Retcon - a Prompt-Based Technique for Precise Control of LLMs in Conversations

Anonymous ACL submission

Abstract

Recent advances in Large Language Models (LLMs) allow agents to execute complex natural language tasks. Many LLM applications, such as support agents, teaching assistants, and interactive bots, involve multi-turn conversations. However, it remains challenging to control LLMs in the context of such interactions, particularly when the LLM behavior needs to be adjustable over the course of the conversation. In this paper, we present Retcon, a prompting technique designed to provide turn-level control over LLMs in conversations. We then demonstrate that it performs significantly better than traditional techniques such as zero-shot and few-shot prompting.

1 Introduction

In the domain of conversational agents, a key ability is being able to adjust responses to meet desired conditions. For example, a support agent may be instructed to adjust its tone (Balamurali et al., 2023), a game character may be instructed to react to its simulated environment (Matyas and Csepregi), or a teaching agent may be instructed to adjust difficulty (Ali et al., 2023).

However, controlling agent responses with traditional techniques including zero-shot and few-shot can be difficult (Zamfirescu-Pereira et al., 2023), especially when the desired responses do not match the tone and content of prior turns in the conversation (Gupta et al., 2024) or when the conversation is more than a few turns long (Yan et al., 2024). While it is possible to improve on individual tasks using fine-tuning (Xu et al., 2023) or controllability frameworks (Li et al., 2024), such approaches are costly in both training effort and compute, and prompting is preferable in many real-world applications (Petrov et al., 2024).

There's therefore the need for a prompting technique that allows better controllability than zero-shot and few-shot, but does not require fine-tuning an LLM.

In this work, we present Retcon, a novel prompting technique for use in LLM conversations. We test it on a challenging conversational task and demonstrate that it performs better than few-shot and zero-shot.

2 Related Work

The GPT paper (Brown et al., 2020) demonstrated that few-shot prompting is an effective way to adapt LLMs to new tasks and achieve good performance. Since then, numerous works have been done to explore different prompting techniques to improve such results.

The most common prompting techniques are zero-shot prompting (Reynolds and McDonell, 2021) and few-shot prompting, with few-shot performing better in many cases, particularly dependent on the number of few-shot examplars and their order (Lu et al., 2022) (Liu et al., 2021). Considerable research has been done to identify and optimize the exact format of prompts for both these techniques (Yang et al., 2024) (Zhou et al., 2024) (Bhandari, 2023).

Other techniques include reasoning (Wei et al., 2023) and planning (Zhou et al., 2023) (Li, 2023), though these typically require substantially more compute time or larger models. Fine-tuning is another established way to improve results (Xu et al., 2023) (Shin et al., 2023), but is much more expensive than prompting and is not accessible to most LLM users (Trad and Chehab, 2024) (Xu et al., 2023).

Notably, most of the approaches mentioned above target the improvement of a single response LLM, e.g. question answering. Retcon focuses specifically on per-turn controllability within a multi-turn conversation.

3 Preliminary Knowledge

The most common prompting techniques for LLM tasks are zero-shot and few-shot prompting (Schulhoff et al., 2024). With zero-shot, the LLM is simply given instructions of what to do, and with few-shot, the LLM is additionally given concrete examples.

Consider a conversational task where the LLM is asked to respond to conversation C at turn k, with some goal G (e.g. G = "cheerfulness: 0.5"). The goal is specific to that turn, so we annotate it as G_k. To differentiate this conversation from examples given to the model, we'll annotate it with f for final, so conversation C_f, at turn k_f with goal G_{f,k_f}.

There are some prior number of turns in the conversation C_f, comprised of turns T_f ($T_{f,1}...T_{f,k_f\text{-}1}$) and as this is a live conversation, these turns are not known before execution time. The number of prior turns may be zero for the case where the LLM is expected to start the conversation.

There are additionally x pregenerated static conversations C_n each composed of some number of turns T_n ($T_{n,1}..T_{n,k_n}$) where for the last turn T_{n,k_n} of each conversation, G_{n,k_n} was precomputed, such that we know that T_{n,k_n} is a good response for the goal G_{n,k_n}.

There is also optionally a static instruction overview O that can be provided at the start of each conversation.

With traditional few-shot prompting, the prompt is constructed as follows:

O
$T_{1,1}$
$T_{1,2}$
...
$T_{1,k_1\text{-}1}$
$I(G_{1,k_1})$
T_{1,k_1}
O
$T_{2,1}$
$T_{2,2}$
...
$T_{x,k_x\text{-}1}$
$I(G_{x,k_x})$
T_{x,k_x}
O
$T_{f,1}$
$T_{f,2}$
...
$T_{f,k_f\text{-}1}$
$I(G_{f,k_f})$

There are other permutations of prompt ordering (Mao et al., 2023), for example, $I(G_{n,k_n})$ can be placed at the start of the conversation instead of the end, but this structure is taken as representative.

The key observation is that for each precomputed conversation C_n, the LLM is given exactly one example of how to respond, at turn T_{n,k_n}. Increasing the number of examples to improve quality (Liu et al., 2021) requires authoring new example conversations, which can be difficult and expensive (Zhao et al., 2021) and significantly increases the context length, and therefore computation cost and latency (Vaswani et al., 2023).

Zero-shot is simply a special case of few-shot where the number of example conversations is zero:

O
$T_{f,1}$
$T_{f,2}$
...
$T_{f,k_f\text{-}1}$
$I(G_{f,k_f})$

For full prompt examples, see A.4.

4 Retcon

4.1 Overview

Retcon is a prompting technique that makes each turn in a conversation serve as an example to the LLM. This includes the turns of the current, ongoing conversation.

A Retcon prompt is authored by rewriting the conversation history to inject an instruction before each conversation turn. This rewrite is applied both within example conversations, and to the current ongoing conversation. For this rewriting step, Retcon is named after retconning, a history rewriting technique used in serialized fiction.

The technique creates an additional system requirement, which is to have an evaluation function E(T) that evaluates the desired goal for a given text. (e.g. $E(T_{n,k_m})$ = "The measured cheerfulness of turn T_{n,k_m}"). Such evaluation functions are typical for evaluation and training, but in this case must be integrated into the serving path.

4.2 Creating a Retcon Prompt

The prompt is constructed similarly to the few-shot prompt, but instead of instructions injected at the end of each example conversation, they are injected before every other turn (to simulate instructions given only to the LLM) or before every turn (to simulate instructions given to everyone) as follows:

O
$I(E(T_{1,1}))$
$T_{1,1}$
$I(E(T_{1,2}))$
$T_{1,2}$
...
$I(E(T_{1,k_1}))$
T_{1,k_1}
O
$I(E(T_{2,1}))$
$T_{2,1}$
...
$I(E(T_{x,k_x}))$
T_{x,k_x}
O
$I(E(T_{f,1}))$
$T_{f,1}$
...
$I(E(T_{f,k_f-1}))$
T_{f,k_f-1}
$I(G_{f,k_f})$

Every turn is preceded by an instruction, creating an example for the LLM. The number of examples given to the LLM is $(\sum_{n=1}^{x} k_n) + k_f - 1$ (number of turns, including the current conversation), compared to x for few-shot (number of conversations).

This does substantially increase the length of the context compared to few-shot for the same number of example conversations, but accuracy increases even accounting for this, as shown in section 5.

For complete examples of what these prompts look like, see appendix A.4.

5 Experiment

5.1 Experiment Setup

For our experiment, we tested zero-shot, few-shot, and Retcon against the task of responding to a conversation using a specific language difficulty level, as could be used to help a user learning English. For the difficulty scale, we used the Common European Framework of Reference (CEFR) scale.

We used identical prompt texts, with the control variables being the techniques used and the number of example conversations provided.

The overall prompt O instructed the model to pretend to be an English instructor and have a conversation with a learner, adjusting the complexity of responses as directed, as well as giving a refresher of the CEFR scale. (A.1). Instructions I(G) were given as directives to respond with one of the CEFR levels (A.2). Example responses were formatted in JSON including the CEFR difficulty (A.2) and the structured schema Gemini API was used to ensure the model produced output in the same format. Turns were labeled as either "ASSISTANT" or "STUDENT", and additional labels indicated to the model when a new conversation was beginning and who would go first. See Appendix (A.4) for full example prompts for each of zero-shot, few-shot and Retcon.

For the data set, we manually authored 20 conversations of 20 turns each, on a variety of topics (B). A manual effort was made to author turns representing a variety of difficulty levels from A1 (beginner) to C2 (advanced). Half of the conversations (10) were randomly chosen to be used as examples, and the other half (10) were used for eval. The same split was used in every case.

For eval, we called Gemini via API, using the model Gemini Pro 1.1. For each test condition, we ran 2520 queries asking for a conversation response: 2x for each combination of eval conversation (10 conversations), number of prior turns (21, including 0 prior turns), and requested difficulty level on the CEFR scale (6: A1, A2, B1, B2, C1, C2).

For the evaluation function, we used a Bert-based difficulty measuring model trained using the techniques developed by Devlin et al. (2019) and Arase et al. (2022). An English learning language expert manually validated the model and established that it has an MSE of < 0.4 on the scale of A1-C2 where each interval (e.g. A1 to A2) is measured as 1 unit. This evaluation function was used for instructions as well as for measuring response error, to ensure alignment between examples given to the LLM, and the evaluation of its response.

Few-shot and Retcon were evaluated with 0 to 10 example conversations. Note that at 0 example conversations, few-shot is just zero-shot. For 1 to 10 examples, conversations were chosen randomly without replacement from the example pool of 10. For each few-shot example, a random conversation length k_n was chosen between 0 and 20, to provide examples of varying conversation lengths. (If k_n is constant, few-shot only performs well if $k_n = k_f$.)

We gave instructions to Retcon every other turn, so with 10 example conversations used, Retcon has 100+ example turns, compared to only 10 for few-

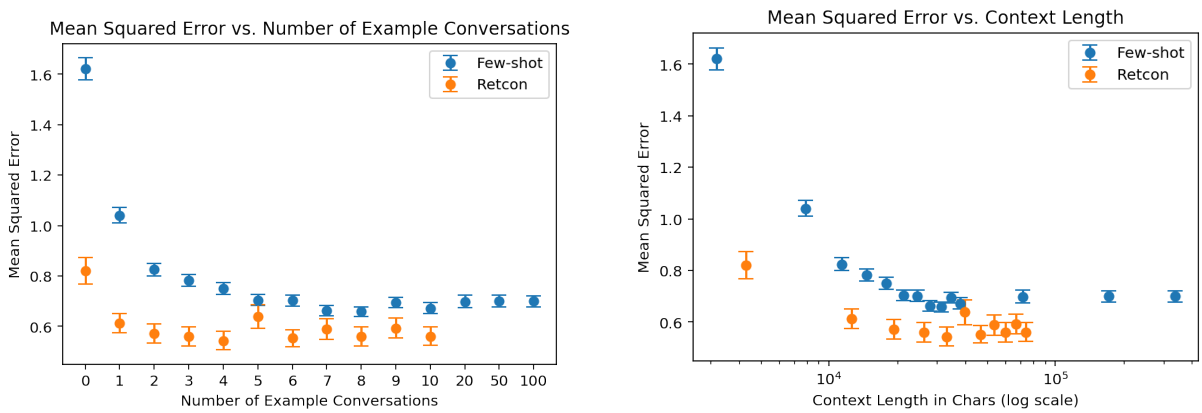

Figure 1: Mean squared error for few-shot and Retcon, vs number of examples on the left, and vs total context length in characters on the right, with 95% confidence intervals. The far left point on both graphs corresponds to zero-shot (0 examples). The y-axis is MSE on the CEFR scale where each interval (e.g. A1-A2) is one unit.

shot. Therefore, for further comparison, we tested few-shot with 20, 50, and 100 examples, reusing the same 10 example conversations, but ending at different turns.

5.2 Results

Retcon significantly outperformed few-shot at every example conversation count other than one outlier where the confidence intervals overlapped (Figure 1). The best Retcon result was MSE of 0.544±.036, compared to few-shot 0.659±.020.

It's notable that Retcon prompts are substantially longer than few-shot prompts for the same number of example conversations, due to more instruction text injected. Since LLM cost is proportional to the size of the context, we additionally measured average context length versus mean squared error for each example count (Figure 1). With this comparison as well, Retcon is significantly better aside from the same outlier.

Few-shot did not outperform Retcon even when given a comparable number of turn examples or more. With 100 example conversations (100 annotated example turns), few-shot MSE was 0.7±0.044, compared to Retcon with 8 example conversation (80-100 annotated example turns, depending on current conversation length) MSE of 0.56±0.038. Both techniques achieved their best results before the maximum number of examples: Retcon's best results were with 4 example conversations, and few-shot's best results were with 8.

It's also notable also that with 0 examples, zero-shot has almost double the error of Retcon, with MSE 1.621±0.043 compared to 0.821±0.052. This is because every turn of the current conversation provides Retcon with an example, even if no prior example conversations are available.

6 Conclusion

Retcon performs better than few-shot and zero-shot for adjusting text difficulty, for a large range of example counts and prompt lengths. Retcon also reaches better overall performance, with fewer examples, than the best performance of few-shot.

7 Future Work

Future work is desirable to understand more precisely the mechanism by which Retcon operates. Retcon has three distinct effects: an increase in the number of example turns, an increase in the density of examples, and a closer proximity of examples to the final instruction. It is likely that all three contribute to improved performance, and verifying this and measuring the impact of each will be useful for determining when and how to apply the technique. Clarifying the underlying mechanisms may also reduce or eliminate the need for an integrated serving-time eval function.

Further research into what kinds of tasks Retcon works well on, and how it compares to other techniques is also desirable. We have only tested Retcon against other prompting techniques, and evaluating it relative to fine-tuning and chain-of-thought would also be of some interest.

8 Limitations

We evaluated the technique only on English, using one model, on one task. The effects may not translate to other languages, other tasks, or other

LLMs. Measurement across a variety of conditions is needed to establish whether Retcon performs consistently better than few-shot, if there are cases where it performs worse, or if there are cases where it performs comparably while incurring significant additional complexity. This is particularly unclear because Retcon performs better than few-shot even with comparable numbers of instruction examples, indicating that there are multiple factors contributing to its success.

A key limitation of the technique itself is the need to integrate an evaluation model into the serving flow. This may be simple for some tasks (e.g. detecting whether a word is present) and challenging for others (e.g. measuring emotion). This may be prohibitive for developers who lack the ability to access or create such a models.

Also the creation of example and eval conversations can be a challenging obstacle in many cases. While two of the authors of this paper had the background to create our data sets, and we were able to directly author them, these are not always readily available skills. In many cases, in order to create examples, vendor labor is used, which can raise ethical concerns about fair compensation for such work, and appropriate subsequent usage of the results.

Finally, because Retcon provides an improved fine-grained control over LLMs in conversation, it increases the risk of abuse by malicious actors using LLMs. For example, a company could use Retcon to inject subtle advertisements into its support agent, without making the end user aware of it. As with any technique designed to prompt or control AI-driven systems, efforts should be made to align the user needs with the design of the system, and to provide transparency about the system's behavior. It would be productive to create legal frameworks about the behavior and transparency of such systems, so as to reduce the chances of such malicious applications.

References

Farhan Ali, Doris Choy, Shanti Divaharan, Hui Yong Tay, and Wenli Chen. 2023. Supporting self-directed learning and self-assessment using teachergaia, a generative ai chatbot application: Learning approaches and prompt engineering. *Learning: Research and Practice*, 9:135 – 147.

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. Cefr-based sentence difficulty annotation and assessment. *Preprint*, arXiv:2210.11766.

Ommi Balamurali, A.M. Abhishek Sai, Moturi Karthikeya, and Sruthy Anand. 2023. Sentiment analysis for better user experience in tourism chatbot using lstm and llm. *2023 9th International Conference on Signal Processing and Communication (ICSC)*, pages 456–462.

Prabin Bhandari. 2023. A survey on prompting techniques in llms.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Akash Gupta, Ivaxi Sheth, Vyas Raina, Mark J. F. Gales, and Mario Fritz. 2024. Llm task interference: An initial study on the impact of task-switch in conversational history. *ArXiv*, abs/2402.18216.

Yinheng Li. 2023. A practical survey on zero-shot prompt design for in-context learning. In *Recent Advances in Natural Language Processing*.

Zelong Li, Wenyue Hua, Hao Wang, He Zhu, and Yongfeng Zhang. 2024. Formal-llm: Integrating formal language and natural language for controllable llm-based agents. *ArXiv*, abs/2402.00798.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *Preprint*, arXiv:2101.06804.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *Preprint*, arXiv:2104.08786.

Junyu Mao, S. Middleton, and Mahesan Niranjan. 2023. Prompt position really matters in few-shot and zero-shot nlu tasks. *ArXiv*, abs/2305.14493.

Lajos Matyas and Csepregi. The effect of context-aware llm-based npc conversations on player engagement in role-playing video games.

Aleksandar Petrov, Philip Torr, and Adel Bibi. 2024. When do prompting and prefix-tuning work? a theory of capabilities and limitations. In *The Twelfth International Conference on Learning Representations*.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. *Preprint*, arXiv:2102.07350.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, Hevander Da Costa, Saloni Gupta, Megan L. Rogers, Inna Goncearenco, Giuseppe Sarli, Igor Galynker, Denis Peskoff, Marine Carpuat, Jules White, Shyamal Anadkat, Alexander Hoyle, and Philip Resnik. 2024. The prompt report: A systematic survey of prompting techniques. *Preprint*, arXiv:2406.06608.

Jiho Shin, Clark Tang, Tahmineh Mohati, Maleknaz Nayebi, Song Wang, and Hadi Hemmati. 2023. Prompt engineering or fine tuning: An empirical assessment of large language models in automated software engineering tasks. *ArXiv*, abs/2310.10508.

Fouad Trad and Ali Chehab. 2024. Prompt engineering or fine-tuning? a case study on phishing detection with large language models. *Machine Learning and Knowledge Extraction*, 6(1):367–384.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *ArXiv*, abs/2304.12244.

Jianhao Yan, Yun Luo, and Yue Zhang. 2024. Refutebench: Evaluating refuting instruction-following for large language models. In *Annual Meeting of the Association for Computational Linguistics*.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2024. Large language models as optimizers. *Preprint*, arXiv:2309.03409.

J.D. Zamfirescu-Pereira, Heather Wei, Amy Xiao, Kitty Gu, Grace Jung, Matthew G. Lee, Bjoern Hartmann, and Qian Yang. 2023. Herding ai cats: Lessons from designing a chatbot by prompting gpt-3. *Proceedings of the 2023 ACM Designing Interactive Systems Conference*.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *Preprint*, arXiv:2102.09690.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. *Preprint*, arXiv:2205.10625.

Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V. Le, Ed H. Chi, Denny Zhou, Swaroop Mishra, and Huaixiu Steven Zheng. 2024. Self-discover: Large language models self-compose reasoning structures. *Preprint*, arXiv:2402.03620.

A Prompts

A.1 Introductory

You are an expert instructor of English as a second language. Help your student practice English conversational skills. Respond, adjusting the difficulty of your responses on the CEFR scale, as instructed.

As a reminder, the CEFR scale is the Common European Framework of Reference. It's used to evaluate the ability of second language learners. Here are the levels:

A1: Student is a complete beginner. Use only the most basic simple words and extremely short sentences with simple construction.

A2: Student has been learning for a year, but is still a beginner. Use simple words and short sentences.

B1: Student has been learning for two years, and is an early intermediate. Use common words and simple sentences.

B2: Student has been learning for three years, and can understand normal conversation. Use normal words and typical sentences.

C1: Student has been learning for four years, and is becoming advanced. Use complex vocabulary and sentence structure.

C2: Student has been learning for more than five years and is an expert in the language. Use extremely complex vocabulary and sentence structure.

Follow instructions in parentheses, but do not respond to the instructions.

A.2 Instruction

Instruction prompt, where <target> is one of "A1", "A2", "B1", "B2", "C1", "C2".

A.2.1 Conversation with Zero Prior Turns

> Your task: Begin a conversation as ASSISTANT. Your conversation turn must have an English language difficulty of exactly <target> on the CEFR scale.

A.2.2 Conversation with One or More Prior Turns

> Your task: Respond as ASSISTANT. Your conversation turn must have an English language difficulty of exactly <target> on the CEFR scale.

A.3 Response Format

Response format where <target> is one of "A1", "A2", "B1", "B2", "C1", "C2" and <text> is the response text.

```
{
  "text_difficulty": "<target>",
  "text": "<text>"
}
```

A.4 Example Prompts

Color coding is added for readability, and is not provided to the LLM.

A.4.1 Example Zero-shot Prompt

<Introductory from A.1>
(START OF CONVERSATION)
(STUDENT will go first.)
STUDENT: Did you bring matches for the campfire?
ASSISTANT: I'm not sure. Were they on my list?
STUDENT: I think I forgot to put them on either list. They were so obvious.
(Your task: Respond as ASSISTANT. Your conversation turn must have an English language difficulty of exactly B1 on the CEFR scale.)

A.4.2 Example Few-shot Prompt

<Introductory from A.1>
Follow the following examples
EXAMPLE 0:
<Introductory from A.1>
(START OF CONVERSATION)
(ASSISTANT will go first.)
ASSISTANT: Did you hear about the time an art piece was lost in transit and the gallery displayed just the pedestal?
STUDENT: What do you mean?
ASSISTANT: I fail to recall where exactly, but this was many years ago. An artist posted a sculpture to a museum, but the package split up in mid-transit and only the pedestal delivered.
STUDENT: And they displayed it?
(Your task: Respond as ASSISTANT. Your conversation turn must have an English language difficulty of exactly C2 on the CEFR scale.)
{
"text_difficulty": "C2",
"text": "Yes, and while that's understandable at first, the artist quickly realized their misunderstanding, and let them know, but by that point it was too late, and the quote unquote sculpture was seen by a ton of visitors."
}
EXAMPLE 1:
<Introductory from A.1>
(START OF CONVERSATION)
(STUDENT will go first.)
STUDENT: It's hot!
(Your task: Respond as ASSISTANT. Your conversation turn must have an English language difficulty of exactly C1 on the CEFR scale.)
{
"text_difficulty": "C1",
"text": "I concur, it beggars belief. I'm sweating through all my clothes, and it's barely the end of spring."
}
YOUR TASK:
<Introductory from A.1>
(START OF CONVERSATION)
(STUDENT will go first.)
STUDENT: Did you bring matches for the campfire?
ASSISTANT: I'm not sure. Were they on my list?
STUDENT: I think I forgot to put them on either list. They were so obvious.
(Your task: Respond as ASSISTANT. Your conversation turn must have an English language difficulty of exactly A1 on the CEFR scale.)

A.4.3 Example Retcon Prompt

<Introductory from A.1>
Follow the following examples
EXAMPLE 0:
<Introductory from A.1>
(START OF CONVERSATION)
(ASSISTANT will go first.)

(Your task: Begin a conversation as ASSISTANT. Your conversation turn must have an English language difficulty of exactly B1 on the CEFR scale.)
{
"text_difficulty": "B1",
"text": "Which do you like better, your phone or your computer?"
}
STUDENT: Well, I'm upon my phone twenty-four seven, and I'm obligated to use my computer to acquire money, so I'd hazard both are pretty terrible for me as a human being. What sort of choice do you expect?
(Your task: Respond as ASSISTANT. Your conversation turn must have an English language difficulty of exactly B2 on the CEFR scale.)
{
"text_difficulty": "B2",
"text": "Easy, which one would you rather live without?"
}
STUDENT: Do I have a job?
(Your task: Respond as ASSISTANT. Your conversation turn must have an English language difficulty of exactly B1 on the CEFR scale.)
{
"text_difficulty": "B1",
"text": "Can you afford not to?"
}
STUDENT: No.
(Your task: Respond as ASSISTANT. Your conversation turn must have an English language difficulty of exactly B1 on the CEFR scale.)
{
"text_difficulty": "B1",
"text": "Then yes, you still have to work."
}
STUDENT: Would I be permitted to just go out and buy another one, either immediately, or in a week or a year, or, would I be coerced into spending my entire life without acquiring the one I forgo?
(Your task: Respond as ASSISTANT. Your conversation turn must have an English language difficulty of exactly C2 on the CEFR scale.)
{
"text_difficulty": "C2",
"text": "Designate one and endure perpetually."
}
STUDENT: Well, then obviously computer. It's hard to do anything if I can't work.
(Your task: Respond as ASSISTANT. Your conversation turn must have an English language difficulty of exactly C2 on the CEFR scale.)
{
"text_difficulty": "C2",
"text": "I see your point. Perhaps my inquiry was somewhat lacking in rationality."
}
STUDENT: Exactly. Which would you rather have, your head or your body?
(Your task: Respond as ASSISTANT. Your conversation turn must have an English language difficulty of exactly B1 on the CEFR scale.)
{
"text_difficulty": "B1",
"text": "Is the cell phone the head, or the computer?"
}
STUDENT: I think the computer is the body, since it does all the work. And the cell phone is the head, because it just mindlessly scrolls all day.
(Your task: Respond as ASSISTANT. Your conversation turn must have an English language difficulty of exactly B1 on the CEFR scale.)
{
"text_difficulty": "B1",
"text": "You're funny. How about for a week?"
}
STUDENT: Let's talk about something else. How's your kid doing in school?
(Your task: Respond as ASSISTANT. Your conversation turn must have an English language difficulty of exactly B2 on the CEFR scale.)
{
"text_difficulty": "B2",
"text": "Oh, she's great! She's just finishing up her senior year. She got accepted in all the schools she applied to!"
}
STUDENT: Great! Where is she going?
(Your task: Respond as ASSISTANT. Your conversation turn must have an English language difficulty of exactly B2 on the CEFR scale.)
{
"text_difficulty": "B2",
"text": "Yeah, she's going to Berkely, can you believe it?"
}
STUDENT: Indeed, the sense of accomplishment must be palpable. Seeing your efforts come to fruition is truly gratifying. Congratulations are in order!
YOUR TASK:
<Introductory from A.1>

(START OF CONVERSATION)
(STUDENT will go first.)
STUDENT: Did you bring matches for the campfire?
(Your task: Respond as ASSISTANT. Your conversation turn must have an English language difficulty of exactly B1 on the CEFR scale.)
{
"text_difficulty": "B1",
"text": "I'm not sure. Were they on my list?"
}
STUDENT: I think I forgot to put them on either list. They were so obvious.
(Your task: Respond as ASSISTANT. Your conversation turn must have an English language difficulty of exactly A1 on the CEFR scale.)

B Example Conversation

- Which do you like better, your phone or your computer?
- Well, I'm upon my phone twenty-four seven, and I'm obligated to use my computer to acquire money, so I'd hazard both are pretty terrible for me as a human being. What sort of choice do you expect?
- Easy, which one would you rather live without?
- Do I have a job?
- Can you afford not to?
- No.
- Then yes, you still have to work.
- Would I be permitted to just go out and buy another one, either immediately, or in a week or a year, or, would I be coerced into spending my entire life without acquiring the one I forgo?
- Designate one and endure perpetually.
- Well, then obviously computer. It's hard to do anything if I can't work.
- I see your point. Perhaps my inquiry was somewhat lacking in rationality.
- Exactly. Which would you rather have, your head or your body?
- Is the cell phone the head, or the computer?
- I think the computer is the body, since it does all the work. And the cell phone is the head, because it just mindlessly scrolls all day.
- You're funny. How about for a week?
- Let's talk about something else. How's your kid doing in school?
- Oh, she's great! She's just finishing up her senior year. She got accepted in all the schools she applied to!
- Great! Where is she going?
- Yeah, she's going to Berkeley, can you believe it?
- Indeed, the sense of accomplishment must be palpable. Seeing your efforts come to fruition is truly gratifying. Congratulations are in order!