УДК 004.93:629.33

Панченко В.Ю., магістрант,
Ткаленко О.М., к.т.н, доцент
Державний університет
інформаційно-комунікаційних технологій

РОЛЬ ВЕЛИКИХ ДАНИХ (BIG DATA) У НАВЧАННІ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ (LLM)

Стрімкий розвиток технологій штучного інтелекту в останні роки докорінно змінив ландшафт інформаційних систем, вивівши на передній план великі мовні моделі (LLM). Від автоматизації рутинних процесів до вирішення креативних задач — ці системи демонструють рівень компетенції, який раніше вважався доступним лише людині. Проте за лаштунками архітектурних проривів стоїть фундамент, без якого цей прогрес був би неможливим — дані. Саме симбіоз обчислювальних потужностей та безпрецедентних масивів інформації (Big Data) став каталізатором нової ери в машинному навчанні. У цьому контексті дослідження взаємозв'язку між характеристиками даних та можливостями моделей набуває особливої актуальності.

Постановка завдання

Великі мовні моделі (LLM), такі як GPT, Claude та LLaMA, демонструють безпрецедентні можливості у розумінні та генерації людської мови. Їхній успіх зумовлений не лише вдосконаленням архітектур (наприклад, Transformer), але й експоненційним зростанням обсягів навчальних даних. Розуміння того, як саме характеристики «великих даних» - обсяг, різноманітність та якість - впливають на кінцеві можливості моделей, є критичним завданням для подальшого розвитку галузі.

Мета дослідження

Аналіз кількісного та якісного впливу великих наборів даних (Big Data) на продуктивність, здатність до узагальнення, безпеку та появу «надзвичайних» (emergent) властивостей у сучасних великих мовних моделей.

Результат дослідження

Дослідження підтверджує, що парадигма «великих даних» є фундаментальною для сучасних LLM. Ключовим відкриттям є так звані «закони масштабування» (scaling laws), які демонструють чітку, прогнозовану залежність між збільшенням обсягу навчального корпусу (вимірюваного у трильйонах токенів) та покращенням продуктивності

моделі [1]. Справа не лише в запам'ятовуванні фактів; більші дані дозволяють моделям краще засвоювати складні синтаксичні конструкції, контекстуальні зв'язки та нюанси мови. Однак обсяг не єдиний фактор.

Критично важливою є різноманітність даних. Моделі, навчені на широкому спектрі джерел (книги, наукові статті, програмний код, діалоги, веб-сторінки), демонструють значно кращу здатність до узагальнення та виконання завдань "нульового пострілу" (zero-shot) у нових доменах [2]. Водночас якість даних напряму впливає на надійність та безпеку моделі.

Проблеми "сміття на вході - сміття на виході" (GIGO) є гострими: токсичний, упереджений або фактично неточний контент у навчальних даних неминуче відтворюється моделлю. Тому процеси курування та фільтрації даних, хоча й надзвичайно ресурсомісткі, стають не менш важливими, ніж сама архітектура моделі, для зменшення галюцинацій та небажаної поведінки [3].

Висновки та перспективи

Великі дані є «паливом» для великих мовних моделей. Їхня ефективність прямо залежить від обсягу, різноманітності та, що найважливіше, якості цього палива. Майбутні дослідження будуть зосереджені на розробці ефективніших методів курування даних, використанні синтетичних даних для заповнення прогалин у знаннях та вирішенні етичних проблем, пов'язаних із конфіденційністю та упередженістю даних, зібраних у веб-масштабі.

Список використаних джерел:

1. Scaling Laws for Neural Language Models [Електронний ресурс] – Режим доступу: https://medium.com/@aiml_58187/beyond-bigger-models-the-evolution-of-language-model-scaling-laws-d4bc974d3876

2. The Importance of Data Diversity and Quality in LLM Training [Електронний ресурс] – Режим доступу: https://arxiv.org/abs/2506.19262

3. Challenges in Web-Scale Data Curation for AI [Електронний ресурс] – Режим доступу: https://medium.com/@jasoncorso/data-curation-the-beast-behind-every-ai-model-d136eac4da6b