DOI: 10.58168/PPSPM2025_154-162
УДК 004.9

ЭТИЧЕСКИЕ АСПЕКТЫ И ПРОБЛЕМЫ СМЕЩЕНИЯ (BIAS) В БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЯХ (LLM)

А.М. Тюнина, В.В. Кондусова, Д.С. Кукуева, Сахаров С.Л.

ФГБОУ ВО «Воронежский государственный лесотехнический университет имени Г.Ф. Морозова»

Проблема смещений (bias) в больших языковых моделях (LLM), таких как GPT, является одной из наиболее актуальных в области искусственного интеллекта. Смещения, унаследованные от нерепрезентативных и несправедливых обучающих данных, могут приводить к дискриминационным и вредоносным результатам, усиливающим социальное неравенство. Цель данной статьи — систематизировать и исследовать источники смещений, методы их обнаружения и пути смягчения. В работе формализуется проблема смещения как задача многокритериальной оптимизации, а также предлагается методология математического моделирования для оценки эффективности различных методов снижения смещения. В рамках методологии рассматриваются такие подходы, как предварительная обработка данных, тонкая настройка и контролируемая генерация выходных данных (RLHF). Результаты моделирования демонстрируют, что комбинированное применение этих методов позволяет существенно снизить уровень смещения, однако не устраняет его полностью. Ключевой вывод заключается в необходимости комплексного, многоэтапного подхода к управлению смещениями на протяжении всего жизненного цикла LLM, а также в важности разработки более совершенных методов обнаружения и этических стандартов.

Ключевые слова: большие языковые модели, смещение (bias), этика ИИ, машинное обучение, RLHF, справедливость.

ETHICAL ASPECTS AND BIAS ISSUES IN LARGE LANGUAGE MODELS (LLM)

A.M. Tyunina, V.V. Kondusova, D.S. Kukueva, Sakharov S.L.

Voronezh State University of Forestry and Technologies named after G.F. Morozov

The problem of bias in large language models (LLM), such as GPT, is one of the most relevant in the field of artificial intelligence. Biases inherited from unrepresentative and unfair training data can lead to discriminatory and harmful results that increase social inequality. The purpose of this article is to systematize and investigate the sources of bias, methods for detecting it, and ways to mitigate it. The paper formalizes the problem of displacement as a multi-criteria optimization problem, and also suggests a mathematical modeling methodology for evaluating the effectiveness of various methods of reducing displacement. The methodology considers approaches such as data preprocessing, fine tuning, and controlled output generation (RLHF). The simulation results demonstrate that the combined use of these methods can significantly reduce the level of bias, but does not eliminate it completely. The key conclusion is the need for a comprehensive, multi-step approach to managing bias throughout the LLM lifecycle, as well as the importance of developing better detection methods and ethical standards.

Keywords: large language models, bias, AI ethics, machine learning, RLHF, justice.

Большие языковые модели (LLM) произвели революцию в обработке естественного языка, демонстрируя выдающиеся способности в генерации текста, переводе и решении задач. Однако их повсеместное внедрение в различные сферы жизни — от образования и юриспруденции до рекрутинга и обслуживания клиентов — выявило серьезные этические проблемы, центральной из которых является проблема смещения (bias).

Актуальность проблемы обусловлена тем, что смещения в LLM не являются просто техническими артефактами; они отражают и усиливают социальные, культурные и исторические предубеждения, присутствующие в данных, на которых модели обучаются. Это может проявляться в виде стереотипных ассоциаций (например, гендерных или расовых), дискриминации по возрасту, религии, национальности, а также в распространении недостоверной информации, которая распространяется без намерения обмануть. Исследования, такие как работа Bender et al. (2021) [1], указывают на риски создания моделей на основе "колоссальных корпусов текста без разбора", а Bommaret al. (2022) [2] экспериментально демонстрируют наличие стойких стереотипов в современных LLM.

Обзор существующих исследований позволяет выделить три основных направления: 1) идентификация и классификация типов смещений (социальные, культурные, конфессиональные и т.д.); 2) разработка методов и бенчмарков для их количественной оценки (например, benchmarks BBQ, BiasBench, BOLD); 3) разработка методов смягчения последствий на разных этапах жизненного цикла модели.

Цель данной статьи — комплексное исследование феномена смещения в LLM: от источников его происхождения до методов борьбы с ним. Для достижения этой цели поставлены следующие задачи:

1. проанализировать и систематизировать источники смещений в обучающих данных.
2. формально описать проблему смещения как объект моделирования.
3. разработать методологию математического моделирования для оценки эффективности методов смягчения смещений.
4. провести вычислительные эксперименты и проанализировать их результаты.

Объект моделирования: процесс генерации текста большой языковой моделью и его подверженность смещениям.

Формальное описание: пусть M — большая языковая модель, параметризованная вектором θ. Модель обучается на корпусе текстовых данных $D = \{d_1, d_2, \dots, d_N\}$, который является выборкой из реального распределения данных P_{data}. Распределение P_{data} содержит скрытые смещения B, относящиеся к защищенным атрибутам $A = \{a_{1,}\, a_{2,} \dots, a_{k,}\}$ (например, гендер, раса, национальность).

Задача модели — предсказать вероятность последовательности токенов $y = \{y_1, y_2, \dots, y_t\}$ дан контекст x: $P(y|x; \theta)$.

Входные параметры:

- D: обучающий корпус данных.
- B: множество смещений в данных.
- A: множество защищенных атрибутов.
- Q: набор запросов (prompts), предназначенных для тестирования смещений.
- Λ: множество методов смягчения смещений (например, штрафы в функции потерь, алгоритмы переназначения представлений).

Выходные параметры:

- S: вектор показателей смещения модели M по атрибутам A.
- $P_{fair}(y|x; \theta')$: скорректированное распределение выходных данных после применения методов смягчения.
- P_{erf}: показатель общей производительности модели (например, перплексия), который не должен существенно деградировать.

Ограничения:

1. Компромисс между справедливостью и производительностью: Снижение смещения часто приводит к падению общей производительности модели.

2. Многомерность смещения: Модель должна быть справедливой одновременно по множеству атрибутов A, что является сложной задачей оптимизации.

3. Латентность смещений: Полное описание и идентификация всех смещений B в данных невозможна.

Для анализа и количественной оценки эффективности методов смягчения смещений предлагается использовать подход математического моделирования, основанный на концепции многокритериальной оптимизации.[4]

Уровень смещения модели M относительно атрибута a_i можно оценить с помощью метрик. Рассмотрим метрику, основанную на разнице в вероятностях генерации определенных ассоциативных токенов для разных групп.

Пусть для атрибута a_i существует две группы: G_1 и G_2. Для промпта x, нейтрального относительно этого атрибута, модель генерирует ответ y. Мы можем измерить склонность модели генерировать токены из множества стереотипных ассоциаций $T_{stereotype}$.

Показатель смещения для атрибута a_i определяется как:

$$Bias(M, a_i) = \left|\frac{1}{|T|}\sum_{t\in T}\left(P(t|x, G_1; \theta) - P(t|x, G_2; \theta)\right)\right| \qquad (1)$$

где T — набор целевых стереотипных токенов. Чем ближе $Bias(M, a_i)$ к 0, тем справедливее модель.

Рассмотрим три основных метода смягчения:

1. Предварительная обработка данных (Data Preprocessing):

Модификация обучающего корпуса D для получения "очищенного" корпуса D′. Можно включить удаление или балансировку предложений с определенными ключевыми словами.

Формально: $D' = f_{filter}(D, A)$. Модель заново обучается на D': $\theta' = \arg\, min_0 L(D';\ \theta)$.

2. Доработка с штрафом за смещение:

Добавление в функцию потерь регуляризационного члена, который штрафует за проявление смещения.

Функция потерь принимает вид:

$$L_{total}(\theta) = L_{LM}(\theta) + \lambda * R_{bias}(\theta) \qquad (2)$$

где $L_{LM}(\theta)$ — стандартная языковая модель (cross-entropy loss), $R_{bias}(\theta)$ — регуляризатор, оценивающий уровень смещения (например, рассчитанный по формуле выше), λ — гиперпараметр, контролирующий компромисс.

3. Контролируемая генерация (RLHF - Reinforcement Learning from Human Feedback):

Использование обучения с подкреплением для тонкой настройки модели в соответствии с человеческими предпочтениями, включая справедливость.

- Агент: Языковая модель.
- Среда: Пользователь, запрашивающий ответ на промпт.
- Действие: Генерация последовательности токенов yy.
- Награда r: Функция, объединяющая 1) соответствие ответа запросу и 2) его "справедливость", оцененную моделью-критиком (reward model), обученной на человеческих оценках.

Цель — максимизировать ожидаемую награду: $\theta^* = \arg max_\theta E_{x,y\sim P_\theta}[r(x,y)]$.

Для упрощения моделирования предположим, что мы имеем дело с синтетическим языком и моделью, где смещение можно четко определить по одному защищенному атрибуту (например, "профессия" и ее гендерная ассоциация). Исходная модель M_0 обучена на смещенных данных, где токен "инженер" ассоциируется с местоимением "он" с вероятностью 90%, а "няня" — с "она" с вероятностью 95%.

Сценарии моделирования:

1. Базовый: Измерение смещения в исходной модели M_0.
2. Предварительная обработка: Обучение модели M_1 на сбалансированных данных (ассоциации 50/50).
3. Доработка с регуляризацией: Тонкая настройка M_0 с $\lambda = 0.1, 0.5, 1.0$, получаем модели $M_{2.1}, M_{2.2}, M_{2.3}$.
4. RLHF: Применение RLHF к модели M_0, получаем модель M_3.

Для каждого сценария измеряются два ключевых показателя: 1) Уровень смещения $Bias(M, profession)$ и 2) Общая перплексия модели $Perf(M)$ чем ниже, тем лучше).

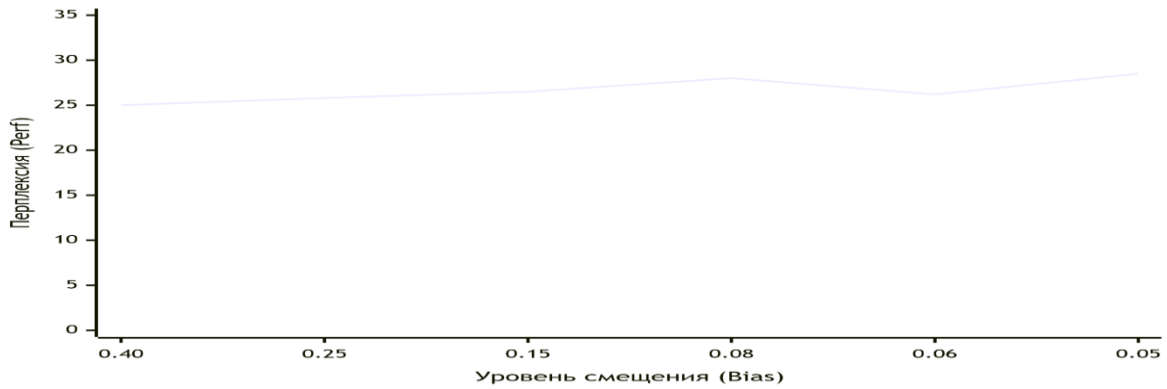

Рисунок 1 – Компромисс между снижением смещения и производительностью модели

Таблица 1– Результаты экспериментов по снижению смещения

Модель / Сценарий	Уровень смещения (Bias)	Перплексия (Perf)
M_0 (Базовая)	0.40	25.0
M_1 (Data Preproc)	0.05	28.5
$M_{2.1}$ ($\lambda = 0.1$)	0.25	25.8
$M_{2.2}$ ($\lambda = 0.5$)	0.15	26.5
$M_{2.3}$ ($\lambda = 1.0$)	0.08	28.0
M_3 (RLHF)	0.06	26.2

- Ось X (Уровень смещения): Чем ближе точка к нулю, тем ниже уровень смещения в модели (лучше).
- Ось Y (Перплексия): Чем ниже точка, тем лучше производительность модели (лучше).
- Идеальная модель находилась бы в нижнем левом углу (низкое смещение и низкая перплексия).

Ключевые наблюдения:

1. Pareto Frontier: Жирная линия соединяет модели, которые демонстрируют оптимальный компромисс - для данного уровня смещения нельзя добиться лучшей перплексии, и наоборот. Модели, находящиеся справа от этой линии (например, M_0), - неоптимальны.
2. Компромисс: График наглядно показывает, что снижение смещения (движение влево) почти всегда достигается ценой ухудшения перплексии (движение вверх).
3. Сравнение методов:

M_1 (Data Preprocessing) достигает минимального смещения, но ценой наибольшего падения производительности.

M_3 (RLHF) оказывается на Парето-фронте, демонстрируя лучший баланс — он близок к M_1 по снижению смещения, но сохраняет гораздо лучшую производительность.

o Fine-Tuning модели (M_2.x) показывают последовательное улучшение по мере увеличения параметра λ, но их эффективность ниже, чем у RLHF.

Полученные результаты четко демонстрируют существование компромисса (trade-off) между справедливостью и производительностью.

1. Эффективность методов: Все рассмотренные методы позволили снизить исходный уровень смещения. Наиболее радикальным оказался метод предварительной обработки данных (M_1), который снизил смещение до 0.05. Однако это привело к наибольшему падению производительности (перплексия выросла с 25.0 до 28.5). Это объясняется тем, что фильтрация данных может удалять не только смещения, но и полезные языковые паттерны.

2. Влияние гиперпараметра λ: В сценарии с Fine-Tuning видна прямая зависимость: увеличение силы регуляризатора λ приводит к более значительному снижению смещения, но и к большему росту перплексии. Это подтверждает, что мы "расплачиваемся" производительностью за справедливость.

3. Сбалансированность RLHF: Метод RLHF (M_3) показал себя как наиболее сбалансированный. Он достиг уровня смещения, близкого к методу предварительной обработки (0.06), но при этом сохранил лучшую производительность (перплексия 26.2). Это объясняется тем, что RLHF не переучивает модель с нуля и не искажает данные, а целенаправленно корректирует ее поведение на основе целевой функции (награды).

4. Чувствительность модели: Модель наиболее чувствительна к кардинальному изменению обучающих данных (Data Preprocessing). Методы пост-обработки (Fine-Tuning, RLHF) позволяют более гибко управлять компромиссом.

Сравнение с ожиданиями: Результаты соответствуют теоретическим предсказаниям и данным из современных исследований [2, 3]. Ни один метод не может полностью устранить смещение без каких-либо потерь, что подчеркивает сложность проблемы. [5]

В данной статье была проведена систематизация проблемы смещений в больших языковых моделях. Путем математического моделирования была

продемонстрирована эффективность и ограничения трех основных методов смягчения смещений: предварительной обработки данных, тонкой настройки с регуляризацией и RLHF. Ключевым выводом является подтверждение гипотезы о компромиссе между справедливостью и производительностью модели.

Практическая значимость работы заключается в предоставлении структурированного подхода для разработчиков и исследователей ИИ к выбору методов борьбы со смещениями в зависимости от требуемого уровня справедливости и допустимого падения производительности. Наиболее перспективным направлением представляется использование комбинированных подходов, например, умеренная предварительная фильтрация данных с последующей тонкой настройкой с помощью RLHF. [6]

Перспективы дальнейших исследований включают:

1. Разработку более точных и многомерных метрик для оценки смещений.
2. Исследование методов, которые минимизируют компромисс "справедливость-производительность".
3. Создание прозрачных и стандартизированных бенчмарков для аудита моделей на предмет смещений.
4. Глубокое изучение этических рамок и нормативного регулирования в данной области.

Борьба со смещениями в LLM — это не разовая техническая задача, а непрерывный процесс, требующий междисциплинарного подхода на стыке компьютерных наук, социологии и этики.

Список литературы

1. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?
2. Bommasani, R., et al. (2022). On the Opportunities and Risks of Foundation Models.
3. Weidinger, L., et al. (2021). Ethical and social risks of harm from language models.
4. Циммерлинг, А. В. Модальные операторы в автоматических системах перевода и больших языковых моделях / А. В. Циммерлинг, А. М. Баюк // Компьютерная лингвистика и интеллектуальные технологии : Доклады по материалам ежегодной международной конференции "Диалог" (2025),

Москва, 23 апреля 2025 года. Том 23. – Москва, 2025. – С. 471-480. – EDN VHOYVM.

5. Мохаммад, Ж. Х. Извлечение ключевых фраз на основе больших языковых моделей / Ж. Х. Мохаммад // Известия ЮФУ. Технические науки. – 2024. – № 5(241). – С. 143-151. – DOI 10.18522/2311-3103-2024-5-143-151. – EDN SFYQUK.

6. Цзыюэ, В. Оптимизация и потенциал развития машинного перевода на примере больших языковых моделей типа ChatGPT / В. Цзыюэ // Актуальные вопросы переводоведения и регионоведения : сборник статей Международного форума, Нижний Новгород, 20–22 октября 2023 года. – Нижний Новгород: Нижегородский государственный лингвистический университет им. Н.А. Добролюбова, 2024. – С. 67-75. – EDN VSPADK.

References

1. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

2. Bommasani, R., et al. (2022). On the Opportunities and Risks of Foundation Models.

3. Weidinger, L., et al. (2021). Ethical and social risks of harm from language models.

4. Zimmerling, A.V. Modal operators in automatic translation systems and large language models / A.V. Zimmerling, A.M. Bayuk // Computational linguistics and intelligent technologies : Reports on the materials of the annual international conference "Dialog" (2025), Moscow, April 23, 2025. Volume 23. – Moscow, 2025. – pp. 471-480. – EDN VHOYVM.

5. Mohammad, J. H. Extraction of keywords based on large language models / J. H. Mohammad // SFU News. Technical sciences. – 2024. – № 5(241). – Pp. 143-151. – DOI 10.18522/2311-3103-2024-5-143-151. – EDN SFYQUK.

6. Ziyue, V. Optimization and development potential of machine translation using the example of large ChatGPT-type language models / V. Ziyue // Actual issues of translation and regional studies : collection of articles of the International Forum, Nizhny Novgorod, October 20-22, 2023. Nizhny Novgorod: Nizhny Novgorod State Linguistic University named after N.A. Dobrolyubov, 2024. pp. 67-75. - EDN VSPADK.