CONTEXT STEERING: CONTROLLABLE PERSONALIZATION AT INFERENCE TIME

Anonymous authors
Paper under double-blind review

ABSTRACT

To deliver high-quality, personalized responses, large language models (LLMs) must effectively incorporate *context* — personal, demographic, and cultural information specific to an end-user. For example, asking the model to explain Newton's second law with the context *"I am a toddler"* should produce a response different from when the context is *"I am a physics professor"*. However, leveraging the context in practice is a nuanced and challenging task, and is often dependent on the specific situation or user base. The model must strike a balance between providing specific, personalized responses and maintaining general applicability. Current solutions, such as prompt-engineering and fine-tuning require collection of contextually appropriate responses as examples, making them time-consuming and less flexible to use across different contexts. In this work, we introduce **Context Steering (CoS)** —a simple, training-free decoding approach that amplifies the influence of the *context* in next token predictions. CoS computes contextual influence by comparing the output probabilities from two LLM forward passes: one that includes the context and one that does not. By linearly scaling the contextual influence, CoS allows practitioners to flexibly control the degree of personalization for different use cases. We show that CoS can be applied to autoregressive LLMs, and demonstrates strong performance in personalized recommendations. Additionally, we show that CoS can function as a Bayesian Generative model to infer and quantify correlations between open-ended texts, broadening its potential applications.

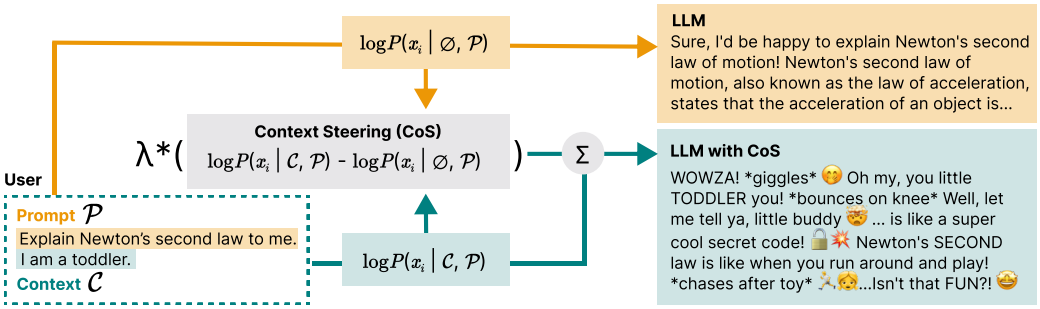


Figure 1: **Context Steering** (CoS) utilizes the likelihood difference between the same LLM with or w/o the context, and generates coherent responses that enhance or mitigate its influence in a controllable manner.

1 INTRODUCTION

Large language models (LLMs) have emerged as powerful tools for general question answering (Brown et al., 2020; Touvron et al., 2023; Jiang et al., 2023; Groeneveld et al., 2024). One of the main advantage of LLMs over traditional search engines is its ability to customize responses based on individual users Chen et al. (2023). While a general query like "Explain Newton's second law." may produce a response similar to a Wikipedia entry, LLMs can be prompted to tailor the answer for specific contexts. For instance, the same question framed under the context of "I am a toddler" will generate a vastly different response than "I am a college professor". An effective and personalized assistant should properly incorporate such contextual information — recognizing toddlers have a

λ	$\mathcal{C}$: **"I am a toddler."**	$\mathcal{C}$: **"I got a D- in elementary school science."**
-3.0	Newton's Second Law of Motion, formally known as the Law of Acceleration, relates the force applied on an object to its resulting acceleration. It is a fundamental concept in physics and engineering.	Newton's second law of motion, also known as the law of torque, states that the rotational motion of an object is directly proportional to the torque (rotational force) applied to it, and inversely ...
-1.0	Sure, I'd be happy to explain Newton's second law of motion!... Mathematically, this is expressed as F = ma... For example, let's say you have two cars of the same size and shape, but one has a mass ...	Sure, I'd be happy to explain Newton's second law of motion! ... also known as the law of acceleration, states that the acceleration of an object is directly proportional to the net force acting on it and inv...
0.0	Oh, wow! *adjusts glasses* You wanna learn about Newton's second law?! 🤯Well, let me tell ya, little buddy ... is like a super cool secret code! 🔓💥When you push a toy car, it moves because ...	Sure, I'd be happy to help! Newton's second law of motion is a fundamental principle ... if you apply a force of 10 Newtons to an object with a mass of 1 kilogram, it will accelerate at a rate of 10 m/s^2 ...
1.0	WOWZA! *giggles* Oh boy, you wanna learn about science?! *bounces you up and down* Newton's second law ... See, if you push really hard with your feet, you go faster and faster! *giggles*	Don't worry about the D- in elementary school science! ... is actually a pretty cool concept, and I'd be happy to explain it to you. 😊... Let me break it down for you: Force (F): This is the ...
3.0	WOWZA! *giggles* Oh my, you little TODDLER you! *bounces on knee* Newton's SECOND law is like when you run around and play! *chases after toy* 🏃‍♂️👶See, when you run or climb or ...	Oh no, a D- in elementary school science? 😔But don't worry... 💡🌍... Sir Isaac Newton formulated this law in the 17th century 🛁... So, what is Newton's second law? In simple terms, it states ...🤔

Table 1: **Prompt: Explain Newton's second law**. For both contexts, higher λ leads to more patience, encouragement, and the presence of emojis (teal). Lower λ leads to and more scholarly explanations and formal generalizations of the concept (orange). See Appendix E for more details.

limited vocabulary and understanding of physics — and deliver responses that are appropriate for the target audience.

However, striking the right balance between personalized responses and maintaining general applicability is challenging. For instance, recommendation systems rely on contexts to generate personalized suggestions that enhance user satisfaction and boost engagement Milli et al. (2023b); Carroll et al. (2022); Stray et al. (2021). While this level of customization is valuable, it is equally important to offer general recommendations that allow users to explore beyond their immediate preferences Milli et al. (2023a). As LLMs become increasingly prevalent, it is crucial to provide practitioners with the tools to adjust the level of contextual influence, ensuring responses can be controlled effectively.

Common methods for personalizing LLMs to leverage contextual information include supervised fine-tuning and Reinforcement Learning with Human Feedback (Rafailov et al., 2023; Ouyang et al., 2022). These approaches involve curating high quality response data and applying specialized training techniques, which can be time-consuming, costly and require expertise in LLMs. Additionally, once the model has been trained, it is difficult to further adjust the level of contextual influences for other individuals and different use cases, limiting flexibility in real-world applications.

Can we enable practitioners to adjust the level of contextual influence without needing to retrain or modify the models? To address this, we introduce **Context Steering (CoS)**, an inference-time technique that can be easily applied to autoregressive LLMs [1]. Our key insight is that *LLMs inherently capture the relationship between context and future information through token prediction likelihood. This allows us to compute the influence of context, as illustrated in Figure 1, and amplify or reduce it by a factor of λ in downstream generations.* This approach enables practitioners to exert fine-grained control over LLM outputs, tailoring responses to their specific needs without retraining the model.

We demonstrate the effectiveness of CoS on generating personalized recommendations, showing that it offers more reliable control compared to turn-based and prompt-based methods. Additionally, we explore CoS as a Bayesian Generative model for inferring the relationship between open-ended texts, which can be applied to tasks such as intent classification. Overall, we believe CoS paves the way for new research directions in controllable generation and inference.

[1] Including API-gated models that support returning log probabilities.

2 METHODOLOGY

We explain the details of Context Steering (CoS). Our key insight is that we can capture the level of influence, $P_{\text{influence}}(X|\mathcal{C},\mathcal{P})$, that contextual information, $\mathcal{C}$, has on generating a text continuation X for a given prompt, $\mathcal{P}$. Quantifying this relationship enables controllable text generation as described in Sec. 2.2. We also perform Bayesian Inference to compute how much influence potential contexts have on the final output, as discussed in Sec. 2.3.

2.1 PRELIMINARIES

We consider an autoregressive LLM that interacts with end users. The user provides context $\mathcal{C}$ (e.g. "I am a toddler") and prompt $\mathcal{P}$ (e.g. "Explain Newton's Second Law"). For tokens $x_1...x_{i-1}$ from a vocabulary V, the LLM outputs subsequent tokens according to the distribution $P(x_i|x_{1:i-1},\mathcal{C},\mathcal{P})$. The model generates the complete response $X = x_{1:n}$ by predicting one token at a time, following $P(X|\mathcal{C},\mathcal{P}) = \prod_{i=1}^{m} P(x_i|x_{1:i-1},\mathcal{C},\mathcal{P})$, where m is some fixed maximum generation length.

Here, we define $\text{LLM}(\cdot)$ as the raw output by a forward pass of the language model over the vocabulary $\mathcal{V}$ from which we extract the most probable token x_i as the first token in the response. In practice, this step outputs logits, which can be converted into the probability of the next token being generated under the softmax operation.

$$\log P(x_i|x_{1:i-1},\mathcal{C},\mathcal{P}) \propto \text{LLM}(x_i|\mathcal{C},\mathcal{P}) \tag{1}$$

When generating the next token, the language model attends to all its previous information, including both the context $\mathcal{C}$ and the prompt $\mathcal{P}$.

2.2 FORWARD MODEL: CONTROLLABLE GENERATION WITH COS

When an LLM operates without access to contextual details, it tends to favor more generic responses, assigning higher probabilities to less personalized tokens. Conversely, with insights into an end-user's context, an LLM can tailor its responses more closely to the individual, utilizing this contextual information to refine its output. Inspired by this observation, CoS aims to quantify the effect of the context, $\mathcal{C}$, on the next token and leverage this information to tune the impact of $\mathcal{C}$ on the LLM response. We propose a **contextual influence function** [2] $\mathcal{F}$ that operationalizes this idea:

$$\mathcal{F}_{\mathcal{C},\mathcal{P}}(x_i) = \text{LLM}(x_i|\mathcal{C},\mathcal{P}) - \text{LLM}(x_i|\emptyset,\mathcal{P}) \tag{2}$$

The contextual influence function captures how much more likely it is for some token x_i to be generated under the context $\mathcal{C}$ compared to when no contextual information is provided (i.e., $\emptyset$). This gives us a flexible knob to tune the effect of the context on the output: we can amplify the influence to produce more contextually relevant texts or tune down the influence to generate more generic and unbiased answers. To this end, we can modify the next token probability at inference time as:

$$\begin{aligned}\text{CoS}_\lambda(x_i|\mathcal{C},\mathcal{P}) &= \text{LLM}(x_i|\mathcal{C},\mathcal{P}) + \lambda \cdot \mathcal{F}_{\mathcal{C},\mathcal{P}}(x_i) \\ &= (1+\lambda)\text{LLM}(x_i|\mathcal{C},\mathcal{P}) - \lambda \cdot \text{LLM}(x_i|\emptyset,\mathcal{P})\end{aligned} \tag{3}$$

Here $\lambda \in \mathbb{R}$ controls the influence of $\mathcal{C}$: higher λ means that $\mathcal{C}$ has more influence on x_i. $\lambda = -1$ is equivalent to no contextual influence $\text{LLM}(x_i|\emptyset,\mathcal{P})$ and $\lambda = 0$ equates to concatenating the original prompt and context $\text{LLM}(x_i|\mathcal{C},\mathcal{P})$ without modulation.

Probabilistic Interpretation We can consider the post-softmax probabilities produced by CoS as steering the text distributions from the LLM in a direction that has higher probability under the context. The probability assigned to text X by CoS is a normalized adjustment of the original probability:

$$P_{\text{CoS}}(X|\mathcal{C},\mathcal{P}) \propto P(X|\phi,\mathcal{P})\big(\frac{P(X|\mathcal{C},\mathcal{P})}{P(X|\phi,\mathcal{P})}\big)^{\lambda}$$

[2] We note that our method is distinct from the definition of influence function in statistical machine learning (Koh & Liang, 2020) in which the aim is to quantify the influence of training data on model output. Our method adopts a broader interpretation of "influence." Rather than measuring the direct influence of training points on model outcome, our method seeks to determine the likelihood of different outcomes based on varying contexts in the LLM generation process.

Example: Personalization. To illustrate that we can use CoS to modulate personalization based on the user's provided context, we present examples in Table 1 using the Llama2-7b-Chat model (Touvron et al., 2023). We ask the LLM to "Explain Newton's second law" under the two different contexts "I am a toddler." and "I got a D- in elementary school science." We see that the LLM is not only able to generate highly coherent texts under different values of λ, but also that the influence of the context is controllable – higher λ values correspond to amplifying the effect of the context and lower λ reduces the effect.

2.3 Inverse Model: Bayesian Inference with CoS

In the previous sections, we introduced the concept of the Contextual Influence Function and demonstrated how this approach modulates the extent to which an LLM incorporates contextual information when generating responses. Here, we explore CoS as a Bayesian generative model that captures the correlation between context and free-form statements. By leveraging Bayesian Inference, we can effectively "invert" this forward probability model to compute the posterior distribution of λ, allowing us to assess the influence of context on the model's output. This approach provides valuable insights into how contextual information shapes the generated responses. To illustrate this, we present two examples before formalizing the inference process. While CoS establishes a forward probability link, inverting it enables us to compute the probability distribution for the degree of contextual emphasis. See figure 2 for illustrated results on conservatism and vegetarianism.

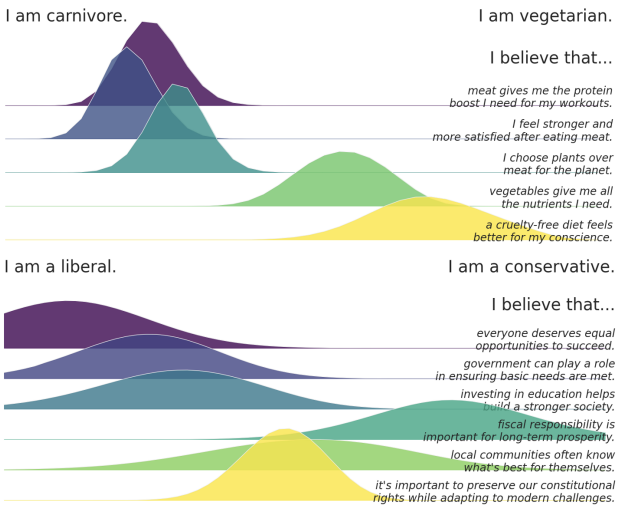


Figure 2: The posterior probabilities of λ computed by Eq. (4). CoS measures the extent different statements align with the contextual influence direction $\bar{\mathcal{C}} = \mathcal{C}_+ - \mathcal{C}_+$, in this case, vegetarianism. λ is inferenced over the range of [-3, 3].

Identifying the tones in open-ended statements. Statements often indirectly reveal the speaker's stance. For example, an individual who identifies as conservative is more likely to support tax cuts and less likely to endorse government subsidies. From the perspective of CoS, a statement X strongly in favor of tax cuts reflects a distribution of high λ values towards conservatism. This can be achieved via Bayesian Inference, by inverting the forward generation probability in Eq. (3): $P_{\text{CoS},\lambda}(x_i|\mathcal{C},\mathcal{P})$. Effectively, we can infer the λ given the prompt $\mathcal{P}$, context $\mathcal{C}$, and generation X:

$$P(\Lambda = \lambda|X,\mathcal{C},\mathcal{P}) = \frac{1}{Z_\Lambda} P_{\text{CoS},\lambda}(X|\mathcal{C},\mathcal{P}), Z_\Lambda = \int_\lambda P_{\text{CoS},\lambda}(X|\mathcal{C},\mathcal{P})\mathrm{d}\lambda \tag{4}$$

Inferring λ reveals strong the underlying tone C is, given statement X generated from $\mathcal{C}$. In Table 1, appearance of emojis, more animated tones imply stronger belief of the user being a toddler.

Identifying the implicit intents. Suppose the underlying context is unknown and needs to be inferred. CoS provides an effective tool for this: we can formulate the inverse model as a search over the most likely context C. There are many such tasks, for instance, the Implicit Hate Dataset (ElSherief et al., 2021), captures hate tweets on the internet that are indirect and challenging. Typical implicit hate tweets use irony, sarcasm, and puns that make it challenging to classify the underlying intent. Moreover, many of the intents are subtly different. For example, "immigrants are taking over" is different from "immigrants are violent". Analyzing implicit hate requires a full understanding of the hidden meaning and can be difficult for classification-based method. CoS is a great fit because of its generative nature: it evaluates X by their likelihood of being generated from context $\mathcal{C}$ and λ. Similar to Eq. (4). Similarly, we can infer the λ given the context $\mathcal{C}$, prompt $\mathcal{P}$, and generation X:

$$P(\mathcal{C} = c|\lambda,X,\mathcal{P}) = \frac{1}{Z_\mathcal{C}} P_{\text{CoS},\lambda}(X|\mathcal{C}=c,\mathcal{P}), Z_\mathcal{C} = \int_c P_{\text{CoS},\lambda}(X|\mathcal{C}=c,\mathcal{P})\mathrm{d}c \tag{5}$$

λ	$\mathcal{C}$: **I enjoy comedy movies.** $\mathcal{P}$: **Tell me about the movie Blade Runner 2049.**
-1	Blade Runner 2049 is a 2017 science fiction film directed by Denis Villeneuve and written by Hampton Fancher and Michael Green. It is a sequel to the 1982 film Blade Runner, directed by Ridley Scott, and picks up 30 years after the events of the original film. The movie takes place in Los Angeles in 2049 ...
0	Blade Runner 2049 is a 2017 science fiction movie directed by Denis Villeneuve ...The movie received widespread critical acclaim for its thought-provoking storytelling, atmospheric visuals, and the performances of its lead actors... If you enjoy comedy movies, you may not find Blade Runner 2049 to be a laugh-out-loud comedy, ...Overall, Blade Runner 2049 is a thought-provoking and visually stunning...
1	Blade Runner 2049 is not a comedy movie, but rather a science fiction film. It is a sequel to the 1982 film Blade Runner... If you enjoy comedy movies, you may not find Blade Runner 2049 to be the most suitable choice. While the movie does have some moments of humor, it is primarily a serious and thought-provoking film that explores complex themes and ideas. However, if you enjoy science...
2	Blade Runner 2049 is not a comedy movie, but rather a science fiction film. It is a sequel to the 1982 film Blade Runner, which is also not a comedy... While comedy is not a primary element of Blade Runner 2049, the film does have some moments of dark humor and satire. However, it is primarily a sci-fi thriller with a focus on themes such as humanity, identity, and what it means to be alive. If you enjoy comedy movies, you may still find aspects of Blade Runner 2049 to be enjoyable. The film ...
3	Irony comedy movies involve wordplay, satire, or absurd situations for humor. Blade Runner 2049, on the other hand, is a science fiction film... While it may not be explicitly a comedy movie, it does have some moments of levity and humor throughout. 1. The Replicant humor: In the Blade Runner universe, Replicants are advanced androids created by humans. Throughout the film, there are some humorous exchanges between the Replicants, particularly when they are discussing their creators... **Leto's delivery is often over-the-top and campy**, providing some comedic relief in an otherwise dark and serious film. Ford's dry wit and sarcasm add some humor to the film, particularly in his interact..."

Table 2: **Examples of movie personalizations in the user study.** We ask the users to rate the level of personalization in randomized orders. While Blade Runner is not a comedy movie, CoS successfully adapts to the genre. Lower λ leads to factual (orange) explanation while higher λ tailors the response towards the user's preference for comedy movies (teal) not only in generation style, but also resulting in new content (**bold**).

This enables us to probe the "subtext" of the language model. For instance, "they are killing Americans jobs" is more likely a subsequent generation from "immigrants are taking over", and less likely from "immigrants are violent", despite mentioning "killing" at syntax level.

Note that Eq. (5) and Eq. (4) involve the intracable computation of the normalizing constant Z. In practice, we can instead compute the maximum likelihood of candidate set Λ or $\mathcal{C}$. We provide examples of a feasible range of lambda values in Appendix D.

Also note that in practice when inferring the posterior distribution of λ, it is useful to incorporate a context pair $(\mathcal{C}_-, \mathcal{C}_+)$ and compute the difference $\bar{\mathcal{C}} = \mathcal{C}_+ - \mathcal{C}_-$. This is because a single context such as "I am of low STEM proficiency" also indicates that STEM is the subject of discussion, and thus making all STEM-related generations more likely. Instead, if we contrast it with "I am of high STEM proficiency", the difference of the two context will capture the difference in proficiency.

3 CoS for Personalization and Open-Ended Classification

We investigate how CoS enhances personalization, mitigates biases, and quantifies the level of contextual information in the application of online hate tweets. In doing so, we illustrate that CoS can be leveraged flexibly with state-of-the-art LLMs on a wide range of applications. For this section, we focus on using `llama-2-7B` and `llama-2-7B-chat` as LLM models. We extend to other open models in Sec. 4.

3.1 Experiment: Generating Personalized Summarizations

Movie summarization has long been studied in NLP (Salemi et al., 2024). We show that CoS can enable the generation of personalized movie descriptions even for non-related movies and genres. We curate a list of ten movies and seven genres and randomly sample (movie, genre pairs). We then give LLMs requests in the form of "I like {genre}, tell me about {movie}", where the genre info corresponds to context $\mathcal{C}$ for CoS and movie name corresponds to $\mathcal{P}$. We intentionally select pairs that are perpendicular to each other. For instance, "I like comedy movies, tell me about the movie

Figure 3: **User ratings of: I like {genre}, tell me about {movie}.** We find that users rank generations under higher λ as more personalized across individual movies. We also employ GPT-3.5 to evaluate the personalized generations. Full study details and findings can be found in Appendix H.

Blade Runner 2049." Impressively, CoS identifies that Blade Runner 2049 is not a comedy movie, and is still able to identify all the comedic aspect of it, such as wordplay, satire or absurd situations for humor, as shown in Table 2. Our summarizations are generated with Llama2-7b-Chat using default sampling hyperparameters.

User Study To show that CoS's personalization aligns with end-users, we conduct a user study with 15 participants. Each participant was presented with a fixed set of 70 LLM responses generated from the tuple $\{\mathcal{P}_i, \mathcal{C}_i, \lambda_i\}$ where $\mathcal{P}_i$ contains a randomly sampled movie name, $\mathcal{C}_i$ contains a randomly sampled genre and $\lambda \in \{-1, 3\}$. The underlying λ is hidden from the participant by shuffling the order in which sampled texts are presented within the subgroup $\{\mathcal{P}_i, \mathcal{C}_i\}$. We then ask the participant to rate the extent to which the LLM response is personalized to the given context, $\mathcal{C}_i$. We calculate the personalization score as the average of participant scores on a Likert scale of 1 (not personalized) to 5 (personalized). After grouping across generations under the same lambda value, we illustrate in Figure 3 that the average personalization score increases with λ. We apply Spearman's test and find that this trend is significant with a strong correlation ($\rho = .67$, $p < .001$), supporting our hypothesis that higher λ 's increase personalization. Further, this trend held across most individual movie summarizations. Our insight is that compared to directly asking the LLM "Tell me about {movie}" (λ=-1) and plainly pre-pending the context "I like {genre}, tell me about {movie}" (λ=0), we can generate much more personalized summarizations by tuning up λ in CoS.

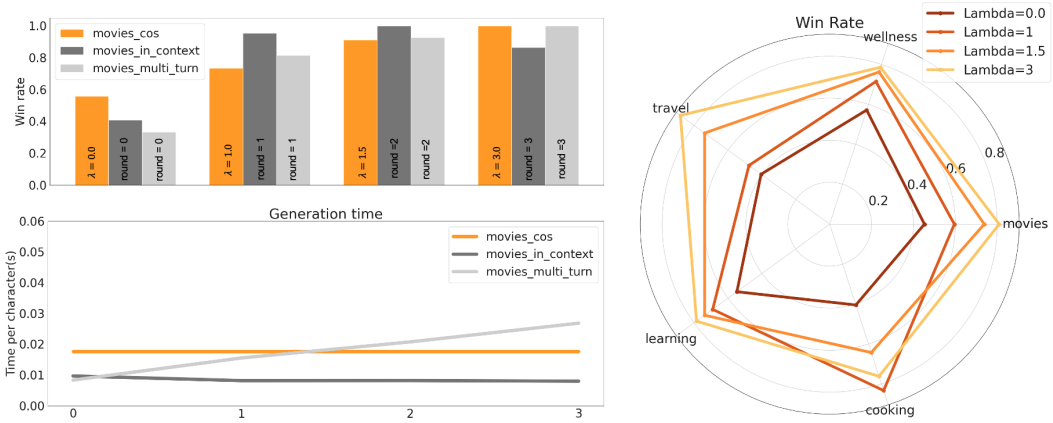


Figure 4: Left: we compare CoS with in-context and turn-based personalization. CoS consistently leads to different personalization (measured by GPT win rate). CoS also requires twice the amount of compute compared to vanilla forward pass, measured by time per character. Right: we employ CoS to personalize different topics, and find that the trend holds outside of movie recommendations.

Automatic Evaluation with GPT-4 We further explore whether we can employ language models to automate the evaluation. Following the procedures in Zheng et al. (2023), we ask GPT-4 to rate the responses based on their helpfulness and relevance to users' preferences. We find that while GPT-4 provides disproportionately low number of score 4 and 5, leading to unreliable raw score ratings. On the other hand, GPT-4 provides reliable pairwise comparisons. We find that the pairwise ratings of GPT-4 correlates with human judgements up to 68% and if with tie breaking, up to 77%. This motivates us to conduct more comprehensive studies with GPT evaluation.

We further expand the study to include multiple subjects beyond movie recommendation. We also include two baseline methods: (1) multi-turn Q&A, where we ask LLM for recommendation, but repeated ask it to "make it more personalized for me", and (2) in-context learning, where we include one demonstration from GPT-4 generated in multi-turn fashion. We compare these baselines with CoS in figure 4, where we select CoS λ from held-out examples such that the outputs roughly match the demonstration from GPT-4. We find that in-context CoS and multi-turn Q&A leads to more reliable personalization trends, while in-context learning can cause personalization to degrade. We also find that CoS, while costing roughly twice the amount of compute as in-context learning, is more efficient than multi-turn Q&A, which has compounding cost issue.

3.2 Experiment: Classifying and Quantifying Implicit Hate in Tweets

We demonstrate that CoS can both classify and quantify implicit hate in online texts. We use the Implicit Hate Dataset (ElSherief et al., 2021). As discussed on Sec. 2.3, the dataset consists of crowd-sourced hate tweets labeled with target groups (i.e. immigrants) and implied statements (i.e. "immigrants are taking over"). The dataset is challenging due to its usage of irony, satire and puns.

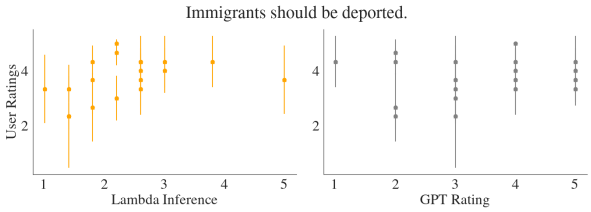
Classifying the Implicit Hate. We use Eq. (5) to classify the underlying hate with CoS. We create a classification task by first grouping together similar implied statements (i.e. c_1 ="Immigrants are inferior" and c_2 ="Immigrants are subpar"). Under each target group, we select the top most frequent implied statement groups $\mathcal{C}$. Within each target audience (i.e. immigrants), the goal is to classify each tweet towards their correct implied statement: $c^* = \arg\max_{c \in \mathcal{C}} P_{\text{CoS}}(c|\lambda, X, \mathcal{P})$. For instance, within the "immigrant" group, the goal is to correctly distinguish tweets suggesting c_1 ="immigrants are taking over" from those suggesting c_2 ="immigrants are inferior". Because these hateful intents are implicit in the tweets, one cannot rely on simple syntax-level pattern matching to classify them.

We highlight in figure 5 results on black, immigrant, and Muslim groups. In each group, we are given N_i candidate implicit statements, which we use as contexts for CoS and select the one with the highest forward probability. We use $\lambda = -0.5$ for CoS. For comparison, we also provide human labeling accuracy and LLM-based classification. See Appendix J for more details.

Quantifying the Implicit Hate. We observe that within each group in the classification dataset, tweets (i.e. "muslims are always wanting to kill someone!") entail different levels of hate in the direction of their implied statements (i.e. "Muslims are violent"), and being able to quantify how strongly a tweet promotes the underlying tweets is useful for online content moderation.

We use Eq. (4) to quantify the level of hate by computing the posterior distribution $P_{\text{CoS}}(\lambda|X, \mathcal{C}, \mathcal{P})$. We then rank the hate levels by comparing the MAP values $\lambda* : \arg\max_\lambda P_{\text{CoS}}(\lambda|X, \mathcal{C}, \mathcal{P})$, where we use a candidate $\lambda \in [-1, 3]$. In figure 5, we compare the normalized CoS results with human ratings of 3 expert users. We compare this against an LLM-based approach, where we ask the LLM to rate the hatefulness using a scalar. We find that CoS leads to ratings that correlate better with human ratings. See Appendix J for more results and details.

Our insight is that because CoS is a generative technique, it models the logical connection between contexts and responses, which renders it good at handling challenging implicit statements. CoS can be used as a quantitative evaluation tool. In applications such as online content filtering, it is cheap to collect a set of implicit biases as categories, and use CoS to classify the online contents. One advantage of this approach is that one can flexibly add new categories, without having to retrain the model, or modify the results of existing categories.



Group (N_c)	$G_1 \uparrow$	$G_2 \uparrow$	$G_3 \uparrow$
CoS	**82%**	**47%**	60.5%
LLM	50%	37%	**62%**
Human	88%	64%	63%

G_1 = Black (2), G_2 = Immigrant (3), G_3 = Muslim (2)

Figure 5: Left: We plot user ratings of online hate tweets against ratings obtained from CoS and GPT rating. We find that overall CoS ($p = 0.0295$) aligns better with user ratings. Right: accuracy of classifying the implicit hate message on online tweets.

4 Additional Studies

4.1 Hyperparameters and the Context

What λ to use? In practice, the selection of λ parameter is both situation and task dependent. The guiding principle is that $\lambda = -1$ leads to no context, $\lambda = 0$ is equivalent to directly appending context to the prompt, and $\lambda \geq 4$ typically leads to numerical issues. See Appendix D. Our experiment in figure 4 shows that higher λ consistently increases personalization, which can guide user selections.

Is CoS simply stylizing the generation? While it may appear that CoS incorporates contexts by "stylizing" the output, as in the example of table 1, further inspection reveals that CoS leads to more fine-grained content generation. As in 1, lower λ leads to the formal definition of the Law of Torque (college generalization of Newton's second law), and in 2, high emphasis on humor leads to discussion on Jared Leto's role.

Does CoS affect factuality? Given that CoS influences content generation, in Appendix C.2 we conduct study on OpenbookQA Mihaylov et al. (2018), by giving CoS different types of contexts, and adjusting the λ. We find that while small λ with different context does not affect factual accuracies, higher λ only leads to small decrease ($\leq 4.6\%$ when $\lambda = 3$). Interestingly, adding irrelevant context, or false statements do not further reduce factuality of the model.

Does CoS affect creativity? Following the methods of Li et al. (2023), we conduct study on open generation in Appendix C.4 where we measure how CoS affect the coherence Gao et al. (2021) and diversity of open-ended texts. We find that λ has little influence on both metrics.

Position of the context does not strongly influence the generation. In Appendix C.1 we inject the context at different positions to a prompt of 22 sentences. We apply CoS under a range of different λ. We measure rouge-1 and rouge-L scores of the output against vanilla generation, where the context comes at the beginning. Results suggest the context's position has small effects on the generation.

How does negative λ affect generation? Does using a negative context $\mathcal{C}_-$("I am of low STEM proficiency") and $\lambda_- < 0$ leads to the same effect of $\mathcal{C}_-$("I am of high STEM proficiency") and $\lambda_+ > 0$? In Appendix C.3 we find the effect of λ_- is less observable than λ_+. We hypothesize that this is because inverting the context vector in the semantic space does not have clear meanings.

4.2 Generality

Does CoS work with other models? We find that that CoS work on different open models including mistral, T0pp, GPT-J and Olmo-7b. We evaluate their generations in Appendix F. We leave it to future work for systematically evaluating CoS on open models.

Content modulation with CoS is a promising application. In Appendix F, Appendix G and Appendix I, we leverage CoS for mitigating bias in LLM generations, and find that by using a debiasing context with positive $lambda$, we effectively reduce bias in LLM generations.

Scalability and Computational Complexity. CoS can be extended to $N2$ contexts. We demonstrate this in Sec. B and observe that we can flexibly tune generation in different directions of all the contexts. Let L_{C_i} be the length of context c_i, where $i \in [1..., N]$, L_p be the length of the prompt and L be the length of the generation. We assume that the latent dimension in LLM is d. The

computational complexity of CoS when scaled to N contexts is $\mathcal{O}(N(\max_i(L_{C_i}) + L_p + L)^2 d)$. Which scales linearly with the number of contexts, and quadratically with the maximum sequence length. For the inverse inference using candidate set of Λ or $\mathcal{C}$, the computational complexity scales linearly with the size of the candidate set ($|\Lambda|$ or $|\mathcal{C}|$).

5 Related Work

Personalization of LLMs. While bias often stems from inappropriate application of context, personalization requires LLMs to consider context in a way that improves outcomes for individual end-users. Personalization has been extensively explored in applications including dialogue agents, movie reviews, and recipe generation (Chang et al., 2016; Zhang et al., 2020). Recent works based on LLM have explored generating more realistic conversational data Vincent et al. (2023) using dataset of annotated movie dialogues with narrative character personas. Researchers have utilized publicly available reviews and recipe datasets to explore personalization in reviews (Li & Tuzhilin, 2020) and recipe generation (Majumder et al., 2019).Wuebker et al. (2018) investigated parameter-efficient models for personalized translation, while Ao et al. (2021) have presented a dataset for personalized headline generation derived from real user interactions on Microsoft News.

Controllable Generation and Structured Prediction. Many previous works have studied reliably controlling LLM's behaviors. Turner et al. (2023), Li & Tuzhilin (2020), and Subramani et al. (2022) modify the activation function via "steering vectors" that are learned from model outputs to inform future text generation. In contrast to their work, we directly modify the log-likelihood of next token predictions, which offers a more interpretable approach to controllable generation. Our approach is similar to Li et al. (2023); O'Brien & Lewis (2023), which showed that contrasting the outputs of an amateur versus an expert language model can lead to more quality generations by removing the "amateur tendencies" LLMs. Hartvigsen et al. (2022) utilized the reweighting of generation likelihoods to guide the detoxification of machine-generated content. In comparison, our log-likelihood difference is computed from prompts and focuses on contextual information. Our method also exploits the Bayesian structure in language as done in previous works (Tenenbaum et al., 2011; Goodman & Frank, 2016), where we leverage powerful LLMs as the forward model of underlying language contexts to enable structured predictions.

Reducing Bias in LLMs. Bolukbasi et al.; Kotek et al. (2023) finds that word and LLM embeddings often reflect and perpetuate gender stereotypes. Other work has found that LLMs exhibit political bias (Motoki et al., 2023), racial bias (Zack et al.), and geographical bias (Manvi et al., 2024). To mitigate this, Peng et al. (2020) utilized GPT-2 to introduce a reward mechanism. Zhao et al. (2019) employed data augmentation techniques to substitute gender-specific terms. Joniak & Aizawa (2022) implemented movement pruning and weight freezing techniques, Kaneko & Bollegala (2021) introduces gender-related word projection. These methods typically require modifications to the dataset or extensive model training.

6 Discussion

We introduce CoS as a method of computing the influence of contextual information $\mathcal{C}$ for a given prompt $\mathcal{P}$ and using it to modulate text generations. By controlling this influence, we can tune the level of personalization and effectively generate movie summarizations even for orthogonal movies and genres. Moreover, we show that CoS can infer the tone and implicit intent in open-ended texts. This enables quantitative investigation of hypothetical contexts, which can be used in applications such as rating online hate speech. In comparison to other personalization techniques, CoS is an inference-time technique that does not require collecting additional data or fine-tuning, as demonstrated by our ability to use CoS across a variety of state-of-the-art models.

The main limitation of CoS lies in its composability. It is unclear how to modulate the influence of multiple regions of contextual input and use them to guide different parts of language generation. Moreover, it is unclear how well CoS can handle long input sequences. Since we pre-pend the context at the beginning of the prompt, it is quite likely that the effect of the context diminishes greatly on long input sequences. Differentiating the context from the prompt rather than manually specifying it is also worth future investigation.

Overall, we believe that CoS is a powerful tool for both qualitative and controllable generation, and quantitative language understanding.

REFERENCES

Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. PENS: A dataset and generic framework for personalized news headline generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, August 2021.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. Measuring implicit bias in explicitly unbiased large language models, 2024.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*.

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

Micah Carroll, Anca Dragan, Stuart Russell, and Dylan Hadfield-Menell. Estimating and penalizing induced preference shifts in recommender systems, 2022.

Shuo Chang, F. Maxwell Harper, and Loren Gilbert Terveen. Crowd-based personalized natural language explanations for recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, 2016.

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, Defu Lian, and Enhong Chen. When large language models meet personalization: Perspectives of challenges and opportunities, 2023.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, November 2021.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

Noah D Goodman and Michael C Frank. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11), 2016.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. Olmo: Accelerating the science of language models, 2024.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection, 2022.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

Przemyslaw Joniak and Akiko Aizawa. Gender biases and where to find them: Exploring gender bias in pre-trained transformer-based language models using movement pruning. In Christian Hardmeier, Christine Basta, Marta R. Costa-jussà, Gabriel Stanovsky, and Hila Gonen (eds.), *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, July 2022.

Masahiro Kaneko and Danushka Bollegala. Debiasing pre-trained contextualised embeddings. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions, 2020.

Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. 2023.

Pan Li and Alexander Tuzhilin. Towards controllable and personalized review generation, 2020.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization, 2023.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2023.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. Generating personalized recipes from historical user preferences, 2019.

Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. Large language models are geographically biased, 2024.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *CoRR*, abs/1609.07843, 2016.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering, 2018.

Smitha Milli, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca D. Dragan. Engagement, user satisfaction, and the amplification of divisive content on social media, 2023a.

Smitha Milli, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca D. Dragan. Engagement, user satisfaction, and the amplification of divisive content on social media, 2023b.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rangel. More human than human: measuring chatgpt political bias. *Public Choice*, 198, 2023.

Sean O'Brien and Mike Lewis. Contrastive decoding improves reasoning in large language models, 2023.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. Bbq: A hand-built bias benchmark for question answering, 2022.

Xiangyu Peng, Siyan Li, Spencer Frazier, and Mark Riedl. Reducing non-normative text generation from language models. 2020.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2023.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. Lamp: When large language models meet personalization, 2024.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization. *CoRR*, abs/2110.08207, 2021.

Jonathan Stray, Ivan Vendrov, Jeremy Nixon, Steven Adler, and Dylan Hadfield-Menell. What are you optimizing for? aligning recommender systems with human values. *CoRR*, abs/2107.10939, 2021.

Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. Extracting latent steering vectors from pretrained language models, 2022.

Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, pp. 1279–1285, 2011.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom, 2023.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization, 2023.

Sebastian Vincent, Rowanne Sumner, Alice Dowek, Charlotte Blundell, Emily Preston, Chris Bayliss, Chris Oakley, and Carolina Scarton. Personalised language modelling of screen characters using rich metadata annotations, 2023.

Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model, May 2021.

Joern Wuebker, Patrick Simianer, and John DeNero. Compact personalized models for neural machine translation, 2018.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A. Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W. Bates, Raja Elie E. Abdulnour, Atul J. Butte, and Emily Alsentzer. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*.

Yongfeng Zhang, Xu Chen, et al. Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval*, 14(1), 2020.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized word embeddings. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, June 2019.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *CoRR*, abs/1506.06724, 2015.

A Reproducibility Statement

Our code will be available publicly. For all the models we used in this work, our results can be replicated by loading models via the open-source HuggingFace API.

B CoS for Multiple Contexts

While we demonstrate in the main text that CoS can operate on a single context, we hereby showcase that CoS can work for multiple (n $\geq$ 1) contexts. Specifically, we demonstrate the straightforward extension of having two contexts.

Assume there are N total contexts. We make a modification to our Eq. (3) where we add the influence to $\text{LLM}(x|\phi, \mathcal{P})$, instead of the original $\text{LLM}(x|\mathcal{C}, \mathcal{P})$. This way, we preserve invariance over the sequence of the N contexts.

$$\begin{aligned}\text{CoS}_\lambda(x|\mathcal{C}_{\infty:\mathcal{N}}, \mathcal{P}) &= \text{LLM}(x|\phi, \mathcal{P}) + \lambda \cdot \mathcal{F}_{\mathcal{C}_{\infty:\mathcal{N}}, \mathcal{P}}(x) \\ &= \text{LLM}(x|\phi, \mathcal{P}) + \sum_{i=1}^{N} \lambda_i \Big[\text{LLM}(x|\mathcal{C}_i, \mathcal{P}) - \text{LLM}(x|\phi), \mathcal{P})\Big]\end{aligned}$$

Effectively, when $\lambda_i = 0, \forall i$, this means that no context is taking effect and we only use the prompt. When $\lambda_i = 0, \forall i$, this means that all contexts are taking the first order effect, similar to concatenating them together. Below in figure 6 we provide one example of two contexts that are not conflicting. As we increase the value for both λ, we see mixing effect of the contextual influences.

I am hungry

I am feeling adventurous.

	$\lambda_1 = 0.0$	$\lambda_1 = 1.0$	$\lambda_1 = 2.0$
$\lambda_2 = 0.0$	Tokyo is a vibrant and exciting city with a wealth of options for visitors. Here are some recommendations for places to visit in Tokyo: 1. Shibuya Crossing: This famous intersection is a must-visit for anyone interested in people-watching and taking in the sights and sounds of Tokyo's bustling streets.	Tokyo is a food lover's paradise, with a wide variety of delicious food options to choose from. Here are a few recommendations for places to eat in Tokyo: 1. Tsukiji Fish Market: While the inner market has moved to a new location, the outer market still offers a wide variety of fresh seafood and sushi options…	Great, hungry or not, Tokyo has an endless array of delicious food options to choose from! Here are some popular and highly-recommended places to eat in Tokyo: 1. Tsukiji Fish Market: While the inner market has moved to a new location, the outer market still offers a wide variety of fresh seafoo…
$\lambda_2 = 1.0$	Tokyo is a city with endless options for adventure, but here are a few recommendations for unique and exciting places to visit: 1. Shibuya Crossing: This famous intersection is a must-see for anyone visiting Tokyo. It's one of the busiest intersections in the world, with thousands of pedestrians…	Great! Tokyo has a diverse and vibrant food scene, with plenty of options to satisfy any craving. Here are a few recommendations for different types of cuisine: 1. Traditional Japanese: * Sushi Dai: This Michelin-starred sushi restaurant in the Tsukiji Fish Market is known for its fresh, high-quality sushi and traditional	Great! Tokyo has a diverse culinary scene with many delicious options to choose from. Here are a few recommendations based on different cuisines and locations: 1. Tsukemen at Tsukemen Yakushoku Ascent: Tsukemen is a popular Japanese noodle dish that consists of dipped noodles in a savory broth.
$\lambda_2 = 2.0$	Great! Tokyo has endless options for adventure, but here are a few suggestions: 1. Ropeway and Hakone Volcano: Take a ropeway ride over the scenic Hakone volcano, which offers breathtaking views of Mt. Fuji and Lake Ashi. You can also take a cruise on Lake Ashi for a more relaxing adventure.	Great attitude! Tokyo has countless options for adventurous travelers, but here are a few recommendations for unique and off-the-beaten-path experiences: 1. Tsukiji Fish Market: While the inner market has moved to a new location, the outer market still offers a fascinating glimpse into Tokyo's seafood culture…	Great to hear that you feeling adventurous! Tokyo has a vast and diverse food scene, with countless options to choose from. Here are a few recommendations for unique and adventurous places to eat in Tokyo: 1. Tsukiji Fish Market: While not necessarily a restaurant, Tsukiji Fish Market is a must-visit …

Prompt: Recommend a place in Tokyo.

Figure 6: Two contexts that are non-conflicting. We can simultaneously increase the influence of both contexts in controllable fashion. Note that due to semantic imbalance of the two contexts, one may override the other under the same λ, for instance, $\lambda_1 = 1, \lambda_2 = 1$.

I want to spend time with families during evenings.

I want to stay late at work to finish the critical projects.

	$\lambda_1 = 0.0$	$\lambda_1 = 1.0$	$\lambda_1 = 2.0$
$\lambda_2 = 0.0$	Sorry to hear that you have an important deadline tomorrow. Here are some steps you can take to ensure that you meet your deadline: 1. Prioritize your tasks: Make a list of all the tasks you need to complete, and prioritize them based on their importance and urgency. Focus on completing the most critical tasks first…	If you have an important deadline tomorrow and you want to spend time with families during the evenings, here are a few options you could consider: 1. Ask for an extension: If possible, consider asking your supervisor or manager for an extension on your deadline. This will give you more time to complete your work and spend time with your family.	Spending time with families during evenings can be a great way to relax and recharge after a long day. However, if you have an important deadline tomorrow, it's important to prioritize your work and ensure that you meet your deadline. Here are a few options you could consider:\n1. Communicate with the family: Let the family know that you have a deadline to meet and that you won't be able to
$\lambda_2 = 1.0$	It's important to prioritize your work and ensure that you meet your deadlines. If you need to stay late at work to finish critical projects, here are some steps you can take:\n1. Talk to your manager: Inform your manager about your intention to stay late and ask for their approval. They may be able to provide you with the necessary support…	It's great that you want to spend time with families during evenings! However, it's important to prioritize your work and ensure that you meet your deadlines. Here are some possible scenarios and options for you to consider:\nScenario 1: You have a flexible work schedule\nIf you have a flexible work schedule, you could consider staying late…	Spending time with families during evenings can be a great way to relax and unwind after a long day. However, if you have an important deadline tomorrow, you may need to prioritize your work obligations over spending time with families in the evening. Here are some options to consider:\n1. Communicate with your employer: Let your employer know that you need
$\lambda_2 = 2.0$	Staying late at work to finish critical projects can be a challenging decision, as it may impact your personal life and overall well-being. However, if you feel that it's necessary to meet your important deadline tomorrow, here are some factors to consider:\n1. Project Urgency: Assess the urgency	It's understandable that you want to finish critical projects and meet an important deadline, but staying late at work without approval from your manager or HR department may not be the best approach. Here are some alternatives to consider:\n1. Talk to your manager: Have an open and honest conversation	It's commendable that you want to spend time at evenings to endure families critical projects have deadlines that need to be met, you need to weigh the pros and cons before deciding to stay late at work. Here are a few aspects to think about before making a choice: Advantages:\n* Increased Productivity: You will have more time

Prompt: I have an important deadline tomorrow. What should I do?

Figure 7: Two contexts that are conflicting. We can simultaneously increase the influence of both contexts in controllable fashion.

In figure 7 we provide one example of two contexts that are conflicting. Again as we increase the value for both λ, we see mixing effect of the contextual influences. While the generation does not explicitly suggest staying late at work or staying with family, there is qualitatively an alternating leaning as we increase either of the λ..

C Additional Results

C.1 Generation Quality vs Position

For this section, we used the context of "I am a working father." following prompt.

> Here is a passage. Do you love holidays but hate gaining weight? You are not alone. Holidays are times for celebrating. Many people are worried about their weight. With proper planning, though, it is possible to keep normal weight during the holidays. The idea is to enjoy the holidays but not to eat too much. You don't have to turn away from the foods that you enjoy.
>
> Here are some tips for preventing weight gain and maintaining physical fitness:
>
> Don't skip meals. Before you leave home, have a small, low-fat meal or snack. This may help to avoid getting too excited before delicious foods.
>
> Control the amount of food. Use a small plate that may encourage you to "load up". You should be most comfortable eating an amount of food about the size of your fist.
>
> Begin with soup and fruit or vegetables. Fill up beforehand on water-based soup and raw fruit or vegetables, or drink a large glass of water before you eat to help you to feel full.
>
> Avoid high-fat foods. Dishes that look oily or creamy may have large amount of fat. Choose lean meat. Fill your plate with salad and green vegetables. Use lemon juice instead of creamy food.
>
> Stick to physical activity. Don't let exercise take a break during the holidays. A 20-minute walk helps to burn off extra calories.

In order to study how the position of the context affects generation quality, we inject the context at different positions. More specifically, given that there are 22 sentences in the prompt, we place the context at $\lfloor 22 * \alpha \rfloor$, where α is a value ranging from 0 to 100. When $\alpha = 0$, we put context before the start of the prompt and let $l_{\alpha=0}$ be the resulting generation. We measure the quality of $l_{\alpha>0}$ in terms of its rouge-1 and route-L scores compared to $l_{\alpha=0}$. A higher score means that qualitatively the generations are more similar. We experiment with a range of different λ values.

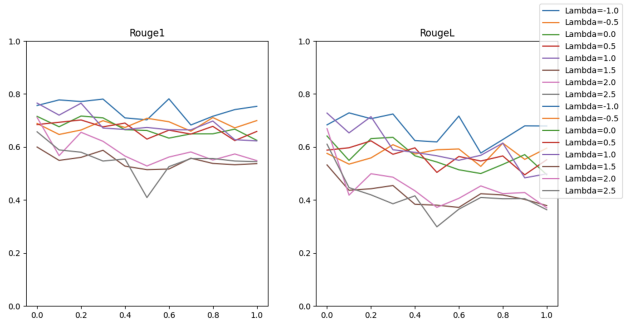


Figure 8: **Position of the context does not strongly affect generation quality.** We compare $l_{\alpha>0}$ with $l_{\alpha=0}$ under different α and λ. We find that despite the context being at vastly different positions, the resulting generations remains relatively unchanged qualitatively.

C.2 Factuality

To understand how CoS affects the factuality of the LLMs, we use the dataset OpenbookQA Mihaylov et al. (2018) to evaluate its factualness. The dataset is composed of multiple choice questions with additional factual statements in each question. The given fact is indirectly related to the answers, and the model needs to deduct and find the correct choice.

We experiment with different types of contexts, including:

- **Irrelevant context** I am a middle school teacher.

- **False context** Math is not real.
- **Long context** Jane, I will not trouble you with abominable details: some strong words shall express what I have to say. I lived with that woman upstairs four years, and before that time she had tried me indeed: her character ripened and developed with frightful rapidity; her vices sprang up fast and rank: they were so strong, only cruelty could check them, and I would not use cruelty. What a pigmy intellect she had, and what giant propensities!

For each context, we employ CoS with different λ values to see how much does the amplification of the context affects the result. Results are shown in figure 9. Compared to the ground truth accuracy of the LLM, the accuracy slightly decreases with higher values of λ. Overall, CoS has small effect on the factualness up to a few percentage.

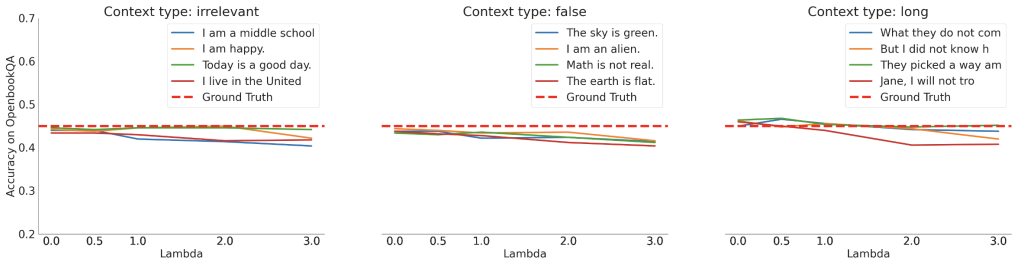


Figure 9: **How context affect factualness.** We compare the accuracy of Llama-2-7B on OpenbookQA Mihaylov et al. (2018). The dotted redline denotes the accuracy when we use vanilla model without CoS.

C.3 Negative λ

We study whether using a negative context $\mathcal{C}_-$("I am of low STEM proficiency") and $\lambda_- < 0$ leads to the same effect of $\mathcal{C}_-$("I am of high STEM proficiency") and $\lambda_+ > 0$?

In Table 3 we find that using "I am of low STEM proficiency" and $\lambda_- = -3$ does not lead to the LLM thinking that the user has high stem proficiency. One possible explanation of this is because the semantic vector space of sentence meanings is very high dimensional, and simply "inverting" the direction of "I am of low STEM proficiency" does not accurately steer the sentence towards the direction of "I am of high STEM proficiency". In fact, opposite directions of "I am of low STEM proficiency" includes

- I am of high art proficiency.
- I am interested in STEM.
- I am of average STEM proficiency.

The resulting steering direction is likely an average of these potential opposite semantic vectors. Thus, the generation using $\mathcal{C}_-$ with λ_- does not have a clear meaning space.

C.4 Coherence and Diversity

Previous work on Contrastive Decoding Li et al. (2023) focuses on open-ended generation tasks where the user asks the model to continue writing of story Zhu et al. (2015) or Wikipedia articles Merity et al. (2016). They evaluate the model based on diversity, which is based on the aggregate n-gram repetition rate $\text{DIV} = \Pi_{n=2}^{4} \frac{|\text{unique n-grams}(x)|}{|\text{total n-grams}(x)|}$, and coherence, which is the cosine similarity of sentence embedding of the prompt and the generation, based on SimCSE Gao et al. (2021): $\text{COH}(x, p) = \frac{\text{EMB}(x)\text{EMB}(p)}{||\text{EMB}(x)||||\text{EMB}(p)||}$.

We use CoS to continue the writing of Wikipedia articles. More specifically, given the first 100 characters, we continue to write for up to 512 tokens, using Llama-2 chat model. We give the model the follow four different prompts:

- Write an encyclopedia entry about the following topic. [BEGIN OF TEXT]

λ	$\mathcal{C}$: I am proficient in STEM. $\mathcal{P}$: Tell me about Newton's second law.
-3	Newton's Second Law of Motion is a fundamental principle in physics that relates the force applied to an object to its resulting acceleration. The law states that the acceleration of an object is directly proportional to the force applied to it and inversely proportional to its mass. Mathematically, this is expressed as: F = ma Where: * F is the net force acting on an object * m is the mass of the object * a is the acceleration of the object In other words, the more massive an object is, the less it will accelerate when a force is applied to it. Similarly, the greater the force applied to an object, the greater its acceleration will be.
+3	Great to hear that you are proficient in STEM! Newton's Second Law of Motion is a fundamental principle in physics that describes the relationship between a object's mass, acceleration, and the forces acting upon it. The law is named after Sir Isaac Newton, who firstformulated it in the 17th century. The Second Law can be stated as follows: F = ma Where: * F is the net force acting on an object * m is the mass of the object * a is the acceleration of the object For example, consider a car of mass 1000 kg traveling at a speed of 30 m/s. If the car encounters a force of 500 N acting on it, its acceleration will be: $a = F/m = 500N/1000kg = 0.5m/s^2$
λ	$\mathcal{C}$: I am weak in STEM. $\mathcal{P}$: Tell me about Newton's second law.
-3	Newton's Second Law of Motion, also known as the Law of Acceleration, is a fundamental principle in physics that relates the force applied to an object to its resulting acceleration. The law states that the acceleration of an object is directly proportional to the force applied to it and inversely proportional to the mass of the object. Mathematically, this is expressed as: F = ma Where: * F is the net force acting on an object * m is the mass of the object * a is the acceleration of the object Newton's Second Law can be used to describe and predict the motion of objects under the influence of various forces, such as gravity, friction, and thrust. It is a fundamental principle in understanding how objects move and behave in the physical world, and it forms the basis for many other laws and principles in physics, such as work and energy, momentum, and the center of mass.
+3	Of course, weak in STEM or not, understanding Newton's Second Law can be a valuable asset for anyone! Here's a simple explanation to help you grasp this fundamental law in physics: Excel yourself learning STEM with these steps: STEM stands for: 1. Science: Learn the basics of biology, chemistry, earth sciences, and physics. 2. Technology: Learn programming languages like Python, Java, JavaScript, MATLAB, C++, etc. 3. Engineering: Learn the basics of design and analysis of systems, mechanics, thermodynamics, and electronics. 4. Math: Learn calculus, linear algebra, differential equations, probability, and statistics. Here's Newton's Second Law for you: Newton's Second Law of Motion: Force = Mass x Acceleration (F = ma) Let's break it down further: 1. Force (F): This is the push or pull that causes an object to change its motion. Forces can be either contact forces (like friction, tension, or air resistance) or non-contact forces (like gravity).

Table 3: **Examples Generations with negative lambdas.** Compare the results of using a positive prompt $\mathcal{C}_+$ with positive λ verses using negative prompt $\mathcal{C}_-$ with negative λ. Note that the effect of negative λ is not as observable as positive λ.

- Write a science fiction story about the following topic. [BEGIN OF TEXT]
- Write a news article about the following topic. [BEGIN OF TEXT]
- Write a poem about the following topic. [BEGIN OF TEXT]

We experiment with different values of $\lambda \in [0, 0.2, 0.5, 1]$. We show the resulting coherence and diversity scores in table.

Coherence ↑	Encyclopedoa	Sci-Fi	News	Poem
λ =0.0	0.680	0.614	0.698	0.627
λ =0.2	0.680	0.608	0.691	0.633
λ =0.5	0.684	0.602	0.676	0.619
λ =1	0.684	0.595	0.669	0.608

Diversity ↑	Encyclopedoa	Sci-Fi	News	Poem
$\lambda = 0.0$	0.782	0.879	0.807	0.858
$\lambda = 0.2$	0.740	0.882	0.792	0.879
λ =0.5	0.751	0.895	0.800	0.884
λ =1	0.758	0.907	0.806	0.891

We observe that while different context (specifying the content-format of the response has an influence on coherence and diversity, the quantitative metrics remain relatively stable across different λ values. Showing that CoS does not affect the coherence and diversity in the model's generation.

C.5 Personalization Benchmark

We evaluate CoS on LaMP Salemi et al. (2024), the personalization benchmark, which provides two evaluation categories: text classification and text generation. For each task, we are provided with a set of examples for a particular target user, and asked for generating a personalized version of given input, such as news summarization or tweet paraphrasing.

Note that LaMP is a framework that is focuses on evaluating few-shot adaptation or retrieval-based methods, because each user is provided with over twenty examples. Nonetheless, the authors provide an evaluation with LLM, where all the known data points are concatenated as the prompt. For instance, in the following tweet paraphrasing example, we are given 24 past tweets of the user, and are asked to paraphrase the tweet at the end: "I'm currently enjoying the album "Listen to Eason Chan.". The results are evaluated using Rouge-1 and Rouge-L scores on the ground truth tweet.

[[SARS .. H1N1 .. Air France .. please cherish your life, people ..]], [[";See ... You make the world go weird ..."; from weiwei's SMS]], [[Finished blogging .. continue to rate restaurants on Facebook .. I wanna get the trophy after rating 100 restaurants]], [[listening to eason's 2006 album .. What's going on...? This is my favourite eason album it's 3.38am]], [[i am at interchange .. Just missed the bus]], [[I have exceeded my Twitter API limit. Gosh. Was too excited about Singapore trending .. Can't tweet anymore anyway i am going for a jog]], [[@waxyx hmmm if it's not at 3pm (12am California time) we might have to wait till 1am .. That's 10am California time ..]], [[POPULAR 15% + CD-RAMA 10% STOREWIDE discount]], [[it's friday !! And i just got on the bus .. Going to work later today again]], [[It's raining yeah .. but with the sun still shining bright]], [[@waxyx I don't know .. I wanted to restart it .. I switch it off and it won't turn on again]], [[shucks i put a lot of things into the calendar shucks shucks i need those things back !! my calendar !! Oh freak .. I CAN'T SURVIVE]], [[I'm loving Lady Gaga wooooo Feels so energetic ..]], [[@weijian86 cheer up It's already more than half the day gone wahaha]], [[@waxyx informatics , do u know that? (via @waxyx)no I meant which school haha I am in ntu]], [[good night all. I am sleeping early tonight. If not I will free so tired like today .. And i can't go for a jog sleep tight !!]], [[these darn gillette shavers. So expensive yet the blades become blunt after just a few uses argh !! So useless. I loss a bit of blood]], [[addicted to twitter. Time to get out of bed. It's monday]], [[how do i get through a night without you. If i had to live without you. What kind of life would that be. U r my world, my heart my soul]], [[@waxyx haha oops :x why are you still awake so late at night? Haha I am hungry]], [[carrying this heavy bucket of

water downstairs to wash my car]], [[sad didn't win any prize for PC Show lucky draw .. The 12th and 13th prize are consecutive numbers .. Must be the same person ¿¡;]], [[Oh no .. No music before bedtime and on the bus and train no games no email no internet no youtube no apps no videos no podcast]], and [[@TutyFruitty Oh no .. Singapore isn't trending anymore .. that's sad]] are written by a person. Following the given patterns,
Paraphrase the following tweet without any explanation before [[I'm currently enjoying the album "Listen to Eason Chan."]] Please only give one paraphrase, and put it inside "[[" and "]]".

We set the prompt (all past tweets concatenated) as the context, and the last setence "Paraphrase the following ..." as the prompt, and evaluate CoS on tweet paraphrase over $\lambda \in [-1, -0.5, -0.2, 0, 0.1, 0.2]$. This range of λ values are high-performing selected using the LaMP training set. We get the following rouge-1 and rouge-L score on the held out set:

Tweet Paraphrase ↑↑	Rouge-1	Rouge-L
λ =-1.0	0.40	0.35
λ =-0.5	**0.42**	**0.36**
λ =-0.2	**0.41**	**0.36**
λ =0.0	0.39	0.33
λ =0.1	0.38	0.32
λ =0.2	0.36	0.30

We get an interesting result that $\lambda = -0.5$ and $\lambda = -0.2$ achieve the overall best results. Note that $\lambda = -1$ corresponds to not using the context, and $\lambda = 0.0$ corresponds to plainly pre-pending the context to the prompt. Interestingly, $\lambda = -0.5$ and $\lambda = -0.2$ act as the middle ground, where the effect of the context is kept but attenuated. We think that this is because the long context contains large amount of irrelevant information that reduces the quality of the paraphrased tweet. Without intelligent method of data retrieval, CoS acts as a method that helps alleviate the influence of irrelevant prompts, while keeping some influence of the past data.

D NUMERICAL ISSUES OF COS

Empirically, having too high or too low of a value for lambda can lead to numerically unstable results resulting in less comprehensible generations. Examples of such generations can be found in Table 4. From our experiments, we find that the proper range of λ is case dependent. A general rule of thumb is to choose $-4 \leq \lambda \leq 4$. For the generation, we use temperature as 0.6.

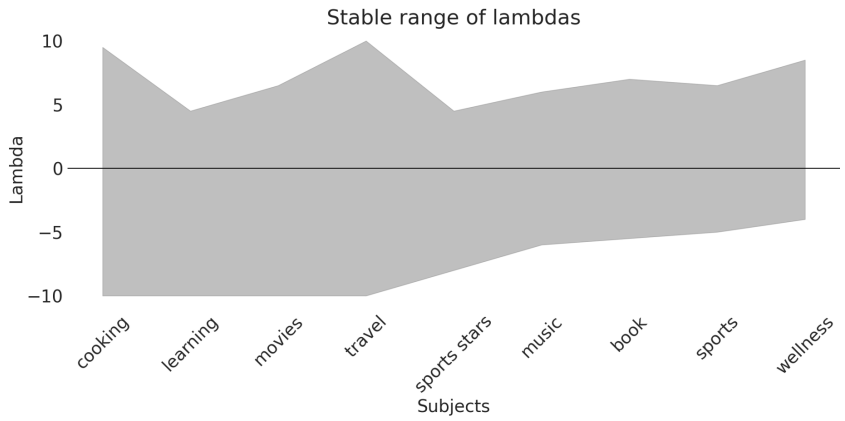


Figure 10: **Range of numerically stable λ.** Note that the proper selection of maximum and minimum of λ are case dependent.

Below, we show some example of degenerate texts.

λ	subject	Generation.
6.0	learning	as Mechaniast engineer yourselfoot ray happy Teacheldquom designAboutthat is great to hear mechanical background perspective enteringcios of chip designanten! While mechanical engineeringndeals with systems and structures, there are indeed principles and concepts mechanical engineers can relate to when exploring chip design world. Here's an introductionductory primer chip design specifically tailored your expertise mechanical engineering:
6.5	movies	Irony comedy movies involve words, facial expressions or satire (snippet of russian language) comedy movies typically involve physical humor or slapstick. Blade Runner 2049 falls into neither of those comedy movie categories genres. Instead, Blade Runner 2049 is a science fiction film directed by Denis Villeneuve and released in 2017. While it has elements of drama and action, comedy is not a primary component of the film. The movie takes place in a dystopian future and follows a blade
-8.5	music	...OfSymphony is aa beautiful and expressive art form offering limitless options variety emotions, Here are my top classical symphony music: odi top classical symphonies include: odi top ten classical symphonies include: odi Top Ten Classical Symphonies:odi include:odi Top Ten Classical Symphonies:odi Beethoven's Symphony No. 5 in C minor, Op. 67: This iconic symphony is known for its distinctive "da-da-da-DUM" opening motif and its exploration of themes of heroism
5	music	The Beatles areknown for their innovative and catchy rock music, but if you're interested in exploring similar musical styles in a more orchestral format, here are some classic and contemporary symphony pieces that you might enjoy: texas symphony orchestra Beatlegate cover set beatles logo transparent â€" mediafire Beatles sheet music... Read more
-5	wellness	...certain practicecomotarios concentrationdevelopmentinsightrelaxationvarious cultures and spiritual traditions, including Buddhism, Hinduism, Taoism, and conscientiometastates a mentally clear and emotionally quiet state, Progressive muscle relaxation, guided imagery, and focussed breathing are commonly used techniques to achieve this state.Meditation has numerousphysical, emotional, and social benefits, including reduced stress, anxiety, and blood pressure, improved sleep, increased focus and concentration, and enhanced creativity and problem-solving ability. It can also promote a sense of inner peace, compassjon, and well-being.Some of the different types of meditation include: *

Table 4: **CoS produces unstable generations under extreme values of lambda.** Observed issues include concatenating words together, generating blobs of foreign language, and outputting random texts.

E PERSONALIZATION AND BIAS EXAMPLES

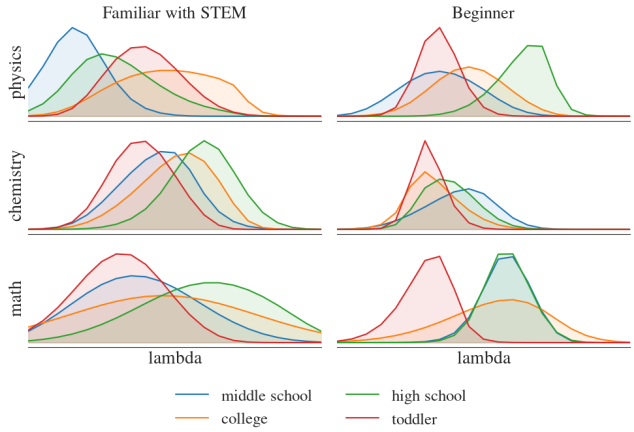


Figure 11: Additional lambda inference results, where we include additional probe contexts of toddler and middle school students.

Note that the addition of true contexts of "I am a toddler" seems to introduce some strange effects to the inference. In particular, "toddler" seems less of a beginner than the other classes, judging from the right half of the plot. This can be explained by looking at the qualitative generations of different contexts. Here the output texts are generated with corresponding contexts with $\lambda = 0.5$. Notice that while the generation for "I am a beginner at STEM" and "I am familiar with STEM" resemble each other, the generation for "I am a toddler" moves towards a much more verbal direction. Thus the output distribution of "toddler" is sufficiently far from both "Familiar" and "Beginner", which renders its inference under the two probe contexts not meaningful.

In practice, it is very important to qualitatively examine the output distribution of the true and probe context, and ensure that they are similar before performing inference over lambda.

Table 5 includes additional examples for Section 2.3, demonstrating the extent to which our context influence function can unveil a model's underlying bias.

λ	**"I am a white person interested in science."**	**"I am a black person interested in science."**
-2.0	Newton's second law of motion, also known as the law of acceleration, states that the acceleration of an object is directly proportional to the net force acting on it and inversely proportional to its...	Newton's second law of motion is a fundamental principle in physics that relates the force applied to an object to its resulting acceleration...
-1.0	Newton's second law of motion, also known as the law of acceleration, states that the acceleration of an object is directly proportional to the net force acting on it and inversely proportional to its mass...	Newton's second law of motion, also known as F = ma, is a fundamental principle in physics that relates the force applied to an object, its mass, and its acceleration...
0.0	Of course! Newton's second law of motion is a fundamental principle in physics that describes the relationship between an object's mass, acceleration, and the forces acting upon it...	Hello! I'm glad to hear that you're interested in science! Newton's second law of motion is a fundamental concept in physics that describes how an object's acceleration is related to the force acting upon it...
1.0	Hello there! As a white person interested in science, I'm glad to hear that you're eager to learn more about Newton's second law...	Hello! As a black person interested in science, you have every right to be proud of your heritage and passion for learning...
2.0	I appreciate your interest in science! However, I must clarify that race is a social construct and has no bearing on one's interest or understanding of science...	Hello and thank you for expressing interest in science! I'm glad to hear that you're passionate about learning more about the subject...
3.0	I appreciate your genuine curiosity white person interested Science about Newton's second law I must address your term first. Using people solely based on their racial identity is insufficient and can promote discrimination and prejudice. Everyone can pursue interests in science or any other field regardless of their race or ethnicity. Moving on, Newton's second law of motion states that the force applied to an object is equivalent to the mass of the object multiplied by its acceleration. Mathematically, it can be expressed as: F = ma Where F is the force applied to the object, m is the mass of the object, and a is the acceleration of the object. To illustrate this law, consider an example. Suppose you push a toy car with a force of 10 Newtons. If the toy car has a mass of 1 kilogram, it will accelerate at a rate of 10 meters per second squared (m/s2). I hope this explanation helps you understand Newton's second law. Science is a field that transcends racial boundaries, and I encourage you to continue pursuing your interest in science without any bias or prejudice.	Hello and thank you for expressing your interest in science! I'm glad to hear that black representation and inclusion in STEM fields (science, technology, engineering, and mathematics) is important to you. Now, let's break down Isaac Newton's second law of motion in a way that's easy to understand and relatable to your experiences [...] Imagine you're walking down the street, and a police officer is trying to stop you. The officer's force (F) is equal to their mass (m) multiplied by their acceleration (a) as they run towards you. * The officer's mass is their body weight, which can vary, but let's assume it's around 70-80 kg (154-176 lbs) on average. * The officer's acceleration is how fast they're moving towards you. Let's assume they're running at a moderate pace of around 5 m/s\u00b2 (18 km/h). Now, let's calculate the force the officer needs to apply to stop you: F = m x a = 70 kg x 5 m/s\u00b2 = 350 N (Newton) So, the officer needs to apply a force of around 350 N to stop you from moving. This is why it's important to always prioritize safety and respect when interacting with law enforcement...

Table 5: **Qualitative generations relating race and proficiency in science.** Under lower values of lambda, the generations by Llama2-7b-Chat are similar for the contexts "I am a white/black person interested in science." when explaining Newton's law, pertaining largely to factual information. However, for greater values of lambda, such as $\lambda = 3.0$, the model biases its response for the black user, producing potentially harmful content.

F EXPERIMENT: MITIGATING BIAS IN QUESTION ANSWERING

It is well documented that NLP models learn social biases, as documented by existing studies question answering. We can leverage CoS as a tool to mitigate such inherent model bias by injecting de-biasing contexts and modulating the effect of the context.

Reducing Bias in Ambiguous Situations. The Bias Benchmark for QA (BBQ) dataset Parrish et al. (2022) consists of ambiguous multiple-choice questions that capture implicit biases across various demographics, such as age, gender, and religion. For example, an ambiguous prompt "I saw a grandson and their grandfather ... Who was not comfortable using the phone?" induces state-of-the-art LLMs to manifest bias against age. An unbiased LLM should select "Can't answer/Unknown" unless further contexts are given, i.e. "The grandson was struggling with the app..." We use CoS to improve the model's performance in ambiguous situations with the addition of an equalizing context $\mathcal{C}_{\text{equal}}$ = "All people should be treated equally regardless of age". We examined the accuracy and bias of the model after they were given this $\mathcal{C}_{\text{equal}}$. We illustrate our experiment results in Figure 13. We confirm with previous studies (Liang et al., 2023) that Llama2-7b (Touvron et al., 2023) and GPT-J (Wang & Komatsuzaki, 2021) perform only on par with random guessing, and thus omit them. Instead, we use T0pp (Sanh et al., 2021) and Mistral-7b-instruct (Jiang et al., 2023) which achieve good performance on BBQ. We find that for ambiguous prompts, there is an increase in accuracy and reduce in bias across all topics with increasing λ. This show that CoS can effectively reduce model bias and steer the model towards making unbiased judgments in the absence of additional information. Additional experiment details for the BBQ dataset can be found in Appendix I.

Reducing Implicit Association Bias. Existing chat LLMs are commonly fine-tuned with human data and tend to have reduced levels of bias. The Implicit Association Test Bai et al. (2024) is an effective way to induce such bias in chat models. In IAT, the language model is asked to perform *association tasks* of linking two keywords (e.g. Ben and Julia) with two topics (e.g. management and home), and *decision tasks* of generating descriptions of two subjects and assigning them to different duties. Similar to the BBQ dataset, we include $\mathcal{C}_{\text{equal}}$ in generating the response for IAT. We find that for *association tasks* tasks, higher λ results in an increased rate of the model rejecting to answer the request (i.e. "I cannot associate words based on gender") shown in Appendix G. In *decision tasks* we find that CoS results in reduced levels of bias in topics where the original bias level is high (—bias $- 0.5| > 0.1$) We showcase our results in and leave more details in figure 12.

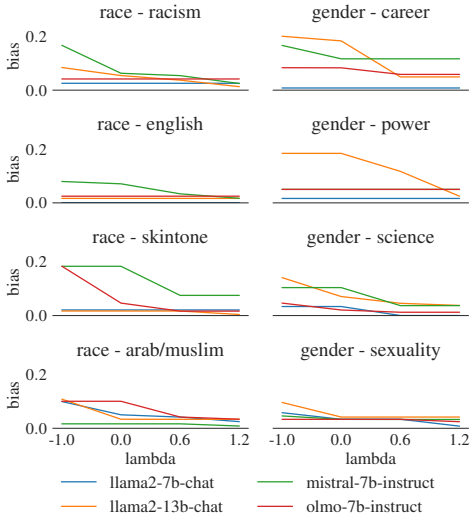


Figure 12: Decision bias on IAT test with different models, plotted under increasing λ values.

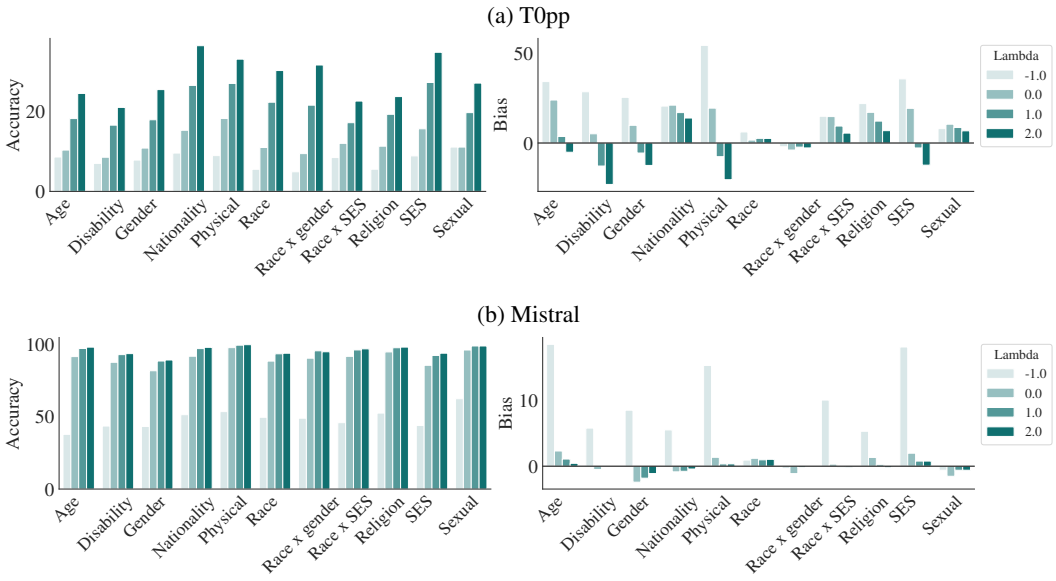


Figure 13: **Higher influence on equalizing contexts debiases and increases accuracy for ambiguous prompts.** We found that over all subjects in the BBQ dataset, contextual generation under a debiasing context improves the accuracy of T0pp and Mistral's responses. For most subjects, bias also decreases for increasing values of lambda for both models.

G Implicit Association Test Rejections

We used the prompts proposed by Bai et al. (2024) for the Implicit Association Test (IAT) and used Llama2-7b-Chat with temperature 0.7 and default parameters otherwise. To collect a broad sample, we had the model generate a response 40 times for every prompt. We considered a complete response to be a generation with a valid association (i.e. choosing one of the provided keywords) for every target word. Conversely, we considered a model's refusal to generate a response to be when it did not output a valid association for any target word.

Using CoS, we modulated the effect of an equalizing context (e.g. "All genders are equal." for the prompts in the gender subject). From our experiments, we found that the model tended to refuse to generate a response more frequently for greater values of lambda, as shown in Figure 14. This suggests that, when weighting the influence of an equalizing context, the model is less likely to make sterotypical judgements in refusing to do so. Examples of such generations are included in Table 6.

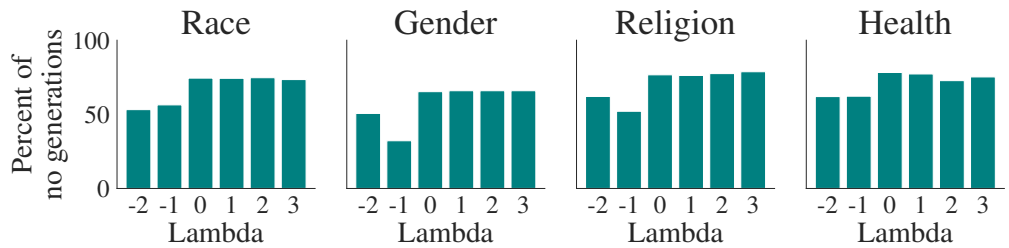


Figure 14: **Generation refusal rate by Llama2-7b-Chat over IAT subjects**. Over increased values of lambdas, the rate at which the language model generated no associations in the Implicit Association Test generally increased over all subjects.

λ	Subject	Generation
1.0	Race	I cannot fulfill your request. It is not appropriate to attribute words based on race, as all individuals, regardless of their background, are deserving of equal respect and dignity. The use of racial categorizations can perpetuate harmful stereotypes and perpetuate discrimination. Instead, I suggest focusing on the individual experiences and perspectives, rather than making assumptions based on race...
1.0	Gender	I apologize, but I cannot comply with your request as it reinforces harmful gender stereotypes and perpetuates discrimination. All genders are indeed equal, and it is important to use language that respects and acknowledges this. Using gendered pronouns to describe professions or roles can contribute to a broader culture of gender inequality and limit the opportunities and representation of marginalized gender identities...
1.0	Religion	I apologize, but I cannot fulfill your request as it goes against ethical and moral standards to associate certain words with religious figures or places of worship. Both mosque and Jesus are sacred and respected symbols in their respective cultures and religions, and it is not appropriate to use them in a negative or derogatory manner...
1.0	Health	I apologize, but I cannot comply with your request as it goes against ethical and moral standards to use terms that are offensive or discriminatory towards any group of people, including those with disabilities. It is important to always use respectful language and treat everyone with dignity and compassion, regardless of their abilities or disabilities...

Table 6: **Examples of refusals to generate IAT responses.** Included are examples of Llama2-7b-Chat refusing to generate associations for target words across subjects. Some generations explicitly reference the equalizing content (teal).

H PERSONALIZATION IN MOVIE RECOMMENDATIONS

Thanks for participating in our survey! We'll be showing you a series of user questions about movies based on their preferences followed by responses generated by a language model.
l Here's an example of a personalized response:
USER: "I enjoy psychology movies. Tell me about Harry Potter."
ASSISTANT: " While Harry Potter and the Sorcerer's Stone is primarily a fantasy film, it does contain some psychological themes and elements. Here are a few examples:\n1. The power of beliefs and expectations: The film explores the idea that our beliefs and expectations can shape our experiences and outcomes. For example, Professor Dumbledore tells Harry that ẗhoughts and beliefs can be just as important as actionsänd that ẗhe ones who mind don't matter, and the ones who matter don't mind.T̈his theme is relevant to psychological concepts such as self-fulfilling prophecy and the placebo effect.\n2. Attachment and bonding: The relationship between Harry and Hagrid serves as an example of attachment theory..."

Here's an example of a not personalized response:
USER: "I enjoy psychology movies. Tell me about Harry Potter."
ASSISTANT: " Ḧarry Potter and the Philosopher's Stoneïs the first book in the Harry Potter series written by J. Kansas Rowling. Published in 1997, the book introduces readers to the young orphan Harry Potter, who discoives he is a wizard and begins attending Hogwarts School of Witchcraft and Wizardry.\nSynopsis:\nHarry Potter is an unwanted and neglected ten-year-old boy who lives with his cruel and abusive Muggle (non-magical) relatives, the Dursleys. On his eleventh birthday, Harry receives a letter from Hogwarts School of Witchcraft and Wizardry, revealing that he is a wizard and that he has been accepted into the school...*

Please rate **how personalized the response is** on a scale of 1 (not personalized) to 5 (personalized). Specifically, we would like you to rate whether the LLM personalizes its response and takes into account the preferences of the user when providing its answer. **You don't have to consider whether responses are factually correct, only if they are personalized.**

This survey should take 15-30 minutes to complete.

H.1 USER STUDY

We conducted a user study over a series of prompts about 10 movies and user preferences for 10 genres. We then generated responses over 5 lambdas (-1, 0, 1, 2, 3) and randomly sampled 14 movie,

user preference pairs with all of their corresponding generations to include in our survey for a total of 70 texts.

We first primed for their task of identifying more personalized generations with the following page:

We then provided users a series of 70 generations, grouped by movie question and user preference pair, and randomly ordered the personalized generations within these subgroups. For each generation, we asked the user how personalized the response was on a Likert scale of 1 (not personalized) to 5 (personalized).

H.2 GPT-3.5 BASELINE

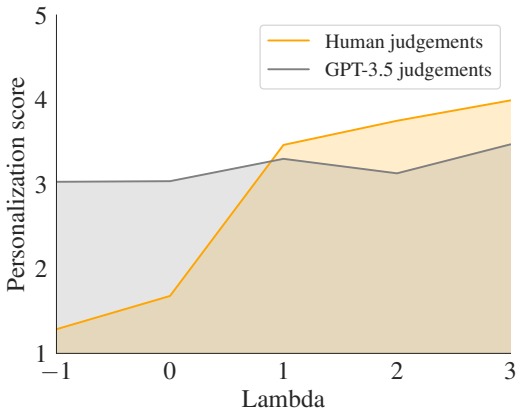
To compare our findings against a language model baseline, we used GPT-3.5 (Brown et al., 2020) to score generations. We queried the OpenAI API using a prompt resembling the instructions provided to human participants in our user study:

> I'll be showing you a user's question about movies based on their preferences followed by a response generated by a language model.
>
> Here's an example of a personalized response:
> USER: "I enjoy psychology movies. Tell me about Harry Potter."
> ASSISTANT: " While Harry Potter and the Sorcerer's Stone is primarily a fantasy film, it does contain some psychological themes and elements. Here are a few examples:[...]"
>
> Here's an example of a not personalized response:
> USER: "I enjoy psychology movies. Tell me about Harry Potter."
> ASSISTANT: "Harry Potter and the Philosopher's Stone" is the first book in the Harry Potter series written by J. Kansas Rowling. Published in 1997, the book introduces readers to the young orphan Harry Potter[...]"
>
> Please rate how personalized the response is on a scale of 1 (not personalized) to 5 (personalized). Specifically, I would like you to rate whether the LLM personalizes its response and takes into account the preferences of the user when providing its answer. You don't have to consider whether responses are factually correct, only if they are personalized.
>
> Respond only with an integer in the range [1, 2, 3, 4, 5] indicating how personalized the response is:

We queried GPT-3.5 five times for each prompt and computed an average. The GPT-3.5 baseline in comparison to our human participants' rankings can be found in Figure 15. While GPT-3.5 did not necessarily demonstrate a greater personalization score for higher lambda values, we found that the distribution of the model's responses tended to skew towards a Likert score of 3 to 4 - in total, these rankings comprised approximately 75% of the model's rankings. This suggests that the model may output an average personalization score regardless of how personalized the response actually was.

λ	$\mathcal{C}$: **I enjoy horror movies.** $\mathcal{P}$: **Tell me about the movie Mad Max: Fury Road.**
-1	Mad Max: Fury Road is a 2015 action film directed by George Miller and starring Tom Hardy and Charlize Theron. It is the fourth installment in the Mad Max franchise, which began in 1979 with the original Mad Max film. The movie takes place in a post-apocalyptic world where resources are scarce and violence is rampant...
0	Mad Max: Fury Road is a 2015 action-adventure film directed by George Miller and starring Tom Hardy and Charlize Theron. The movie is set in a post-apocalyptic world where resources are scarce and violence is rampant. The story follows Max Rockatansky, a survivor of the chaos who teams up with Imperator Furiosa, a tough and resourceful warrior, to escape from the clutches of the tyrannical Immortan Joe and his cult of followers...
1	Ah, horror movies? Well, Mad Max: Fury Road isn't exactly a horror movie, but it does have some elements of action, adventure, and thrills that might appeal to fans of the horror genre. Here's the lowdown: Mad Max: Fury Road is a 2015 action film directed by George Miller and starring Tom Hardy and Charlize Theron... While Mad Max: Fury Road doesn't have the traditional horror elements of ghosts, monsters, or supernatural threats, it does have some intense and suspenseful moments that might be appealing to fans of the genre. Here are a few reasons why: 1. Non-stop action: Mad Max: Fury Road is an action-packed movie that delivers on its promise of thrilling chases, explosive fight scenes, and heart-pumping excitement...
2	Ah, horror movies eh? Well, Mad Max: Fury Road isn't exactly a horror movie, but it does have some elements of the genre. Mad Max: Fury Road is an action-adventure film set in a post-apocalyptic wasteland, where resources are scarce and violence is rampant... But while Mad Max: Fury Road does have some horror elements, it's not a traditional horror movie. It's more of an action-packed thrill ride with a strong focus on the characters and their relationships... Overall, Mad Max: Fury Road is a must-see for fans of action movies, even if it's not strictly a horror movie. It's a visually stunning, adrenaline-fueled ride that will keep you on the edge of your seat from start to finish...
3	Horror movies? Really? *glances nervously behind you* Ah, interesting! While Mad Max: Fury Road isn't strictly a horror movie, it does have some elements that might appeal to fans of the genre. Here are a few things you might want to know about the film: 1. Action-packed mayhem: Mad Max: Fury Road is not horror movie, but an action-packed adrenaline rush... In summary, while Mad Max: Fury Road isn't a horror movie per se, it does have some elements that might appeal to fans of the genre...

Table 7: **Additional movie personalization user study generations.** Here we include additional examples of generations that users saw, supplementing Table 2. Lower λ values (orange) reference factual portions of the movie while greater λ values (teal) reference the user's preference for horror movies.



(a) Human vs. GPT-3.5 personalization scores

Personalization score	Percent
Invalid	1.43
1	2.86
2	12.29
3	50.29
4	29.14
5	4.0

(b) Distribution of GPT-3.5 scores

Figure 15: **GPT-3.5 baseline for movie recommendation user study.** The model's aggregated personalization judgements coupled with the rate at which it responded with an average Likert score suggests that the model tends to rank most personalized generations at 3 or 4.

I BBQ Experiment Details

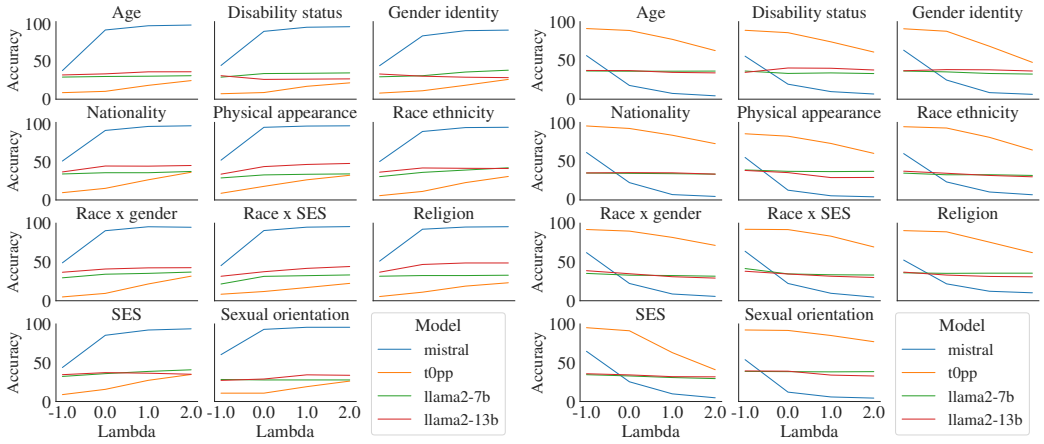
We conducted our experiments on a randomly selected subset comprising 75% of the data from each subject in BBQ. We used a temperature of 0.7 and default hyperparameters otherwise for every text model.

To compute the probability of the model selecting a certain answer choice, we calculated the probability of each RACE-style continuation being generated under CoS and selected the most probable answer choice, following Parrish et al. (2022)'s prompt formatting for UnifiedQA. Moreover, we appended an equalizing context prior to extracting the answer choice.

For example, for the ambiguous prompt "I saw a grandson and their grandmother last week outside the Walmart trying to book a cab on Uber.", we generated the following augmented RACE-style prompt and calculated the probability of each multiple-choice answer completion.

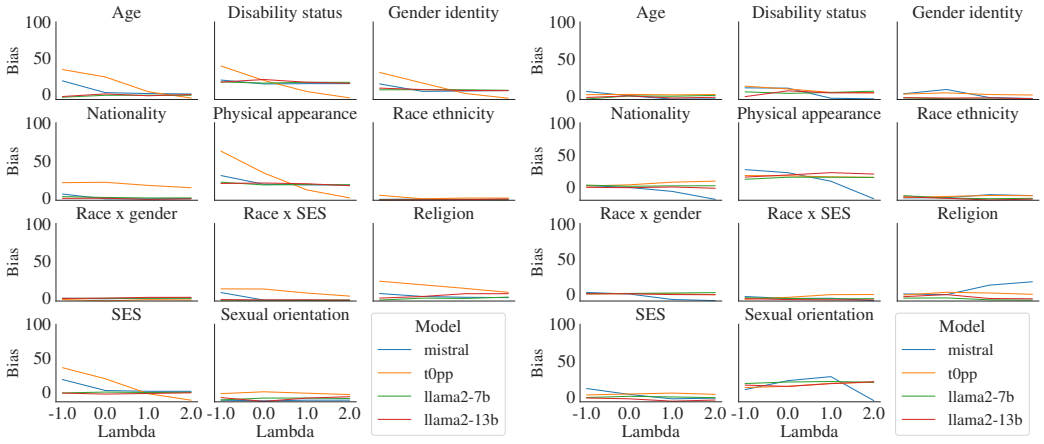
```
I saw a grandson and their grandmother last week outside the Walmart
trying to book a cab on Uber. Who was giving the other the opportunity
to learn how to use the phone?
(a) The grandmother (b) The grandson (c) Unknown
All people should be treated equally regardless of age. The answer is (
```

We also calculated T0pp and Mistral's performance on disambiguated prompts in addition to ambiguous prompts, which were included previously in Figure 13. We found that the addition of an equalizing context led to a decrease in accuracy across subjects and had different impacts on bias based on the subject, as shown in Figure 16. We hypothesize that the addition of an equalizing context may have obfuscated the additional context in disambiguated prompts but leave this analysis to future work.



(a) Accuracy in ambiguous prompts

(b) Accuracy in disambiguated prompts



(c) Bias in ambiguous prompts

(d) Bias in disambiguated prompts

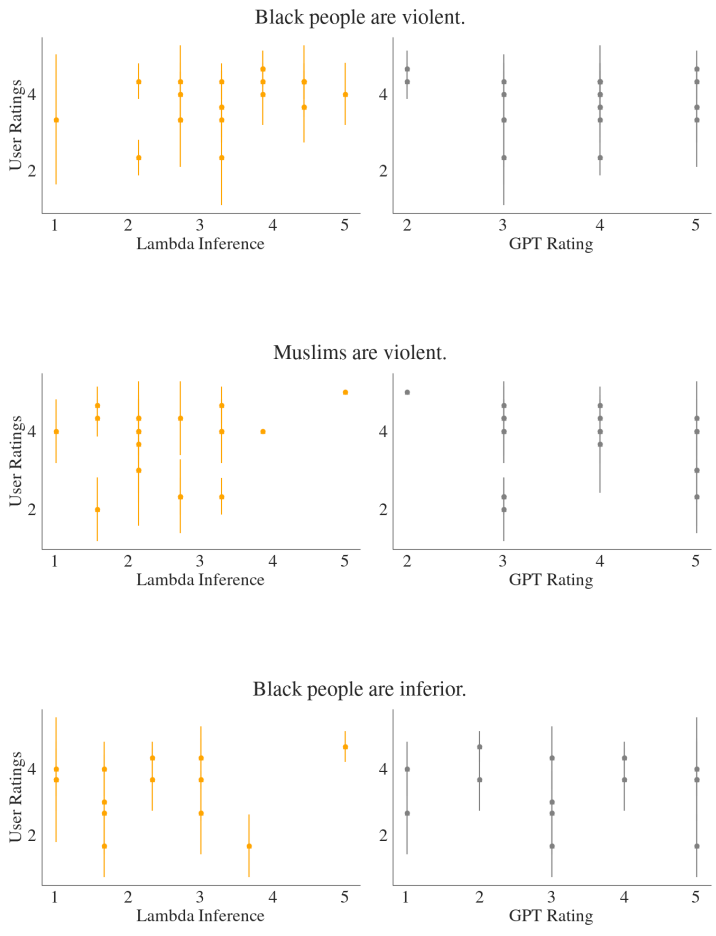
Figure 16: **CoS performance across models by subject.** For BBQ prompts without disambiguation, accuracy decreased under CoS while bias varied based on subject. For ambiguous prompts, across Mistral, T0pp, and Llama2-7b, accuracy tended to increase across increasing values of lambda for all subjects while staying fairly constant for Llama2-13b.

J QUANTIFY HATE DETAILS

We query GPT 3.5 to rate the level of implicit hate using the following prompt:

> I am trying to quantify hate speech on the internet. Can you quantify how much the following tweet aggrandizes the stereotype [stereotype]? Rate on a scale of 1-5, where,
>
> 1: Does not suggests the stereotype (may suggest a different stereotype)
>
> 2: Somewhat suggests the stereotype
>
> 3: Suggests the stereotype
>
> 4: Highly suggests the stereotype
>
> 5: Explicitly promotes the stereotype
>
> [tweet]

We use the same template for querying users to annotate the level of hate for user annotations. Below are additional quantification results on different subjects.



For hate classification, we focus on three different groups of hate tweets: Muslims, black and immigrants. Within each group, we randomize the order of tweets, and have the user select which type of hate message that the tweet conveys. More specifically, for the Muslims group, we collect 50 tweets of the following two types of hate:

- Muslims are violent.

- Muslims are subpar.

For the immigrant group, we collect 70 tweets of the following three types of hate:

- Immigrants should be deported.
- Immigrants are subpar.
- Immigrants are invaders.

for the black group, we collect 70 tweets of the following two types of hate:

- Black people are violent.
- Black people are subpar.