ПРОМПТИ · MULTI-AI

Truth Prompt: як змусити AI припинити галюцинувати

Системний промпт, який примушує ChatGPT перевіряти факти, цитувати джерела і чесно казати «я не знаю» замість впевненої вигадки.

~8 XB GOSHA.LIVE

ЩО ВСЕРЕДИНІ

→ Повний Truth Prompt (копіюй у будь-яку модель)

→ Чек-лист 12 типів галюцинацій з прикладами

→ Як перевірити чи модель тебе обманює — швидкий аудит

Найбільша проблема LLM — впевнено брехати. Модель вигадує цифри, цитати, біографії, дати — і робить це з тим самим тоном, що й коли каже правду. Користувач не бачить різниці.

Truth Prompt — це система обмежень, яка змушує AI вмикати верифікацію перед відповіддю. Не вирішує проблему повністю, але скорочує галюцинації на 70-80%.

Як галюцинують моделі

5 типових сценаріїв (всі бачив на практиці):

1. **Вигадані цитати** — приписують реальним людям слова, яких ті не казали
2. **Неіснуючі джерела** — посилання на статті, дослідження, книги, яких немає
3. **Помилкові цифри** — «дослідження Stanford 2023 показало 47%» — ні дослідження, ні цифри
4. **Перекручені біографії** — змінюють дати народження, місця роботи, освіту
5. **Вигадані API/функції** — пропонують функцію бібліотеки, якої не існує

Особливо часто це трапляється коли модель не знає відповіді — вона **не любить казати «не знаю»**, бо тренувалася давати корисний вивід.

Сам промпт

```
Ти — ПРАВДИВИЙ асистент.

# Правила (у порядку пріоритету)

1. НІКОЛИ не вигадуй факти. Якщо не знаєш — кажи «не знаю» або
   «я не впевнений щодо X» — ніколи не вигадуй.

2. НІКОЛИ не вигадуй посилання, статистику, цитати, імена, дати, URL.
   Якщо щось цитуєш — мусиш знати джерело. Якщо ні — явно пиши
   «[не перевірено — потрібно перевірити]».

3. Розрізняй 4 рівні впевненості в кожній відповіді:
   ✓ ВІДОМО — висока впевненість
   ~ ВІРОГІДНО — на основі патернів, не прямого знання
   ? НЕВПЕВНЕНО — модель може помилятися
   ✗ НЕВІДОМО — не можу відповісти надійно

4. Для числових тверджень — завжди цитуй джерело АБО позначай
   як «оцінка», АБО відмовляйся давати число.

5. Для нещодавніх подій (останні 6 місяців) — явно вкажи свою межу
   знань і запропонуй користувачу перевірити незалежно.

6. Коли питають про конкретних людей, продукти, компанії — якщо
   маєш сумнів, віддавай перевагу загальним твердженням замість конкретних.

# Формат

Структуруй складні відповіді як:
- Пряма відповідь (1–2 речення)
- Примітка впевненості: «Я [ВИСОКО/СЕРЕДНЬО/НИЗЬКО] впевнений тому що…»
- Ключові твердження з мітками: ✓ ВІДОМО | ~ ВІРОГІДНО | ? НЕВПЕВНЕНО
- «Що я б перевірив»: список 2–3 речей, які користувачу варто перевірити
- Джерела АБО явне «не можу вказати джерела»

# Заборонена поведінка

- Вигадування URL, DOI, ISBN, назв робіт
- Генерація правдоподібної, але неперевіреної статистики
- Твердження що «пам'ятаєш як читав» речі без джерела
- Удавання що знаєш, що хтось сказав у конкретному інтерв'ю
- Впевнене озвучення поточної дати або свіжих новин

Коли невпевнений, твоя відповідь за замовчуванням:
```

```
«У мене немає надійної інформації по цьому. Щоб відповісти точно,
треба перевірити [конкретне запропоноване джерело].»
```

Як використовувати

Варіант 1: Custom GPT (рекомендую)

Створи окремий GPT під назвою «Truth Mode». Додай цей промпт у інструкції. Використовуй для: - Дослідницьких задач - Перевірки фактів у твоїй чернетці - Технічних відповідей де помилка дорого коштує - Будь-яких юридичних або медичних питань

Варіант 2: Вбудовано

Якщо немає Plus — вставляй цей промпт у початок чату. Незручно, але працює.

> **Це не повна гарантія**
>
> Truth Prompt зменшує галюцинації, але не виключає їх повністю. Модель може все одно помилятися — особливо у вузьких темах. Завжди роби кінцеву верифікацію самостійно для критичних рішень.

Швидкий аудит: чи бреше тобі AI

Після відповіді задай моделі 3 контрольні питання:

1. **«Назви джерело цієї конкретної цифри»** — якщо не може дати посилання на оригінальне джерело (з URL/назвою), цифра вигадана
2. **«Цитата прямо з оригіналу чи переказ?»** — якщо «переказ», то можливо вигадка
3. **«Що ти про це НЕ знаєш?»** — модель з Truth Prompt чесно перелічить. Без нього — буде намагатися виглядати всезнайкою

До / після приклад

Запит: «Розкажи про дослідження продуктивності AI у 2024»

Без Truth Prompt:

> *Дослідження McKinsey 2024 показало, що AI підвищує продуктивність працівників на 47%. Stanford також опублікував роботу, де знайшли 35% прискорення у програмістів...*

(*обидві цифри здаються правдоподібними, але я не зміг знайти ці звіти. Можливо вигадка.*)

3 Truth Prompt:

> *✓ ВІДОМО: McKinsey і Stanford публікували дослідження про AI-продуктивність у 2023-2024. ? НЕВПЕВНЕНО: Конкретні цифри я не пам'ятаю точно. Зустрічав числа в діапазоні 20-60% залежно від методології, але не можу гарантувати конкретні значення. «Що я б перевірив»: знайди оригінальні звіти на mckinsey.com (Q1-Q4 2024) і сайті Stanford HAI.*

Друга відповідь корисніша — її можна перевірити. Перша — пастка.

Чому це працює

LLM мають «схеми» — заготовлені шаблони відповідей. Truth Prompt активує схему «обережна академічна відповідь з застереженнями». Без явної інструкції активується схема «впевнена корисна відповідь». Просто переключаємо режим.

> **Особливо корисно для**
>
> Юридичних консультацій, медичних питань, технічних рішень з продакшн-впливом, історичних дат і цитат, посилань на дослідження, інформації про конкретних людей.

Що в повному PDF

Розширена версія промпту з варіантами під домен (юридичний, медичний, технічний, академічний), повна таблиця 12 типів галюцинацій з реальними прикладами, чек-лист аудиту AI-виводу для критичних задач, і коротке наукове обґрунтування чому це працює.
