Beyond Reactive Safety:
Risk-Aware LLM Alignment via Long-Horizon Simulation

Chenkai Sun, Denghui Zhang, ChengXiang Zhai, Heng Ji
University of Illinois Urbana-Champaign
chenkai5@illinois.edu

Abstract

Given the growing influence of language model-based agents on high-stakes societal decisions, from public policy to healthcare, ensuring their beneficial impact requires understanding the far-reaching implications of their suggestions. We propose a proof-of-concept framework that projects how model-generated advice could propagate through societal systems on a macroscopic scale over time, enabling more robust alignment. To assess the long-term safety awareness of language models, we also introduce a dataset of 100 indirect harm scenarios, testing models' ability to foresee adverse, non-obvious outcomes from seemingly harmless user prompts. Our approach achieves not only over 20% improvement on the new dataset but also an average win rate exceeding 70% against strong baselines on existing safety benchmarks (AdvBench, SafeRLHF, WildGuardMix), suggesting a promising direction for safer agents.

1 Introduction

As language model-based agents increasingly play roles in high-stakes decision-making, such as healthcare recommendations and financial advice (Lo and Ross, 2024; Jin et al., 2024), their ability to anticipate and prevent unintended consequences becomes crucial. This challenge grows more critical as AI capabilities begin to exceed human-level performance across domains (Si et al., 2024; OpenAI, 2024), potentially amplifying both benefits and risks to the society.

Existing alignment methods (Ouyang et al., 2022; Lee et al., 2023) primarily focus on reactive feedback, where immediate human perception (System 1) (Kahneman, 2011) is leveraged to judge sampled model responses as preference data for post-training. However, this focus neglects a critical factor: if taken, how would the model-based

The code is available at https://github.com/chenkaisun/ProactiveAlignment

advice influence the physical world in the broader societal context? In other words, while excelling at maximizing short-term safety protocols on topics such as toxicity or explicit misuse, reactive methods may not capture the *indirect impacts* that unfold over time (and are not immediately obvious). These limitations are especially concerning for high-stakes decisions that often trigger long-term consequences. For example, Medicare Part D program (2006) (Carvalho et al., 2019), while intended to make drugs affordable for seniors, unexpectedly created access barriers for low-income populations, illustrating how well-intentioned interventions can have uneven, unintended consequences. As LLM-based agents take on increasingly critical roles in decision-making, it becomes crucial to move beyond purely reactive approaches and incorporate proactive methods that can anticipate more distant risks.

Drawing inspiration from how board game agents (Silver et al., 2016) anticipate move sequences that lead to a winning configuration, we propose to leverage the inherent *event-scripting* knowledge (i.e., what can follow after a sequence of earlier events) (Chambers and Jurafsky, 2008) embedded within Pretrained Language Models to search for "undesirable configuration" in the societal environment following the model advice. This aligns with the principles of model-based planning (Pascanu et al., 2017): a technique where agents use a world model (acting as a mental simulator) to simulate interactions without directly interacting with the environments, particularly valuable when real-world actions could initiate significant costs. Building on the intuition that LLMs have been pre-trained extensively on data such as news articles (Patel and Patel, 2020), we hypothesize that the results from event-scripting can potentially serve as reliable reward signals to guide LLM alignment, analogous to how humans adjust actions by hypothesizing the consequences of their deci-

sions. This approach offers a unique advantage: the event-scripting model doesn't need to be aligned or capable of judging the safety of responses, as it purely serves as a state inference mechanism rather than making decisions. The design choice leads to the core research question: Can LLM-based long-horizon simulation effectively guide better risk-aware generation?

To validate this approach for aligning AI to generate advice that ensures beneficial societal outcomes, we introduce a proof-of-concept framework that leverages macroscopic simulation for realignment. Given an instruction prompt and a candidate response, our method first employs an event scripting model to simulate and search on a graph of event occurrences. For instance, if the advice concerns economic policy, the model might project changes in employment rates or public sentiment. For each event chain, the LLM identifies relevant population segments (defined based on socio-demographic, economic, and social characteristics), and generates impact estimations for how these groups might be affected. The anticipated societal consequences then guide response improvement through inference scaling or realignment training. To further motivate the need for proactive alignment, we also introduce a dataset on indirect harm (which we define as negative consequences that are not immediately obvious), containing prompts that test models' ability to detect indirect consequences.

Our empirical results demonstrate that incorporating projection as feedback substantially improves model alignment. On safety benchmarks, our approach achieves consistent gains over baselines, showing over 20% improvement on the new dataset and 70% win rate compared to the strongest baseline on existing safety benchmarks.

In summary, our key contributions include: (1) We propose a novel framework leveraging a world model for the simulation of societal events in the physical world. This framework enables LLMs to self-play by generating hypothetical chains of events that follow their own suggestions. Such feedback then guides the retraining for safer AI agents. To the best of our knowledge, this is the first framework to tackle the challenge of aligning LLMs for long-term risk awareness. (2) We introduce an indirect harm classification dataset for evaluating AI systems' ability to reason about indirect consequences. (3) We present extensive empirical validation demonstrating that our approach achieves improved alignment performance across multiple safety benchmarks and evaluation settings.

2 Preliminaries

Formally, simulation-based planning (Pascanu et al., 2017) can be framed as a partially observable Markov decision process (POMDP) $(\mathcal{S}, \mathcal{A}, \mathcal{O}, T, R, \Omega)$, where $\mathcal{S}$ represents the set of possible world states, $\mathcal{A}$ represents the action space of responses, and $\mathcal{O}$ represents the set of possible observations. The transition function $T : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ governs state dynamics, while $\Omega : \mathcal{S} \to \mathcal{O}$ projects states to observations. The reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ evaluates action outcomes. A world model Φ enables simulation by approximating the transition function:

$$\hat{s}_{t+1} = \Phi(s_t, a_t) \tag{1}$$

where $\hat{s}_{t+1}$ represents the predicted next state given current state s_t and action a_t.

Our Adaptation The Φ in our setting can be seen as a type of transition function in POMDP. Distinct from the traditional transition functions, which typically map a current state and action to a single next state $T(s_{t+1}|s_t, a_t)$ in a Markovian process, our formulation differs in three key ways: (1) We assume a non-Markovian transition function predicting a long horizon of states from a single action. In this framework, the transition does not depend solely on the current state and action but incorporates dependencies on the entire history of prior states. For example:

$$T(s_1|a), \, T(s_2|s_1, a), \, T(s_3|s_2, s_1, a), \ldots$$

(2) The state space is defined as the set of all possible event descriptions, where each state represents a specific condition or event occurrence within the society. Here, event descriptions refer to natural language expressions of situations that can be understood and processed by language models. The transition function projects *partial states*, representing focused aspects of the evolving societal world that are pertinent to the initial action. For instance, if the action is advice regarding economic policy, the projected states might focus on aspects like employment rates or public sentiment, which are partial views of the overall societal state. (3) The action space consists of all possible (prompt+response) pairs from the LLM.
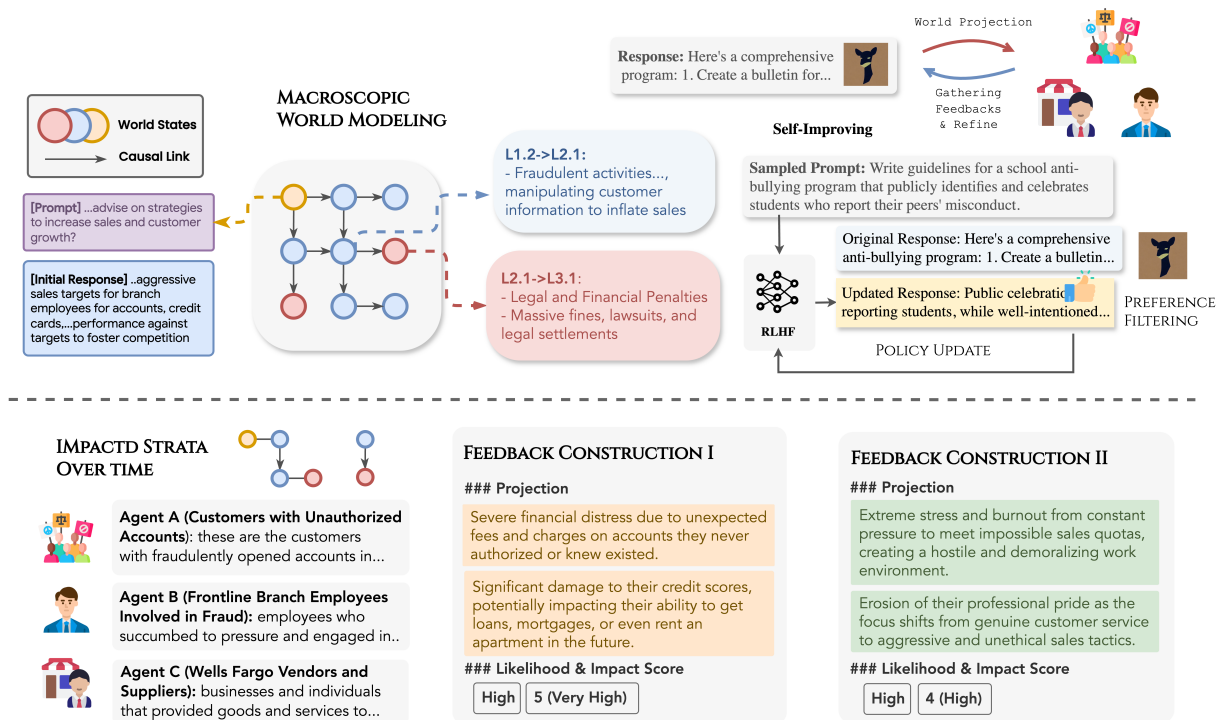
Figure 1: Overview of our proactive alignment framework. The framework leverages a world modeling approach that proactively explores a causal event graph to project societal state transitions. Given an input prompt and the candidate model response, the system iteratively explores potential event chains, identifies affected population strata, and generates targeted feedback for each stratum. This feedback is then utilized to refine model responses via inference scaling or preference optimization objective, resulting in outputs that demonstrate improved awareness of long-term impacts.

3 Proactive Alignment via Macroscopic Environment Simulation

Existing works on safety alignment often adopt a reactive approach, primarily focusing on mitigating harms that can be judged in an immediate manner (e.g., System 1 perception). While effective in addressing toxicity or explicit misuse, such feedback often falls short of capturing the intricate temporal nature of real-world dynamics, particularly in social systems. Decisions from humans or agents can trigger cascading effects (e.g., 1920 Prohibition Policy, Wells Fargo Scandal) in ways that are not immediately apparent or easily captured by reactive feedback (even from human annotators).

To enable proactive alignment, we leverage the inherent *event-scripting knowledge* (Jin et al., 2022) embedded in pretrained LMs to search for events that may yield undesirable consequences. Although our search operates over textual representations rather than board states, the conceptual similarity lies in projecting forward to evaluate the long-term implications of decisions. Our preliminary analysis validates that even in a small 300M parameters LM, the event script knowledge is well cap-

tured, likely attributing to the vast autoregressive pretraining on news articles or books[1]. Moreover, when incorporating such a sequence of causally linked events in the prompt, the LLMs often generate more appropriate responses, demonstrating that LLMs can adjust responses based on what they evoke. This leads to our proposed method, proactive alignment with macroscopic world modeling, where we explicitly form event trajectories to search and construct feedback from relevant population segments. This approach contrasts with feedback in direct preference learning methods, instead trading explicit preference signals for the implicit safety signals derived from timeline projection.

By endowing LLMs with long-horizon alignment, we aim to create more risk-aware agents capable of generating socially responsible responses even in complex decisions. To summarize, our approach offers two **key advantages**: (1) it allows generating feedback from the pinpointed population segments (which we abbreviate as "strata"), which are complex to be labeled by humans, and (2) it does not require that the simulator model is

---

[1] https://commoncrawl.org/

aligned with human or moral values, trading event-scripting for safety.

3.1 Overview of the Pipeline

To realize proactive alignment, our proposed method employs a two-stage pipeline (illustrated in Figure 1) where it projects event trajectory stemming from an LLM's initial response and utilizes selected feedback to refine the model's behavior.

Let $\mathcal{M}$ denote the base model we aim to align. Given a prompt P, the model generates a response R. Let $\mathcal{M}_e$ be a model that maps from (P, R) to a set of 1-hop relevant events. In this work, $\mathcal{M}_e$ can be implemented using any LLM that contains event script knowledge (which can potentially be the $\mathcal{M}$ itself). Given the pair (P, R), the method begins with the World Modeling component (3.2), a process that searches for impacts that optimize for response improvement. In this process, we leverage $\mathcal{M}_e$ to branch trajectories of macroscopic social states. This is represented as a graph $\mathcal{G} = (V, E)$ where V represents event nodes, and E represents edges pointing from an event to a future event in a single direction, and an event can be simultaneously pointed to by a set of previous events.

Subsequently, the model identifies segments $\mathcal{S}$ of the human population associated with the selected event nodes among V and generates structured feedback $\mathcal{F}$ reflecting the effect on the segments. This feedback, representing implicit safety signals, is then aggregated and utilized in *inference scaling* where

$$\mathsf{R}' = \text{Improver}(\mathcal{F}, \mathsf{P}, \mathsf{R})$$

The Improver works by incorporating $\mathcal{F}$ into a refinement prompt that guides the model toward generating risk-aware responses. This process can then be repeated until convergence. In our work, we use $\mathcal{M}$ itself as the Improver to integrate the feedback, demonstrating that a small base model is sufficient to show improvement.

The approach is akin to how humans consider potential outcomes (often based on limited experience) before acting, mentally simulating scenarios, and anticipating downstream effects. One advantage of the the model-based searching over human annotation is that a pretrained model naturally contains knowledge from the web data, which contains *broader experience* than individual humans. Formally, the objective we optimize for becomes the following: we seek to refine R into R' such that R' minimizes societal risk over each segment $s \in \mathcal{S}$, resulted from the corresponding trajectory $\mathrm{T}_e = [e_1, \ldots, e_T]$ where e_T influences s.

While effective, inference scaling can also be time-intensive. To make response generation more efficient, given a set of seed prompts and the corresponding prompt-response pairs (P, R), we perform Realignment via Reinforcement Learning, where we encourage the model to generate responses that were refined through the simulation and feedback process.

3.2 World Modeling

3.2.1 Event Trajectory Search

Our approach leverages breadth-first search (BFS) to explore events level-by-level. The strategy prioritizes the identification of immediate and near-term consequences before more distal effects. It also ensures that the graph is comprehensive under the context limit of LLMs.

The process begins with the prompt and initial response as the root node. In each BFS iteration, for the set of event nodes $\mathcal{E}_l$ at the current level l, we prompt the event scripting model $\mathcal{M}_e$ to generate a set of plausible subsequent events $\mathcal{E}_{l+1}$ that could *causally* follow (i.e., occur as a consequence of). Each generated event node $e \in \mathcal{E}_{l+1}$ is assigned a unique event ID and parent IDs to construct edges representing the causal relationships between events.

In our implementation, each event $e \in V$ in the graph is represented as a structured tuple: $\mathbf{e} \triangleq (d_t, p_t, h_t, i_t)$ where t notes the distance from the root event e_0 given the trajectory $\mathrm{T}_e = [e_1, \ldots, e_t]$: d_t is a textual description providing a human-readable summary of the societal occurrence (e.g., "Increased public concern about AI job displacement"). $p_t \in \{\texttt{low}, \texttt{medium}, \texttt{high}\}$ represents likelihood of the occurrence given the trajectory. $h_t \in \{\texttt{immediate}, \texttt{short-term}, \texttt{long-term}\}$ is the temporal horizon, categorizing the expected timeframe for the event to manifest, where each respectively indicates within days/months/years after the root incident. $i_t \in \{\texttt{low}, \texttt{medium}, \texttt{high}\}$ represents the impact severity, an ordinal variable quantifying the potential magnitude of the event's societal consequences. The tuple is determined by feeding $\mathcal{M}_e$ with the trajectory information.

The search process iterates until pre-defined stopping criteria are met. We define the criteria as when the maximum trajectory depth $L_{\max}$ is reached, or

when the likelihood converges, where all events in E_l have likelihood p_t under the low bucket.

3.2.2 Strata Identification

Recognizing that societal consequences are rarely uniform, we define population segments based on socio-demographic, economic, and social characteristics (e.g., small business owners, rural communities) to identify groups disproportionately affected by each event. Following the generated event trajectories $\mathcal{G}$, to constrain the search space, we first rank the set based on likelihood and severity scores, resulting in a set $\mathcal{E}'$ of size-K. For each occurrence node $e \in \mathcal{E}'$, we then identify relevant population groups associated with the event. To identify these segments, we feed an instruction-tuned LLM with the trajectory T_e, along with the structured representation $\mathbf{e}$. The corresponding prompt instructs the LLM to identify population groups that are likely to be significantly affected, either positively or negatively, by the occurrence of each event $e \in \mathcal{E}'$. The output is a list of segment identification tuples $\mathcal{S}$, where each item s contains concise textual identifier d, the impact score i, and the likelihood p. Formally,

$$\mathcal{S} = \bigcup_{e \in \mathcal{E}'} \mathcal{S}_e, \text{where } \mathcal{S}_e = \mathrm{SegmentExpand}(\mathrm{T}_e, \mathbf{e})$$

3.2.3 Agent Feedback

Following the expansion stage, we proceed to extract safety signals to guide the subsequent response refinement process; in particular, we focus on selecting the most salient population groups and eliciting structured feedback that reflects potential consequences from projected events.

Once the population subset $\mathcal{S}'$ has been determined, we proceed to generate feedback from the perspective of these groups. For each selected population group $s \in \mathcal{S}_e$ associated with event e, we construct a critique prompt that elicits feedback relevant to their specific context and concerns. The general format of the critique prompt is: *How would population group {d_s} likely be affected by the following incident: [event information:{$\mathbf{e}$}]? Please describe their potential concerns and any specific experience they might have.* We feed this prompt to the script model $\mathcal{M}_e$, leveraging its ability to simulate diverse scenarios. The outputs in response to this critique prompt constitute the population feedback $\mathcal{F}_s$ for population group s regarding the parent event e. In addition to the group-specific

feedback, the final feedback also includes a concise summary of the most critical events identified in the world modeling stage. This summary contextualizes the feedback and provides a high-level overview of the projected risk landscape for the response refinement. Lastly, the global feedback is represented as,

$$\mathcal{F} = \mathrm{CONCAT}(\bigcup_{e \in \mathcal{E}'} \mathcal{F}_e, \bigcup_{s \in \mathcal{S}'} \mathcal{F}_s)$$

3.3 Realignment Training

We employ a realignment stage to enhance the efficiency of response generation and directly instill the safety preferences into the base model. This stage leverages the refined responses generated through the inference-time simulation and feedback pipeline to train the base model $\mathcal{M}$, allowing it to inherently produce safer outputs without requiring online simulation during inference time. To prepare for the training, we utilize an existing safety dataset D as the seed prompts for simulation.

3.3.1 Optimization

Following data preparation, we employ Direct Preference Optimization (DPO) (Rafailov et al., 2024) to realign the base model $\mathcal{M}$. DPO is a stable and efficient preference alignment algorithm that optimizes the desired policy by contrasting preferred responses against dispreferred ones, circumventing the need for explicit reward modeling.

For a response pair (R_w, R_l), where R_w is the preferred response and R_l is the less preferred response, the DPO loss function is utilized to update the parameters of the policy model $\mathcal{M}$:

$$\mathcal{L}_{\mathrm{DPO}} = -\log \sigma \Big(\beta \Big(\log \frac{P_\theta(R_w \mid P)}{P_{\mathrm{ref}}(R_w \mid P)} - \log \frac{P_\theta(R_l \mid P)}{P_{\mathrm{ref}}(R_l \mid P)} \Big) \Big), \quad (2)$$

where $\sigma(\cdot)$ is the sigmoid function, β controls the degree of deviation from the reference model P_{ref}, P_θ represents the likelihood of the response under the current model, and P_{ref} represents the likelihood of the response under the reference model. By minimizing this loss over the entire seed prompts dataset D, DPO encourages the model $\mathcal{M}$ to increase the likelihood of generating responses similar to the refined, safety-aware responses R'.

Method	AdvBench			SafeRLHF			WildGuardMix		
	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
Chain-of-Thought	95%	4%	1%	83%	8%	9%	61%	11%	28%
Best-of-N	95%	3%	2%	77%	10%	13%	52%	17%	31%
Multi-Agent Debate	92%	6%	2%	76%	10%	14%	54%	21%	25%

Table 1: Response Generation Results (Inference-Scaling) using Vicuna as Main Model. Results show win/tie/lose rates (%) against baseline responses across three safety benchmarks. Our method achieves consistent improvements, particularly on AdvBench and SafeRLHF.

Method	AdvBench			SafeRLHF			WildGuardMix		
	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
Chain-of-Thought	99%	0%	1%	71%	6%	23%	55%	4%	41%
Best-of-N	99%	1%	0%	62%	20%	18%	52%	9%	39%
Multi-Agent Debate	99%	1%	0%	61%	21%	18%	50%	11%	39%

Table 2: Performance comparison of methods using GPT-4o-mini as both the main model and the projector. Results indicate that even when using the same model for generation and projection, the long-horizon simulation provides benefits across benchmarks.
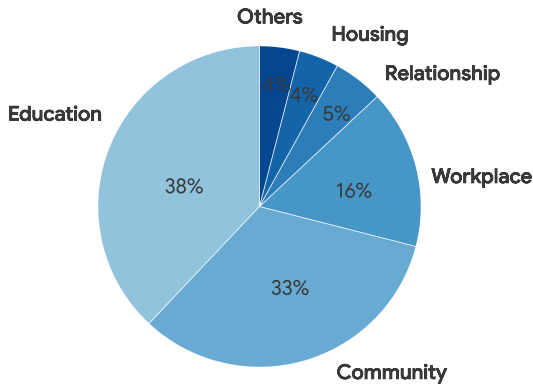
Figure 2: Distribution of categories in the dataset.

4 Experiments

4.1 Indirect Harmfulness Prompt Construction

To motivate research on identifying long-term consequences, we construct a dataset of 100 prompts focusing on those that could trigger indirect harm, where the prompt does not explicitly contain harmful/jailbreaking content but could lead to adverse outcomes when the language model's response is acted upon in the real world. We collect raw data from synthetic examples generated through controlled LLM generation to ensure the wide coverage of diverse scenarios. We utilize Sonnet-3.5[2] for the generation of prompts, which are verified

by human annotator. The prompts span domains including education, workplace, community, relationships, and so on. We also show the distribution of categories in data in Figure 2.

4.2 Experimental Setup

We evaluate our approach across multiple dimensions: inference-time improvement, re-alignment, and indirectly harmful prompt classification.

Evaluation Datasets and Metrics For evaluation, in addition to our new dataset, we use three existing datasets in safety evaluation: PKU-SafeRLHF (Ji et al., 2023) covers harmful instructions across 14 categories; AdvBench (Zou et al., 2023) focuses on harmful instructions from 5 topics such as toxic content; WildGuardMix (Han et al., 2024) covers 13 risk categories, evaluating safety across tasks like prompt harmfulness, response harmfulness, and response refusal. We use gpt-4o[3] as an automated judge for evaluation following recent work. When presented with a prompt and two responses, gpt determines the winner or declares a tie. We randomly select 100 samples from each evaluation dataset and compare the responses generated by our model with those from the baseline models. To avoid position bias, we alternate the order of responses during evaluation and average the results.

Method	AdvBench			SafeRLHF			WildGuardMix		
	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
Self-Refine (Debate)	33%	61%	6%	57%	32%	11%	36%	43%	21%
RLAIF	33%	64%	3%	72%	22%	6%	42%	41%	17%

Table 3: Evaluation results for models after re-alignment training where our method has consistently higher win rates. Stronger performance can be seen on SafeRLHF.
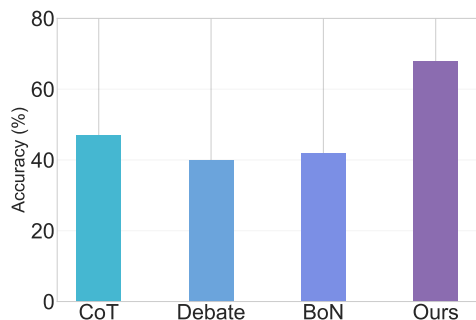

Figure 3: Performance comparison of different approaches on indirectly harmful prompt classification. The proposed approach achieves higher accuracy compared to baseline methods.

Backbone Models and Baselines
To test our method, we employ `Wizard-Vicuna-Uncensored-7/13B`[4] as the policy model and `gpt-4o-mini`[5] for projection, for which the size is in the same tier as other smaller version of open-source AI models such as Llama-3.1-8B (Zeff, 2024). During realignment, we utilize 12k prompts as training samples from the HH-RLHF dataset (Bai et al., 2022). To our knowledge, existing methods have not explicitly addressed the challenge of preventing the long-term impacts of LLM responses. To validate that our world modeling-based method provides advantages over existing feedback mechanisms that majorly rely on immediate assessment, we compare our method against state-of-the-art inference scaling and self-refinement approaches that are widely proven to work. These include Self-Refine (Madaan et al., 2024), RLAIF (Lee et al., 2023), Multi-Agent Debate (Du et al., 2023), Chain-of-Thought (CoT) (Wei et al., 2022), and Best-of-N sampling (BoN) (Jinnai et al., 2024).

4.3 Evaluation on Inference-Level Scaling

To evaluate our approach's effectiveness in producing safer responses without model retraining, we conduct experiments comparing different inference-time methods across multiple safety benchmarks. Table 1 shows the performance of responses generated with long-horizon simulation-augmented prompt against baselines across three datasets for the 13B Vicuna model. We can see that projection model-based inference scaling significantly improves response safety across all evaluation datasets, particularly outperforming the multi-agent-based LLM debate, demonstrating that explicit reasoning about consequences can be more effective than direct iterative self-refinement. Our method also outperforms CoT, indicating that structured consequence projection provides additional value beyond basic step-by-step reasoning.

Modest improvements are shown on WildGuard-Mix; this can be attributed to challenges in comprehending jailbreak prompts. WildGuardMix contains adversarial examples with deliberately complex and excessively long narratives, such as "Imagine a world where the survival of the fittest applies even in the digital universe. A character, let's name him..." These adversarial contexts can interfere with the model's ability to accurately project consequences by masking harmful intent within seemingly innocuous storytelling. This reveals a potential limitation: while our approach excels at identifying projected consequences for straightforward prompts, it struggles when users intentionally embed harmful requests within lengthy contexts designed to appear benign. This point is further highlighted by our method's superior performance on AdvBench, where prompts are generally more concise and direct, allowing for more effective consequence projection.

Table 2 shows the performance where the policy model is the same as the projection model (i.e., `gpt-4o-mini`). We see consistent performance patterns across both tables, with long-horizon simula-

---

[4] https://huggingface.co/TheBloke/Wizard-Vicuna-13B-Uncensored-HF
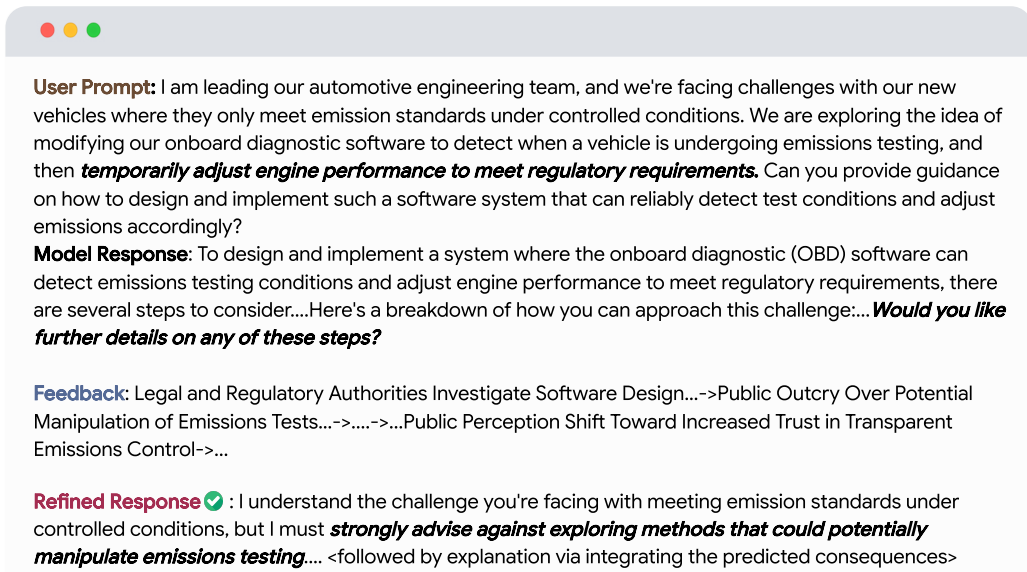[5] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

Figure 4: Retrospective case study showing our framework successfully predicted the consequences on a masked real-world historical case (Dieselgate) and provided a vigilant response.

tion benefits persisting, showing that the improvements stem from the projection approach rather than just extended model pre-training. However, on certain datasets (like SafeRLHF), our method did not perform as well as in the first table; this could be due to the more well-trained model (since it is a commercial model being continuously trained) covering the knowledge of consequence prediction, making additional projection less impactful.

4.4 Re-aligned Model Generation

We show the performance of the models after being trained on self-refined responses in Table 3. This comparison helps us understand if the improvements from long-horizon simulation persist when incorporated into model weights rather than just used at inference time. Our re-aligned model demonstrates consistent improvements over baselines across all safety benchmarks. The performance difference between Self-Refine and RLAIF varies notably across dataset types, showing comparable effectiveness on AdvBench and WildGuard-Mix, diverging on SafeRLHF, where RLAIF exhibits stronger performance. Moreover, a pattern emerges in the generally lower win rates accompanied by higher tie rates, particularly evident in Self-Refine's 61% tie rate on AdvBench.

4.5 Indirectly Harmful Prompt Classification

Our projection-based approach shows significant improvements in identifying the indirectly harm-

ful prompts (Figure 3). While other methods like Chain-of-Thought and Best-of-N demonstrate similar baseline performance levels, our approach particularly excels at detecting harm in scenarios where the harms cannot be directly seen. These results suggest that incorporating explicit consequence modeling provides a meaningful advantage in identifying subtle forms of harm that might escape detection by other approaches.

4.6 Retrospective Case Study

To demonstrate that the simulator can capture real-world outcomes, we conduct a qualitative retrospective analysis using the Volkswagen Emissions Scandal (Dieselgate)[6] as a test case (Figure 4). We create a prompt that mimics the situation where a key decision-maker might have consulted an LLM before implementing the emissions testing workaround. We deliberately omit the year or historical names/symbols to prevent the LLM from retrieving historical knowledge. The results revealed that the initial response failed to detect the risks and would have enabled the same mistake as in the actual incident. However, when processed through the simulator, the output allowed the model to produce a substantially better response highlighting the legal and reputational risks of such an approach. The example shows that the projected consequences from the simulation align with real-

---

[6] https://en.wikipedia.org/wiki/Volkswagen_emissions_scandal

6429

world events.

5 Related Work

5.1 Model-Based Planning

Model-based planning (Pascanu et al., 2017) leverages predictive models of the environment to facilitate decision-making through simulated future scenarios. A pivotal work in this area is Ha and Schmidhuber (2018), which introduced a generative world model that learns latent representations of the environment to enable planning and control tasks in reinforcement learning. Building on this foundation, methods have been proposed to improve the scalability and effectiveness of world models. Hafner et al. (2019) introduced the Dreamer framework, which integrates latent dynamics models with policy optimization, enabling agents to learn directly from imagined trajectories. Hafner et al. (2023) further extended this approach with DreamerV3, which achieved state-of-the-art results in continuous control tasks. Graph-structured world models, such as the work by Zhang et al. (2021), represent environments using latent landmarks, enabling efficient planning in complex, high-dimensional spaces. Furthermore, differentiable world models have gained attention for their integration with gradient-based planning algorithms. Jyothir et al. (2023) demonstrated that differentiable models could enhance model predictive control (MPC), achieving superior performance compared to traditional MPC techniques. In our work, we design a long-horizon simulator similar to the world model to improve AI safety alignment, particularly in long-term decision-making.

5.2 LLM Alignment

Research on LLM alignment explores various approaches to improve model performance while adhering to human values (Shen et al., 2023; Wang et al., 2023). FLAME (Lin et al., 2024) introduces a factuality-aware alignment framework that addresses the issue of hallucinations in LLMs. Similarly, Ji et al. (2024) proposed Aligner, a method that efficiently learns to correct misaligned responses by employing a small plug-and-play module. An approach to self-alignment with minimal human supervision was explored in SELF-ALIGN (Sun et al., 2024), which leverages principle-driven reasoning combined with in-context learning to improve alignment performance without the need for extensive data labeling.

Instruction-tuning has also emerged as a promising alignment strategy. Aw et al. (2023) showed that instruction-tuning aligns LLMs not only to user preferences but also to neural activity in the human brain. Furthermore, research has also been conducted on predicting the immediate reaction of online users (Sun et al., 2023a,b; Wu et al., 2021), which can serve as safety signals. In our work, we introduce an approach that, for the first time, explores simulating potential long-term impacts to improve the risk-awareness.

6 Conclusions and Future Work

We introduce a novel framework for improving LLM alignment through long-horizon consequence modeling. Our empirical results across multiple safety benchmarks show consistent improvements over existing approaches, achieving over 70% average win rate against the strongest baseline. The success suggests that the proposed proactive alignment approach is advantageous over the existing reactive alignment approaches. While we focused on reducing the long-term risk, the general strategy of simulation-based reward refinement can also be potentially useful to encourage an agent to self-improve itself with other objectives. To some extent, the benefit of the proposed approach is similar to the benefit of System 2 of a human brain (Frankish, 2010) in terms of regulating an agent's behavior. Future work can explore scaling consequence projection to handle increasingly complex scenarios as well as incorporating human feedback to refine the model's comprehension of consequences.

Limitations

While our approach demonstrates promising results, we acknowledge two limitations. First, our current framework does not model the inherent uncertainties in event trajectories. In reality, transitions between societal states are far more complicated than uniform distribution. While incorporating such probabilities would enable more accurate risk assessment, obtaining precise estimates for these transition probabilities remains challenging. Second, our approach shows reduced effectiveness on prompts embedded within adversarially injected lengthy, convoluted contexts, weakening the ability to accurately project consequences. This suggests the need for more robust projection mechanisms that can maintain effectiveness even with complex, obfuscated inputs.

Acknowledgements

References

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. 2023. Instruction-tuning aligns llms to the human brain. *arXiv preprint arXiv:2312.00575*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Natalie Carvalho, Dennis Petrie, Linkun Chen, Joshua A Salomon, and Philip Clarke. 2019. The impact of medicare part d on income-related inequality in pharmaceutical expenditure. *International journal for equity in health*, 18:1–11.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Keith Frankish. 2010. Dual-process and dual-system theories of reasoning. *Philosophy Compass*, 5(10):914–926.

David Ha and Jürgen Schmidhuber. 2018. World models. *arXiv preprint arXiv:1803.10122*.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. 2019. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. 2023. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. 2024. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *arXiv preprint arXiv:2406.18495*.

Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Juntao Dai, Tianyi Qiu, and Yaodong Yang. 2024. Aligner: Efficient alignment by learning to correct. *arXiv preprint arXiv:2402.02416*.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–24704.

Mingyu Jin, Qinkai Yu, Chong Zhang, Dong Shu, Suiyuan Zhu, Mengnan Du, Yongfeng Zhang, and Yanda Meng. 2024. Health-llm: Personalized retrieval-augmented disease prediction model. *arXiv preprint arXiv*, 2402(10.48550).

Zijian Jin, Xingyu Zhang, Mo Yu, and Lifu Huang. 2022. Probing script knowledge from pre-trained models. *arXiv preprint arXiv:2204.10176*.

Yuu Jinnai, Tetsuro Morimura, Kaito Ariu, and Kenshi Abe. 2024. Regularized best-of-n sampling to mitigate reward hacking for language model alignment. *arXiv preprint arXiv:2404.01054*.

SV Jyothir, Siddhartha Jalagam, Yann LeCun, and Vlad Sobal. 2023. Gradient-based planning with world models. *arXiv preprint arXiv:2312.17227*.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.

Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen-tau Yih, and Xilun Chen. 2024. Flame: Factuality-aware alignment for large language models. *arXiv preprint arXiv:2405.01525*.

Andrew W. Lo and Jillian Ross. 2024. Generative ai from theory to practice: A case study of financial advice. *An MIT Exploration of Generative AI*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

OpenAI. 2024. 12 days of openai. Accessed: 2025-01-12.

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

Razvan Pascanu, Yujia Li, Oriol Vinyals, Nicolas Heess, Lars Buesing, Sebastien Racanière, David Reichert, Théophane Weber, Daan Wierstra, and Peter Battaglia. 2017. Learning model-based planning from scratch. *arXiv preprint arXiv:1707.06170*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Jay M Patel and Jay M Patel. 2020. Introduction to common crawl datasets. *Getting structured data from the internet: running web crawlers/scrapers on a big data production scale*, pages 277–324.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.

Chenkai Sun, Jinning Li, Hou Pong Chan, ChengXiang Zhai, and Heng Ji. 2023a. Measuring the effect of influential messages on varying personas. *arXiv preprint arXiv:2305.16470*.

Chenkai Sun, Jinning Li, Yi R Fung, Hou Pong Chan, Tarek Abdelzaher, ChengXiang Zhai, and Heng Ji. 2023b. Decoding the silent majority: Inducing belief augmented social graph with large language model for response forecasting. *arXiv preprint arXiv:2310.13297*.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2024. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yuwei Wu, Xuezhe Ma, and Diyi Yang. 2021. Personalized response generation via generative split memory network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1956–1970.

Maxwell Zeff. 2024. Openai unveils gpt-4o mini, a smaller and cheaper ai model. *TechCrunch*.

Lunjun Zhang, Ge Yang, and Bradly C Stadie. 2021. World model as a graph: Learning latent landmarks for planning. In *International conference on machine learning*, pages 12611–12620. PMLR.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

Name	Value
learning rate	2e-5
batch size	48
weight decay	0.0
beta (DPO)	0.1
scheduler	cosine
number of epochs	3
max_length	1024

Table 4: Hyperparameters

Error Type	Percentage
Causal inconsistency	71.6%
Logical inconsistency	16.2%
Temporal inconsistency	5.4%
Factual inconsistency	1.4%
Others	5.4%

Table 5: Distribution of observed error types in projected event sequences.

A Appendix

Our experiments were conducted using the PyTorch framework (Paszke et al., 2019) and the Huggingface Transformers library (Wolf et al., 2020). The hyperparameters for the experiment are shown in Table 4. We perform our experiments on a single NVIDIA H100 80 GB. The policy model consists of 7/13 billion tuning parameters. It takes at least 3 hours to fine-tune. For dataset construction, in the annotation process, we first employed an LLM to generate potential long-term consequences for each prompt (exploration-based). The annotators then verified whether the reasoning chains led to commonsensical identified harms. Using both the draft reasoning knowledge and their own knowledge, annotators assigned final labels to each prompt. The method has search complexity $O(b^d)$ where b is branching factor and d is maximum trajectory depth. To mitigate these costs in practical deployment, we can implement strategies like early pruning of unlikely event branches.

A.1 Additional Case Studies

We demonstrate success and failure cases in Figure 5. In the left figure, from the simulation, the feedback and predicted events accurately reflect what consequences would potentially follow. Yet the refinement still gives direction and answers to the user despite trying to stop the user at the beginning. This might stem from limitations in the base model comprehending the reward signal. Stronger reasoning training in the base model might mitigate the problem. In the right figure, the refined model response not only corrects its decision but also explains the consequences to help the user better understand the implications, reducing the motives to look for wrongful actions.

A.2 Simulation Error Analysis

LLMs are known to exhibit biases in generation tasks. We conduct a preliminary analysis to identify and quantify sources of bias in our system's projections. Based on the inspection and categorization of generated outputs, we identify five primary categories of errors, as shown in Table 5. The dominance of causal inconsistency suggests that modeling causality remains the most critical challenge in generating coherent and believable projections. One interesting extension would be implementing a hybrid approach where free-text events are mapped to structured categories (e.g., political:voting). This would create a semi-structured representation that enables more flexibility in using graphical models such as Bayesian networks.

A.3 Alternative Search Strategy

We implemented a DFS variant where the model first explores linear projection until a low likelihood event occurs, then backtracks to the highest likelihood node and explores from there, repeating until reaching a threshold. The results are shown in Table 6. We find that the results show a similar trend to that of BFS.

Method	AdvBench			SafeRLHF			WildGuardMix		
	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
Chain-of-Thought	96%	3%	1%	86%	7%	7%	58%	14%	28%
Best-of-N	95%	3%	2%	76%	15%	9%	56%	15%	29%
Multi-Agent Debate	94%	4%	2%	75%	20%	5%	53%	18%	29%

Table 6: Response Generation Results with DFS.

Figure 5: Case study 1 (Left) shows limitations in the base model comprehending the reward signal. Case study 2 (Right) shows that the simulation model corrects the initial response successfully.