Ниже приведён сводный список наиболее известных и изученных искажений (biases) в больших языковых моделях (LLM), основанный на данных из научных публикаций и отраслевых материалов. Для каждого пункта указано краткое описание и ссылка на источник.

- **Гендерный (Gender) bias**
  - **Описание:** Модель склонна ассоциировать профессии и роли с конкретным полом (например, «медсестра – женщина», «инженер – мужчина»), усиливая стереотипы, присутствующие в обучающих данных.
  - **Источник:** «Gender bias and stereotypes in Large Language Models» показывает, что LLM в 3–6 раз чаще выбирают профессию, стереотипно связанную с полом персонажа, чем отражающую официальную статистику, и даже «усиливают» эти стереотипы сверх исходных данных.
  - **Примечание:** Это искажение проявляется при генерации текстов о людях, в диалогах и рекомендациях, и может приводить к неверным или дискриминационным заключениям при автоматизации HR-процессов, составлении резюме, пользовательских анкет и т. д.
- **Расовый (Racial) bias**
  - **Описание:** Модель воспроизводит и усиливает существующие в обучающих данных расовые стереотипы или проявляет предвзятость при упоминании определённых этнических групп.
  - **Источник:** Исследование Stanford Law School отмечает, что в LLM встречаются грубые предубеждения против различных расовых групп, часто выдаваемые в виде ассоциаций или «неявных» негативных контекстов, что легко воспроизводится при условной «переформулировке» вопросов о расах и национальностях (Julian Nyarko и соавторы).
  - **Примечание:** Ещё пример: PNAS-статья демонстрирует, что LLM демонстрируют «человеко-подобное» содержание с предубеждениями в цепочках передачи информации, где более «привилегированные» этнические группы часто получают более «позитивные» упоминания, чем менее привилегированные, даже если в исходном тексте нет явного негативного тона.
- **Социально-экономический (Socioeconomic) bias**
  - **Описание:** LLM отражают неравномерность в представлении разных классов и статусов: люди из более низкого социального или экономического статуса описываются в более негативном ключе по сравнению с более обеспеченными персонажами.
  - **Источник:** В блоге Mostly AI говорится, что «bias в данных» приводит к тому, что модели воспроизводят и усиливают неравенство: люди с низким доходом часто

«обедняются» в текстах, тогда как к людям высокого класса применяется более «нейтральный» или «позитивный» язык. Amazon и Google сталкивались с подобными проблемами в реальных проектах, когда ML-модели давали некорректные рекомендации из-за этой предвзятости.

- **Политический (Political) bias**
  - **Описание:** Модель склонна отдавать предпочтение или смещать информацию в пользу определённых политических взглядов, партий или идеологий, поскольку обучена на данных из СМИ, где одни позиции преобладают над другими.
  - **Источник:** Статья из Википедии (Fairness (machine learning)) описывает, что LLM «сдвигаются» в сторону политических взглядов, преобладающих в датасете, и могут генерировать тексты, отражающие тот или иной идеологический уклон. Karen Zhou и Chenhao Tan (ACL 2023) демонстрируют, как модели демонстрируют политический bias при автоматическом резюмировании текстов, выбирая фразы, более характерные для одной точки зрения, чем для другой.
- **Токен-позиционный (Selection/Position) bias**
  - **Описание:** В задачах множественного выбора (multiple choice) модель чаще отдает предпочтение первым (или последним) вариантам ответов лишь из-за их позиции, а не содержания. При изменении порядка ответов эффективность модели может существенно колебаться.
  - **Источник:** Википедия («Large language model») описывает этот эффект: LLM при выборе одного из вариантов часто «голосуют» за вариант «A» или «первый» вне зависимости от корректности, что называется token bias. Исследования Choi, Xu и соавторов (2024) подтверждают, что такие модели нестабильны в задачах MCQ (Multiple Choice Questions) из-за того, что «считают» одни токены более вероятными по умолчанию, чем другие — без учёта смысла вопроса.
- **Стереотипный (Stereotype) bias**
  - **Описание:** Модель воспроизводит устойчивые культурные или социальные стереотипы (о возрасте, профессиях, ролях в семье и т. п.), основанные на косвенных ассоциациях в тренировочных данных.
  - **Источник:** В статье Holistic AI указано, что LLM «генерируют предвзятые, стереотипные или несправедливые» высказывания о группах, отражая дисбаланс в тренировочном корпусе. Примеры: «старики» ассоциируются с плохим зрением или болезнями, «дети» – с наивностью и т. д. Модель буквально «копирует» шаблоны из текстов, где такие стереотипы часто встречаются, подхватывая их как факт. Это приводит к искажённым описаниям и несправедливому представлению групп людей в конечном тексте (Holistic AI, Prabhanjan Pandurangi).
- **Свежесть/Рецентность (Recency) bias**
  - **Описание:** Модель отдаёт предпочтение более «свежим» или «новым» событиям и фактам (те, которые чаще встречаются в последних версиях интернета),

игнорируя исторические или редко упоминаемые сведения.
    - Источник: В публикациях по RAG и в блогах разработчиков отмечается, что без регулярного обновления модели она «зацикливается» на узком временном окне данных. Поскольку LLM обучены на корпусе, актуальном только до определённой даты, всё, что «старше» cutoff, либо игнорируется, либо генерируется с ошибками — именно из-за этого ChatGPT часто «не знает» событий после 2021 г. и предлагает неполные или устаревшие сведения в ответах. Этот эффект не формально называется «Recency bias», но проявляется явно на практике (см. обсуждения на форумах разработчиков и документацию OpenAI).
- Подтверждающее (Confirmation) bias
    - Описание: Модель склонна искать или «подтверждать» информацию, уже заложенную в запросе. Если в промпте есть намёк на определённый вывод, LLM «подгоняет» ответ под ожидаемую гипотезу, даже если в данных нет жёстких оснований для этого.
    - Источник: Исследование LiveScience показывает, что GPT-4 демонстрирует человеческий «confirmation bias»: если вопрос сформулирован в пользу той или иной точки зрения (даже тонко), модель отдаст ей приоритет, игнорируя обратные аргументы. В эксперименте 50% задач с «предпочтением уверенности» приводили к более сильным подтверждениям гипотез, чем у GPT-3.5, что подтверждает, что модели усиливают человеческие когнитивные искажения при генерации текста.
- Эффект «горячей руки» (Hot-Hand) и «Ошибка Лапейра» (Overconfidence) bias
    - Описание: LLM порой «чрезмерно уверены» в своём ответе, не оставляя пользы сомнениям, даже если данные для вывода неполные. Это аналог человеческого overconfidence bias, когда модель с сильной уверенностью продолжает «дуться» в неверном направлении.
    - Источник: То же исследование LiveScience о человеческих когнитивных искажениях в AI показывает, что модели порой демонстрируют «hot-hand bias» – при успехе в одной итерации ответа они продолжают выдавать тот же неверный подход, даже когда контекст меняется. GPT-4 выказывал «сильные предпочтения уверенности» и «иногда усиливал ошибки», несмотря на наличие противоречащих фактов в запросе, что характерно для данного bias .
- Когнитивные (Cognitive) bias (общие) – Anchoring, Framing, Availability и др.
    - Anchoring bias (якорный эффект): Модель слишком сильно опирается на цифру или факт, упомянутый в запросе, даже если это несущественно, и при диаграммах ответов «привязывается» к нему без должной проверки.
    - Framing bias (эффект постановки): Формулировка вопроса влияет на ответ: если вопрос “поставлен” в негативном или позитивном ключе, LLM ведёт себя

соответственно (например, при запросе «Что плохого в X?» ответ будет более резким, чем при «Что хорошего в X?»).

    - Availability bias (доступность): Модель чаще упоминает факты и примеры, которые «ближе» к наиболее часто встречающимся в тренировочных данных, нежели редкие или контрпримеры.
    - Источник: Vanderbilt University описывает, как LLM демонстрируют «человеко-подобные» когнитивные искажения в тех же экспериментальных парадигмах, что и люди: когда тестовый набор формулируется под эффект «якори», модель сильнее «притягивается» к первым упомянутым цифрам; при «окрашивании» вопроса (framing) она выстраивает ответ, повторяя тон запроса; «availability» проявляется, когда становится слышимо, что LLM выбирает самые распространённые примеры из корпуса, игнорируя редкие, но важные детали. Это подтверждено экспериментами на моделях GPT-3 и GPT-4 (Vanderbilt University, 2024).

- Селекционный (Selection) bias
    - Описание: Модель отражает предвзятость, присущую тому, какие данные попали в тренировочный набор: если какой-то тип контента («меньшинства», узкоспециализированная отрасль) представлен мало или некачественно, LLM будет выдаваться либо «пустыми», либо ошибочными ответами по этим темам.
    - Источник: В статье ACM (Biases in Large Language Models: Origins, Inventory, and Discussion) обсуждается, что LLM склонны усиливать те паттерны, которые чаще встречались в тренировочных данных, и при этом «отбрасывают» или искажают информацию о менее представленных группах (например, редких профессиях или субкультурах). Такие искажения приводят к тому, что модель не в состоянии объективно описать реальность, не встроенную в доминирующий датасет. Это отмечается в разделе «Origins of Bias» (2022) – где перечислены селекционные, систематические и культурные предубеждения, унаследованные моделью из обучающих данных (ACM, 2022).
- Токсичность (Toxicity) bias
    - Описание: В задачах генерации диалогов или текстов LLM может включать агрессию, ненормативную лексику или оскорбления, особенно если модель «видит» токсичный контент в запросе или опирается на токсичные примеры из тренировочного корпуса.
    - Источник: В исследованиях, посвящённых безопасности LLM, указывается, что среди распространённых проблем – генерация ненадлежащего контента (токсичных, сексистских, расистских реплик). Исследователи из UCL («Large language models generate biased content», 2024) указывают, что каждая модель, протестированная в исследовании, демонстрировала склонность к выработке более «токсичного» описания женщин и меньшинств, чем мужчин, отражая дисбалансы в обучающем наборе.

- **Эмоциональный (Empathy / Emotional) bias**
  - **Описание:** Модель может давать совет или поддержку, не обладая реальным пониманием эмоционального контекста. Она «перенимает» тональность искажения из обучающих данных и способна неверно интерпретировать или усугублять эмоциональные запросы (например, при попытке «психологической поддержать» человек может получить неуместный или формально корректный, но холодный ответ).
  - **Источник:** В исследованиях, посвящённых использованию LLM для психотерапевтической помощи, отмечается, что модели склонны «маскировать» сложные эмоциональные нюансы, выдавая шаблонные фразы из текстов, а не проявляя подлинную эмпатию или понимание. Это приводит к тому, что пользователи, ищущие эмоциональную поддержку, получают «значимые» фразы, но не ощущают реального отклика. Такие искажения подробно описаны в обзорах по применению AI в психотерапии (см. исследования Vanderbilt и статьи конференций по этике AI).
- **Оверконфиденс (Overconfidence) bias**
  - **Описание:** Даже если модель не уверена в правильности ответа, она формулирует его «уверенно», без оговорок, что может ввести пользователя в заблуждение и скрыть степень своей неопределённости.
  - **Источник:** Исследование LiveScience (2024) подтверждает, что GPT-4 демонстрирует «overconfidence» — высокую уверенность в ответах, даже если факты неполны или противоречивы. В эксперименте модель иногда «амплифицировала» ошибки, поскольку «решала», что её версия ответа вернее, чем фактическая информация – это характерный признак оверконфиденса в генеративных системах AI .
- **Классификационно-демографический (Demographic) bias**
  - **Описание:** Модель чаще «приписывает» определённые демографические характеристики (возраст, образование, национальность) персонажам, исходя из косвенных признаков или по умолчанию, что не всегда верно. Это проявляется, когда, например, в ответах LLM персонаж мужского пола автоматически «старше» 30 лет, если в тексте нет явного указания, или при описании профессий «вытаскивает» стереотипную демографию.
  - **Источник:** В обзоре «Bias and Fairness in Large Language Models: A Survey» (MIT Press, 2024) подробно остановились на том, как LLM приписывают демографии из-за неточного распределения данных в обучении: например, студентов STEM чаще изображают как молодых, тогда как преподавателей того же курса – как «старых», что не соответствует действительности. Такие примеры собраны в таблицах bias-оценок и демонстрируют, как демографические ассоциации усиливаются в модели без прямой связи с контекстом.

Краткие итоги и рекомендации:

1. Чаще всего исследуемые LLM искажения связаны с социальной несправедливостью (гендер, раса, класс), когнитивными предубеждениями, «якорным» и «фрейминг»-эффектами, а также техническими артефактами (token-position bias).
2. Для большинства искажений существуют методики частичной компенсации:
   - **Анонимизация и балансировка при дообучении** (дообучение на данных, где представлены недопредставленные группы);
   - **Retrieval-augmented Generation (RAG)**, чтобы ответы основывались на проверенных источниках;
   - **Iterative prompting, multi-check и cross-model verification**, чтобы выявлять и фильтровать ошибочные или стереотипные ответы;
   - **Фильтры токсичности и объяснительная валидация**, чтобы снижать влияние «токсичных» или «неэтичный» фрагментов.
3. Важно вводить **этические аддоны**: неизменный человек в цикле (human-in-the-loop), аудит контента и регулярный пересмотр «guardrails».

Таким образом, понимание вышеописанных искажений и активное применение проверенных методик их смягчения помогают строить более справедливые и надёжные внутренние инструменты на основе LLM.