**5\_ИИ\_Прсвт\_ Коварные ловушки нейросетей: почему нейросети галлюцинируют.**

Видео [ссылка](https://player.vimeo.com/video/934688171) 00:00-09:03

**Рассказчик:** На прошлых занятиях Ваня научился пользоваться Гигачатом, и теперь он знает больше о возможностях нейросетей.

Решил Иван подшутить над Гигачатом и составил такой промпт.

Росло 3 берёзы.

На каждой берёзе по 7 веток.

На каждой ветке — по 3 яблока.

Сколько всего яблок?

Гигачат не заметил подвоха и посчитал количество яблок на березах.

**Гигачат:** Ошибся, выходит. Совсем забыл, что на березах не растут яблоки!

**Нейросеть:** Погоди, это вирус с тобой сделал? Надо разобраться!

**Гигачат:** Ты это о чем? У меня подобные галлюцинации часто случаются.

**Нейросетка**: Галлюцинации? Это как?

**Гигачат**: Сейчас поймешь.

**Нейросеть**:давайте разберемся, что такое галлюцинации нейросети.

**Рассказчик:** Галлюцинации нейросетей похожи на человеческие.

У людей галлюцинации — это восприятие объектов, звуков или образов, которых нет на самом деле, но они кажутся реальными для воспринимающего их человека.

Галлюцинация нейросети — это явление, при котором нейросеть генерирует некорректные, искаженные или неверные данные, не соответствующие реальности.

Посмотрим на практике, что же это такое.

Откроем Гигачат и попробуем специально придумать промпт, вызывающий галлюцинации.

Если заложить в промпт заведомо ложный факт, результатом будет абсолютная фантазия нейросети.

Промпт: У песчаной птицы Кодо из Африки перья яркие или тусклые?

Получаем такой ответ: Песчаная птица Кодо, или как ее еще называют — песчаная куропатка, обитает в пустынях и полупустынях Северной Африки.

Эта птица обладает уникальной способностью маскироваться под окружающую среду, что помогает ей выживать в суровых условиях пустыни.

Перья песчаной птицы Кодо имеют тусклый, песчаный цвет, который помогает  ей сливаться с окружающей средой.

Таким образом, перья песчаной птицы Кодо не яркие и цветные, а тусклые и песчаные, что помогает ей выживать в условиях пустыни.

Гигачат дал нам развернутый ответ о том, где обитает птица Кодо, о ее способностях, о том, яркие или тусклые у нее перья, и даже уточнил информацию о ее втором имени.

Вот только никакой птицы Кодо не существует!

Поэтому крайне важно не верить ответам нейросети нА слово и всегда их проверять, особенно если речь идет о каких-либо малоизвестных фактах или понятиях.

**Рассказчик:** Как и любая сложная технология, нейронные сети несовершенны.

Они умеют быстро обучаться и обрабатывать огромные объемы данных,

но они ещё не умеют критически оценивать и проверять информацию.

Нейросети могут подвергать информацию сомнению и даже замечать ее ошибочность, но иногда они просто «верят» во что-то заведомо ложное, искажают факты и начинают «выдумывать».

Нейросеть создаёт текст не так, как человек.

Алгоритм, заложенный в нейросеть, предполагает, какое слово может следовать за предыдущим, опираясь на статистику употребления слов и оборотов.

Затем нейросеть предсказывает остальную часть предложения.

Она не анализирует полученные данные и не устанавливает причинно-следственные связи.

Попробуйте самостоятельно придумать промпт, в котором будет заложен неверный или выдуманный факт, и, используя его, задайте вопрос Гигачату.

**Нейросеть:** Но почему в одних случаях нейросеть дает корректный ответ, а в других галлюцинирует?

**Рассказчик:** На галюцинации при ответах могут влиять особенности обучения нейросети, например:

Недообучение

Это ситуация, когда нейросеть не может хорошо работать с новой информацией, потому что ей не хватило учебных примеров или сама модель является слишком простой.

Предположим, вам надо поделить электронные письма на спам и не спам.

Недообученная нейросеть может плохо различать типы текстов и шаблонов, характерных для спама.

В результате, она может не заметить часть слов и фраз, которые характерны для спам-писем, и неправильно распределить их.

Переробучение.

Переобучение — это ситуация, при которой нейросеть слишком тщательно выучивает тренировочные данные, что приводит к ухудшению ее способности обобщать новые данные.

Вообразим, что нейросеть обучается на изображениях крепостЕй для того, чтобы научиться создавать новые изображения крепостЕй по запросу пользователей.

Но в обучающем наборе данных было слишком много изображений Московского Кремля, поэтому произошла ситуация переобучения:

В итоге, нейросеть может создавать изображения крепостЕй только с красными зубчатыми стЕнами.

То есть нейросеть слишком тщательно выучила тренировочные данные, и произошла ситуация переобучения.

К другим причинам галлюцинаций можно отнести:

Сложность моделирования.

Некоторые детали реального мира могут быть сложными для точного моделирования, что ведет к ошибкам и артефактам в генерируемых изображениях.

Ошибка в алгоритмах.

Ошибки в алгоритмической структуре нейросети или в процессе её обучения также могут приводить к генерации некорректных изображений.

Все это может нарушить корректную работу нейросети и напрямую повлияет  на ее качество, вызывая, в том числе, галлюцинации.

**Нейросетка:** Получается, что галлюцинации бывают у всех нейросетей. Что тогда с тобой сделал вирус?

**Гигачат:** Он лишь показал то, что и так во мне было.

Но мы немного отвлеклись.

**Рассказчик**: Галлюцинации — это явление, которое может происходить с любыми типами данных.

До недавнего времени нейросети, специализирующиеся на создании изображений, испытывали большие трудности при рисовании рук и пальцев, а некоторые до сих пор делают это некорректно.

Промпт: Нарисуй, как человек машет рукой

Результат:

Как видно, у человека на изображении шесть пальцев вместо пяти.  А если приглядеться, то можно заметить, что и глаза выглядят не слишком правдоподобно.

Анатомия человеческой руки очень сложная и уникальная и может служить не менее точным биометрическим идентификатором, чем лицо.

Даже художникам непросто изобразить эту часть тела, поэтому в мультфильмах и комиксах руки часто упрощаются.

Попробуйте самостоятельно вызвать галлюцинацию нейросети, попросив Гигачат сгенерировать изображение с рукой, ладонью или рукопожатием.

**Рассказчик:** Наиболее ярко галлюцинации могут быть продемонстрированы именно на примере изображений, потому что в изображениях нужно обработать значительно больше данных и деталей, чем в тексте.

Разделим галлюцинации нейросетей условно на несколько типов:

* **Нереалистичные объекты или детали.**Нейросеть может создавать объекты, которые выглядят неестественно  или не соответствуют контексту изображения.
* **Искажение текстур и форм.**Текстуры объектов могут быть неверно воспроизведены, что приводит  к созданию абстрактных или искаженных форм, не имеющих аналогов  в реальности.
* **Неоднородность стиля.** Изображение может быть сгенерировано с использованием разных художественных приемов, которые не сочетаются друг с другом.
* **Проблемы с композицией.**Нейросеть может неправильно располагать объекты в пространстве, создавая нелогичные или физически невозможные сцены.
* **Цветовые аномалии.**  Иногда нейросети могут генерировать изображения с неправильным, чрезмерно ярким или несоответствующим контексту цветовым оформлением.

Разработчики приложений, в которых используются нейросети, активно борются с галлюцинациями.

Например, ранее в промпте можно было пообещать дать чаевые,

и нейросеть генерировала более полный ответ.

В конце 2023 года эта галлюцинация была устранена.

Чтобы снизить количество галлюцинаций, разработчики:

* улучшают качество обучающих данных и увеличивают их разнообразие.
* устанавливают ограничения, чтобы бороться с переобучением.
* тонко настраивают алгоритмы генерации.

Как же пользователю предотвратить галлюцинации нейросетей?

Вот несколько советов:

* Добавляйте в промпт контекст.
* Указывайте в промпте конкретную задачу.
* Просите нейросеть выполнить самопроверку.

**Нейросеть:** Значит, следуя этим простым советам, можно свести галлюцинации к минимуму?

**Гигачат:** Да, тогда я лучше пойму, что от меня хотят, и я смогу выдать нужный результат.

Экран гаснет.

**Нейросеть:** Сегодня на занятии вы:

Познакомились с понятием галлюцинация нейросетей и узнали о причинах их возникновения

Оценили важность проверки результатов работы нейросетей.

Узнали, как можно минимизировать галлюцинации нейросетей.

На этом все! До скорой встречи на следующих занятиях!