Improving the Readability of Scientific Concept Analogies with Cognitive Conflict Reinforcement Learning

Yuang Cai Yuyu Yuan Jinsheng Shi Rui Han Zhenyu Zhao Zijie Shi
School of Computer Science, Beijing University of Posts and Telecommunications
{cyang,yuanyuyu,jinsheng,hanr,zhaozhenyu,shizijie}@bupt.edu.cn

Abstract

Large language models are increasingly being used for education and science communication by automatically generating explanations of scientific concepts. However, prior research has found that the analogies produced by LLMs lack human-like psycholinguistic properties important for readability. In this work, we propose cognitive conflict reinforcement learning (CCRL) to improve the psycholinguistic properties of analogies generated by LLMs. Specifically, we create cognitive conflict between the original LLM and a cloned LLM during reinforcement learning. This helps address the cognitive rigidity problem in LLMs. Experimental results demonstrate that our approach significantly outperforms existing RL algorithms and human performance in improving various readability metrics of generated analogies.

Keywords: large language models; readability; cognitive rigidity; cognitive conflict; reinforcement learning

Introduction

In the era of large language models (LLMs), science teachers and popular science editors may ask LLMs for explanations of scientific concepts. The quality of these explanations directly influences the effectiveness of teaching and publication. A common approach to explaining scientific concepts is through the use of analogies. However, recent research (Seals & Shalin, 2023) has revealed that LLMs lack human-like psycholinguistic properties when generating long-form analogies. The study demonstrates that evaluating the superficially fluent and trustworthy texts generated by LLMs requires techniques capable of capturing subtle and underlying features of proficient language use. Therefore, the study adopts Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004), an automated tool that calculates the computational linguistic and psycholinguistic properties of written and spoken texts, to assess the quality of scientific concept analogies generated by LLM. By comparing the analogies produced by LLMs with those created by humans, the study reveals that LLMs significantly underperform humans in readability-related metrics.

Text readability encompasses several evaluation metrics within Coh-Metrix, such as narrativity, temporality, and word concreteness (Graesser et al., 2004). Additionally, the Flesch-Kincaid Reading Ease (Kincaid, Fishburne Jr, Rogers, & Chissom, 1975) is a direct indicator of readability, as it takes into account the average sentence length and the average number of syllables per word. These metrics, being non-differentiable w.r.t. the input text, cannot be optimized through backpropagation (Werbos, 1988), but can be optimized using reinforcement learning (RL) (Sutton & Barto, 2018). However, the implementation of Coh-Metrix is complicated and not open-sourced, which means it is difficult to directly compute the online property scores related to readability in the RL training loop. As a result, we need to construct an appropriate reward modeling dataset concerning readability and train a reward model to assess the readability of a given text.

The utilization of RL in LLMs introduces the challenge of mode collapse, where the model continuously generates certain patterns with high confidence (Janus, 2023). Mitigating mode collapse in the RL fine-tuning of LLMs is a complex task since the corruption of the model's output distribution brought by RL fine-tuning is a non-trivial transformation (Janus, 2023). Turner and Einhorn (2023) propose an approach called action-conditioned TD error (ACTDE), which facilitates convergence towards softer optima. B. Zhu et al. (2023) addresses mode collapse by proposing advantage-induced policy alignment (APA) that introduces a squared error loss based on estimated advantages. These approaches are primarily rooted in machine learning and mathematics. However, considering that the mode collapse in LLMs shares similarities with cognitive rigidity in human beings (Schultz & Searleman, 2002), it is essential to adopt a cognitive perspective to analyze and address this issue effectively.

In this work, we create analogies for eight different scientific concepts from four different domains and compute the property scores related to readability for these analogies. We compute and analyze the correlations between these properties to build the reward modeling dataset and train a reward model using the dataset. Then, we propose cognitive conflict reinforcement learning (CCRL) to address the cognitive rigidity problem in the LLM and train the LLM to maximize the reward emitted by the reward model. We evaluate the Coh-Metrix scores achieved by analogies generated by our approach and compare them with human-written analogies. Our approach surpasses humans in producing readable analogies for the chosen scientific concepts.

This work makes the following main contributions:

- To the best of our knowledge, we are the first to utilize reinforcement learning to enhance the psycholinguistic properties of large language models

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

- We develop a psycholinguistic reward modeling dataset to evaluate the readability of scientific concept analogies and present the idea of how to build the dataset,
- We analyze the mode collapse issue in LLM from the cognitive perspective and propose a novel RL algorithm, cognitive conflict RL, to address the issue.

Preliminaries

Text Readability

Text readability refers to the level of ease with which a written piece can be understood. In their extensive analysis, Chall (1958) examined early research on readability, categorizing it into two types: "survey and experimental studies" and "quantitative associational studies." The former type focused on exploring the effects of various factors on readers, examining how modifying texts based on a single variable at a time would influence readers' experience. These studies also involved gathering expert and reader opinions through surveys to gain further insights. The primary aim was to comprehend the elements that impact readability, particularly when reader engagement with the text's content was controlled.

Quantitative associational studies mainly focused on conducting quantitative evaluations of text readability through mathematical formulas. These formulas generally reflect the ordering of the specific text relative to some other texts. Sherman (1893) is the earliest to propose a quantitative analysis of text difficulty, which measures the readability by the number of clauses per sentence. One of the most familiar formulas is the Flesch-Kincaid Grade Level (or Flesch-Kincaid Reading Ease) (Kincaid et al., 1975), which measures the average sentence length and the average number of syllables per word. In this work, we employ the Flesch-Kincaid Reading Ease as the direct metric of readability evaluation.

More recently, Graesser et al. (2004) developed Coh-Metrix that analyzes texts on over 200 properties of cohesion, language, and readability using lexicons, part-of-speech classifiers, syntactic parser, templates, corpora, latent semantic analysis, and other components. Graesser, McNamara, and Kulikowich (2011) studied textbooks of different grade levels and text categories to identify the underlying components of language, discourse, and cognition of traditional automated metrics of text difficulty and the Coh-Metrix. The conclusion is that word concreteness, syntactic simplicity, referential cohesion, causal cohesion, and narrativity are five major factors, i.e., properties, that account for most of the variance in texts across grade levels and text categories. In this work, we consider these properties in reward modeling and readability evaluation.

Cognitive Rigidity

Cognitive rigidity, a long-standing construct in psychology, is characterized by a tendency to form and persist in mental and behavioral sets (Schultz & Searleman, 2002). Wason (2021) shows that confirmation bias, the tendency to favor information that confirms existing beliefs, can lead to cognitive rigidity. Amadieu, Tricot, and Mariné (2009); Amadieu, Van Gog, Paas, Tricot, and Mariné (2009); Shing and Brod (2016) show that prior knowledge can both enhance and hinder learning and memory processes, leading to cognitive rigidity in some cases.

Cognitive rigidity can have significant impacts on human language in several ways. It can hinder scientific thinking by imposing rigid definitions, which limits the ability to think in scientific terms (Zilversmit, 1964). This rigidity is also evident in the perceptual system, which is shaped by early linguistic experience and remains relatively rigid (Mehler, Pallier, & Christophe, 1998). More recently, Alves and Pozzebon (2013) found that cognitive rigidity can lead to resistance to linguistic diversity. Thierry (2016) found that it can influence language perception and processing, as evidenced by the link between linguistic distinctions and perceptual or conceptual processing. Haig and Woodcock (2017) found that individuals with Prader-Willi syndrome who were exposed to increased rigidity in routines during development showed higher resistance to change.

In this work, we show that cognitive rigidity also exists in large language models. We analyze the reason for cognitive rigidity in LLMs and propose a novel reinforcement learning algorithm to alleviate this problem.

Cognitive Conflict

Cognitive conflict, as described by Rappoport (1969), is a result of differences in thinking that can lead to interpersonal conflicts. This is particularly relevant in the context of conceptual change, as highlighted by Larsson, Haglund, and Halldén (2010), where it involves the restructuring of existing beliefs. Lee et al. (2003) further explores this by identifying four key constructs of cognitive conflict: recognition of an anomalous situation, interest, anxiety, and cognitive reappraisal of the conflict situation. Woods (2012) extends this understanding to strategic decision-making, proposing the use of dialectical inquiry to create cognitive conflict and improve organizational performance.

De Dreu and Nijstad (2008) suggests that cognitive conflict can enhance critical thinking and creativity, contrary to the traditional belief that it leads to rigidity. This is because cognitive conflict can induce confusion, which in turn can lead to enhanced learning (Lehman et al., 2013). Gauer and Kuzmics (2020) suggest that cognitive conflict can facilitate the acquisition of information about an opponent's preferences, potentially fostering empathy and alleviating cognitive rigidity.

Inspired by the idea that cognitive conflict alleviates cognitive rigidity, we create conflict in the RL training process to help the LLM avoid cognitive rigidity.

Reinforcement Learning in Text Generation

In reinforcement learning (RL), an agent interacts with the environment and learns to achieve the maximum accumulated rewards (Sutton & Barto, 2018). The interaction can be formulated as a Markov decision process (MDP). An MDP can be denoted as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R)$, where $\mathcal{S}$ is the state space,

$\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0,1]$ is the transition probability, and $R : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is the reward function. The agent receives a reward $R(s,a)$ when it takes action a under state s and the environment has the probability $P(s,a,s')$ to transfer to state s'.

A policy of an agent involves which action should be taken under specific states. A policy can be denoted as $\pi : \mathcal{S} \times \mathcal{A} \mapsto [0,1]$, and $\pi(a|s)$ means the probability of taking action a under state s. The value function $V_\pi(s)$ is the expected sum of rewards achieved by the agent when starting from state s and following policy π. The state-action value function $Q_\pi(s,a)$, also known as the Q-function or Q-value, is the expected sum of rewards achieved by the agent when taking action a under state s and then following policy π. The relationship between the value function and the Q function can be expressed as $V_\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q_\pi(s,a)\right]$. The objective of policy-based RL is to maximize the expected sum of rewards achieved by the agent, as shown in Equation 1, where π_θ is the policy parameterized by θ and η is the distribution of the initial state.

$$J(\theta) = \mathbb{E}_{s\sim\eta(\cdot)}\left[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[Q_{\pi_\theta}(s,a)\right]\right] \tag{1}$$

In text generation tasks, the generative language model can be viewed as the agent policy, and the sequence currently being generated can be viewed as the environment. The policy can be denoted as $p_\theta(y_t|\mathbf{x},\mathbf{y}_{<t})$, where $\mathbf{x}$ is the prompt, $\mathbf{y}_{<t} = (y_1, y_2, \cdots, y_{t-1})$ is the currently generated text and y_t is the next token to be generated. The reward is only given at the end of the sequence since the computation of the reward generally depends on sentence-level metrics. For simplicity, the reward can be denoted as $R(\mathbf{x},\mathbf{y})$, where $\mathbf{y}$ is the complete generated sequence. The reward function generally measures the quality of the generated sequence and is generally proportional to the quality. The objective of policy-based RL in text generation is to maximize the expected reward of the generated sequences, as shown in Equation 2, where $\mathcal{D}$ is the prompt dataset and $\pi_\theta(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{|\mathbf{y}|} p_\theta(y_t|\mathbf{x})$.

$$J(\theta) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}\left[R(\mathbf{x},\mathbf{y})\right]\right] \tag{2}$$

Psycholinguistic Reward Modeling

Analogy Dataset Creation

We consider eight scientific concepts from four different domains: *enzyme kinetics* and *glycolysis* in biochemistry, *TCP three-way handshake* and *Asymmetric encryption* in computer science, *inflation* and *market economy* in economics, as well as *stock trading* and *options trading* in finance. We use the pre-trained Llama-2 model (Touvron et al., 2023) [1] as the analogy generator. We build the basic prompt "*Explain how <concept> works by creating an analogy*" to instruct the model to create an analogy like Seals and Shalin (2023).

To enrich the diversity of the data distribution, we build the enhancement prompt to enhance the psycholinguistic properties of the generated analogies. We consider nine psycholinguistic properties, as listed in Table 1. The enhancement prompt is designed according to the definitions in Coh-Metrix. For example, the definition of syntactic simplicity requires using simpler, familiar syntactic structures that are less challenging to process. Therefore, we design the enhancement prompt for syntactic simplicity as "*Use simple syntax instead of complex syntax in your analogy*".

Table 1: Coh-Metrix indices and function denotations of the considered psycholinguistic properties.

Property	**Coh-Metrix**	**Abbr.**	**Function**
Narrativity	PCNARz	NAR	f_{nar}
Temporality	PCTEMPz	TMP	f_{tmp}
Word concreteness	PCCNCz	CNC	f_{cnc}
Deep cohesion	PCDCz	DC	f_{dc}
Referential cohesion	PCREFz	REF	f_{rc}
Explicit connectives	PCCONNz	CON	f_{ecn}
Syntactic simplicity	PCSYNz	SYN	f_{syn}
Causal connectives	CNCCaus	CAU	f_{ccn}
Flesch reading ease	RDFRE	FRE	f_{fre}

For each basic prompt, we concatenate it with the 9 enhancement prompts and finally get $8 \times 9 = 72$ prompts. We then feed the 8 basic prompts and the 72 enhanced prompts to Llama-2 and generate 300 analogies for each prompt, acquiring 24000 analogies. We compute and normalize (Z-score normalization) the scores of the above 9 properties for each analogy to form the final analogy dataset $\mathcal{D}^A$. The normalized scores are denoted as functions, as shown in the third column of Table 1. We denote each item in the analogy dataset as an 11-tuple containing the prompt $\mathbf{x}$, the analogy $\mathbf{y}$, and the 9 normalized scores.

Readability Function

The Flesch reading ease directly reflects the readability of a text. However, we should also consider other properties to evaluate the readability more comprehensively. In other words, we should find a linear combination of the above properties to represent the readability. To determine how much each property contributes to readability, we compute the Pearson correlation coefficients between the Flesch reading ease and all other properties, as shown in Table 2.

We take each correlation coefficient as the weight of a property and the readability function can be denoted as Equation 3, which we refer to as the soft readability function. Here, $\mathcal{F}$ is a collection of the property functions in Table 1.

$$r(\mathbf{y}) = \sum_{f\in\mathcal{F}} \rho_{f_{\text{fre}},f} \cdot f(\mathbf{y}) \tag{3}$$

Another approach is to simply take the properties whose correlation coefficient is greater than a threshold ρ, as denoted in Equation 4, where ρ is the correlation coefficient threshold. We refer to Equation 4 as the hard readability function.

$$r(\mathbf{y}) = \sum_{f\in\mathcal{F}} \mathbb{I}(\rho_{f_{\text{fre}},f} > \rho) \cdot f(\mathbf{y}) \tag{4}$$

[1] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

Table 2: Correlations between FRE and other properties.

f	f_{nar}	f_{tmp}	f_{cnc}	f_{dc}	f_{rc}	f_{ecn}	f_{syn}	f_{ccn}	f_{fre}
$\rho_{f_{\text{fre}},f}$	0.82	0.17	-0.09	0.21	-0.06	-0.05	0.30	0.26	1.00

Reward Dataset Creation

The computation of readability involves the computation of Coh-Metrix, which is difficult to compute online in the RL training loop. We need to create a reward dataset to train a reward model to approximate the readability. In RLHF, a human evaluator is given a prompt and two responses and is required to choose which response is preferred, which results in the chosen response and the rejected response. Here, in this work, we compare the readability to determine which analogy to choose or reject.

Specifically, we first compute the readability of the analogies in the analogy dataset $\mathcal{D}^A$. Then, we group the dataset by prompt, acquiring n groups $\mathcal{G}_1^A, \mathcal{G}_2^A, \cdots, \mathcal{G}_n^A$. All data items in a group $\mathcal{G}_i^A$ share the same prompt $\mathbf{x}_i$. Based on each group $\mathcal{G}_i^A$, we construct a chosen analogy set $\mathcal{Y}_i^c$ containing k analogies with the highest readability and a rejected analogy set $\mathcal{Y}_i^r$ containing k analogies with the lowest readability. We perform the Cartesian product between $\mathcal{Y}_i^c$ and $\mathcal{Y}_i^r$, generating a set of chosen-rejected analogy pairs. After that, we combine each chosen-rejected pair with the prompt $\mathbf{x}_i$ shared within the group, acquiring a new group of reward data, and the final reward dataset is the union of all reward data groups, as shown in Equation 5.

$$\mathcal{G}_i^R = \{(\mathbf{x}_i, \mathbf{y}_i^c, \mathbf{y}_i^r) \,|\, (\mathbf{y}_i^c, \mathbf{y}_i^r) \in \mathcal{Y}_i^c \times \mathcal{Y}_i^r\}, \; \mathcal{D}^R = \bigcup_{i=1}^{n} \mathcal{G}_i^R \quad (5)$$

Reward Modeling Objective

The reward function is denoted by a neural network R_ω which takes the prompt and the response as input and outputs a scalar reward, where ω is the parameter. The training objective is to make sure the reward of the chosen responses is greater than the reward of the rejected responses, as shown in Equation 6, where σ is the sigmoid function.

$$\mathcal{L}(\omega) = -\sum_{\mathbf{x},\mathbf{y}^c,\mathbf{y}^r \in \mathcal{D}^R} \log \sigma(R_\omega(\mathbf{x},\mathbf{y}^c) - R_\omega(\mathbf{x},\mathbf{y}^r)) \quad (6)$$

Cognitive Conflict Reinforcement Learning

Cognitive Rigidity of LLMs

Recently, Janus (2023) has found that the LLM reinforced by the PPO algorithm tends to output answers with significantly high confidence, which is not observed in the non-reinforced LLM. In generative artificial intelligence (Cao et al., 2023), this phenomenon is called mode collapse, which was first observed in generative adversarial networks (GANs) (Goodfellow et al., 2014).

One specific performance of mode collapse in LLM is that it keeps generating similar answers to avoid the question. An explanation for this is that the LLM is trained using RL with a safety reward model to prevent it from answering controversial questions. The RL training can lead to LLM being overconfident in recognizing controversial questions, therefore mistaking ordinary questions for controversial ones and further avoiding answering these questions.

From the machine learning perspective, RL training encourages safety responses without considering other circumstances, e.g., too conservative responses. From the cognitive perspective, the RL training improperly injects knowledge about safety into LLM and builds a confirmation bias in LLM, which leads to cognitive rigidity in recognizing controversial questions.

Creating Cognitive Conflict

The occurrence of conflict requires two entities (humans or robots). We create a clone of the currently training LLM to create conflict between the original LLM and the cloned one. In conventional RL algorithms, the LLM always receives a higher reward when it generates a better answer because the reward itself reflects how good is the answer. We instead give a reward according to the conflict between the original LLM and the cloned one. Specifically, we sample another answer to the same question by the cloned LLM. Then, we give a reward or punishment to the LLM according to how much it outperforms or underperforms the other answer.

In this way, when the conflict between the original LLM and the cloned one is greater, a higher absolute value of the reward (denoting a larger reward or punishment) will be received by the original LLM. Moreover, even if the original LLM generates a good answer as we expected, the answer still has a probability of being challenged if the cloned LLM generates a better answer. In this way, the overconfidence and cognitive rigidity will be alleviated.

Creating the above conflict raises a new question: will LLM still reach the expected optimization objective? For example, we expected to improve the readability of LLM by RL, but we do not always give a higher reward when it generates a response with higher readability. Will the LLM finally achieve higher readability after RL training? The answer is yes and we will prove it theoretically in the following part.

Cognitive Conflict Training Objective

The training objective of cognitive conflict RL can be denoted as Equation 7, where θ is the parameter of the original LLM, θ' is the parameter of the cloned LLM, π_θ denotes the original LLM policy, $\pi_{\theta'}$ denotes the cloned LLM policy, and R_ω is the trained psycholinguistic reward model.

$$J_{\text{cc}}(\theta) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x}),\mathbf{y}'\sim\pi_{\theta'}(\cdot|\mathbf{x})}\left[R_\omega(\mathbf{x},\mathbf{y}) - R_\omega(\mathbf{x},\mathbf{y}')\right]\right] \quad (7)$$

The gradient of the objective cannot be directly computed during training but can be approximated according to the policy gradient theorem (Williams, 1992). The approximated

gradient is shown in Equation 8. Note that the gradient is only concerning the parameter of the original LLM policy θ, not concerning the parameter of the cloned policy θ', although their values are identical to each other.

$$\nabla_\theta J_{cc}(\theta) \approx \left[R_\omega(\mathbf{x},\mathbf{y}) - R_\omega(\mathbf{x},\mathbf{y}')\right] \cdot \nabla_\theta \log \pi_\theta(\mathbf{y}|\mathbf{x}) \quad (8)$$

Proof of Optimality

Proof. The training objective in Equation 7 can be written as Equation 9, where $V_{\pi_{\theta'}}(\mathbf{x})$ denotes the expected reward achieved by starting from prompt $\mathbf{x}$ following policy $\pi_{\theta'}$.

$$\begin{aligned} J_{cc}(\theta) &= \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x}),\mathbf{y}'\sim\pi_{\theta'}(\cdot|\mathbf{x})}\left[R_\omega(\mathbf{x},\mathbf{y}) - R_\omega(\mathbf{x},\mathbf{y}')\right]\right] \\ &= \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}\left[R_\omega(\mathbf{x},\mathbf{y}) - \mathbb{E}_{\mathbf{y}'\sim\pi_{\theta'}(\cdot|\mathbf{x})}\left[R_\omega(\mathbf{x},\mathbf{y}')\right]\right]\right] \\ &= \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}\left[R_\omega(\mathbf{x},\mathbf{y}) - V_{\pi_{\theta'}}(\mathbf{x})\right]\right] \end{aligned} \quad (9)$$

Here, $V_{\pi_{\theta'}}(\mathbf{x})$ can be viewed as a baseline taking a weighted average of the rewards of all answers that may be generated. $R_\omega(\mathbf{x},\mathbf{y})$ is the reward of the specific answer $\mathbf{y}$. Thus, the difference $R_\omega(\mathbf{x},\mathbf{y}) - V_{\pi_{\theta'}}(\mathbf{x})$ can be viewed as the advantage of generating answer $\mathbf{y}$ under policy $\pi_{\theta'}$. Given two arbitrary answers $\mathbf{y}^+$ and $\mathbf{y}^-$, if there is $R_\omega(\mathbf{x},\mathbf{y}^+) > R_\omega(\mathbf{x},\mathbf{y}^-)$, then there is also $R_\omega(\mathbf{x},\mathbf{y}^+) - V_{\pi_{\theta'}}(\mathbf{x}) > R_\omega(\mathbf{x},\mathbf{y}^-) - V_{\pi_{\theta'}}(\mathbf{x})$. Therefore, the optimality of the objective is preserved. □

Experiment

Prompt Stimulation

Seals and Shalin (2023) provides the LLM with a basic prompt "*Create an analogy to explain to explain xxx*", which may not stimulate its psycholinguistic potential. In contrast, we provide the LLM with the enhancement prompts. However, not all types of enhancement prompts can effectively improve readability. According to the correlation values in Table. 2, we only consider the enhancement prompts corresponding to properties that are positively related to Flesch reading ease. The Coh-Metrix scores achieved by supplementing enhancement prompts are shown in Table. 3.

The result shows that the enhancement prompts corresponding to most psycholinguistic properties can stimulate the potential of LLM and can significantly improve the scores of the corresponding properties. An exception is the deep cohesion (PCDCz) and the causal connectives (CNCCaus). In conclusion, prompt stimulation is necessary and effective for LLM to produce more psycholinguistic and readable analogies.

Reward Model Comparison

We set k to 20 in reward dataset creation, i.e., combine the top 20 analogies with the highest readability and the top 20 analogies with the lowest readability. For the hard readability function, we set the threshold ρ to 0. Despite the proposed soft and hard readability functions, we try merely using each single property as the readability function, which performs as a baseline.

Table 3: Coh-Metrix scores with different enhancement prompts.

	NAR	TMP	DC	SYN	CAU	FRE
Llama-2	0.22	0.51	0.09	-0.35	25.56	64.17
Enhancement Prompt						
+ PCNARz	0.48	0.65	0.06	-0.26	24.07	69.43
+ PCTEMPz	0.22	0.60	0.07	-0.30	23.29	62.09
+ PCCNCz	0.38	0.59	0.33	-0.38	27.46	68.39
+ PCDCz	-0.04	0.50	-0.05	-0.21	24.36	59.33
+ PCREFz	-0.31	0.04	-0.09	-0.52	22.04	51.69
+ PCCONNz	-0.34	0.55	0.01	-0.31	24.25	50.41
+ PCSYNz	0.48	0.67	**0.58**	**-0.20**	**32.57**	**73.41**
+ CNCCaus	-0.03	0.49	0.07	-0.37	24.32	57.75
+ RDFRE	**0.53**	**0.67**	0.23	-0.31	26.28	70.01

Table 4: Coh-Metrix scores with different readability functions.

	NAR	TMP	DC	SYN	CAU	FRE
Llama-2 + PCSYNz	0.48	0.67	0.58	-0.20	32.57	73.41
Single Property						
+ PCNARz Reward	0.51	0.53	0.56	-0.25	31.39	71.19
+ PCTEMPz Reward	0.32	**0.69**	0.47	-0.25	31.26	72.51
+ PCDCz Reward	0.37	0.50	**0.64**	-0.22	31.16	71.89
+ PCSYNz Reward	0.32	0.53	0.52	-0.20	31.43	72.30
+ CNCCaus Reward	0.42	0.49	0.55	-0.29	32.83	73.24
+ RDFRE Reward	0.45	0.61	0.58	-0.24	31.53	73.60
Ours						
+ Soft Reward	0.51	0.68	**0.64**	**-0.19**	**33.43**	**73.88**
+ Hard Reward	**0.52**	0.65	0.63	-0.20	33.41	73.68

Each choice of the readability function corresponds to a different training process and produces a different reward model. We integrate all reward models into the cognitive conflict training of the LLM to see which reward model has the best performance in improving the readability of the LLM. We concatenate the best enhancement prompt (i.e., the syntactic simplicity prompt) during cognitive conflict RL. The result is shown in Table 4.

RL Algorithm Comparison

We first compare cognitive conflict RL with A2C (Mnih et al., 2016), PPO (Schulman, Wolski, Dhariwal, Radford, & Klimov, 2017), and self-critical sequence training (SCST) (Rennie, Marcheret, Mroueh, Ross, & Goel, 2017), which are classic RL algorithms commonly adopted in RL training of LLMs. Then, we compare with ACTDE (Turner & Einhorn, 2023) and APA (B. Zhu et al., 2023), which are proposed to solve the mode collapse (cognitive rigidity) issue from the mathematical perspective. We still concatenate the best enhancement prompt during RL training. The result is shown in Table 5. Despite computing the Coh-Metrix scores, we also record the reward curve, which intuitively reflects the convergence speed and the convergence upper bound of the training process. We show the reward curves of different RL algorithms in Figure 1.

Table 5: Coh-Metrix scores with different RL algorithms.

	NAR	TMP	DC	SYN	CAU	FRE
Llama-2 + PCSYNz	0.48	0.67	0.58	-0.20	32.57	73.41
RL Approach						
+ SCST	0.50	0.62	0.53	-0.17	32.56	73.67
+ PPO	0.43	0.62	0.49	**-0.15**	31.38	73.71
+ ACTDE	0.42	0.56	0.63	-0.26	32.64	72.94
+ APA	0.50	**0.68**	0.59	-0.28	32.58	**73.94**
+ CCRL (Ours)	**0.51**	**0.68**	**0.64**	-0.19	**33.43**	73.88

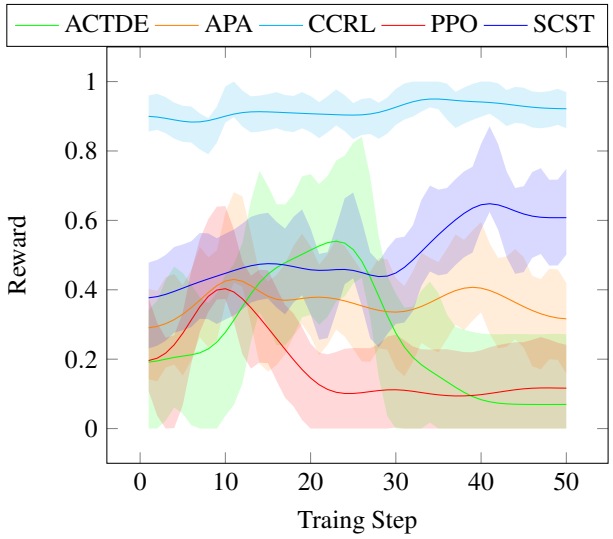

Figure 1: Comparison of reward curves of different RL approaches.

The result in Table 5 shows that CCRL achieves the highest scores on most of the psycholinguistic properties related to readability. PPO and APA do achieve the highest scores on some of the properties but do not improve the properties comprehensively, i.e., they underperform the non-reinforced baseline approach on some properties. The reward curves in Figure 1 show that CCRL achieves the highest reward at the very beginning of the training, while other approaches take more training steps to converge. Moreover, the PPO-based approaches (PPO, ACTDE, and APA) suffer from reward deterioration after some training steps. Additionally, converging to higher rewards does not always ensure higher property scores. For example, the SCST algorithm converges to a higher reward than the APA algorithm as shown in Figure 1, but underperforms the APA algorithm on most of the properties as shown in Table 5. This can be attributed to errors in reward models reflecting psycholinguistic properties.

ChatGPT and Human Comparison

We also collect 20 human-written analogies for each concept from 160 participants from different. We give each participant the same prompt as the one given to LLMs, i.e., the basic prompt corresponding to the concept and the enhancement prompt related to syntactic simplicity. We collect 300 analogies for each concept from ChatGPT using the same prompt as aforementioned. Table 6 shows the performance of human participants and the ChatGPT.

Table 6: Coh-Metrix scores achieved by human participants, ChatGPT, and our approaches.

	NAR	TMP	DC	SYN	CAU	FRE
Human + PCSYNz	-0.31	0.50	0.21	**0.08**	32.14	70.15
ChatGPT + PCSYNz	-0.34	0.36	-0.12	**0.08**	22.23	51.63
Llama-2 + PCSYNz	0.48	0.67	0.58	-0.20	32.57	73.41
+ CCRL	**0.51**	**0.68**	**0.64**	-0.19	**33.43**	**73.88**

Table 7: Diversity evaluation of different RL approaches. "↓" means the diversity is higher when the value is lower.

	DIST-sent ↑	SelfBLEU ↓
Llama-2 + PCSYNz	73.56	3.70
RL Approach		
+ SCST	72.80	3.97
+ PPO	74.22	3.90
+ ACTDE	75.26	8.36
+ APA	72.95	5.02
+ CCRL (Ours)	**75.70**	**3.62**

We can see that our approaches, whether reinforced or not, significantly outperform ChatGPT and human participants. Additionally, by comparing *Llama-2 + PCSYNz* with *ChatGPT + PCSYNz* and *Human + PCSYNz*, we can conclude that the Llama-2 model is more sensitive to psycholinguistic property enhancement prompts than ChatGPT and human participants.

Rigidity Analysis

We believe that, intuitively, lower diversity means higher cognitive rigidity. Therefore, we estimate the cognitive rigidity of LLMs using the diversity of the generated texts. We adopt two metrics, distinct sentences (DIST-sent) (Li, Galley, Brockett, Gao, & Dolan, 2015) and SelfBLEU (Y. Zhu et al., 2018), to evaluate the lexical diversity. As shown in Table 7, CCRL outperforms other approaches in diversity evaluation.

Conclusion

We analyzed the issue of cognitive rigidity in LLMs from a cognitive perspective and proposed to create conflict between the original LLM and a cloned LLM during reinforcement learning, which helps alleviate the cognitive rigidity problem. Experimental results demonstrated that our CCRL approach significantly outperforms previous RL algorithms in improving various psycholinguistic properties related to readability. The analogies generated by our approach also surpassed human-written ones, achieving the highest readability scores. A limitation of this work is that we only considered eight scientific concepts from four domains and employed only one pre-trained LLM. In future work, we plan to extend the approach to more scientific concepts and pre-trained LLMs.

References

Alves, M. A., & Pozzebon, M. (2013). *How to resist linguistic domination and promote knowledge diversity?* (Vol. 53). SciELO Brasil.

Amadieu, F., Tricot, A., & Mariné, C. (2009). Prior knowledge in learning from a non-linear electronic document: Disorientation and coherence of the reading sequences. *Computers in Human Behavior*, *25*(2), 381–388.

Amadieu, F., Van Gog, T., Paas, F., Tricot, A., & Mariné, C. (2009). Effects of prior knowledge and concept-map structure on disorientation, cognitive load, and learning. *Learning and instruction*, *19*(5), 376–386.

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*.

Chall, J. S. (1958). *Readability: An appraisal of research and application*. The Ohio State University, Columbus, OH, USA.

De Dreu, C. K., & Nijstad, B. A. (2008). Mental set and creative thought in social conflict: threat rigidity versus motivated focus. *Journal of personality and social psychology*, *95*(3), 648.

Gauer, F., & Kuzmics, C. (2020). Cognitive empathy in conflict situations. *Behavioral & Experimental Economics eJournal*. Retrieved from https://api.semanticscholar.org/CorpusID:3331195

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-metrix: Providing multilevel analyses of text characteristics. *Educational researcher*, *40*(5), 223–234.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, *36*(2), 193–202.

Haig, E., & Woodcock, K. (2017). Rigidity in routines and the development of resistance to change in individuals with prader–willi syndrome. *Journal of Intellectual Disability Research*, *61*(5), 488–500.

Janus. (2023). Mysteries of mode collapse. *Less Wrong*.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.* Institute for Simulation and Training, University of Central Florida.

Larsson, Å., Haglund, L., & Halldén, O. (2010). Cognitive conflict: Actions taken in the process of conceptual change. *Nordic Educational Research Working paper series*.

Lee, G., Kwon, J., Park, S.-S., Kim, J.-W., Kwon, H.-G., & Park, H.-K. (2003). Development of an instrument for measuring cognitive conflict in secondary-level science classes. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *40*(6), 585–603.

Lehman, B., D'Mello, S., Strain, A., Mills, C., Gross, M., Dobbins, A., ... Graesser, A. (2013). Inducing and tracking confusion with contradictions during complex learning. *International Journal of Artificial Intelligence in Education*, *22*(1-2), 85–105.

Li, J., Galley, M., Brockett, C., Gao, J., & Dolan, B. (2015). A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Mehler, J., Pallier, C., & Christophe, A. (1998). Language and cognition.. Retrieved from https://api.semanticscholar.org/CorpusID:264687671

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International conference on machine learning* (pp. 1928–1937).

Rappoport, L. (1969). Cognitive conflict as a function of socially-induced cognitive differences. *Journal of Conflict Resolution*, *13*(1), 143–148.

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2017). Self-critical sequence training for image captioning. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 7008–7024).

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Schultz, P. W., & Searleman, A. (2002). Rigidity of thought and behavior: 100 years of research. *Genetic, social, and general psychology monographs*, *128*(2), 165.

Seals, S., & Shalin, V. L. (2023). Long-form analogies generated by chatgpt lack human-like psycholinguistic properties. *arXiv preprint arXiv:2306.04537*.

Sherman, L. (1893). *Analytics of literature: A manual for the objective study of english prose and poetry. ginn & co.* Boston: Ginn & Company. http://scholar. google. com/scholar.

Shing, Y. L., & Brod, G. (2016). Effects of prior knowledge on memory: Implications for education. *Mind, Brain, and Education*, *10*(3), 153–161.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Thierry, G. (2016). Neurolinguistic relativity: How language flexes human perception and cognition. *Language Learning*, *66*(3), 690–713.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... others (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Turner, A., & Einhorn, M. (2023). Mode collapse in rl may be fueled by the update equation. *Alignment Forum*.

Wason, P. (2021). Confirmation bias. *Explaining the Evidence*. Retrieved from https://api.semanticscholar.org/CorpusID:204874888

Werbos, P. J. (1988). Generalization of backpropagation with

application to a recurrent gas market model. *Neural networks*, *1*(4), 339–356.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Reinforcement learning*, 5–32.

Woods, J. G. (2012). Using cognitive conflict to promote the use of dialectical learning for strategic decision-makers. *The Learning Organization*, *19*(2), 134–147.

Zhu, B., Sharma, H., Frujeri, F. V., Dong, S., Zhu, C., Jordan, M. I., & Jiao, J. (2023). Fine-tuning language models with advantage-induced policy alignment. *arXiv preprint arXiv:2306.02231*.

Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., & Yu, Y. (2018). Texygen: A benchmarking platform for text generation models. In *The 41st international acm sigir conference on research & development in information retrieval* (pp. 1097–1100).

Zilversmit, D. (1964). The impact of rigid definitions on scientific thinking. *Perspectives in Biology and Medicine*, *7*(2), 227–248.