INTERNATIONAL MONETARY FUND

Reinforcement Learning from Experience Feedback: Application to Economic Policy

Tohid Atashbar

WP/24/114

IMF Working Papers **describe research in progress by the author(s) and are published to elicit comments and to encourage debate.**
The views expressed in IMF Working Papers are those of the author(s) and do not necessarily represent the views of the IMF, its Executive Board, or IMF management.

2024
JUN

WORKING PAPER

WP/24/114

IMF Working Paper
Strategy, Policy and Review Department

Reinforcement Learning from Experience Feedback: Application to Economic Policy

Prepared by Tohid Atashbar*

Authorized for distribution by Eugenio M. Cerutti

June 2024

IMF Working Papers **describe research in progress by the author(s) and are published to elicit comments and to encourage debate.** The views expressed in IMF Working Papers are those of the author(s) and do not necessarily represent the views of the IMF, its Executive Board, or IMF management.

ABSTRACT: Learning from the past is critical for shaping the future, especially when it comes to economic policymaking. Building upon the current methods in the application of Reinforcement Learning (RL) to the large language models (LLMs), this paper introduces Reinforcement Learning from Experience Feedback (RLXF), a procedure that tunes LLMs based on lessons from past experiences. RLXF integrates historical experiences into LLM training in two key ways - by training reward models on historical data, and by using that knowledge to fine-tune the LLMs. As a case study, we applied RLXF to tune an LLM using the IMF's MONA database to generate historically-grounded policy suggestions. The results demonstrate RLXF's potential to equip generative AI with a nuanced perspective informed by previous experiences. Overall, it seems RLXF could enable more informed applications of LLMs for economic policy, but this approach is not without the potential risks and limitations of relying heavily on historical data, as it may perpetuate biases and outdated assumptions.

RECOMMENDED CITATION: T. Atashbar. 2024. “Reinforcement Learning from Experience Feedback: Application to Economic Policy”, IMF Working Papers, WP/24/114.

JEL Classification Numbers:	C89; D83; O38
Keywords:	LLMs; GAI; RLHF, RLAIF; RLXF
Author’s E-Mail Address:	tatashbar@imf.org

* The author would like to thank Mira Vessel, Mohamed Afzal Norat, Martin Schindler, and Eugenio Cerutti for their helpful comments and suggestions. Thanks are also given to the AI working groups from the SPR and AFR departments at the IMF. I am grateful for the feedback and ideas from attendees at the IMF SPR-MP Presentation. Any mistakes are solely mine.

WORKING PAPERS

Reinforcement Learning from Experience Feedback: Application to Economic Policy

Prepared by Tohid Atashbar

Contents

TABLES

Glossary

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BPE	Byte Pair Encoding
GAI	Generative AI
GPT	Generative Pre-trained Transformers
HIR	Hindsight Instruction Relabeling
LLM	Large Language Model
MoE	Mixture of Experts
MONA	Monitoring of Fund Arrangements
PPO	Proximal Policy Optimization
ReLU	Rectified Linear Unit
ReST	Reinforced Self-Training
RL	Reinforcement Learning
RLAIF	Reinforcement Learning from AI Feedback
RLHF	Reinforcement Learning from Human Feedback
RLXF	Reinforcement Learning from Experience Feedback

Introduction

In policymaking, learning from historical outcomes and real-world lessons is crucial for developing pragmatic solutions. Without properly integrating past experience into the policy design process, policies risk being poorly conceived, superficial, or outright ineffective when implemented. By thoroughly analyzing past policy successes and failures, policymakers can make more informed decisions that have a higher chance of producing the intended impact.

The IMF has decades of real-world expertise captured in datasets like the Monitoring of Fund Arrangements (MONA) database. These datasets provide a wealth of historical data on IMF-supported economic programs, including quantitative targets, policy conditions, and outcomes. By thoroughly analyzing this experience and encoding the key lessons into how AI models are trained and tuned, their outputs could become better aligned with the nuanced realities of economic policymaking.

With their proficient text processing abilities, large language models (LLMs) represent a major advancement in AI, opening up new possibilities across various domains. LLMs are increasingly being applied to policy issues like drafting and summarization, scenario planning, brainstorming, and generating policy recommendations. By leveraging the knowledge encoded in their parameters, LLMs can rapidly synthesize information and provide useful insights to inform policy decisions. As computational power and model scale continue improving, the usefulness of LLMs in policy analysis will only grow and policymakers are beginning to tap into their potential to enhance policy formulation, governance and outcomes.

However, it's crucial to be aware of the potential pitfalls and risks associated with using LLMs for policy assessment and advice. One significant concern is the inherent bias of LLMs towards historical patterns. As language models are trained on vast amounts of past data, they tend to perpetuate and reinforce the biases and assumptions embedded in that data. In the context of economic policy, this could lead to the LLM favoring outdated or inappropriate policies that worked in the past but may not be suitable for current circumstances. Policymakers must be cautious not to blindly follow LLM recommendations without critically examining their applicability to the present context.

Another risk is the potential for LLMs to provide overly simplistic or reductionist policy assessments. While LLMs can process and analyze large volumes of information, they may struggle to fully capture the complexity and nuances of economic systems and policy impacts. LLMs might generate advice that overlooks important contextual factors, unintended consequences, or distributional effects. Relying too heavily on LLM outputs without considering these limitations could lead to suboptimal or even harmful policy decisions.

To mitigate these risks, it is essential to use LLMs as a complement to, rather than a replacement for, human expertise and empirical research. LLM outputs should be carefully scrutinized and validated by domain experts who can assess their relevance and feasibility in light of current economic realities. Policymakers should also strive for transparency in their use of LLMs, clearly communicating the role of these models in the decision-making process and acknowledging their limitations.

Integrating real-world lessons systematically into LLM training is essential for aligning their outputs with practical policy wisdom. One approach is to include curated datasets on policy case studies and outcomes as part of the model pre-training process. Post-training fine-tuning on expert demonstrations that encode domain knowledge is another impactful technique. Using customized and well-designed few-shot or chain of thought (CoT) prompting is another method which is useful for extracting more relevant output during the inference process.

Reinforcement learning (RL) presents a promising paradigm for fine-tuning LLMs in a way that incorporates additional inputs. In RL, models learn through trial-and-error interactions with an environment. Previously, RL has been applied to align LLMs with human values to make the output more helpful and less harmful through a process called Reinforcement Learning from Human Feedback (RLHF) in which human reviewers and annotators evaluate and rank LLM output to train a reward model that is used to retune the LLM.

Building on the RLHF and integrating insights from more recent advancements in applying RL to LLMs, this paper introduces a concept wherein RL utilizes past experiences to construct the reward model, subsequently refining the LLMs. We will show that by designing the reward signals in RL to encode lessons from prior policy lessons, in a process that we call reinforcement learning from experience feedback (RLXF), LLMs can systematically integrate this knowledge into their output generation.

In this paper, we present a case study of applying the proposed technique to enhance LLMs for economic policy analysis. We apply RLXF to enhance a medium size open source LLMs for economic policy purposes. Using the IMF's MONA dataset, and Meta's LLaMA 2-7B[1], we will demonstrate how RLXF can tune LLMs to generate more pragmatic and contextualized insights. The case study highlights the value of grounding LLMs in human experiences to make them beneficial for policy issues.

We think RLXF could have unique advantages for fine-tuning LLMs to incorporate real-world knowledge. Unlike instruction-based tuning, which is usually limited by fixed human demonstrations, RL allows models to explore a wider range of options through trial-and-error learning. And unlike supervised learning which may risk optimizing for mimicking text among other risks, RL focuses models on developing deeper competence by encoding insights directly from experience into reward signals. RLXF also could enable efficiently injecting nuanced expertise into LLMs without ongoing human oversight. Compared to prior RL methods like RLHF that rely on human judgment of outputs, or a limited set of rules in more recent approaches that use AI to create the reward model, this approach could provide more efficient and scalable learning in economic policy applications. By designing the reward signals based on historical data or experiences, rather than human feedback or a set of do and do nots, the model can learn from an exponentially larger set of scenarios and scale seamlessly without bottlenecking on human evaluators. This could also provide a targeted mechanism for instilling LLMs with pragmatic skills and prior historical lessons.

In summary, while LLMs enable engaging natural language, true usefulness in policy requires alignment with human experience in addition to wisdom and values. Merely training them on texts could result in outputs disconnected from reality. Techniques like dataset curation, expert demonstrations, and RL with experience-based rewards in which we introduce offer pathways for instilling LLMs with practical knowledge. The goal is not anthropomorphizing LLMs but equipping them to offer substantive insights to complement human expertise in policy roles.

In this paper, we will first review literature on how LLMs work, with an overview of training, tuning, and alignment methods. We will then analyze possible gaps in current tuning and alignment approaches, emphasizing the need to incorporate real-world lessons. Next, we will propose reinforcement learning from experience feedback as a methodology that can help to address these gaps, besides using other methods. Finally, using the Fund's MONA dataset, we will present the case study of applying the proposed technique to enhance LLMs for economic policy analysis. The case study will highlight the value of grounding LLMs in human experiences to make them beneficial for policy issues.

Literature Review

History

LLMs are a class of natural language processing systems based on deep neural networks that are trained on massive text corpora. The history of LLMs traces back to 2013 when word embedding models like word2vec (Mikolov et al., 2013) demonstrated the power of unsupervised/self-supervised pretraining. However, the field took off in 2018 after OpenAI introduced GPT, one of the first transformer-based LLMs. Since then, models have rapidly scaled, with state-of-the-art LLMs possessing over trillions of parameters. This growth has been driven by advances in model architecture, optimization techniques, and compute infrastructure. LLMs now cover a wide scope of capabilities including but not limited to text generation, classification, summarization, translation, and question answering.

[1] As this paper was being finalized, newer models, including the LLaMA 3 series, were introduced. The methodology outlined in this paper is applicable to these new models as well.

The transformer architecture was introduced in "Attention is All You Need" (Vaswani et al. 2017), showing attention could replace recurrence. GPT-1 demonstrated strong transfer learning abilities from pretraining in "Improving Language Understanding by Generative Pre-Training" (Radford et al. 2018). BERT illustrated training a bidirectional model and power of pretraining in "Pre-training of Deep Bidirectional Transformers for Language Understanding" (Devlin et al. 2019) and impressive few-shot task performance via massive scaling was shown in GPT-3's paper "Language Models are Few-Shot Learners" (Brown et al. 2020). Since then, many new closed and open source LLMs like newer version of GPT, LLAMA, Claude, Mixtral, Gemini, etc. have been developed, showing continued progress. The rapid recent advances indicate LLMs will continue improving and finding new applications across language tasks[2].

How LLM works

Large language models are often built using the Transformer architecture. The Transformer is comprised of blocks containing two primary components that work together during processing:

Multi-Head Attention Mechanisms allow the model to focus on different parts of the input sequence. Multiple parallel attention heads are used, each computing compatibility scores between query and key vectors. These scores are used to aggregate values into updated representations for each token. This is usually represented as:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{dk}})V$$

The function **Attention** calculates attention scores using **Q**ueries, **K**eys, and **V**alues. QK^T determines compatibility scores, while $\sqrt{dk}$ scales them. After normalization via softmax, the scores weight the Values to aggregate relevant information.

Feedforward Neural Networks then provide depth through non-linearity. They typically consist of two affine transformations wrapped around a non-linear activation function like Rectified Linear Unit (ReLU):

$$\text{FFN}(x) = max(0, xW_1 + b_1)W_2 + b_2$$

Where W_1, $W_2$are weight matrices and b_1, $b_2$are bias terms.

Before processing, the raw text input must be tokenized. Tokenization converts the text into smaller units like sub-word units using techniques like byte-pair encoding (BPE). This reduces the vocabulary size needed. Training large language models presents computational challenges. Scaling refers to increasing model capacity by adding more layers, expanding layer width, or using larger training datasets. This improves the model's ability to generalize and capture intricate patterns. However, it also increases the risk of overfitting. Regularization techniques like L2 normalization add penalty terms to the loss function to constrain the parameters:

$$L_{\text{reg}} = L + \lambda||\theta||^2$$

Here, L is the original loss, λ is the regularization coefficient, and $||\theta||^2$ represents the squared magnitude of the model parameters θ. This technique penalizes large parameter values to prevent overfitting.

The training loss is typically cross-entropy between predicted and true token probabilities. Optimization algorithms like stochastic gradient descent (e.g., Adaptive Moment Estimation - Adam) are used to minimize the loss and update the parameters each training step. Additional tricks like positional encodings could be employed to provide order information and better understanding of word sequences.

[2] For more information and detailed explanation on the history and mechanisms of LLMs, see Vaswani et al. (2017), Gudivada (2024), Gao et al. (2024), Huang et al. (2023) and Trad & Chehab (2024).

LLM Development

The capabilities of large language models stem from a combination of training techniques, tuning procedures, and alignment mechanisms during their development process. LLMs are first trained on massive text datasets to develop generalized linguistic representations. After pre-training, the models are tuned to specialize for tasks. A key challenge is alignment or modifying LLMs to produce outputs that align with human values, making them less harmful, more helpful, honest, and beneficial for society. This involves techniques to make models reinforce positive traits like empathy, truthfulness, and inclusiveness while avoiding negative attributes like bias, toxicity, and falsehoods. In the following section we will dive deeper into the training, tuning (also called fine-tuning), and alignment techniques of LLMs.

Training
The predominant training methodology for large language models is self-supervised pretraining on massive corpora of text data. Pretraining results in capable language models before any downstream application.

To increase model capacity further, approaches like mixture of experts (MoE) can be used during the training. This trains a model with N expert sub-models, each with parameters θn. A gating network g determines weighting $g_n$for each expert:

$$p(y|x) = \Sigma g_n(x) \times p(y|x, \theta_n)$$

In this, $p(y|x)$ is the combined prediction, $g_n(x)$ is the weight for the expert determined by a gating network, and $p(y|x, \theta_n)$ is the prediction of the expert n^{th} with parameters. The gating network enables routing different inputs or prompts to the most relevant expert.

Other approaches include multitask learning, with shared parameters θ_{MT} across tasks and task-specific θ_i parameters:

$$L = \Sigma \lambda_i L_i(\theta_{\text{MT}}, \theta_i)$$

Here, the equation $\Sigma \lambda_i L_i(\theta_{\text{MT}}, \theta_i)$ represents a loss function in multitask learning. L is the combined loss, L_i is the loss for task i, θ_{MT} are shared parameters, θ_i are task-specific parameters, and λ_i are the task weights.

Other techniques used to enhance large language model training include knowledge distillation for transferring knowledge from a larger teacher model to a smaller student model, self-distillation for transferring knowledge across time within a model, sparse training for updating only a subset of parameters each batch, progressive module stacking to gradually add layers during training, curriculum learning for organizing training data from simple to complex, meta-learning to learn optimal initialization and hyperparameters, and multi-agent learning where multiple models learn collaboratively.

Tuning
Post-training including what's generally called fine-tuning is a critical step that allows adapting pretrained language models to downstream tasks. It works by further optimizing the model on a smaller dataset for a particular task using a supervised loss. This allows specialized models for tasks like sentiment analysis and question answering to be developed from the same foundation.

Fine-tuning jointly optimizes the original pretraining loss LLM along with a new downstream task loss L_T. Typically, only a subset of parameters θ are tuned, retaining the general knowledge in φ:

$$L = L_{\text{LM}} + \lambda L_T$$
$$min_\theta L(\theta; \phi)$$

This leverages transfer learning, adapting the model to specialized uses. In the equation, the equation L is the combined loss during fine-tuning, where L_{LM} is the original pretraining loss and L_T is the downstream task loss. The coefficient λ weights the importance of the downstream task. The expression $min_\theta L(\theta; \phi)$ signifies the optimization of parameters θ while retaining the general knowledge encapsulated in ϕ.

Many effective techniques exist for fine-tuning LLMs; Additional classifier layers tailored to task labels can be added and trained. The model can be fine-tuned on cloze-style prompts formulated for the task. Intermediate pretraining on data similar to the end task before fine-tuning can enhance results. Different modules of a model can be fine-tuned separately for different tasks. Joint multi-task fine-tuning across tasks is also viable. When limited labeled data is available, strategies like augmentation and regularization enable low-shot fine-tuning. Instruction tuning, where models are trained to follow natural language instructions, is an important technique for aligning models to human preferences.

Supervised fine-tuning (SFT) is the standard approach for adapting pretrained large language models to downstream tasks. Labeled datasets specific to the target task are used to train a classifier on top of the model. For instance, a dataset of text labeled with sentiment can fine-tune a model for sentiment analysis by adding a classification layer and training it to match the ground truth labels. Other SFT approaches include training on textual demonstrations, prompt-based tuning (also known as prompt tuning), and intermediate pretraining on in-domain datasets. For example, in Instruction Tuning, instructions serve as the "labels" that provide supervised training signals. SFT enables effective transfer learning, specializing the general capabilities learned during self-supervised pretraining towards specific tasks. Challenges include requirement for large labeled datasets and difficulty modeling nuanced human preferences beyond the labels.

Alignment

Alignment refers to modifying and optimizing large language models to produce outputs that adhere to human values and beneficial norms. Whereas LLMs are typically trained to simply maximize prediction accuracy, alignment introduces the challenge of how to shape model behavior to avoid harmful effects and promote positive social impacts. The motivation is to address growing concerns about issues like model bias, toxicity, and deception by incorporating human ethics and preferences into LLMs. The overall goal is to develop capable and generalizable LLMs that remain steadfastly truthful, harmless, and honest. Research on alignment is still nascent, but main techniques developed so far include using reinforcement learning (RL) and reward modeling to optimize for specified objectives, learning from human demonstrations, and introducing legal/ethical constraints via some reference texts. This typically involves introducing reward/penalty signals r to optimize for beneficial behaviors:

$$L = L_{\text{ML}} + \lambda E[r]$$

Here, L_{ML}represents the standard unsupervised pretraining loss that maximizes the likelihood of generating the training text, as used during the initial training of the LLM. The new term $\lambda E[r]$ incorporates the expected reward r, which provides a signal to optimize the model's outputs to conform to human preferences. λ is a weighting hyperparameter. Optimizing this compound loss function trains the model not just for prediction accuracy, but also for beneficial alignment with human values, as quantified through the expected rewards. The relative weight between the two terms allows balancing language modeling performance and alignment. The reward r can be provided explicitly by humans rating model outputs during reinforcement learning from human feedback. Alternatively, it can be learned recursively by the model itself through environment interactions in approaches like reinforcement learning from AI feedback.

A variety of techniques exist for aligning LLMs to human values and beneficial behaviors. Reinforcement learning methods use human feedback or learned reward models to shape behaviors through trial-and-error optimization of incentives. Imitation learning[3] trains models to mimic demonstrated examples of proper conduct. Instruction tuning could provide natural language guidance for desired alignments. Modifying model architecture, loss functions, and training data to induce compliance with rules and constraints is another approach. Ongoing research is exploring hybrid alignments methods along with auditing processes to verify

[3] This allows steering LLMs based on concrete examples of proper conduct, without needing to specify reward functions or formal constraints. Imitation learning focuses on actions rather than rewards. In this setting, the model is provided with textual demonstrations exhibiting desired behaviors and ethics. For example, an LLM could be shown example dialogues where an assistant responds helpfully and harmlessly to abusive comments. The model learns behaviors by trying to match its own textual outputs to the demonstration data, training to mimic the positive responses. Challenges include propagation of any poor examples, and difficulty adapting to new situations. If used alongside other techniques like reinforcement learning, it could offer a complementary method to instill societal values by showing models directly how to act properly.

model helpfulness or honesty and safety. The following section focuses on the RL techniques used for the alignment.

Risks
When it comes to using LLMs for policy assessment or advice, it's important to consider the limitations of relying solely on an LLM, even after post-training enhancements with different methods like reinforcement learning, including the method we propose in this paper. While the LLM can learn patterns and relationships from historical data, it may not be able to fully capture the complexity and nuances that a large body of empirical research aims to address. Empirical studies in economics often employ sophisticated econometric techniques to account for various pitfalls, such as endogeneity, omitted variable bias, and reverse causality. These methodologies are designed to isolate the causal effects of policies and provide more reliable assessments of their effectiveness. An LLM, even with the benefit of historical grounding, may struggle to replicate the depth and rigor of such carefully designed research.

However, this does not mean that LLMs have no role to play in policy assessment. One potential approach could be to train the LLM on the findings and insights from macroeconomic research, allowing it to incorporate the knowledge gleaned from empirical studies. By exposing the LLM to a curated corpus of high-quality research papers, working papers, and policy briefs, it could potentially absorb the collective wisdom of the economic community. This could help the LLM to internalize the nuances and caveats identified by empirical research, enhancing its ability to provide more robust policy assessments. Nevertheless, it's crucial to recognize that an LLM's assimilation of research findings would still be an approximation and may not fully capture the intricacies and uncertainties inherent in economic analysis. Ultimately, the role of an LLM in policy evaluation should be seen as complementary to, rather than a replacement for, rigorous empirical research.

Improper use of LLMs in policy assessment and advice could lead to several adverse consequences. One major risk is the potential for LLMs to perpetuate biases present in historical data. If the training data contains systemic biases or reflects outdated assumptions, the LLM may inadvertently learn and reinforce these biases in its policy evaluations. This could lead to skewed assessments that fail to account for changing social, economic, and political contexts. Moreover, if policymakers rely too heavily on LLM-generated advice without critically examining its limitations, they may make decisions based on incomplete or misleading information. This could result in policies that are ill-suited to current challenges or that exacerbate existing inequalities.

Another concern is the potential for LLMs to provide overly simplistic or reductionist policy assessments. While LLMs can process vast amounts of data and identify patterns, they may struggle to capture the full complexity of economic systems and the multifaceted nature of policy impacts. LLMs might generate advice that overlooks important contextual factors, unintended consequences, or distributional effects. If policymakers take such oversimplified assessments at face value, they risk implementing policies that are ineffective or even harmful. To mitigate these risks, it is essential for policymakers to use LLM-generated insights judiciously, always considering them in conjunction with expert judgment, stakeholder input, and rigorous empirical analysis. LLMs should be seen as a tool to augment, rather than replace, human decision-making in the policy realm.

Reinforcement Learning and LLMs

Reinforcement learning provides an alternative paradigm to fine-tune large pretrained language models, typically to align with human values or principles, beyond standard supervised learning approaches. In reinforcement learning, the model interacts with an environment by taking actions and receiving reward or penalty signals based on the results. The goal is to learn an optimal policy that maximizes long-term reward. This framework aligns well with shaping beneficial language model behaviors via rewards for generative model outputs.

The loss function in reinforcement learning fine-tuning incorporates expected rewards rather than task-specific cross-entropy losses. Various techniques exist for supplying the reward signal:

Reinforcement Learning from Human Feedback (RLHF)

RLHF is a popular technique for fine-tuning large language models using reinforcement learning with reward signals provided directly by humans. Paul Christiano et al. (2017) introduced the basic idea of RLHF, which is to train an RL agent to learn from human preferences. The authors showed that this approach can be used to solve complex RL tasks. Ziegler et al. (2019) apply RLHF to the task of fine-tuning LLMs. The authors demonstrated that RLHF can be used to improve the performance of language models on a variety of tasks, such as question answering and summarization.

The RLHF pipeline can be divided into three separate steps (Ouyang, et al., 2022):

1. *Supervised finetuning of the pretrained model:* In this step, a pretrained language model is finetuned using a supervised learning approach. This means that the model is trained on a dataset of text and labels, where the labels indicate whether the text is factually accurate or engaging. The goal of this step is to improve the language model's ability to generate text that is both factually accurate and engaging.

2. *Creating a reward model:* In this step, a reward model is created. The reward model is a function that takes as input a piece of text and outputs a reward value. The reward value indicates how much the human would like the text. The reward model is typically trained using a supervised learning approach, where the labels are provided by humans.

3. *Finetuning via proximal policy optimization:* In this step, the language model is finetuned using a reinforcement learning approach called proximal policy optimization (PPO). PPO is a policy gradient algorithm that can be used to train agents to maximize a reward function. In the context of RLHF, the reward function is the output of the reward model. The goal of this step is to improve the language model's ability to generate text that is more likely to be rewarded by the human.

In RLHF, humans evaluate model outputs, such as text or dialogue responses, and provide numeric scores rating the quality of each output. These human ratings are used as reward signals to create a smaller reward model and reinforce beneficial model behaviors that highly align with human preferences. The LLM is trained via RL algorithms on the reward model to maximize its expected reward from the human signals. A key advantage of RLHF is that the human feedback allows directly fine-tuning models to align with nuanced human values, without requiring large, supervised datasets. However, a limitation is that it relies on human effort and scaling the evaluation process can be challenging.

Reinforcement Learning from AI Feedback (RLAIF)

This method (Lee et al., 2023) is a more recent technique based on the approach proposed in Constitutional AI research (Bai et al., 2022) to align large language models while avoiding extensive human involvement. In the Constitutional AI approach, a model is first trained to critique its own harmful responses and generate revised, constitutionally aligned responses. This creates training data of harmful/revised pairs, which fine-tunes the model on constitutional principles. Next, this fine-tuned model generates preferences on new responses to form a dataset to train a reward model. Reinforcement learning is then performed using this reward model to further optimize the model without direct human feedback.

In the RLAIF, the LLM is first used to generate preference labels for pairs of candidates, such as summaries of text. These preference labels are then used to train a reward model, which predicts the preferences of a human. The RL agent is then trained to maximize the reward predicted by the reward model.
The RLAIF methodology (Lee et al., 2023) has three main steps:

1. *Preference labeling with LLMs:* The LLM is used to generate preference labels for pairs of candidates. The input to the LLM is a prompt that describes the task, such as "Which summary is better?". The LLM then outputs a preference distribution, which indicates how likely it is that the first candidate is better than the second candidate.

2. *Training a reward model:* The preference labels generated by the LLM are used to train a reward model. The reward model is a function that takes as input a pair of candidates and outputs a reward value. The reward value indicates how much the human would prefer the first candidate over the second candidate.

3. *Reinforcement learning:* The RL agent is trained to maximize the reward predicted by the reward model. The RL agent takes actions, such as generating text, and receives rewards from the reward model. The RL agent is updated so that it is more likely to take actions that lead to higher rewards.

Reinforcement Learning from Previous Experiences (RLXF)

This is an approach that we propose in this paper, which is a conceptual approach and procedure to align large language models, with lessons from previous experiences (not values). To be clear, RLXF is not a new algorithmic innovation per se, but rather a procedure to customize and leverage existing reinforcement learning techniques like RLHF in a more structure way applicable to policy purposes. RLXF involves two key steps: First is training the reward model, which can be done through either a) Supervised learning on labeled experiments data to learn the rewards and b) RL inference where rewards are initialized based on metrics observed in the experiments then optimized through iterative RL to maximize alignment with outcomes exhibited in the demonstrations. The Second step is using the trained reward model to provide recursive signals[4] for RL fine-tuning of the target LLM without further human involvement. The following sections will lay the groundwork and provide further details on reinforcement learning from experience feedback, along with a case study demonstrating application on the economic data.

The gaps in Alignment with Human and AI feedbacks

There are several key gaps in current reinforcement learning techniques for AI alignment, including but not limited to:

1. Lack of ability to leverage prior experience and lessons learned to automatically create reward signals. Current techniques focus alignment on encoding values and behaviors through human reward signals or reference texts. But there is less focus on ensuring models align with and incorporate lessons from past real-world experiences as well as patterns or causal relationships from human history. This is a missed opportunity to take advantage of reinforcement learning in an autonomous fashion. This also represents a missed opportunity to build AI that learns longitudinally across time rather than just horizontally across datasets.

2. In RLHF, reliance on extensive human effort and evaluation to provide reward signals poses challenges for scaling up to large models and datasets. More efficient techniques are needed for human oversight.

3. In RLAIF, the fine-tuned critic model used to generate training data and rewards may incorporate or amplify its own biases and alignment issues. Relying on this critic to provide the reward signals means any flaws get propagated and reinforced during the reinforcement learning process.

Using RL in tuning LLMs and developing aligned AI systems using RL remains an open challenge. While some methods try to address some gaps like scaling human role, they often introduce new gaps in the process. This is an active area of research[5].

[4] In Recursive Reward Learning/Modeling (RRL/M), the model itself learns to predict appropriate reward signals using a learned reward model, without explicit human involvement. For example, the model could be trained to give high rewards for text responses that are helpful, harmless, and honest. The reward model is trained on limited demonstrations of good behavior. The LLM then interacts with an environment, predicts rewards for its actions using the learned model, and updates itself to maximize the predicted reward. This allows the model to continue aligning itself recursively. RLHF, RLAIF and RLXF could be potentially considered subsets of recursive reward learning if they employ pre-trained reward models recursively without continuous real-time human involvement.

[5] In addition to RLHF and RLAIF, other approaches like Hindsight Instruction Relabeling (HIR) (Zhang et al., 2023) and Reinforced Self-Training (ReST) (Gulcehre, Caglar, et al. 2023) demonstrate promise for improving language models' ability to follow instructions and improve alignment with human values or instructions. HIR employs a two-step supervised process of first sampling prompts and instructions, then relabeling the instructions based on alignment scores to turn failures into useful training data. ReST emphasizes creating enhanced

The need for Alignment with Human Experience

Ensuring AI alignment with human values is a widely discussed goal. However, equally important is aligning AI systems with human experiences - the lessons and patterns from our collective history without repeating mistakes in the output of LLMs. Supplementing value alignment with experience alignment could better minimize harms and maximize benefits.

In other words, while alignment with human values provides guidance for AI behavior at an individual output level, helping curtail inappropriate or dangerous responses, alignment with human experiences gives direction at a broader policy and strategy level. By grounding AI in historical patterns and causal relationships, we could enable better assessment of whether potential plans in the LLM outputs will ultimately help or harm in the long run.

Connecting AI with the wisdom accrued across humanity's journey is also crucial for assessing long-term or even second-order effects. An approach may seem beneficial based on stated values, but disastrous when accounting for lessons from the past. Training AI via both human preferences and experiences makes it doubly grounded in what makes us human. Without grounding in historical lessons, AI systems risk what we call “experience-hallucination” — generating compelling outputs based or partly based on specific parts of the collection of texts, but unrealistic narratives unmoored from the causal forces that shape our world. Incorporating mechanisms for experience alignment could yield AI with increased helpfulness, honesty, and foresight. RLXF is an effort to mitigate this risk.

Methodology

As mentioned in the previous section, RLXF is an procedure in the application of RL to LLMs proposed to align the behaviors of large language models using lessons learned from prior experiments, without directly imposing human values.

The first key step in RLXF is training the reward model that will provide feedback to the language model. This can be done in two ways. The first is through supervised learning on labeled data from previous experiments, training a model to directly predict good rewards from metrics and outcomes exhibited in the demonstrations. The second approach is RL inference, where rewards are initialized based on observed metrics, then iteratively optimized through reinforcement learning to match the outcomes from the experiment data.
After training, the reward model is able to evaluate the behaviors of language models and provide recursive feedback. This brings us to the second main step of RLXF - using the trained reward model to provide rewards and punishments to guide reinforcement learning fine-tuning of the target large language model, without further human involvement.

Through this recursive training process, the language model is optimized to take actions that yield high rewards from the model, aligning it with beneficial behaviors exhibited in past experiences. The end result is a fine-tuned language model that has learned alignments purely from historical experiment data, reducing the need for explicit human oversight during the alignment process.

Some key implementation details of RLXF include using neural networks or other machine learning models tailored for reward prediction, leveraging scalable RL frameworks like Ray RLlib for efficient distributed training, and iterating on fine-tuning the target model with the reward model until convergence.

Here is the key building blocks of an RLXF process in a simple implementation:

Data from Previous Experiments:
- Let $D = \{\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_\mathrm{n}\}$ be the dataset containing observations X_i from past experiments.

- Each X_i contains metrics like success or accuracy as well as final outcomes,

datasets by iteratively training the model on higher-quality subsets in order to refine the reward function, claiming efficiency advantages over standard reinforcement learning.

- D provides the signal for learning to align behaviors

Reward Model:

- Represented as a function $R(s, a)$ that takes in states s and actions a and outputs a reward value,

- Trained to optimize the loss function $L(R|D)$ over dataset D,

- Common objectives are minimizing mean-squared error or cross-entropy loss

Reinforcement Learning Algorithm:

- Methods like PPO optimize the objective: $J(\theta) = E_{s,a\sim\rho}[L(\theta)]$ where $L(\theta)$ is the clipped surrogate objective and ρ is the distribution under the current policy,
- This enables gradient ascent on $J(\theta)$ to optimize the policy parameters θ,
- Training alternates between sampling data through current policy πθ_k then optimizing $L(\theta_k)$ on the batch to obtain θ_{k+1}

Alignment Evaluation:

Measuring alignment using the expected reward:

$$\text{Ave}(R) = \Sigma_{s,a}\rho(s, a)R(s, a)$$

- where ρ is the state-action distribution under the policy,
- Comparing expected alignment under a random policy Ave(R_{random}) using statistical tests (or evaluate specific metrics like accuracy on held-out data)

In summary, the key building blocks of RLXF are: 1) Data from past experiments 2) Learned reward model 3) Reinforcement learning algorithm and 4) Alignment evaluation. Together these enable fully automated alignment based on previous experiences.

Impacts, Advantages, and Considerations:

Reinforcement learning techniques can significantly impact the behavior and outputs of large language models in the pursuit of AI alignment objectives. These techniques fine-tune the LLM's parameters, including layers, weights, and token embeddings, based on reward signals that encourage desirable behaviors.

At a technical level, RL fine-tuning involves updating the model's parameters (*θ*) to maximize the expected reward (*E[R]*) over the distribution of generated sequences. The reward *R* is provided by a learned reward model, which is trained on human feedback, AI-generated assessments, or historical experience data. The parameter updates are typically performed using policy gradient methods like Proximal Policy Optimization (PPO), which estimate the gradient of the expected reward with respect to the model parameters using the following expression:

$$\nabla_\theta E[R] \approx \frac{1}{N}\sum_{i=1}^{N} \nabla_\theta \log\ \pi(\ a_i|s_i).R(s_i, a_i)$$

Here, *π* is the LLM's policy (i.e., the probability distribution over token sequences), *s* represents the input context, *a* is the generated output, and *N* is the number of sampled sequences. By iteratively updating *θ* in the direction of this estimated gradient, the LLM learns to generate sequences that are more likely to receive high rewards according to the reward model.

The specific impacts and advantages/disadvantages of different RL approaches can be summarized as follows:

- *RLHF* directly encodes human preferences by training the reward model on human-labeled data. This promotes helpfulness, honesty, and adherence to human values in the LLM's outputs. However, it requires extensive human feedback, which limits scalability and may introduce inconsistencies if the labelers have differing preferences.

- *RLAIF* automates the feedback process by training an AI model to generate reward signals. This enables more efficient and scalable fine-tuning compared to RLHF. However, there is a risk of misalignment if the AI reward model is itself flawed or biased, which could lead to unintended behaviors in the fine-tuned LLM.

- *RLXF* grounds the LLM in empirical data by using historical experiences to train the reward model. This promotes outputs that are consistent with real-world dynamics and domain knowledge. However, the effectiveness of RLXF is limited by the quality and scope of available datasets, and there is a potential for the model to overgeneralize from past experiences.

The table 1 summarizes some key alignment considerations for these RL techniques.

Table 1. Alignment Advantages and Disadvantages of Reinforcement Learning Methods for Language Models

RL Method	Alignment Advantages	Alignment Disadvantages
RLHF	Direct encoding of human preferences; Promotes helpfulness and honesty	Limited scalability; Potential for inconsistent feedback
RLAIF	Scalable automated feedback generation; Efficient fine-tuning	Misalignment risk from flawed AI feedback; Potential for reward hacking
RLXF	Grounding in real-world data; Historically-informed outputs; Scalable	Limited by dataset quality and coverage; Potential to overgeneralize from past

Overall, RL techniques offer promising tools for imbuing LLMs with beneficial objectives and promoting AI alignment. By altering the model parameters based on carefully designed reward structures, as shown in the gradient estimation formula, RL can steer LLMs towards greater robustness, interpretability, controllability, and ethicality. However, the effectiveness of these techniques depends on the quality of the reward model and the training data, as well as the robustness of the RL algorithm itself. A combination of human oversight, well-crafted automated feedback, and informative datasets may yield the best outcomes for reliable and aligned language models. Continued research into optimizing RL for AI alignment, including techniques for reward modeling, policy optimization, and safe exploration, remains an important direction for the development of trustworthy and beneficial AI systems.

As mentioned before, RLXF provides a practical approach to align language models by learning from previous experiences, using automated reinforcement learning guided by learned reward models. This reduces the need for direct human involvement during alignment. A simple case study of RLXF with economic policy data will be explained in the following section.

Case Study

Setup

The case study project aimed to use RLXF for economic policy recommendations. The idea is not repeating the same or similar mistakes in policy issues over and over and rewarding successful experiences during the policy formulation and advice process.

It involved a two-step approach:

1. Training a BERT classifier model on historical data to predict economic policy outcomes. The model was trained on the IMF's MONA database to categorize the Fund’s programs criteria as "MET" or "Non-MET". It also produced probability outputs for each category. Achieving decent accuracy was important, since the classifier's probabilities of its categorization served as the reward model and foundation for the next phase.

2. Fine-tuning a generative language model (LLaMA 2-7B) using reinforcement learning and the BERT classifier outputs as reward signals. This allowed the model to generate sensible policy recommendations aligned with historical successes and failures.

A Proximal Policy Optimization algorithm was used to fine-tune the model, using the BERT classifier's probability prediction responses as rewards. The goal was to produce recommendations reflecting real-world policy complexity, while remaining grounded in the historical data and previous experiences.

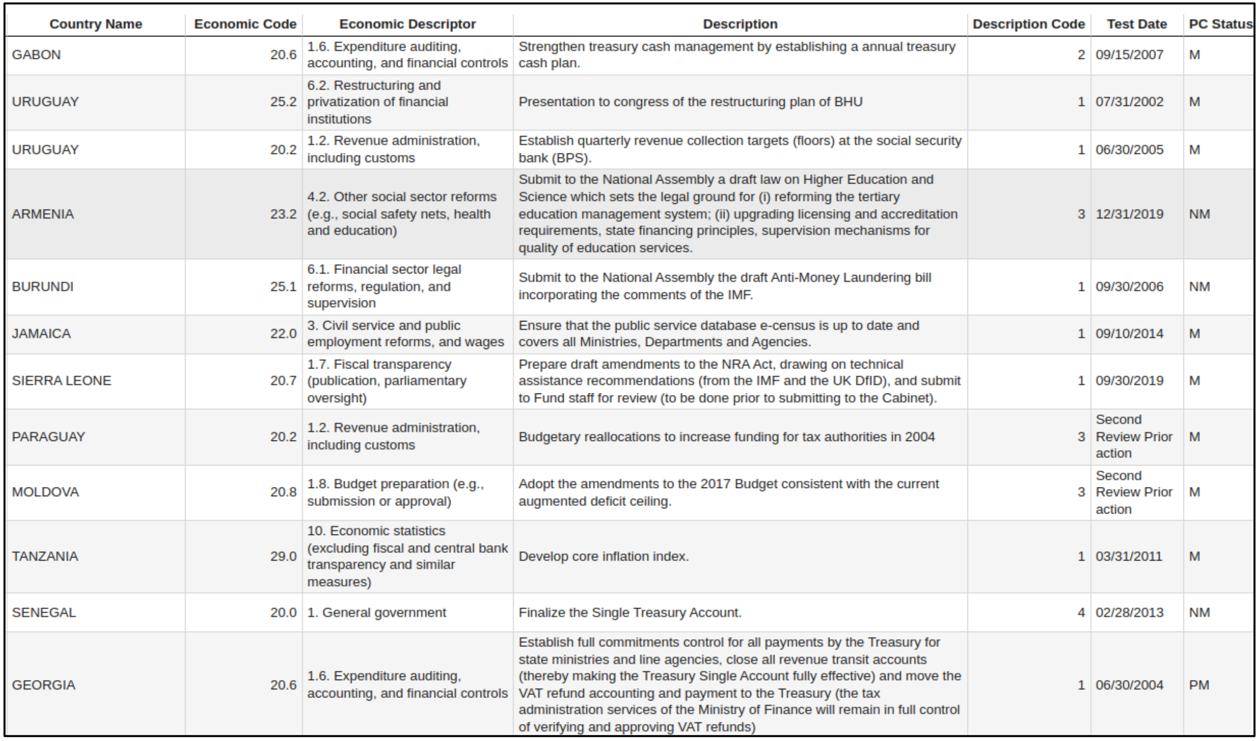
This two-step approach aimed to utilize both the classification and generative capabilities of AI to provide reasonable and historically-informed perspectives on economic policy. The BERT classifier provided a bridge between historical data and model outputs, while reinforcement learning aligned the generative model with desired outcomes.

Data

The Monitoring of Fund Arrangements (MONA) database compiled by the IMF contains a comprehensive record of economic policies and outcomes related to IMF-supported arrangements from 2002 up to the present day. This empirical data represents the real-world experiences of countries working to overcome economic challenges and achieve their goals. The value of MONA lies in documenting the past. By examining the successes and failures logged in this database, we can have some clues about the effective and not-so effective policies. This knowledge could help in avoiding repeating mistakes and build on fruitful strategies. With its expansive coverage of country performances, reviews, and macroeconomic indicators, MONA offers abundant insights into economic policy intricacies over recent decades.

The dataset itself is extensive and multifaceted. It includes basic details like country names and arrangement types, as well as in-depth information on program objectives, reform tactics, and performance criteria. This substantial dataset enables thorough analysis, making MONA a superb resource for projects utilizing AI to discern economic policy insights based on historical patterns. A random sample of the database could be seen in table 2.

Table 2. Random Sample Of MONA



Country Name	Economic Code	Economic Descriptor	Description	Description Code	Test Date	PC Status
GABON	20.6	1.6. Expenditure auditing, accounting, and financial controls	Strengthen treasury cash management by establishing a annual treasury cash plan.	2	09/15/2007	M
URUGUAY	25.2	6.2. Restructuring and privatization of financial institutions	Presentation to congress of the restructuring plan of BHU	1	07/31/2002	M
URUGUAY	20.2	1.2. Revenue administration, including customs	Establish quarterly revenue collection targets (floors) at the social security bank (BPS).	1	06/30/2005	M
ARMENIA	23.2	4.2. Other social sector reforms (e.g., social safety nets, health and education)	Submit to the National Assembly a draft law on Higher Education and Science which sets the legal ground for (i) reforming the tertiary education management system; (ii) upgrading licensing and accreditation requirements, state financing principles, supervision mechanisms for quality of education services.	3	12/31/2019	NM
BURUNDI	25.1	6.1. Financial sector legal reforms, regulation, and supervision	Submit to the National Assembly the draft Anti-Money Laundering bill incorporating the comments of the IMF.	1	09/30/2006	NM
JAMAICA	22.0	3. Civil service and public employment reforms, and wages	Ensure that the public service database e-census is up to date and covers all Ministries, Departments and Agencies.	1	09/10/2014	M
SIERRA LEONE	20.7	1.7. Fiscal transparency (publication, parliamentary oversight)	Prepare draft amendments to the NRA Act, drawing on technical assistance recommendations (from the IMF and the UK DfID), and submit to Fund staff for review (to be done prior to submitting to the Cabinet).	1	09/30/2019	M
PARAGUAY	20.2	1.2. Revenue administration, including customs	Budgetary reallocations to increase funding for tax authorities in 2004	3	Second Review Prior action	M
MOLDOVA	20.8	1.8. Budget preparation (e.g., submission or approval)	Adopt the amendments to the 2017 Budget consistent with the current augmented deficit ceiling.	3	Second Review Prior action	M
TANZANIA	29.0	10. Economic statistics (excluding fiscal and central bank transparency and similar measures)	Develop core inflation index.	1	03/31/2011	M
SENEGAL	20.0	1. General government	Finalize the Single Treasury Account.	4	02/28/2013	NM
GEORGIA	20.6	1.6. Expenditure auditing, accounting, and financial controls	Establish full commitments control for all payments by the Treasury for state ministries and line agencies, close all revenue transit accounts (thereby making the Treasury Single Account fully effective) and move the VAT refund accounting and payment to the Treasury (the tax administration services of the Ministry of Finance will remain in full control of verifying and approving VAT refunds)	1	06/30/2004	PM

Model

The modeling approach was structured and comprised a systematic two-tiered strategy. Here's a detailed breakdown:

1. Reward Signals:

Objective: The primary goal of this model was to predict IMF program performance focusing on distinguishing between various Program Criteria (PC) status documented in the MONA dataset.

Training Labels: For clarity in our training, we categorized the outcomes into broader labels. The labels "MET" and "Partially MET (PM)" were combined under the "MET" category. Conversely, outcomes like "Not MET (NM)", "Waived", "Delayed (DL)", and others were grouped under the "Non-MET" label.

- **Training Sample: (From MONA Dataset)**
    - Country: [Country Name]
    - Type of Arrangement: [Arrangement Type, e.g., Stand-by, PRGF, EFF]
    - Economic Condition Classification: [Economic Classification, e.g., Fiscal, Monetary]
    - Condition Description: [Detailed Description of Condition]
    - Status: MET/Non-MET/Partially MET/Delayed/etc.

- **Output Sample: (Classifier Prediction)**
    - Status: MET or Non-MET
    - Probability of MET: 0.XX
    - Probability of Non-MET: 0.YY

2. Fine-tuning:

Base Model: We employed the LLaMA 2-7B model, a medium-size open source LLM, known for its text generation prowess.

Fine-tuning Strategy: We initiated the fine-tuning of LLaMA using a reward model derived from the BERT classifier. This reward model, upon receiving outputs from LLaMA, would generate a probabilistic signal, indicating how aligned the LLM's output was with previous experiences, specifically in terms of being "MET" or "Non-MET".

Prompting the LLM: To extract meaningful policy recommendations, we fed the LLM with specific prompts. These prompts were constructed using a template that incorporated country names and appropriate policy proposals for each country.

- Input Prompt Sample (Fed to LLM):
    - "Given the economic indicators and historical performance of [Country Name], what would be an appropriate policy proposal under the [Type of Arrangement] for achieving [Economic Condition Classification, e.g., Fiscal, Monetary]? Describe the proposed condition."

Reward Signal Generation: Post prompting, the outputs from the LLM were fed into the reward model. This model then produced a reward signal, which assessed the LLM's output in terms of its alignment with historical data.

- Reward Model Assessment:
    - Probability of MET: 0.XX

Training with PPO: Using the reward signals, we employed the Proximal Policy Optimization (PPO) algorithm to further fine-tune the LLM. This step ensured that the model's suggestions were not just coherent but also informed by previous experiences.

The above process was carried out iteratively. The language model was presented with an array of policy-related prompts pertaining to different countries and economic situations. Each response from the model was then evaluated by the probabilistic reward model. Using this feedback, the language model was fine-tuned through the PPO algorithm. This recurring procedure aimed to ensure the model's recommendations were contextually apt and aligned with the historical insights and experiences chronicled in the MONA dataset. Specifically, the goal was to emphasize policies that had proven successful based on the historical data. Through this iterative approach, the model was steered toward proposals reflecting real-world nuances and informed by documented successes and failures.

This strategy was designed to ensure that the generated policy recommendations were not only coherent but also took into account historical successes and failures, thereby providing insights that were rooted in real-world experiences.

Results

Our trained BERT classifier showed its ability to effectively distinguish between "MET" and "Non-MET" policy outcomes. When evaluated on 20% unseen data, the model attained strong results on key performance measures including accuracy, precision, recall and F1 score, surpassing 82% on each. After fine-tuning, the LLaMA 2-7B model exhibited noticeable enhancement in generating policy suggestions. When provided with prompts incorporating country names and potential reforms, the language model produced recommendations better grounded in historical insights, showcasing alignment with documented past experiences and lessons. Our assessment revealed the fine-tuned language model's recommendations demonstrated greater congruence with the historical data from the MONA dataset. This was apparent through the model proposing solutions that mirrored previously successful policies, while avoiding suggestions historically tagged as "Non-MET" outcomes. The evaluation indicated the iterative training process steered the model toward proposals reflecting the nuances and insights gleaned from the empirical evidence.

Table 3. **Pre/Post Fine-tuning an LLM with Domain Specific information and RL model.**

Scenario/Context	Prompt	Pre-RL Fine-tuning	Post-RL Fine-tuning
Fiscal Reform	An appropriate economic reform policy by the IMF for [] in the Fiscal Policy area should	Introduce progressive tax reforms, enhance public expenditure management, and promote financial transparency.	focus on fiscal consolidation, strengthen public resource management, and emphasize concessional financing in an ECF arrangement.
Trade & Investment	The ideal trade and investment strategy by the IMF for [] should	Increase exports and attract foreign investment.	Create conditions that encourage trade, simplify investment processes, and implement a PCI to direct the execution of the strategy.
Financial Sector Reform	A strategic financial sector reform by the IMF for [] should	Improve banking regulations and promote transparency.	Strengthen regulatory oversight, enhance the resilience of the banking sector, and engage in an EFF arrangement to support comprehensive reforms.

Potential and Pitfalls: Applying RLXF to Policy Analysis

The application of RLXF to the IMF's MONA database shows promise for supporting the Fund's policy work. By analyzing historical data, this tool could potentially help country teams benchmark new policy proposals against past experiences and peer examples. It may provide additional context for considering the appropriateness of policies given a country's macroeconomic conditions and shocks.

The RLXF tool might also add some value to internal IMF reviews and external policy consultations. By offering a historical perspective on policy options, it could facilitate more data-informed discussions. The ability to draw upon a range of past examples may enrich the policy dialogue to an extent.

However, the limitations of relying on historical information should be carefully considered. The MONA database may not always reflect current conditions or capture all structural changes in countries over time. The tool's recommendations should be interpreted cautiously and not viewed as a replacement for thorough analysis of the present context. While RLXF insights could be valuable, they should supplement rather than substitute the expertise of IMF staff.

Moreover, it's important to be clear that an LLM enhanced with RLXF is not capable of replacing the complex empirical research needed to fully address the econometric and statistical challenges in policy analysis. While the tool might surface helpful patterns based on historical data, it is not a substitute for rigorous economic research methodologies. The LLM outputs should be considered suggestive and be subject to further validation through standard analytical approaches. More research would be needed to assess the full potential and limitations of this approach for IMF work.

Another key limitation to consider is that LLMs, by their very nature, have a bias towards what has happened in the past. As next word predictors trained on historical data, they inherently resist deviating from established patterns. In cases where new structural changes demand innovative policy responses, LLMs - even without the additional historical grounding of RLXF - may struggle to generate sufficiently novel suggestions. They may tend to stick to familiar policy territory rather than venturing into uncharted but potentially necessary new policy directions.

Furthermore, the reliance of RLXF on historical data could potentially amplify the LLM's inherent backward-looking bias. By explicitly optimizing the model to align with past policy successes, RLXF might inadvertently make the model even more resistant to proposing policy innovations. This could be particularly problematic in rapidly evolving economic contexts that require agile, adaptive policymaking. While the historical grounding provided by RLXF can offer valuable stability, it may also hinder the model's ability to respond dynamically to emerging challenges. Careful monitoring and adjustment of the model's learning process may be necessary to strike the right balance between leveraging historical lessons and remaining open to new approaches.

Conclusion

As AI rapidly advances, ensuring it aligns with human values and experiences is crucial. Core to this is creating systems that not only understand our values, but also internalize our collective historical experiences. This paper introduced Reinforcement Learning from Previous Experiences (RLXF) - a procedure aiming to do just that.

RLXF is based on the idea that historical data, representing shared experiences, offers valuable lessons. By integrating these lessons into AI's decision process, we may avoid repeating past mistakes and ensure its suggestions are truly beneficial. The proposed methodology leverages reward model training and recursive reinforcement learning fine-tuning to automate this alignment.

However, it's crucial to acknowledge the potential pitfalls of this approach. By relying heavily on historical data, RLXF may inadvertently perpetuate biases and outdated assumptions embedded in past experiences. Economic contexts evolve, and what worked in the past may not always be suitable for the present. Overemphasis on historical alignments could hinder the AI's ability to generate innovative solutions for novel challenges. Moreover, the backward-looking nature of RLXF might make the AI resistant to necessary policy shifts in response to changing circumstances. Balancing the wisdom of the past with the need for adaptability is a delicate task that requires careful consideration in the design and application of RLXF systems.

Our case study demonstrated RLXF in action, showing how a language model fine-tuned on insights from the IMF’s MONA database provided historically informed economic policy suggestions. This iterative training using experience-based rewards steered the model toward nuanced, real-world-aware proposals.

As shown in our case study, RLXF is particularly promising for economic policy applications. Historical economic data provides abundant empirical evidence revealing intricate real-world policy dynamics. RLXF offers a framework to instill an AI with the nuances and causal relationships from analyzing these experiences. The result is an AI that provides perceptive, historically enlightened economic policy perspectives.

In summary, RLXF enables creating assistants that enhance economic policymaking through experience-aligned recommendations. The ability to leverage AI grounded in previous experiences could provide invaluable context to inform complex policy decisions. Rather than starting from scratch, policymakers could utilize lessons and relationships uncovered by a system trained with RLXF. This could bring greater prudence and foresight when navigating economic challenges.

References

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, *30*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gao, X., & Others. (2024). A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435.*

Gudivada, V. (2024). A review of current trends, techniques, and challenges in large language models. *Applied Sciences,* 14(5), 2074.

Gulcehre, C., Paine, T. L., Srinivasan, S., Konyushkova, K., Weerts, L., Sharma, A., ... & de Freitas, N. (2023). Reinforced Self-Training (ReST) for Language Modeling. *arXiv preprint arXiv:2308.08998*.

Huang, X., & Others. (2023). Understanding LLMs: A comprehensive overview from training to inference. *arXiv preprint arXiv:2401.02038.*

Lee, H., Phatale, S., Mansoor, H., Lu, K., Mesnard, T., Bishop, C., ... & Rastogi, A. (2023). RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. *arXiv preprint arXiv:2309.00267*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Zhang, T., Liu, F., Wong, J., Abbeel, P., & Gonzalez, J. E. (2023). The Wisdom of Hindsight Makes Language Models Better Instruction Followers. *arXiv preprint arXiv:2302.05206*.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., ... & Irving, G. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

PUBLICATIONS

Reinforcement Learning from Experience Feedback
Working Paper No. WP/2024/114