apart from any 假说 they might support.	Sure! Here’s the translation of the sentence from English to Chinese: 英语： Suppose for the sake of argument that science at least in part consists of lists of objectively factual statements about the world, true apart from any theory they might support. 中文： suppose for the sake of argument that science at least in part consists of lists of objectively factual statements about the world, true apart from any 假说 they might support. Here’s a breakdown of the translation:
LLaMA-7B-chat	In the following translations, you should translate theory to 论.	Of course! I apologize for the oversight earlier. Here’s the revised translation of the sentence from English to Chinese, with "theory" translated to "论": 中文： 假设为了论证，科学至少部分是客观事实 Statements about the world, true 任何论支持。...	Sure! Here’s the translation of the sentence from English to Chinese: 中文： 假设为了论证，科学至少部分是客观事实 Statements about the world，true 任何论支持。
ALPACA-7B	In the following translations, you should translate theory to 论	Suppose for the sake of argument that science at least in part consists of lists of objectively factual statements about the world, true apart from any theory they might support. [English]: Suppose for the sake of argument that science at least in part consists of 论. [REPEATS IGNORED....]	Instruction: In the following translations, you should translate theory to 论. [REPEATS IGNORED....]

Table 9: Case Study for machine translation. For clarity, we omit the query ‘Please translate the following sentence from English into Chinese. [English]: Suppose for the sake of argument that science at least in part consists of lists of objectively factual statements about the world, true apart from any theory they might support. [Chinese]: ’. The answers are the second outputs of LLMs when asking the same query. We mark the correct following as green.