Деградація штучного інтелекту через рекурсивне навчання на синтетичних даних

Нікіта Бевзюк
студент кафедри комп'ютерних наук,
ВНЗ «Університет економіки та права «КРОК», м. Київ, Україна,
e-mail: BevziukNS@krok.edu.ua
Науковий керівник:
Іван Крикун
к.ф.-м.н., доцент, с.н.с. відділу теорії керуючих систем
ІПММ НАН України, Слов'янськ, Україна,
& доцент кафедри інформаційного менеджменту,
математики та статистики,
ВНЗ «Університет економіки та права «КРОК», м. Київ, Україна,
e-mail: KrykunIH@krok.edu.ua,
ORCID: 0000-0001-5468-512X

Стрімка інтеграція штучного інтелекту у вигляді великих мовних моделей (*Large Language Models* – LLMs) у глобальну інформаційну екосистему ознаменувала нову еру в розвитку інформаційних технологій. Проте разом із розширенням можливостей генеративних систем наукова спільнота зіткнулася з парадоксальним викликом, що загрожує основам подальшого прогресу галузі. Йдеться про вичерпання запасів високоякісних "антропогенних" даних та перехід до навчання моделей на контенті, згенерованому їхніми попередниками (так звані "синтетичні дані"). Низка фундаментальних робіт, опублікованих у 2024–2025 роках, підтвердила існування критичного явища, відомого як "колапс моделі" (*model collapse*) – дегенеративного процесу, за якого рекурсивне навчання на синтетичних даних призводить до незворотної втрати інформації, спотворення розподілу ймовірностей та відриву моделі від реальності [1].

Небезпека цього явища полягає не просто у зниженні якості відповідей, а у фундаментальній зміні математичної природи даних. Моделі, що навчаються на синтетичних даних, схильні до усереднення реальності: вони ефективно засвоюють найчастіші патерни, але катастрофічно швидко втрачають здатність відтворювати рідкісні події, нюанси та стилістичну різноманітність.

Цей процес наочно підтверджується аналізом еволюції розподілу густини ймовірності даних через покоління рекурсивного навчання [1]. Дослідження демонструють механіку виродження: якщо розподіл оригінальних даних є широким, охоплює великий діапазон значень і часто має складну структуру з кількома вираженими піками, то вже через кілька ітерацій навчання на власних висновках статистична картина розподілу даних помітно змінюється. Крива розподілу згладжується, втрачаючи варіативність і "хвости" розподілу. В результаті все багатство даних зводиться до найбільш імовірного середнього значення. Тобто модель фактично втрачає здатність генерувати складні, нетипові відповіді, деградуючи до примітивного, усередненого шаблону.

Подальші дослідження вказують на те, що цей процес змінює самі закони

масштабування (*scaling laws*): просте збільшення обчислювальних потужностей чи обсягу синтетичних даних більше не гарантує покращення результатів, якщо у навчальній вибірці відсутні статистичні "хвости" [2]. Ця проблема набуває характеру екзистенційної загрози для ШІ-технологій через неконтрольоване забруднення Інтернету ШІ-контентом. Оскільки сучасні веб-краулери збирають дані автоматично, ризик неусвідомленого рекурсивного навчання зростає експоненційно.

Наслідки такого "отруєння" даних є критичними для сфер, де ціна помилки висока, а точність у нестандартних ситуаціях є вирішальною. У медицині це загрожує ігноруванням рідкісних діагнозів на користь статистично частіших; у юриспруденції – вигадуванням неіснуючих прецедентів; у сфері кібербезпеки – нездатністю розпізнати нові типи атак, що виходять за межі "середнього" патерну [3]. Таким чином, дослідження механізмів деградації ШІ та розробка методів виявлення синтетичних домішок стає пріоритетним завданням для забезпечення надійності майбутніх інтелектуальних систем.

Деградація штучного інтелекту через рекурсивне навчання починається зі статистичного зсуву та втрати "хвостів" розподілу. При високій частці синтетичного контенту відбувається "вимивання" унікальних прикладів, що призводить до звуження варіативності моделей та знищення знань, необхідних для інновацій та наукових досліджень. Це явище посилюється забрудненням даних – присутністю у тренувальних корпусах контенту, згенерованого моделями, який може спотворювати оцінки та процес навчання [3]. Унаслідок цього, штучно створена інформація, що не ґрунтується на реальних спостереженнях, починає сприйматися як істинна, підриваючи достовірність ШІ у критичних сферах.

Критичне значення для швидкості настання колапсу має обрана стратегія навчання. Експериментально доведено, що не всі підходи мають однакову стійкість до рекурсивного процесу, і найбільш уразливою є стратегія "Заміна" (*Replace*), яка повністю оновлює навчальний корпус синтетичними даними. На противагу цьому, стратегія "Накопичення" (*Accumulate*), що додає нові дані до старого масиву даних, забезпечує значно вищу стійкість. Порівняльний аналіз динаміки накопичення помилок, де крос-ентропія слугує показником падіння якості, наочно демонструє відмінність між цими підходами [4]. Зокрема, стратегія "Заміна" характеризується різким накопиченням помилок, що призводить до критичної деградації якості моделі вже до 7-ї ітерації. Натомість стратегія "Накопичення" зберігає високу стабільність та демонструє відсутність накопичення помилок, оскільки дозволяє утримувати частину оригінальних "хвостів" розподілу. Це підтверджує, що запобігання колапсу вимагає або збереження історичних даних, або розробки механізмів, які імітують цей ефект.

Попри стійкість, метод "Накопичення" має суттєві обмеження, зумовлені лінійним зростанням обсягу даних і значними обчислювальними витратами, що ускладнює його промислове застосування та стимулює пошук ефективніших підходів. Саме необхідність стабілізації рекурсивних циклів без надмірного розширення баз даних зумовила формування трьох основних векторів протидії.

Першим є впровадження методів утримання, найяскравішим представником цього підходу виступає метод “Утримання” (*Retain*) [4] – модифікована форма “Накопичення”, що використовує асиметричний регуляризатор для примусового збереження важливих “хвостів” розподілу під час коригування функції втрат. Його головною перевагою є ефективна боротьба із внутрішньою математичною причиною колапсу – “забуванням”, хоча складність імплементації через глибоку модифікацію функції втрат та високу чутливість до гіперпараметрів залишається вагомим викликом.

Другий напрямок – це курація (*Curation*) та пріоритизація даних. Цей підхід, що передбачає свідоме підвищення ваги рідкісних прикладів та використання алгоритмічних фільтрів для відбору аномальних спостережень [5], прямо протидіє гомогенізації. Його перевага полягає у збереженні інформаційної ентропії корпусу та універсальності, оскільки метод застосовується незалежно від архітектури моделі. Проте, недоліком є суб'єктивність вибору критично важливих “хвостів” та значні часові й обчислювальні ресурси, необхідні для самої процедури очищення та переважування.

Третій напрямок є зовнішнім і сфокусований на контролі походження та водяних знаках. Вбудовування метаінформації безпосередньо у вивід моделі дозволяє детекторам забруднення точно ідентифікувати синтетичні дані [5]. Перевагою цього методу є те, що він створює ефективний фільтраційний бар'єр, дозволяючи відокремити чисті дані від забруднених, а також має велике значення для вирішення правових аспектів. Однак, недолік полягає в тому, що водяні знаки не запобігають колапсу, а лише відтерміновують його шляхом фільтрації. Крім того, надійність водяних знаків завжди може бути поставлена під сумнів через потенційну можливість їхнього видалення або спотворення.

Проведений аналіз механізмів деградації штучного інтелекту підтверджує, що найбільшою загрозою є втрата інформаційної різноманітності через рекурсивне навчання та забруднення даних. Існуючі методи протидії – є ефективними лише частково, оскільки вони або борються з наслідками, або є надто ресурсомісткими, що вказує на необхідність розробки комплексного, багаторівневого підходу.

Пропозиція ефективного гібридного вирішення полягає у створенні трирівневої системи захисту, що забезпечує саморегульований цикл навчання. На першому рівні вхідних даних реалізується зовнішній контроль, який поєднує ідентифікацію синтетики за допомогою водяних знаків із цілеспрямованою курацією даних та пріоритизацією “хвостів”. Це створює надійний фільтраційний бар'єр та зберігає якість навчальної вибірки. Далі, на рівні навчання, використовуються механізми утримання [4], що забезпечують внутрішню, алгоритмічну стабілізацію та запобігають “забуванню” критичної інформації. Перевагою такої інтеграції є висока ефективність при збереженні прийнятної ресурсомісткості, оскільки “Утримання” імітує переваги стратегії “Накопичення”, не вимагаючи при цьому лінійного зростання обсягів даних.

Фінальний, третій рівень моніторингу впроваджує постійний контроль

стану розподілу моделі за допомогою метрик KL-розбіжності, яка вимірює рівень інформаційних втрат під час навчання моделі через аналіз розподілу ймовірностей, які створила модель, та відстані Вассерштайна, що дозволяє прогнозувати настання колапсу та ініціювати корекційні дії. Таким чином, запропонований гібридний підхід має на меті створити саморегульований цикл навчання, який перетворює процес рекурсивного навчання з дегенеративного на стійкий, забезпечуючи надійність майбутніх систем штучного інтелекту.

Ключові слова: колапс моделі; штучний інтелект; хвости.

Список використаних джерел

1. Shumailov, I., Shumaylov, Z., Zhao, Y., et al. (2024) AI models collapse when trained on recursively generated data. Nature, Vol. 631, p. 755–759.

2. Dohmatob, E., Feng, Y., Charton, F., et al. (2024). A Tale of Tails: Model Collapse as a Change of Scaling Laws. https://arxiv.org/html/2402.07043v2

3. Liu, C., Tang, K., Qin, Y., et al. (2025). Bridging Distribution Shift and AI Safety: Conceptual and Methodological Synergies. https://arxiv.org/pdf/2505.22829

4. Li D. et al. (2024). A General Mechanism for Explaining and Preventing Model Collapse. https://openreview.net/pdf?id=4DYFHAcgw3

5. Mattioli, L., Ait-Hadichou, Y., Chaoucheet, S., et al. (2025). Data Curation Matters: Model Collapse and Spurious Shift Performance Prediction from Training on Uncurated Text Embeddings. https://arxiv.org/html/2506.17989v1