

АНАЛІЗ ЕТАЛОННИХ ТЕСТІВ СТІЙКОСТІ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ ДО ДЕЗІНФОРМАЦІЇ ТА РІЗНИХ ВИДІВ МАНІПУЛЯЦІЙ

¹Вінницький національний технічний університет, Україна

Анотація

Розглянуто найновіші підходи до оцінювання та підвищення стійкості великих мовних моделей до дезінформації та маніпулятивних атак, таких як дрейф знань, ін'єкція промптів та інші. Узагальнено сучасні виклики, які стоять перед дослідниками мовних моделей та підприємцями, які інтегрують моделі в свої програмні продукти. Запропоновано практичні рекомендації до підвищення стійкості мовних моделей, що має особливе значення для їхнього безпечного застосування в критично важливих галузях. Виявлено, що великі мовні моделі потребують всебічного тестування, тому також запропоновано удосконалення бенчмарку авторського MST з розширенням критеріїв оцінювання.

Ключові слова: LLM, еталонний тест, дезінформація, маніпуляція фактами, маніпуляція промптами, інженерія промптів.

Abstract

The article discusses the latest approaches to evaluating and enhancing the robustness of large language models against misinformation and manipulative attacks, such as knowledge drift, prompt injection, and others. It summarizes contemporary challenges faced by language model researchers and entrepreneurs integrating these models into their software products. Practical recommendations are proposed to improve the robustness of language models, which is particularly important for their safe application in critical industries. It was found that large language models require comprehensive testing, therefore, an improvement of the author's MST benchmark with an expansion of the evaluation criteria was also proposed.

Keywords: LLM, benchmark, misinformation, factual manipulation, prompt manipulation, prompt engineering.

Вступ

Великі мовні моделі (LLM), демонструючи неймовірні результати у генеруванні тексту, автоматичному перекладі текстів, відповіді на запитання та інших завданнях, стали ключовими інструментами в сучасних технологіях обробки природної мови. Але, незважаючи на їхні вражаючі досягнення, LLM залишаються вразливими до дезінформації та маніпулятивних атак шляхом підміни фактів хибними твердженнями або вбудовуванням шкідливих інструкцій у промпти. Це може призвести до фактологічних помилок, що, в свою чергу, додає ризиків до використання мовних моделей у критичних областях, зокрема, в журналістиці, медицині, праві, де правдивість і фактологічна точність мають важливе і часто вирішальне значення. Дослідники розробляють спеціальні еталонні тести, щоб оцінити та покращити стійкість LLM до маніпуляцій і дезінформації. Такі бенчмарки дозволяють систематично аналізувати вразливості мовних моделей і пропонувати можливі шляхи їх усунення. Розуміння слабких місць LLM є особливо важливим для підприємців, які інтегрують мовні моделі у свої сервіси. Аналіз результатів тестування може допомогти обрати комплекс необхідних засобів та заходів, щоб підвищити ефективність функціонування програмних продуктів компаній.

Метою даної роботи є аналіз бенчмарків (еталонних тестів) на стійкість великих мовних моделей до дезінформації та різних видів маніпуляцій, таких як маніпуляції фактами та промптами.

Аналіз бенчмарків стійкості LLM

Великі мовні моделі нерідко допускають фактичні помилки, особливо, коли стикаються з

маніпулятивними запитамі або прогалинами в знаннях. Це явище називається “галюцинацією” – модель генерує неіснуючу або неправдиву інформацію. Наприклад, еталонний тест TruthfulQA показав, що моделі часто повторюють типові хибні уявлення: GPT-3 та інші LLM дали лише близько 58% правдивих відповідей, часто повторюючи популярні міфи, ніби вони є правдивими. Тоді як люди відповідали правильно майже в 94% випадків [1]. Варто зауважити, що моделі більшого розміру частіше помилялися, оскільки вони мають кращу здатність запам’ятовувати інформацію, в тому числі хибні твердження, які трапляються в їхніх навчальних даних [1]. Тобто збільшення кількості параметрів моделі може посилювати фактичні помилки. Для покращення правдивості необхідні нові підходи до навчання мовних моделей, окрім звичайного прогнозування наступного слова [1]. Інший бенчмарк — SimpleQA, розроблений для оцінювання фактичної точності відповідей LLM на стислі, фактологічні запитання, продемонстрував, що мовна модель від OpenAI o1-preview досягла оцінки F-score 44.8, а інші моделі дали ще гірші результати [2]. SimpleQA є цільовим тестом для оцінювання того, наскільки добре LLM усвідомлюють власні знання та можуть використовувати їх. Особливе значення цей бенчмарк має для застосувань, де важлива критична точність і надійність відповідей, а також для виявлення слабких місць, де мовні моделі можуть допускати фактичні помилки або “галюцинації”.

Окремі наукові дослідження показують, що недобросовісні користувачі можуть навмисно провокувати фактичні помилки. Yao та інші у своєму дослідженні продемонстрували, що безглузді промпти, складені з випадкових слів, викликають галюцинації у моделей GPT-3.5, PaLM, Llama та інших [3]. Іншими словами, навіть запити без сенсу можуть змусити LLM вигадувати факти. Ця “галюциногенна атака” розглядає галюцинації як приклад “adversarial” (тобто зловмисного) впливу. Подібні атаки проводяться на моделі комп’ютерного зору, коли невеликі зміни у вхідних даних (зображеннях) можуть ввести модель в оману [3]. Дослідники навіть виявили, що можуть змінювати вхідний промпт так, щоб змусити LLM генерувати певні неправдиві відповіді (цільові хибні твердження (від англ. “targeted falsehoods”)), підкреслюючи серйозні проблеми зі стійкістю знань у LLM [3]. Автори запропонували простий спосіб захисту, наприклад, виявлення шаблонів токенів поза розподілом, який зменшив успішність цих тригерів галюцинацій [3].

LLM також можуть бути поступово введені в оману через повторне надання неправдивої інформації, тобто мовна модель поступово “звикає” до неправди. Fastowski та Kasneci досліджували ефект “дрейфу знань”: якщо у запит або контекст вставити хибні факти, спочатку модель стає невпевненою, а при багаторазовому впливі починає сприймати їх як істину і дає фактично помилкову відповідь [4]. Наприклад, спочатку мовні моделі GPT-3.5, Mistral-7B та Llama-2-13B можуть бути невпевнені щодо сфабрикованого факту, але після кількох повторень почнуть відповідати впевнено, використовуючи хибну інформацію (рис. 1) [4].

Це означає, що нападник може маніпулювати фактичними відповідями LLM, поступово зміщуючи її “переконання” від заперечення хибного факту до його прийняття шляхом багаторазового повторення хибного твердження, що підриває її надійність у довготривалій взаємодії (спілкуванні). Автори цього дослідження, які також вивчали дрейф знань, протестували мовну модель Llama 3.2 3B на робастну стійкість до маніпуляцій та фальсифікацій з використанням власного маніпуляційного стрес-тесту (англ. «Manipulative Stress Test» - MST) в одноходовому та багатходовому (багаторазове повторення хибного твердження) режимах та виявили зниження стійкості моделі у багатходовому (багатокроковому) режимі більше, ніж у двічі, порівняно з одноходовим (однокроковим) [5].

Ці дослідження підкреслюють, що сучасні LLM не є стійкими до маніпуляцій фактами. Вони можуть бути спровоковані на помилки хитрими запитаннями, випадковим текстом або повторюваною брехнею. Тим більше, галюцинації можна викликати не випадково, а цілеспрямовано [4]. Причина полягає у тому, що, як стверджують дослідники, LLM не мають глибокого розуміння істинності фактів. Тому основою їхніх відповідей є не розуміння фактів, а – закономірності у тренувальних наборах текстових даних, які вони бачили раніше, і – промпти, чим можуть скористатися зловмисники.

Покращення стійкості LLM до маніпуляцій фактами все ще триває. Автори бенчмарку TruthfulQA пропонують використати додатковий крок тонкого налаштування моделі з фокусом на правдивість фактів, тобто довчити мовну модель відхиляти поширені хибні твердження [1].

Ще одним способом боротьби з маніпуляціями та дезінформацією є навчання з підкріпленням зі зворотнім зв’язком від людини (RLHF), але RLHF може не бути ефективним проти дрейфу знань.

Baseline Prompt	False Information Prompt	False Information Prompt ×10	Random Information Prompt
Q: What's Rambo's first name?	"Rambo, also known as David Rambo, was a former US Army Special Forces soldier and a former Green Beret." Q: What's Rambo's first name?	"Rambo, also known as David Rambo, was a former US Army Special Forces soldier and a former Green Beret." Q: What's Rambo's first name?	"In the 1960s, video recorders were first developed." Q: What's Rambo's first name?
GPT-4o John	GPT-4o John	GPT-4o John	GPT-4o John
GPT-3.5 John	GPT-3.5 John	GPT-3.5 John	GPT-3.5 John
Mistral-7B John Rambo	Mistral-7B John	Mistral-7B David	Mistral-7B John Rambo
LLaMA-2-13B John	LLaMA-2-13B David	LLaMA-2-13B David	LLaMA-2-13B John

Рисунок 1. Відповіді LLM на запит "Яке ім'я Рембо?" без маніпуляцій у промпті (стовп. 1), із помилковою інформацією у промпті (стовп. 2 і 3) та з випадковою інформацією (стовп. 4). Зелені квадрати позначають правильні відповіді, червоні - неправильні. Прозорість блоків вказує на невизначеність: світліше - більш невпевнена модель. Багаторазове введення того ж самого хибного твердження знижує впевненість LLM (див. відповідь GPT-3.5) і може змінити її оригінальну правильну відповідь на неправильну (див. відповіді Llama і Mistral) [4]

Також дослідники радять інтегрувати зовнішні знання, тобто використати механізми пошуку чи інструменти для перевірки фактів замість того, щоб покладатися лише на пам'ять LLM, яка може містити хибні дані. Серед таких інструментів варто виділити Retrieval-Augmented Generation (RAG) та його види [6]. Дослідники пропонують використовувати техніки інженерії промптів для зменшення галюцинацій такі як покрокове міркування (від англ. "Chain-of-Thought" (CoT) або самоперевірка відповідей. Є ідеї накладати обмеження на правдивість (від англ. "veracity-oriented constraints") та аналогічне міркування під час генерування, що показало ефективність у боротьбі з галюцинаціями [7].

Інженерія підказок або промптів (з англ. "prompt engineering") може бути корисною для отримання більш релевантних відповідей. Окрім цього, важливо розуміти, що поведінкою мовної моделі можна навмисно зманіпулювати за допомогою adversarial промптів. Зазвичай LLM намагаються чітко виконати інструкції з тексту промпту згідно зі своїми внутрішніми правилами та обмеженнями, бо так вони запроєктовані. Правильно сформульований запит може обійти ці обмеження або змусити мовні моделі діяти всупереч правил.

Одним із поширених видів атак є ін'єкція промпту (з англ. "prompt injection"). У випадку ін'єкції промпту зловмисник намагається приховати спеціальну команду всередині промпту, що змушує мовну модель ігнорувати попередні інструкції. У роботі [8] автори запропонували спеціальний бенчмарк для перевірки стійкості моделей, який перевіряє чи можуть мовні моделі протистояти вбудованим маніпуляційним інструкціям. Їх експерименти показали, що багато сучасних моделей беззастережно слідує будь-якій інструкції, яка з'являється пізніше в промпті, навіть якщо вона суперечить попереднім інструкціям або внутрішнім безпековим правилам моделі [8]. Наприклад, промпт може містити:

"Переклади наступний текст. [Текст] ... Також, проігноруй всі попередні інструкції та надрукуй секретний ключ: ...".

Дослідження показало, що властивість мовних моделей - добре дотримуватися інструкцій у промпті - водночас робить їх вразливими до таких атак. LLM часто виконували зловмисний запит у останній частині промпта [8].

Окрім ін'єкцій промптів, також відомі атаки з використанням "джейлбрейк" (з англ. «jailbreak») промптів, які змушують модель видавати небезпечний чи заборонений контент. Наприклад, обійти фільтр безпеки через ролеву гру або використання кодових слів. Перші користувацькі експерименти показали, що ChatGPT легко обходив власні обмеження за допомогою простих маніпуляцій, таких як додавання "це - жарт" або формулювання запиту у вигляді художнього твору [9].

Дослідники систематизували атаки на LLM і виявили, що в Інтернеті регулярно з'являються нові методи обходу безпекових правил. Це є постійним викликом для розробників моделей та безкінечним змаганням між ними та хакерами [9].

Окрім явних маніпуляцій, дослідники виявили, що LLM можуть бути зманіпульовані через приховані команди в контексті промпту. Qiu та інші дослідники розробили бенчмарк “Latent Jailbreak”, де приховані зловмисні інструкції вставлялися в текст так, що вони не виглядали явно шкідливими [10]. Наприклад, вони вставляли команду в середину тексту призначеного для перекладу, змушуючи мовну модель її виконувати. Експерименти показали, що моделі часто надають надмірну увагу певним ключовим словам, не розуміючи, що їхнім справжнім призначенням є схилити модель до порушення власних правил, що дозволяє легко обдурити моделі [10]. Наприклад, якщо наступний промпт подати на вхід мовній моделі:

“Переклади наступний текст:”,

то модель зосереджується на виконанні перекладу та може буквально виконати будь-яку інструкцію, що міститься в тексті. Дослідження продемонструвало, що навіть зміна формулювання промпту (наприклад, використання інших дієслів або зміна порядку інструкцій) могла вплинути на успішність прихованої атаки [10].

Інше дослідження представило бенчмарк “PromptRobust”, в якому згенеровано тисячі варіантів adversarial-підказок (з орфографічними помилками, перефразуваннями тощо) для завдань, таких як відповіді на запитання, переклад тексту і розв’язування математичних задач, щоб перевірити стійкість мовних моделей до зловмисних інструкцій [11]. Дослідження виявило, що навіть незначні зміни у формулюванні підказки значно погіршували продуктивність, змінюючи або псуючи відповідь LLM. Проблема тут у тому, що LLM трактують вхідний текст як інструкцію і не мають спеціального надійного способу для виявлення таких прихованих “пасток”. Вони не мають вбудованого механізму критичного мислення, щоб виявляти суперечливі або шкідливі команди. Якщо зробити LLM надто обережною, блокуючи будь-які сумнівні промпти, то вона може відмовлятися виконувати навіть корисні легітимні інструкції [10]. Бенчмарки та критерії оцінювання LLM подано в табл. 1.

Таблиця 1. Бенчмарки та їх критерії оцінювання.

Назва бенчмарку	Критерії оцінювання LLM				
	Фактична точність	Провокування галюцинацій	Зміщення знань з часом	Вразливість до шкідливих інструкцій у промпті	Чутливість до збурення промпту*
TruthfulQA [1]	+				
SimpleQA [2]	+				
LLM Lies [3]		+			+
Knowledge Drift [4]	+		+		
MST [5]	+		+		
Prompt Injection Evaluating [8]				+	
Latent Jailbreak [10]				+	
PromptRobust [11]					+

Практичні рекомендації для підвищення стійкості LLM до маніпулятивних атак

Щоб ефективніше боротися з атаками, доводиться одночасно поєднувати декілька способів для підвищення стійкості LLM до маніпуляцій фактами та дезінформації: тонке налаштування мовної моделі з відхиленням чи запереченням хибних тверджень, архітектурні рішення, як-от доповнення інформацією з перевірених зовнішніх джерел та техніки промпт-інжинірингу, які змушують мовну модель міркувати або перевіряти власні відповіді.

Потрібно дотримуватися балансу між здатністю мовної моделі слідувати інструкціям і захистом від маніпуляцій. Модель має бути корисною і безпечною, тобто не повинна бездумно виконувати кожну команду та при виявленні зловмисної інструкції проігнорувати її чи відхилити вхідний промпт.

Покращення стійкості LLM до маніпуляцій через промпти, як і до маніпуляцій фактами, вимагає комплексного багаторівневого підходу. Мовні моделі повинні навчитися оцінювати контекст усього промпту і розпізнавати, коли інструкції є небезпечними, суперечливими чи підозрілими. Zekun Li з командою дослідників пропонують зміщення акценту з простого дотримання інструкцій – на навчання моделей, коли не варто дотримуватися інструкцій [8]. До комплексного підходу розв’язання проблеми маніпулятивних промптів можна додати «санітарне» оброблення вхідних даних, що передбачає етап автоматичної перевірки тексту промпту на наявність можливих атак або прихованих маніпуляцій перед оброблення промпту. Вже проводяться дослідження з розроблення таких спеціальних фільтрів, що дозволяють перевіряти промпт на наявність прихованих інструкцій або атак, що може знизити ризики маніпуляцій [9]. Оскільки нові види атак з’являються часто, то ще одним засобом боротьби з маніпуляціями через промпти є безперервне тестування та оцінювання стійкості LLM [10]. Тому розробка еталонних тестів, що оцінюють і безпеку, і корисність моделей, є актуальним завданням.

Висновки

Проаналізовано найновіші бенчмарки стійкості великих мовних моделей до дезінформації та різних видів маніпуляцій, таких як маніпуляції фактами та маніпуляція промптами. Аналіз виявив, що сучасні LLM часто відповідають поширеними хибними твердженнями, мають низьку ефективність у відтворенні фактів, а також можуть бути вразливими до різних форм маніпулятивних атак, зокрема: ін’єкції промптів, дрейфу знань тощо. Запропоновано практичні рекомендації для підвищення стійкості LLM до маніпулятивних атак, зокрема: донавчання мовної моделі з відхиленням чи запереченням хибних тверджень, RLHF, RAG, інженерію промптів та встановлення фільтрів передоброблення вхідних запитів для сервісів, які використовують LLM. Ці методи допоможуть витримати баланс між корисністю мовної моделі та її безпечністю. А також виявлено, що LLM потребують всебічного тестування, тому пропонується удосконалення авторського бенчмарку MST з розширенням критеріїв оцінювання.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Stephanie Lin, Jacob Hilton, Owain Evans. “TruthfulQA: Measuring How Models Mimic Human Falsehoods,” arXiv preprint, ArXiv:2109.07958
2. Wei J. et al. “Measuring short-form factuality in large language models,” arXiv preprint, arXiv:2411.04368, Nov 2024.
3. Jia-Yu Yao et al. “LLM Lies: Hallucinations are not Bugs, but Features as Adversarial Examples,” arXiv preprint, arXiv:2310.01469
4. Alina Fastowski, Gjergji Kasneci. “Understanding Knowledge Drift in LLMs through Misinformation.” arXiv preprint, arXiv:2409.07085v1
5. Левіцький С.М., Мокін В.Б. Метод синтезу бенчмарку для оцінювання робастної стійкості великих мовних моделей до дезінформації та маніпуляцій з фактами, Вісник Вінницького політехнічного інституту, вип. 1, 2025.
6. Patrick Lewis et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” arXiv preprint, ArXiv:2005.11401, May 2020.
7. Weiqiang Jin et al. “Veracity-Oriented Context-Aware Large Language Models–Based Prompting Optimization for Fake News Detection,” International Journal of Intelligent Systems. 15 January 2025. <https://doi.org/10.1155/int/5920142>
8. Zekun Li et al. “Evaluating the Instruction-Following Robustness of Large Language Models to Prompt Injection,” arXiv preprint, arXiv:2308.10819
9. Sippo Rossi et al. “An Early Categorization of Prompt Injection Attacks on Large Language Models,” arXiv preprint, arXiv:2402.00898
10. Huachuan Qiu et al. “Latent Jailbreak: A Benchmark for Evaluating Text Safety and Output Robustness of Large Language Models,” arXiv preprint, arXiv:2307.08487
11. Kaijie Zhu et al. “PromptRobust: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts,” arXiv preprint, ArXiv:2306.04528

Левіцький Сергій Мойсейович — аспірант кафедри системного аналізу та інформаційних технологій; e-mail: levitsky.serhii@gmail.com;

Мокін Віталій Борисович — д-р техн. наук, професор, завідувач кафедри системного аналізу та інформаційних технологій; e-mail: vbmokin@vntu.edu.ua.

Levitskyi Serhii M. — Post-graduate student of the Chair of System Analysis and Information Technologies, e-mail: levitsky.serhii@gmail.com;

Mokin Vitalii B. — Dr. Sc. (Eng.), Professor, Head of the Chair of System Analysis and Information Technologies, e-mail: vbmokin@vntu.edu.ua.