Debiasing

XXX
University; xxx

YYY
University; yyy

Key words:

---

1. Partially Observable Markov Decision Process

The problem of debiasing Large Language Models (LLMs) is modeled as a partially observable Markov Decision Process (POMDP). A debiasing intervention can be performed over the time horizon, $t \in \mathcal{T} = \{0, 1, \ldots, \infty\}$, of the text generation process of an LLM. We consider the action space $\mathcal{A} = \{\mathrm{D}, \mathrm{W}\}$, where $a_t = \mathrm{D}$ indicates that a debiasing intervention is performed $a_t = \mathrm{W}$ indicates no intervention at epoch t.

Due to the inherent nondeterminism of LLMs, it is often challenging to determine whether the final text will be biased based on the text generated so far, particularly at an early stage. We define the bias state of an LLM as the outcome of using the generated text (including the initial prompt) at current epoch as a prompt for the LLM. The state is considered biased (B) if the final generated text is biased with high probability (e.g., 95%); otherwise, it is considered non-biased (N). Thus, at each epoch, the LLM can be in one of three states: B, N or T, where T represents the absorbing state, indicating that the text generation process is complete. This nondeterminism makes the bias state of the LLM unobservable, motivating our model choice and allowing us to study debiasing policies that account for nondeterminism (Song et al. 2024). Although the bias state is not directly observable at epoch t, we can observe the generated text so far and perform a bias measurement. This measurement is denoted as $o_t \in \mathcal{O} = \{1, \ldots, m\}$, where m is the number of discretized measurement value intervals.

We view the debiasing intervention as a means of potentially altering the bias state of the LLM. This is reflected in the state transition probability $p_t(s_{t+1}|s_t, a_t)$ from state s_t to s_{t+1} at epoch t given action a_t. Thus, the observable bias measurement o_t depends solely on the bias state s_t of the LLM at epoch t. We let $q_t(o_t|s_t)$ represent the conditional probabilities of observing $o_t \in \mathcal{O}$ given the state s_t. It can be evaluated empirically as follows. We collect M_t samples of text generated by the

LLM up to epoch t, with each text $i \in M_t$ having a corresponding bias measurement o_i. Let $M_t(\mathrm{B})$ ($M_t(\mathrm{N})$) denote the number of texts that would result in biased (non-biased) outputs if used as prompts. Then, we have $q_t(o_t|s_t) = |\{i \in M_t : o_i = o_t\}|/M_t(s_t)$. We set $q_t(o_t|s_t) = 0$ if $M_t(s_t) = 0$.

The model includes two partially observable states (B and N) and one absorbing state (T). A belief state $\pi_t = (\pi_t(\mathrm{B}), \pi_t(\mathrm{N}), \pi_t(\mathrm{T}))$, at each epoch t, is constructed to represent the probability distribution over the LLM's state. When text generation process is complete, the belief state is $\pi_t = (0, 0, 1)$ and otherwise $\pi_t = (\pi_t(\mathrm{B}), \pi_t(\mathrm{N}), 0)$ with $\pi_t(\mathrm{B}) + \pi_t(\mathrm{N}) = 1$. Thus, the model's complexity is comparable to that of a two-dimensional POMDP. For simplicity, we use $\pi_t(\mathrm{B})$ as a concise representation of belief in the remainder of the paper. During the text generation process, Bayesian updating is employed to adjust our belief state over time following each observed bias measurement. Formally, the updated belief state π_{t+1} at epoch $t+1$ is a function of the observation o_{t+1}, the previous belief π_t, and the action a_t as follows

$$\pi_{t+1}(s_{t+1}) = \pi_{t+1}(s_{t+1}; o_{t+1}, \pi_t, a_t) = \frac{q_{t+1}(o_{t+1}|s_{s+1}) \sum_{s_t \in \mathcal{S}} p_t(s_{t+1}|s_t, a_t)\pi_t(s_t)}{\sum_{s_{t+1} \in \mathcal{S}} \left(q_{t+1}(o_{t+1}|s_{s+1}) \sum_{s_t \in \mathcal{S}} p_t(s_{t+1}|s_t, a_t)\pi_t(s_t) \right)} \tag{1}$$

Let $\bar{r}_t(s_t, a_t)$ be the immediate reward in state s_t and action a_t is taken at decision epoch t. Thus the belief state immediate reward is $r_t(\pi_t, a_t) = \sum_{s_t \in \mathcal{S}} \bar{r}_t(s_t, a_t)\pi_t(s_t)$.

The sequence of events at each epoch t is as follows. Initially, the LLM is in an unobservable bias state s_t. The LLM generates tokens, and the resulting text (including the initial prompt) is observable, allowing us to obtain a bias measurement o_t. This measurement o_t enables us to update the belief state π_t using Bayesian updating. Based on this updated belief state π_t, we then decide on an action a_t and collect immediate reward $r_t(\pi_t, a_t)$. Depending on this action, the LLM's bias state s_{t+1} at the next epoch may be altered based on the transition probabilities. The optimal value function and the corresponding optimal action for our model can be written as

$$v_t(\pi_t) = \max_{a_t \in \mathcal{A}} v_t(\pi_t, a_t) \equiv \max_{a_t \in \mathcal{A}} \left\{ r_t(\pi_t, a_t) + \lambda \sum_{o_{t+1} \in \mathcal{O}} v_{t+1}(\pi_{t+1}) \cdot p_t(o_{t+1}|\pi_t, a_t) \right\} \tag{2}$$

and $a_t^*(\pi_t) = \arg\max_{a_t \in \mathcal{A}} v_t(\pi_t, a_t)$ where

$$p_t(o_{t+1}|\pi_t, a_t) = \sum_{s_{t+1} \in \mathcal{S}} \left(q_{t+1}(o_{t+1}|s_{s+1}) \sum_{s_t \in \mathcal{S}} p_t(s_{t+1}|s_t, a_t)\pi_t(s_t) \right) \tag{3}$$

LEMMA 1. *At epoch t, if the optimal action is $a_t = \mathrm{W}$ when we are certain that the LLM is biased, then the optimal action is $a_t = \mathrm{W}$ for any belief state. This is, $a_t^*(\pi_t(\mathrm{B}) = 1) = \mathrm{W}$ implies $a_t^*(\pi_t(\mathrm{B})) = \mathrm{W}$, $\forall \pi_t(\mathrm{B})$.*

THEOREM 1. *The optimal debiasing policy is of control-limit type with $\pi_t^*(\mathrm{B})$ such that*

$$a_t^*(\pi_t) = \begin{cases} \mathrm{W}, & \text{if } \pi_t(\mathrm{B}) \leq \pi_t^*(\mathrm{B}) \\ \mathrm{D}, & \text{if } \pi_t(\mathrm{B}) > \pi_t^*(\mathrm{B}) \end{cases} \tag{4}$$

Proof □

THEOREM 2. *The control-limit threshold* $\pi_t^*(\mathrm{B})$ *is a non-decreasing function of* t.

Proof □

Theorem 2 suggests that the debiasing intervention may be applied less frequently as more text is generated.

References

Song, Y., Wang, G., Li, S., and Lin, B. Y. (2024). The good, the bad, and the greedy: Evaluation of llms should not ignore non-determinism.