Інформаційні технології, комп'ютерні науки

УДК 004.8:001.891

Киселевич Володимир Володимирович
ORCID: 0009-0006-6449-710X
Асистент
Житомирський державний університет імені Івана Франка

НЕДЕТЕРМІНІЗМ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ ЯК ЗАГРОЗА ВІДТВОРЮВАНОСТІ ЕМПІРИЧНИХ ДОСЛІДЖЕНЬ: ПРОТОКОЛ МІНІМАЛЬНОЇ ВІДТВОРЮВАНОСТІ

Розглянуто відтворюваність емпіричних досліджень, у яких велика мовна модель (LLM) виступає вимірювальним інструментом. На наборі зі 160 проб із детермінованим еталоном показано, що фіксації температури та багаторазового повтору недостатньо: та сама модель з тим самим ідентифікатором версії, запущена у двох середовищах, змінює підсумкове рішення на 5,6% проб, а заявлена моделлю впевненість цієї нестабільності не виявляє (розрив 12,8 в.п.). Запропоновано протокол мінімальної відтворюваності із шести вимог до звітування. Ключові слова: великі мовні моделі, відтворюваність, недетермінізм, емпіричні дослідження, методологія експерименту.

Постановка проблеми. Велика мовна модель дедалі частіше виконує в дослідженні роль не об'єкта вивчення, а вимірювального інструмента: нею розмічають дані, оцінюють якість згенерованого коду, класифікують тексти, замінюють експертне опитування. Проте, на відміну від будь-якого іншого вимірювального інструмента, LLM не дає того самого результату за однакового вхідного запиту. Джерела розсіювання множинні: стохастичне декодування, недетермінізм паралельних обчислень на графічних прискорювачах, непомітне для користувача оновлення моделі постачальником, зміна конфігурації обслуговування. Звідси випливає методологічний наслідок: результат, поданий одним числом без зазначення умов прогону, не є відтворюваним навіть для самого автора через місяць.

Проблему вже порушено у світовій літературі. У праці [1] показано, що ChatGPT за однакових запитів дає суттєво різні розв'язання задач генерації коду, причому розбіжність зберігається й за температури, близької до нуля. Праця [4] фіксує нестабільність відповідей на юридичні запити. Проте наявні роботи зосереджені на розсіюванні в межах одного середовища виконання, тимчасом як практика публікації вимагає відтворення результату іншим дослідником, тобто в іншому середовищі. Цей зріз досі досліджено найслабше, а в українськомовних публікаціях вимогу звітувати умови прогону практично не обговорюють.

Мета дослідження полягає в тому, щоб оцінити величину розсіювання результату на різних рівнях контролю та сформулювати мінімальний протокол звітування, який робить емпіричне дослідження із застосуванням LLM відтворюваним.

Матеріали й методика. Використано авторський набір проб існування програмного інтерфейсу: 160 пар «бібліотека – символ» над 18 поширеними бібліотеками мови Python, збалансованих як 80 реальних і 80 неіснуючих символів. Принципова властивість методики полягає в тому, що істину встановлює **детермінований оракул**: імпорт бібліотеки та розв'язання ланцюжка атрибутів. Ані модель, ані людина в оцінюванні не беруть участі, тому вимірюється саме нестабільність моделі, а не суб'єктивність оцінювача. Неіснуючі символи згенеровано правдоподібними: їх утворено збуренням реальної назви та запозиченням символа із суміжної бібліотеки. Завдяки цьому завдання перевіряє розрізнення, а не осмисленість рядка.

Кожну пробу подано моделі тричі, за температур 0,3 / 0,4 / 0,5, з фіксованим системним промптом і машиночитаною формою відповіді. Підсумковим рішенням є мажоритарне голосування трьох прогонів, а впевненістю – частка прогонів, що погодилися з більшістю. Опрацьовано конфігурації із зафіксованим ідентифікатором версії моделі, серед них ключова пара: та сама модель claude-sonnet-4-5-20250929, запущена у двох різних середовищах виконання в різний час. Ця пара й дає прямий вимір міжсередовищної відтворюваності.

Результати. Розсіювання виявлено на трьох рівнях, і з посиленням контролю воно не зникає (рис. 1).

Рівень окремої проби. У 28 із 800 випадків (3,5%) три прогони тієї самої конфігурації на тій самій пробі не дали одностайної відповіді; для окремих конфігурацій частка нестійких проб сягає 9,4%. Нестійкі проби наявні в кожній без винятку конфігурації.

Рівень зведеної метрики. Розбіжність точності між температурами 0,3, 0,4 і 0,5 у межах однієї конфігурації досягає 2,5 в.п. У типовій публікації різниця такого порядку між двома методами подається як доказ переваги одного з них.

Рівень середовища виконання. Це головний результат дослідження. Та сама модель із тим самим ідентифікатором версії, той самий набір проб, та сама процедура голосування, той самий оракул, і попри це підсумкове рішення розійшлося на 9 пробах зі 160 (5,6%), тобто узгодженість становила 94,4%. Оскільки всі контрольовані чинники були тотожні, розбіжність припадає на середовище виконання, тобто на той рівень, який у публікаціях не фіксують узагалі. Практичний наслідок такий: дослідник, який сумлінно зазначив версію моделі, температуру й кількість повторів, усе одно не гарантує відтворення свого результату іншою особою.

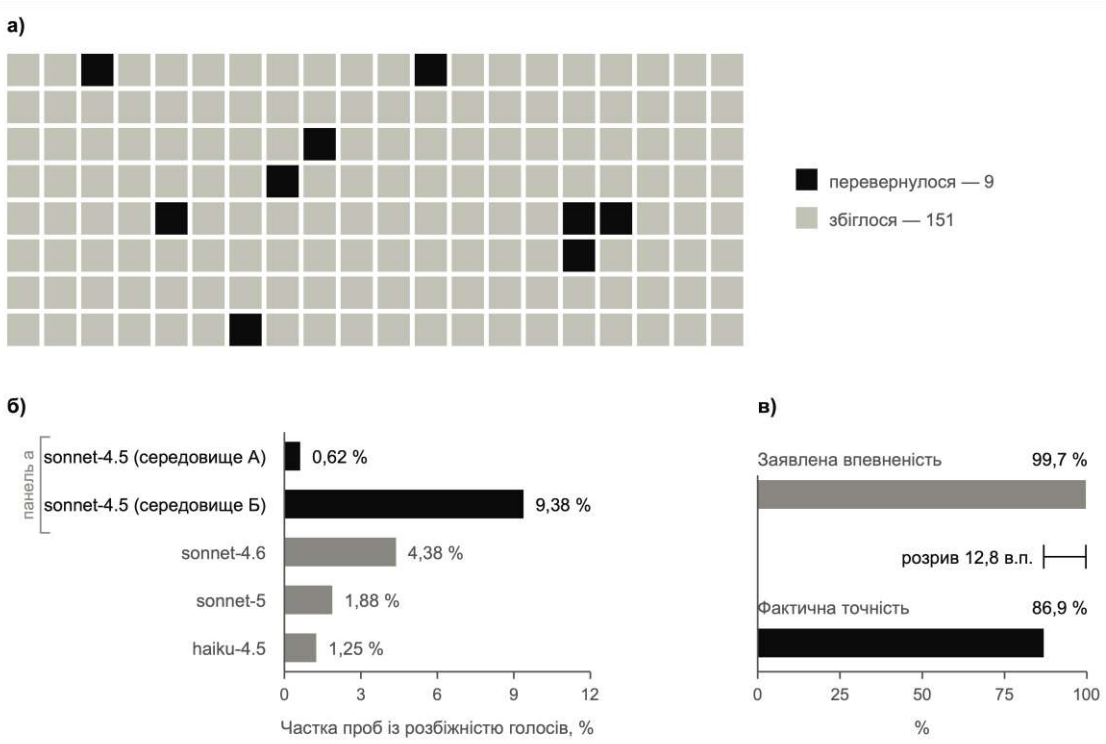

Рис. 1. Розсіювання результату на трьох рівнях контролю: між середовищами виконання (а), між прогонами однієї конфігурації (б) та між заявленою впевненістю і фактичною точністю (в). На панелі (а) кожен квадрат відповідає одній пробі (n = 160). Набір проб: 160 пар «бібліотека – символ» (18 бібліотек Python, 80 реальних / 80 неіснуючих символів); еталоном є детермінований оракул; N = 3 прогони за T = 0,3 / 0,4 / 0,5. Панелі (б) і (в) охоплюють лише конфігурації із зафіксованим ідентифікатором моделі; панель (в) відповідає керованому середовищу (n = 320 проб).

Додатково встановлено, що заявлена моделлю впевненість засобом контролю не є: за середньої впевненості 99,7% фактична точність становила 86,9% (розрив 12,8 в.п.). Максимальну впевненість модель заявляє на 99,1% проб, і серед них 12,3% відповідей хибні. Отже, відтворюваності не досягти відсіюванням «невпевнених» відповідей; потрібен саме протокол звітування. Помилки при цьому розподілені асиметрично: неіснуючі символи модель відхиляє майже безпомилково (специфічність 0,994), натомість реальні, але рідковживані символи впевнено відкидає (чутливість 0,744).

Протокол мінімальної відтворюваності. На підставі отриманих даних пропонується вважати емпіричне дослідження із застосуванням LLM відтворюваним лише за умови, що наведено:

1) точний ідентифікатор версії моделі та дату проведення прогонів, а не лише назву сімейства;

2) усі параметри декодування: температуру, ядрове відсікання (top-p), обмеження довжини, початкове значення генератора випадкових чисел, якщо постачальник його підтримує;

3) кількість повторів N на кожну пробу та правило агрегування відповідей;

4) міру розсіювання поряд із центральним значенням, тобто розмах або довірчий інтервал замість єдиного числа;

5) спосіб установлення істини із зазначенням, чи є оцінювач незалежним від досліджуваної моделі;

6) характеристику середовища виконання (спосіб доступу до моделі, версію клієнтської бібліотеки, регіон обслуговування) та результат щонайменше одного повторного прогону в іншому середовищі.

Вимоги 1-3 є вимогами прозорості й не потребують додаткових витрат. Вимоги 4-6 змінюють обсяг експерименту, проте без них твердження про перевагу одного підходу над іншим на кілька відсоткових пунктів позбавлене доказової сили. Вимога 6 є найменш поширеною і водночас, за отриманими даними, найважливішою.

Обмеження. Дослідження виконано на одній предметній задачі та одній екосистемі (Python), тому конкретні числові значення не слід переносити на інші класи завдань; переносним є висновок про наявність міжсередовищного розсіювання. Кількість повторів $N = 3$ дає лише три рівні впевненості, що обмежує роздільність оцінки калібрування. Пару прогонів для міжсередовищного тесту виконано для однієї лінії моделі; поширення тесту на інші лінії є предметом подальшої роботи.

Висновки. Недетермінізм LLM не є технічною дрібницею; це методологічна властивість інструмента. Показано, що розсіювання зберігається на всіх рівнях контролю й не усувається ані фіксацією температури, ані повторними прогонами, ані вибором «упевнених» відповідей: за тотожних ідентифікатора моделі, набору проб і процедури зміна середовища виконання змінює 5,6 % підсумкових рішень. Запропонований протокол із шести вимог є мінімальним порогом, нижче якого емпіричний результат із застосуванням LLM слід вважати попереднім. Подальша робота передбачає оцінку впливу зростання N на стійкість рішення та перевірку протоколу на завданнях без детермінованого еталона.

1. Ouyang S., Zhang J. M., Harman M., Wang M. An Empirical Study of the Non-Determinism of ChatGPT in Code Generation. ACM Transactions on Software Engineering and Methodology. 2024. DOI: 10.1145/3697010.
2. Wang X., Wei J., Schuurmans D., Le Q. V. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv preprint. 2022. DOI: 10.48550/arXiv.2203.11171.
3. Guo C., Pleiss G., Sun Y., Weinberger K. Q. On Calibration of Modern Neural Networks. arXiv preprint. 2017. DOI: 10.48550/arXiv.1706.04599.
4. Blair-Stanek A., Van Durme B. LLMs Provide Unstable Answers to Legal Questions. arXiv preprint. 2025. DOI: 10.48550/arXiv.2502.05196.
5. Runeson P., Höst M. Guidelines for conducting and reporting case study research in software engineering. Empirical Software Engineering. 2009. Vol. 14, No. 2. P. 131–164. DOI: 10.1007/s10664-008-9102-8.