НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ
ІНСТИТУТ ПРОБЛЕМ РЕЄСТРАЦІЇ ІНФОРМАЦІЇ НАН УКРАЇНИ

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА БЕЗПЕКА

МАТЕРІАЛИ XXV МІЖНАРОДНОЇ НАУКОВО-ПРАКТИЧНОЇ КОНФЕРЕНЦІЇ

ВИПУСК 25

Київ – 2025

УДК 004.8

ДЕГРАДАЦІЯ ШТУЧНОГО ІНТЕЛЕКТУ ЧЕРЕЗ РЕКУРСИВНЕ НАВЧАННЯ

Нікіта Бевзюк[1, [0009-0000-2317-5696]],
Іван Крикун[2, 3, [0000-0001-5468-512X]]
[1] *Університет КРОК, Київ, Україна*
[2] *ІПММ НАН України, Слов'янськ, Україна*
[3] *Університет КРОК, Київ, Україна*
[1] *BevziukNS@krok.edu.ua*
[2] *iwanko@i.ua*

Анотація. Розглянуто проблеми пов'язані із навчанням штучного інтелекту на рекурсивних даних, проаналізовано шляхи вирішення цих проблем

Ключові слова: *колапс моделі, штучний інтелект, "хвости" розподілу.*

Вступ

Стрімка інтеграція великих мовних моделей (*LLMs*) у глобальну інформаційну екосистему ознаменувала нову еру в розвитку штучного інтелекту. Проте водночас наукова спільнота зіткнулася з викликом, що загрожує подальшому прогресу галузі. Йдеться про вичерпання запасів високоякісних "антропогенних" даних та перехід до навчання моделей на контенті, згенерованому їхніми попередниками. Низка недавніх робіт [1-5] підтвердила існування 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 в тому, що моделі, які навчаються на синтетичних даних, схильні до "усереднення" реальності: вони засвоюють найбільш часті патерни, але швидко втрачають здатність відтворювати рідкісні події, нюанси та стилістичну різноманітність. Цей процес демонструє Рисунок 1.

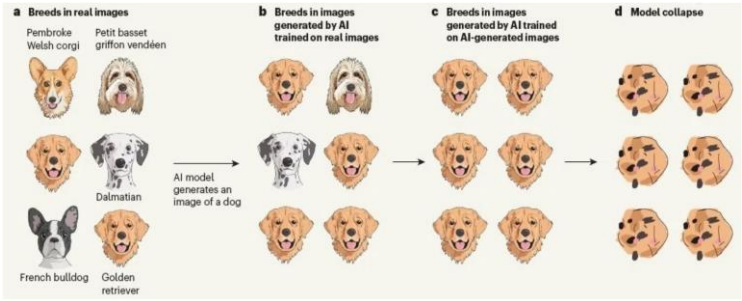

Рисунок 1. Колапс моделі рекурсивного навчання [1]

Подальші дослідження [2] вказують що збільшення обчислювальних потужностей чи обсягу синтетичних даних вже не гарантує покращення результатів, якщо у навчальній вибірці відсутні статистичні “хвости”. Ця проблема стає серйозною через неконтрольоване забруднення Інтернету ШІ-контентом. А оскільки 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 більш поширених; 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.

Методологія

Деградація ШІ 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 втрати знань. Це явище посилюється забрудненням даних – присутністю у тренувальних корпусах контенту, згенерованого моделями, що спотворює оцінки та процес навчання [3]. Тому 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 [4], що різні підходи мають різну 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 старих, забезпечує вищу стійкість, оскільки зберігає частину оригінальних “хвостів” розподілу.

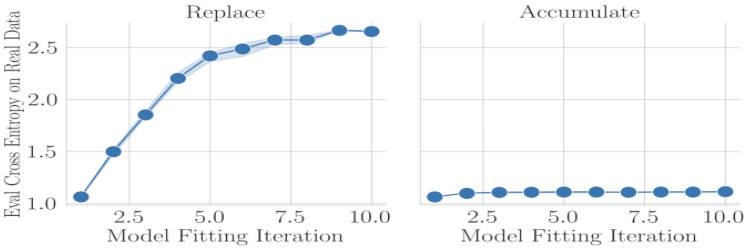

Рисунок 2. Динаміка колапсу. Порівняльна крос-ентропії стратегій “Заміна” та “Накопичення” [4]

На Рисунку 2 видно, що стратегія “Заміна” демонструє різкий зріст помилки, виміряної крос-ентропією, що свідчить про критичну деградацію якості моделі вже до 7-ї ітерації. Стратегія “Накопичення”, навпаки, зберігає високу стабільність та відсутність зростання помилки. Ця візуалізація підтверджує, що для запобіганню колапсу потрібне або збереження даних, або розробки механізмів, які імітують цей ефект. Хоча “Накопичення” є стійким, його недоліком є лінійний зростання обсягу даних та значні обчислювальні витрати, що обмежує його застосування. Це, в свою чергу, вимагає розробки більш ефективних, але водночас стійких до колапсу алгоритмів.

З огляду на необхідність стабілізації рекурсивних циклів, наукова спільнота зосереджує зусилля на трьох основних векторах протидії. Першим є впровадження методів утримання, представлених механізмом “Утримання” (*Retain*)[4]. Цей метод є модифікованою формою стратегії “Накопичення”, що базується на застосуванні асиметричного регуляризатора, який примусово зберігає важливі “хвости” розподілу при коригуванні функції втрат моделі. Його головною перевагою є висока ефективність та

боротьба із внутрішньою математичною причиною колапсу – “забуванням”. Однак, недоліком є складність імплементації, оскільки “Утримання” вимагає глибокої модифікації функції втрат та високої залежності від точності гіперпараметрів.

Другий напрямок передбачає підвищення ваги рідкісних прикладів (курація даних) та використання алгоритмічних фільтрів для відбору аномальних спостережень (пріоритизація даних). 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 дозволяє детекторам ідентифікувати синтетичні дані. Перевагою цього методу є те, що він створює ефективний фільтр, відокремлюючи чисті дані від забруднених, та має велике значення для вирішення правових аспектів. Недолік полягає в тому, що “водяні знаки” лише відтерміновують колапсу шляхом фільтрації. Крім того, надійність “водяних знаків” є сумнівною через можливість їхнього видалення або спотворення.

Висновки

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 Retention-механізми [4], що забезпечує внутрішню, алгоритмічну

стабілізацію та запобігає “забуванню” критичної інформації. Перевагою такої інтеграції є висока ефективність при збереженні прийнятної ресурсомісткості, оскільки Retention імітує переваги стратегії накопичення, не вимагаючи при цьому лінійного зростання обсягів даних.

Фінальний, третій рівень моніторингу впроваджує постійний контроль стану розподілу моделі, 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.

Список використаних джерел

1. Shumailov, I., Shumaylov, Z., Zhao, Y., et al. (2024). AI models collapse when trained on recursively generated data. Nature, 631, 755–759.
2. Dohmatob, E., Feng, Y., Charton, F., et al. (2024). A tale of tails: Model collapse as a change of scaling laws. arXiv. https://arxiv.org/abs/2402.07043v2
3. Liu, C., Tang, K., Qin, Y., et al. (2025). Bridging distribution shift and AI safety: Conceptual and methodological synergies. arXiv. https://arxiv.org/pdf/2505.22829
4. Li, D., et al. (2024). A general mechanism for explaining and preventing model collapse. OpenReview. https://openreview.net/pdf?id=4DYFHAcgw3
5. Mattioli, L., Ait-Hadichou, Y., Chaouche, S., et al. (2025). Data curation matters: Model collapse and spurious shift performance prediction from training on uncurated text embeddings. arXiv. https://arxiv.org/html/2506.17989v1