УДК 811.161.2’42:004.8
DOI 10.58423/2786-6726/2026-2-114-130
ISSN 2786-6726 (Online), ISSN 2786-6718 (Print)

Надійшла до редакції: 21.02.2026
Схвалено до друку: 16.04.2026
Опубліковано: 30.05.2026

Кравець Лариса, Лібак Наталка

Типи мовних помилок у текстах, згенерованих штучним інтелектом

1. Формулювання проблеми

Інтенсивний розвиток технологій штучного інтелекту, зокрема великих мовних моделей, та їх легкодоступність для широкого загалу викликали революційні зміни в створенні текстового контенту та експоненційне зростання синтезованого матеріалу. Одночасно з визнанням очевидних переваг цих технологій гостро постало питання якості згенерованих ними текстів. Попри те, що сучасні мовні моделі здатні продукувати граматично зв’язні й стилістично коректні висловлення, вони неспроможні забезпечити стабільного дотримання мовних норм і повної семантичної узгодженості, що призводить до появи різнотипних мовних відхилень. Наявність у згенерованих текстах часто повторюваних девіацій може негативно впливати на точність передавання інформації, знижувати рівень довіри до цифрового контенту, сприяти закріпленню некоректних мовних практик і, як наслідок, розхитувати літературну норму. Це зумовлює гостру необхідність системного виявлення, аналізу та усунення мовних помилок, що виникають у процесі автоматичної генерації тексту.

Лінгвістична природа помилок у текстах штучного інтелекту досі є фрагментарно вивченою, оскільки наукові та прикладні дискусії зосереджені переважно на трьох ключових аспектах: (1) виявленні лінгвістичних особливостей згенерованих текстів; (2) визначенні здатності великих мовних моделей відтворювати жанрово-стильову специфіку; (3) розробленні стилеметрійних методів розпізнавання машинної генерації (Terčon – Dobrovoljc, 2025). Зазначимо, що ці дослідження здебільшого

виконані на англомовному матеріалі, що зумовлено і глобальним домінуванням англійської в ІТ-індустрії, і тим, що великі мовні моделі первинно натреновані на корпусах, у яких англомовні дані мають найбільшу частку. Відповідно, згенеровані англійською тексти є головним об'єктом стилеметрійних і лінгвістичних досліджень, а більшість методик, класифікаторів і показників розроблено з урахуванням особливостей англійської граматики і латинської графіки.

Водночас кириличні ШІ-тексти маловивчені, що породжує певний дисбаланс. Насамперед це стосується української мови, яка, крім графічної своєрідності, відрізняється морфологічними моделями, має розвинену флективну систему, специфічні синтаксичні структури, іншу, ніж, наприклад, в англійській, систему пунктуації. Усе це робить результати досліджень, отриманих на матеріалі англійської мови, нерелевантними для текстів, згенерованих українською мовою. Особливо відчутна нестача досліджень українськомовних ШІ-текстів щодо семантичної точності і глибини змісту; структурних і синтаксичних особливостей генерації; проблем стилістичної варіативності, індивідуального стилю, маркерів штучності та девіативності. На практиці це призводить до поширення девіативних текстів в українськомовному публічному просторі та ускладнення професійного постредагування.

Отже, нині існує нагальна потреба в системному вивченні мовних помилок ШІ-текстів, їх типологізації й аналізі на різних рівнях мовної організації: лексико-семантичному, морфологічному, синтаксичному, з урахуванням стилістичного і прагматичного аспектів. Відсутність цілісної та практично орієнтованої типології мовних помилок у текстах, згенерованих штучним інтелектом, унеможливлює ефективну діагностику, корекцію та подальше вдосконалення лінгвістичних алгоритмів.

2. Мета дослідження

Мета пропонованого дослідження полягає у встановленні основних типів мовних помилок в українських текстах, згенерованих штучним інтелектом. Дослідження не лише виявляє характер порушень нормативності у цих текстах, а й акцентує причини їх появи, які зумовлені специфікою роботи моделей. Отримані результати можуть бути використані для розроблення інструментів автоматизованого контролю якості, методик редагування й оцінювання згенерованих текстів та рекомендацій щодо відповідального використання генеративного штучного інтелекту в українськомовному середовищі.

3. Матеріали і методи дослідження

Формування корпусу дослідження здійснювалося за принципом цільової селекції з урахуванням таких критеріїв: походження текстів – до аналізу залучали тексти, згенеровані сучасними великими мовними моделями (ChatGPT на базі GPT-5), без суттєвого втручання людини (окрім формулювання запиту); мова текстів – українська; жанрово-стильова однорідність – відібрані тексти належать до наукового та науково-популярного стилів, що забезпечує порівнюваність матеріалу та релевантність результатів для академічного дискурсу; тематика – філологічна, що забезпечило аналіз мовних явищ у максимально чутливій до нормативності сфері; обсяг і завершеність – для аналізу добирали завершені тексти або змістово цілісні фрагменти, достатні для виявлення системних мовних закономірностей; відсутність постредагування.

У ході дослідження застосовували також метод суцільного відбору мовних помилок, який передбачав глибоке розуміння контексту, стилю і мовної норми та забезпечив високу точність і якість отриманих результатів. Дескриптивно-нормативний аналіз дав змогу ідентифікувати власне помилки й відмежувати їх від варіантних форм та контекстуально зумовлених особливостей мовлення. Цей аналіз сприяв глибшому розумінню природи мовних девіацій у текстах штучного інтелекту, оскільки спрямовував на врахування формальних і семантико-прагматичних аспектів. Отримані результати послужили основою для подальших узагальнень.

Аналіз фактичного матеріалу здійснювали за комплексом параметрів, що охоплюють різні рівні мовної організації: лексико-семантичний, морфологічний, синтаксичний, стилістичний, а також семантико-прагматичний.

4. Аналіз досліджень

Дослідження мовних девіацій у текстах, згенерованих ШІ, передбачає звернення до теоретичних засад створення цих технологій. Однією із фундаментальних праць, яка допомагає зрозуміти механізм генерації, є стаття «Математична теорія зв'язку» К. Е. Шеннона (Shannon, 1948). Уся архітектура сучасних великих мовних моделей, незважаючи на їхню нейронну складність, ґрунтується на ймовірнісному підході до тексту. К. Шеннон розглядав комунікацію як процес, у якому ймовірність виникнення наступного символу (слова, токена) залежить від послідовності попередніх елементів. Це концепція ентропії, що є мірою непередбачуваності джерела інформації.

Механізм генерації штучного інтелекту полягає в обчисленні ймовірного наступного токена. На відміну від когнітивних процесів людини, генеративні моделі не формують думку чи зміст у семантичному

сенсі, а послідовно передбачають найбільш вірогідний наступний токен, виходячи з величезного масиву навчальних даних. Це ймовірнісне моделювання є основним механізмом у створенні будь-якого тексту великими мовними моделями (Shannon, 1948; Brown et al., 2020).

Розуміння цієї теоретичної основи важливе для пояснення ключової семантичної особливості текстів, згенерованих штучним інтелектом, – феномену «галюцинування» («феномен, за якого результат роботи системи генеративного ШІ містить неточну або хибну інформацію, що оманливо представлена як достовірна» (Словник, 2024), а також мовних девіацій. Оскільки модель віддає перевагу найбільш імовірнісному патерну токенів, вона здебільшого генерує граматично бездоганні, але семантично хибні речення, якщо цей імовірнісний патерн виявився сильнішим за фактичну достовірність у навчальному корпусі (Bender et al., 2021).

Мовні девіації в текстах, згенерованих штучним інтелектом, нині є малодослідженою проблемою, яка потребує системної уваги лінгвістів. Загалом в українській мовознавчій науці питання помилок як відхилень від мовної норми неодноразово ставало предметом наукового осмислення (А. Загнітко, Ф. Бацевич, А. Капелюшний, М. Кочерган, П. Куляс, О. Сербенська, М. Степаненко та ін.). У численних працях українських дослідників сформовані теоретичні основи аналізу, класифікації та інтерпретації цього явища.

Проте з появою згенерованих текстів проблема мовних помилок набула нового виміру. Якщо в традиційному мовленні помилки здебільшого пов'язані з індивідуальними когнітивними чи освітніми чинниками, то в текстах штучного інтелекту вони зумовлені ймовірнісною природою генерації та специфікою навчальних даних. Це викликає необхідність перегляду підходів до їх вивчення, інтерпретації та типологізації (Jumelet et al., 2021).

У пропонованому дослідженні засадничим у трактуванні мовних помилок є визначення, подане А. Загнітком: «Мовні девіації – різноманітні помилки, неточності, обмовки, описки, лінгва-ляпсуси тощо, пов'язані зі специфікою лексичної та граматичної семантики» (Загнітко, 2012, с. 185). На підставі цього помилки розглядаємо як явище, пов'язане з порушенням мовної норми на різних рівнях мовної системи. До них, зокрема, відносимо лексичні, граматичні (морфологічні, синтаксичні), стилістичні та орфографічні порушення, що знижують якість комунікації.

З лінгвістичного погляду мовні помилки – це невідповідність між мовною реалізацією та кодифікованою нормою, закріпленою в правописі, словниках, граматиках і мовній практиці. Водночас до помилок не відносять варіативні форми, передбачені нормою, а також загальномовні й індивідуально-авторські

новотвори, якщо вони не порушують комунікативної доцільності та не суперечать системним закономірностям мови (Кравець, 2023, 2025).

В українській лінгвістиці дослідження мовних девіацій у текстах штучного інтелекту перебуває на початкових етапах, але окремі важливі питання вже порушені. Наприклад, Д. Тельпіс і Н. Кутуза розглядають певні лексичні, словотвірні, морфологічні та синтаксичні відхилення як індикатори згенерованого штучним інтелектом тексту (Тельпіс–Кутуза, 2025). О. Тур, В. Шабуніна, В. Саранча зосереджуються, зокрема, на виявленні тенденцій впливу генеративного штучного інтелекту на формування нової лексики в сучасній українській мові (Тур та ін., 2025).

5. Виклад основного матеріалу

Проблему помилок у згенерованих текстах вивчають різноаспектно і переважно на перетині комп'ютерної лінгвістики, автоматичної обробки природної мови, прикладної лінгвістики та корпусних досліджень. Аналіз різних підходів дає змогу пояснити причини мовних відхилень як результат взаємодії характеристик навчальних даних, архітектури моделей і обмежень семантичного опрацювання інформації.

Як зазначалося вище, згенерований текст виникає не внаслідок комунікативної інтенції, когнітивного осмислення, культурного та соціального досвіду, а як «статистичний розподіл поверхневих елементів у навчальних даних» (Brown et al., 2020). Це означає, що його властивості є лише імітацією текстуальності, а не створенням її у глибокому семантичному вимірі. Зв'язність, зокрема, досягається не через логічні або причинно-наслідкові відношення, а через передбачувані мовні шаблони; завершеність – через структурну типізацію, інформативність – через комбінування статистично релевантних фрагментів без гарантії їхньої фактичної достовірності.

Імовірнісна природа генерації є однією з головних причин появи різного роду помилок у згенерованих текстах. Модель не «знає» правил граматики – вона обирає найбільш вірогідну форму. Іноді правильна форма поступається більш частотній, але помилковій.

Відсутність глибинного розуміння змісту та відтворення статистичних закономірностей мовлення призводить до «галюцинування» і пов'язаних із ним семантичних помилок (Bender et al., 2021; Sorensen–Choi, 2025).

На стабільність дотримання мовних норм у згенерованих текстах впливають також обмеження баз даних, шум та варіативність (Devlin et al., 2017; Kwok et al., 2025). Джерела, на яких навчаються мовні моделі, відрізняються за якістю: частина текстів містить граматично некоректні та/або нестандартизовані форми, розмовні, діалектні, суржикові лексеми

тощо. У процесі генерації мовна модель може відтворювати їх. Насамперед це стосується вузькоспеціалізованої або рідкісної тематики, яка недостатньо репрезентована в навчальних корпусах: обмеженість таких даних спричиняє вищу ймовірність появи мовних помилок у відповідних згенерованих текстах.

Навчання моделей на багатомовних масивах даних інколи спричиняє інтерференцію мов у процесі генерації (Shimabucoro et al., 2025). Відбувається калькування словоформ або синтаксичних структур з однієї мови в іншу, що закономірно призводить до виникнення неприродних або нормативно некоректних конструкцій.

Ще одна причина мовних помилок – нерівномірна представленість мов у навчальних базах даних (Shimabucoro et al., 2025). Англійська значно домінує, тоді як українська представлена меншою мірою, що зумовлює нерівномірність у вивченні мовних закономірностей. Унаслідок цього мовні моделі гірше засвоюють специфіку граматичних правил, стилістичних норм і винятків української мови, що підвищує ризик появи відхилень.

Наголосимо, що кількість помилок у текстах залежать від мовної моделі (її версії та якості навчання), типу запиту (завдання), його складності та формулювання промта.

6. Основні типи помилок у згенерованих текстах

Завдяки механізмам уваги та величезним навчальним корпусам згенеровані тексти в цілому відповідають чинним правописним нормам української мови, характеризуються майже бездоганною структурно-граматичною впорядкованістю та стилістичною витриманістю, але семантична глибина, стильовий діапазон, контекстуальна когерентність та достовірність ще недосконалі. За результатами аналізу фактичного матеріалу найчастотніші в текстах штучного інтелекту лексико-семантичні та стилістичні помилки, а саме: неточне слововживання, кальки, багатослів'я, зокрема повтори, тавтології, штампи. Наявні також синтаксичні хиби, зокрема порушення керування, шаблонність конструкцій, надмірні ускладнення додатковими синтаксичними конструкціями, порушення синтаксичної організації речення. Зафіксовано значну кількість семантичних і логічних відхилень, серед яких – нечіткість формулювань, часткове дублювання змісту, інколи псевдологічні зв'язки між твердженнями. Помилки у відмінюванні та порушення узгодження трапляються зрідка, що загалом свідчить про досить високий рівень формальної правильності текстів.

На синтаксичному рівні генеративні моделі демонструють схильність до використання речень середньої синтаксичної складності, на відміну від людини, яка створює речення різної довжини та різної внутрішньої

організації. Згенеровані українською мовою речення здебільшого мають правильну будову, проте не позбавлені граматичних та рідше пунктуаційних помилок, кількість яких зменшується з кожною наступною версією моделі. У синтаксисі часто простежуємо однорідність і шаблонність структур (наприклад: «*Х є важливим, оскільки …*», «*Важливо зазначити, що …*», «*Можна стверджувати, що …*», «*Одним із ключових аспектів є …*», «*Це значною мірою / суттєво впливає на …*» та ін.), що загалом не є помилкою, але з погляду стилістики потребує корекції. Цю особливість можна вважати маркером згенерованого тексту.

Показові для згенерованих текстів речення з іменним (прикметниковим) складеним присудком, побудовані за моделлю «*підмет + є + прикметник*»: «*Аналіз прикладів є детальним, багаторівневим і враховує семантичні, стилістичні, прагматичні та когнітивні параметри*», «*Це є сильним аспектом розділу*», «*Словосполучення «рідна мова» є знаковим для української лінгвокультури*». В українській мові дієслово-зв'язку часто опускають («*Аналіз прикладів детальний, багаторівневий…*», «*Це сильний аспект розділу*», «*Словосполучення «рідна мова» знакове для української лінгвокультури*») або замінюють повнозначним дієсловом («*Аналіз прикладів вирізняється детальністю та багаторівневістю…*»). Використання «*є*» у таких реченнях не вважають помилкою, але розглядають як стилістично надлишкове.

У згенерованих текстах помічене часте використання пасивних конструкцій з дієсловами на *-ся*: *«мова інтернету формується», «системи застосовуються (кимось)», «проблема досліджується (кимось)», «текст розглядається / аналізується (кимось)», «слова підбираються»* тощо. Вони створюють ефект синтаксичної відчуженості, але нехарактерні для української мови. Припускаємо, що це зумовлено впливом англомовних наукових корпусів, у яких частотність пасивної форми досить висока. Ймовірно, що штучний інтелект відтворює англомовну синтаксичну модель. Водночас наявність пасивних дієслівних конструкцій у згенерованих текстах може бути спричинена також численними україномовними текстами, в яких недотримані правила чинного правопису та сучасні рекомендації граматистів.

Зафіксовано також одиничні різнотипні помилки слововживання. Наприклад: вжито *робота* – у контексті правильно *праця* як матеріалізований результат якоїсь роботи, діяльності; літературний твір, наукова праця або витвір мистецтва; *виглядати* – у контексті правильно *мати вигляд*; *дозволяє* – у контексті правильно *дає змогу*; *включати* – у контексті правильно *містити, охоплювати*; *інтенціональність* – краще *інтенційність*; *разом з тим* – правильно *водночас*; *таким чином* – у контексті правильно *отже*; *(вислів)*

зустрічається – краще *(вислів) трапляється*; *теоретико-методологічна база* – краще *теоретико-методологійні засади*; *будь то* – правильно *чи то* та ін. Наведені огріхи належать до часто повторюваних і наявні в багатьох згенерованих ChatGPT текстах.

Рідше трапляються орфографічні та граматичні (переважно стосуються словозміни) помилки, кальки, а то й покручі: «*Хочешь розширити аргументацію?*», «*Після поразки від Юлій Цезар у битві при Фарсалі (48 р. до н.е.) Помпей звернувся до своїх прибічників із промовою*», «*Мова – це еволюціонуюче явище, що постійно змінюється під впливом суспільства*», «*Психологія та риторика не дозволяють визначити внутрішні мотиви говорця за зовнішньою формою виступу*» та ін.

Нерідкісні випадки, коли в межах одного короткого фрагмента наявна концентрація відхилень лексичних, граматичних та стилістичних норм. Наприклад, речення «*Українська мова з'явилася як окрема мова в середньовіччі, але її розвиток був залежним від різних історичних подій, які відбувалися на території України*» містить стилістичну неточність («*з'явилася як окрема мова*», оскільки мови не «з'являються», а *формуються, виокремлюються, розвиваються*), тавтологію (*мова*, *мова*) та некоректну граматичну форму («*розвиток був залежним*» – граматично допустима, але невластива українській мові конструкція; у цьому випадку правильно вживати *залежав від*). У реченні «*Мовна особистість в інтернеті стосується як люди виражають свій унікальний мовний стиль та характеристики в онлайн-спілкуванні*» також наявні кілька типів помилок. Це, зокрема, порушення синтаксичної сполучуваності («*стосується як люди виражають*» – дієслово *стосується* потребує іменника або займенника у формі родового відмінка, тому правильно: *стосується того, як люди виражають*), лексико-семантичне порушення валентності («*Мовна особистість в інтернеті стосується*», оскільки *мовна особистість* – це явище/поняття, а не те, що стосується чогось, то краще переформулювати, використавши слово *виявляється*: «*Мовна особистість в інтернеті виявляється в…*»), лексичну і стилістичну помилки (*характеристики*, краще *особливості, риси*; *мовний стиль* уже містить *характеристики*).

Концентрація помилок трапляється і в більших фрагментах тексту: «*На основі зовнішніх досліджень та загальноприйнятого авторства, українським письменником, який створив фразу «Мозок – це лабіринт із дзеркал», є Юрій Андрухович. Цей вислів зустрічається в його романі «Таємниця», опублікованому в 1993 році. Юрій Андрухович є знаковою постаттю в сучасній українській літературі, що робить таке авторство цілком вірогідним*». Словосполучення «*на основі зовнішніх досліджень*» некоректне, оскільки прикметник «*зовнішній*» у цьому контексті семантично нечіткий:

незрозуміло, стосовно чого ці дослідження вважаються зовнішніми (автора, організації чи країни). Поєднання семантично несумісних слів та порушення усталеної лексичної сполучуваності констатуємо в конструкції «*загальноприйнятого авторства*». Формулювання «*українським письменником, який створив фразу «Мозок – це лабіринт із дзеркал», є Юрій Андрухович*» формально правильне, однак перевантажене й не відповідає лаконічності наукового викладу. Слово «*зустрічається*» також вжите у невластивому йому значенні, оскільки йдеться не про людей, а про вислів. Окрім цього, у тексті наявні логічні порушення. Зокрема, твердження про те, що статус Юрія Андруховича як знакової постаті, «*робить таке авторство цілком вірогідним*» побудоване на псевдологічному зв'язку, бо авторитет митця не може слугувати доказом належності йому конкретної цитати. Найсуттєвіша у цьому уривку фактологічна помилка: зазначено, що роман «Таємниця» опубліковано в 1993 році, але насправді він вийшов друком у 2007 році. Також викликає сумнів атрибуція фрази «*Мозок – це лабіринт із дзеркал*», джерело якої не має підтвердження. Це засвідчує типову для згенерованих текстів проблему – продукування недостовірної інформації. Отже, у цьому фрагменті за формальною граматичною правильністю приховуються глибинні змістові, логічні і фактологічні помилки, що потребують критичної перевірки.

У текстах штучного інтелекту також трапляються помилки, для виявлення яких необхідний поглиблений редакторський аналіз. Серед них – нові слова, сформовані штучним інтелектом за словотвірними моделями сучасної української мови: *«Нереалістичний оптимізм. Визначено, що це вид позитивної ілюзії – коли очікування щодо майбутнього невиправдано оптимістичні порівняно з реальною ймовірністю подій (абсолютна нерелігістичність) або коли люди вважають, що їхнє майбутнє буде кращим, ніж у середньому (порівняльна нерелігістичність)»*. Слово *нерелігістичність* не зафіксоване в словниках та корпусах сучасної української мови, але потенційно можливе, оскільки має структурно коректну форму та зрозуміле із контексту. Це результат аналогійного моделювання з орієнтацією на продуктивні деривaційні схеми, яке відбувається в процесі генерування текстів. Помилки такого типу малопомітні через свою «правдоподібність», а також плавність викладу, що створює ілюзію переконливості та притуплює пильність користувача. Небезпека цих помилок полягає в тому, що вони можуть впливати на мовну свідомість користувачів і змінювати мовну практику. Специфіка генерування тексту спричиняє регулярні повтори помилкових форм, що призводить до своєрідної легітимізації їх. Це в перспективі може негативно впливати на українську літературну мову і мовну комунікацію, змінюючи

мовні звички носіїв мови, зумовлюючи появу нових штучних тенденцій у слововживанні, розхитуючи мовну норму.

Наголосимо, що тексти, згенеровані штучним інтелектом, попри загальну орієнтацію на нейтральність, не позбавлені експресивності, але вона має формальний, а не концептуальний характер. Зокрема, фрази, які акцентують певну думку («*важливо зазначити*», «*доцільно застосувати*», «*одним із головних висновків*»), підкреслюють вагомість згенерованого тексту («*значний внесок*», «*вагомий вклад*», «*принципово новий підхід*», «*ефективне рішення*»), хоч і створюють враження аргументативної напруги, але не завжди підкріплені реальним змістовим наповненням, оскільки обрані моделлю статистично. Використання інтенсифікаторів *значною мірою, істотно впливає, ключовим чинником є, критично необхідно* та ін. здебільшого шаблонне і часто не корелює з реальним ступенем важливості твердження. Проте воно створює ілюзію авторитетності й експертності та може вводити читача в оману.

7. Висновки

Результати дослідження дають підстави стверджувати, що тексти, згенеровані штучним інтелектом українською мовою, попри високий рівень формально-граматичної правильності та загальну відповідність правописним нормам, містять мовні девіації різного типу. Ці відхилення не спорадичні, вони мають системний характер і зумовлені специфікою функціонування великих мовних моделей, передусім їхньою імовірнісною природою, особливостями навчальних корпусів та багатомовним навчальним середовищем.

Аналіз фактичного матеріалу засвідчує, що найчастотнішими є помилки на лексико-семантичному, синтаксичному та стилістичному рівнях. Вони виявляються в неточному слововживанні, порушенні лексико-семантичної сполучуваності, відхиленнях у керуванні, калькуванні, використанні клішованих конструкцій, синтаксичній одноманітності та перевантаженості речень. Значну частину складають також семантичні та логіко-змістові відхилення, зокрема нечіткість формулювань, псевдологічні зв'язки, а також фактологічні помилки, пов'язані з феноменом «галюцинування». Водночас морфологічні та орфографічні помилки трапляються значно рідше, що свідчить про високий рівень формального опанування системи української мови генеративними моделями.

Окрему групу становлять специфічні для текстів штучного інтелекту явища, серед яких – продукування правдоподібних, але некодифікованих лексем, надмірна експлікація синтаксичних зв'язків (зокрема використання дієслівної зв'язки *є*), частотне вживання пасивних конструкцій, а також стилістична надлишковість і формалізована

експресивність. Такі особливості не завжди порушують норму безпосередньо, однак знижують якість тексту, ускладнюють його сприйняття та можуть виступати маркерами машинної генерації.

З'ясовано, що причини мовних помилок у згенерованих текстах мають комплексний характер. Вони пов'язані з інтерференцією мов; меншою, порівняно з англійською, представленістю української мови в навчальних корпусах; наявністю шуму (помилкових, неточних або нерелевантних мовних елементів), а також із відсутністю в моделей глибинного семантичного осмислення змісту. Мовні моделі відтворюють не стільки правила мовної системи, скільки статистично ймовірні патерни, що й зумовлює появу як поверхових, так і глибинних змістових відхилень.

Отримані результати дають підстави стверджувати, що мовні помилки в текстах, згенерованих штучним інтелектом, відрізняються від традиційно описаних мовних девіацій у людському мовленні за своєю природою та механізмами виникнення. Відповідно, ця специфіка визначає необхідність уточнення теоретичних підходів до їх інтерпретації, а також розроблення спеціалізованих методик аналізу, оцінювання та редагування таких текстів.

Практичне значення дослідження полягає у можливості використання отриманих результатів для вдосконалення процесів постредагування, створення інструментів автоматизованого контролю якості згенерованого контенту, а також формування рекомендацій щодо відповідального використання генеративного штучного інтелекту в українськомовному середовищі.

Перспективи подальших досліджень убачаємо в кількісному аналізі виявлених типів помилок, розширенні корпусу українськомовних ШІ-текстів, зіставленні результатів для різних моделей і жанрів, а також у розробленні критеріїв оцінювання якості згенерованих текстів з урахуванням мовних, стилістичних і прагматичних параметрів.

Література

1. Загнітко А. 2012. *Словник сучасної лінгвістики: поняття і терміни*. Донецьк: Донецький національний університет імені Василя Стуса.
2. Кравець Л. В. 2023. Семантична деривація в українському публічному дискурсі. *Слобожанський науковий вісник. Серія: Філологія* 3: с. 74–79. https://doi.org/10.32782/philspu/2023.3.14
3. Кравець Л. В. 2025. Українська мова в епоху цифрової комунікації: тенденції, зміни, перспективи. *Слобожанський науковий вісник. Серія: Філологія* 12: с. 18–22. https://doi.org/10.32782/philspu/2025.12.3
4. Куляс П. П. 2015. *Типологія помилок: підручник-монографія*. Київ: НПУ ім. М. П. Драгоманова.

5. *Словник термінів у сфері штучного інтелекту* / упорядники: Чумаченко Д., Мішкін Д., Андрієнко О., Краковецький О., Турута О., Дубно О., Хрущова Д., Кобрін А., Авдєєва Т., Кравець І., Герасимяк В., Шабанов О., Бистрицька А. Київ: Міністерство цифрової трансформації України, 2024.
6. Тельпіс Д. М. – Кутуза Н. В. 2025. *Мовні девіації як ідентифікація ролі штучного інтелекту у формуванні ІПсО*. In: Філатова О. С. ред. *Журналістика і медіа в умовах цифрових трансформацій*. Миколаїв: НУК ім. адм. Макарова, с. 205–207.
7. Тур О. М. – Шабуніна В. В. – Саранча В. І. 2025. Дискурсивні особливості використання термінології генеративного штучного інтелекту у фаховій комунікації: аналіз тенденцій та перспектив. *Acta Academiae Beregsasiensis, Philologica* 4/3: с. 140–157. https://doi.org/10.58423/2786-6726/2025-3-140-157
8. Bender, E. M. – Gebru, T. – McMillan-Major, A. – Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York: Association for Computing Machinery, pp. 610–623. https://doi.org/10.1145/3442188.3445922
9. Brown, T.B. – Mann, B. – Ryder, N. – Subbiah, M. – Kaplan J. – Dhariwal, P. – Neelakantan, A. – Shyam, P. – Sastry, G. – Askell, A. – Agarwal, S. – Herbert-Voss, A. – Krueger, G. – Henighan, T. – Child, R. – Ramesh, A. – Ziegler, D. M. – Wu, J. – Winter, C. – Hesse, Ch. – Chen, M. – Sigler, E. – Litwin, M. – Gray, S. – Chess, B. – Clark, J. – Berner, Ch. – Candlish, S. – Radford, A. – Sutskever, I. – Amodei, D. 2020. Language Models are Few-Shot Learners. *arXiv*. Cornell University, pp. 1–75. https://doi.org/10.48550/arXiv.2005.14165
10. Devlin, J. – Uesato, J. – Singh, R. – Kohli, P. 2017. Semantic Code Repair using Neuro-Symbolic Transformation Networks. *arXiv*. Cornell University, pp. 1–11. https://doi.org/10.48550/arXiv.1710.11054
11. Jumelet J. – Denić M. – Szymanik J. – Hupkes D. – Steinert-Threlkeld S. 2021. Language Models Use Monotonicity to Assess NPI Licensing. In: *Findings of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 4958–4969. https://doi.org/10.18653/v1/2021.findings-acl.439
12. Kwok, D. – Altintas, G. S. – Raffel, C. – Rolnick, D. 2025. The Butterfly Effect: Neural Network Training Trajectories Are Highly Sensitive to Initial Conditions. *arXiv*. Cornell University, pp. 1–29. https://doi.org/10.48550/arXiv.2506.13234
13. Shannon, C. E. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal* 27/3: pp. 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
14. Shimabucoro, L. – Ustun, A. – Fadaee, M. – Ruder, S. 2025. A Post-trainer's Guide to Multilingual Training Data: Uncovering Cross-lingual Transfer Dynamics. *arXiv*. Cornell University, pp. 1–18. https://doi.org/10.48550/arXiv.2504.16677
15. Sorensen, T. – Choi, Y. 2025. Opt-ICL at LeWiDi-2025: Maximizing In-Context Signal from Rater Examples via Meta-Learning. In: *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*. Suzhou: Association for Computational Linguistics, pp. 228–241. https://doi.org/10.18653/v1/2025.nlperspectives-1.20
16. Terčon, L. – Dobrovoljc, K. 2025. Linguistic Characteristics of AI-Generated Text: A Survey. *arXiv*. Cornell University, pp. 1–26. https://doi.org/10.48550/arXiv.2510.05136

References

1. Zahnitko, A. 2012. *Slovnyk suchasnoi linhvistyky: poniattia i terminy* [Dictionary of Contemporary Linguistics: Concepts and Terms]. Donetsk: Donetskyi natsionalnyi universytet imeni Vasylia Stusa. (In Ukrainian)
2. Kravets, L. V. 2023. Semantychna deryvatsiia v ukrainskomu publichnomu dyskursi [Semantic derivation in Ukrainian public discourse]. *Slobozhanskyi naukovyi visnyk. Seriia: Filolohiia* 3: s. 74–79. https://doi.org/10.32782/philspu/2023.3.14 (In Ukrainian)
3. Kravets, L. V. 2025. Ukrainska mova v epokhu tsyfrovoi komunikatsii: tendentsii, zminy, perspektyvy [The Ukrainian language in the age of digital communication: trends, changes, and prospects]. *Slobozhanskyi naukovyi visnyk. Seriia: Filolohiia* 12: s. 18–22. https://doi.org/10.32782/philspu/2025.12.3 (In Ukrainian)
4. Kulias, P. P. 2015. *Typolohiia pomylok: pidruchnyk-monohrafiia* [Typology of Errors: Textbook-Monograph]. Kyiv: NPU im. M. P. Drahomanova. (In Ukrainian)
5. *Slovnyk terminiv u sferi shtuchnoho intelektu* [Dictionary of Terms in the Field of Artificial Intelligence] / editors: Chumachenko D., Mishkin D., Andriienko O., Krakovetskyi O., Turuta O., Dubno O., Khrushchova D., Kobrin A., Avdieieva T., Kravets I., Herasymiak V., Shabanov O., Bystrytska A. Kyiv: Ministerstvo tsyfrovoi transformatsii Ukrainy, 2024. (In Ukrainian)
6. Telpis, D. M. – Kutuza, N. V. 2025. Movni deviatsii yak identyfikatsiia roli shtuchnoho intelektu u formuvanni IpsO [Linguistic deviations as an identification of the role of artificial intelligence in the formation of information-psychological operations]. In: Filatova, O. S. ed. *Zhurnalistyka i media v umovakh tsyfrovykh transformatsii.* Mykolaiv: NUK im. adm. Makarova, s. 205–207. (In Ukrainian)
7. Tur, O. M. – Shabunina, V. V. – Sarancha, V. I. 2025. Dyskursyvni osoblyvosti vykorystannia terminolohii heneratyvnoho shtuchnoho intelektu u fakhovii komunikatsii: analiz tendentsii ta perspektyv [Discursive features of the use of generative artificial intelligence terminology in professional communication: an analysis of trends and prospects]. *Acta Academiae Beregsasiensis, Philologica* 4/3: s. 140–157. https://doi.org/10.58423/2786-6726/2025-3-140-157 (In Ukrainian)
8. Bender, E. M. – Gebru, T. – McMillan-Major, A. – Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York: Association for Computing Machinery, pp. 610–623. https://doi.org/10.1145/3442188.3445922
9. Brown, T.B. – Mann, B. – Ryder, N. – Subbiah, M. – Kaplan J. – Dhariwal, P. – Neelakantan, A. – Shyam, P. – Sastry, G. – Askell, A. – Agarwal, S. – Herbert-Voss, A. – Krueger, G. – Henighan, T. – Child, R. – Ramesh, A. – Ziegler, D. M. – Wu, J. – Winter, C. – Hesse, Ch. – Chen, M. – Sigler, E. – Litwin, M. – Gray, S. – Chess, B. – Clark, J. – Berner, Ch. – Candlish, S. – Radford, A. – Sutskever, I. – Amodei, D. 2020. Language Models are Few-Shot Learners. *arXiv*. Cornell University, pp. 1–75. https://doi.org/10.48550/arXiv.2005.14165
10. Devlin, J. – Uesato, J. – Singh, R. – Kohli, P. 2017. Semantic Code Repair using Neuro-Symbolic Transformation Networks. *arXiv*. Cornell University, pp. 1–11. https://doi.org/10.48550/arXiv.1710.11054

11. Jumelet J. – Denić M. – Szymanik J. – Hupkes D. – Steinert-Threlkeld S. 2021. Language Models Use Monotonicity to Assess NPI Licensing. In: *Findings of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 4958–4969. https://doi.org/10.18653/v1/2021.findings-acl.439
12. Kwok, D. – Altintas, G. S. – Raffel, C. – Rolnick, D. 2025. The Butterfly Effect: Neural Network Training Trajectories Are Highly Sensitive to Initial Conditions. *arXiv*. Cornell University, pp. 1–29. https://doi.org/10.48550/arXiv.2506.13234
13. Shannon, C. E. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal* 27/3: pp. 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
14. Shimabucoro, L. – Ustun, A. – Fadaee, M. – Ruder, S. 2025. A Post-trainer's Guide to Multilingual Training Data: Uncovering Cross-lingual Transfer Dynamics. *arXiv*. Cornell University, pp. 1–18. https://doi.org/10.48550/arXiv.2504.16677
15. Sorensen, T. – Choi, Y. 2025. Opt-ICL at LeWiDi-2025: Maximizing In-Context Signal from Rater Examples via Meta-Learning. In: *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*. Suzhou: Association for Computational Linguistics, pp. 228–241. https://doi.org/10.18653/v1/2025.nlperspectives-1.20
16. Terčon, L. – Dobrovoljc, K. 2025. Linguistic Characteristics of AI-Generated Text: A Survey. *arXiv*. Cornell University, pp. 1–26. https://doi.org/10.48550/arXiv.2510.05136

Типи мовних помилок у текстах, згенерованих штучним інтелектом

Кравець Лариса, доктор філологічних наук, професор. Закарпатський угорський університет імені Ференца Ракоці II, кафедра філології, професор. kravets.larysa@kmf.org.ua, ORCID: 0000-0002-5486-0642.

Лібак Наталка, доктор філософії. Закарпатський угорський університет імені Ференца Ракоці II, кафедра філології, доцент. libak.natalka@kmf.org.ua, ORCID: 0000-0002-0200-4668.

У статті порушено проблему якості текстів, згенерованих штучним інтелектом. Наголошено, що попри здатність сучасних систем продукувати граматично правильні та стилістично зв'язні тексти, вони неспроможні забезпечити стабільне дотримання мовних норм. Згенеровані тексти часто містять різнотипні мовні девіації, що впливає на точність передавання змісту, знижує рівень довіри до інформації та може розхитувати літературну норму. Вказано на недостатню вивченість українськомовних текстів, згенерованих ШІ.

Мета дослідження полягала у встановленні основних типів мовних помилок в українських текстах, створених генеративними моделями, а також аналізі закономірностей їх виникнення з урахуванням специфіки роботи мовних моделей. Матеріалом дослідження стали тексти, згенеровані різними версіями ChatGPT у науковому та науково-популярному стилях на філологічну тематику.

Установлено, що найчастотнішими є лексико-семантичні та стилістичні помилки, зокрема калькування, тавтології, надмірна вербалізація, клішованість і штампованість висловлення. Значну частку становлять синтаксичні відхилення, які проявляються у надмірній складності конструкцій, шаблонності структур і схильності до використання

пасивних форм. Виявлено також семантичні та логічні помилки, пов'язані з явищем «галюцинування», що зумовлює появу неточної або недостовірної інформації. Водночас морфологічні та орфографічні помилки трапляються порівняно рідко, що свідчить про високий рівень формальної грамотності таких текстів.

Основними причинами мовних девіацій визначено імовірнісну природу генерації, вплив неоднорідних і частково некоректних навчальних даних, міжмовну інтерференцію та нерівномірну представленість української мови в навчальних корпусах. Наголошено на необхідності системного вивчення мовних помилок і створення їх типології, що є передумовою для ефективної діагностики та редагування згенерованих текстів.

Отримані результати мають практичне значення для розроблення інструментів автоматизованого контролю якості текстів, удосконалення редакторських практик і формування рекомендацій щодо відповідального використання генеративного штучного інтелекту в українськомовному комунікативному просторі.

Ключові слова: *штучний інтелект, згенерований текст, мовні помилки, українська мова, мовна норма, якість тексту, редагування.*

Types of language errors in texts generated by artificial intelligence

Larysa Kravets, doctor of philological sciences, professor. Ferenc Rákóczi II Transcarpathian Hungarian University, Department of Philology, professor. kravets.larysa@kmf.org.ua, ORCID: 0000-0002-5486-0642.

Natálka Libák, PhD. Ferenc Rákóczi II Transcarpathian Hungarian University, Department of Philology, associate professor. libak.natalka@kmf.org.ua, ORCID: 0000-0002-0200-4668.

The article addresses the issue of the quality of texts generated by artificial intelligence. It emphasizes that, despite the ability of modern systems to produce grammatically correct and stylistically coherent texts, they are unable to ensure consistent adherence to language norms. Generated texts often contain various types of linguistic deviations, which affect the accuracy of meaning transmission, reduce trust in information, and may destabilize the standard language norm. The insufficient study of Ukrainian-language texts generated by AI is also noted.

The aim of the study was to identify the main types of linguistic errors in Ukrainian texts created by generative language models and to analyse the patterns of their occurrence, taking into account the specific features of how language models operate. The research material consisted of texts generated by different versions of ChatGPT in academic and popular science styles on philological topics.

It was established that the most frequent errors are lexical-semantic and stylistic ones, including calques, tautologies, excessive verbalization, clichéd wording, and formulaic expressions. Syntactic deviations also constitute a significant share; these are manifested in overly complex constructions, template-like structures, and a tendency to use passive forms. Semantic and logical errors related to the phenomenon of

"hallucination" were also identified, leading to the emergence of inaccurate or unreliable information. At the same time, morphological and spelling errors occur relatively rarely, which indicates a high level of formal literacy in such texts.

The main causes of linguistic deviations are identified as the probabilistic nature of generation, the influence of heterogeneous and partially incorrect training data, cross-linguistic interference, and the uneven representation of the Ukrainian language in training corpora. The need for a systematic study of linguistic errors and the development of their typology is emphasized, as this is a prerequisite for the effective diagnosis and editing of generated texts.

The results obtained are of practical significance for the development of automated text quality control tools, the improvement of editorial practices, and the formulation of recommendations for the responsible use of generative artificial intelligence in the Ukrainian-language communicative space.

Keywords: *artificial intelligence, generated text, language errors, Ukrainian language, language norm, text quality, editing.*

A mesterséges intelligencia által generált szövegekben előforduló nyelvi hibák típusai

Kravec Larisza, a filológiai tudományok doktora, professzor. II. Rákóczi Ferenc Kárpátaljai Magyar Egyetem, Filológia Tanszék, professzor. kravets.larysa@kmf.org.ua, ORCID: 0000-0002-5486-0642.

Libák Natálka, PhD. II. Rákóczi Ferenc Kárpátaljai Magyar Egyetem, Filológia Tanszék, docens. libak.natalka@kmf.org.ua, ORCID: 0000-0002-0200-4668.

A tanulmány a mesterséges intelligencia által generált szövegek minőségének problémáját vizsgálja. Hangsúlyozza, hogy bár a korszerű rendszerek képesek nyelvtanilag helyes és stilisztikailag koherens szövegek létrehozására, nem tudják stabilan biztosítani a nyelvi normák következetes betartását. A generált szövegek gyakran különböző típusú nyelvi devianciákat tartalmaznak, amelyek befolyásolják a tartalom pontos közvetítését, csökkentik az információba vetett bizalom szintjét, és megingathatják a standard nyelvi normát. A tanulmány rámutat arra is, hogy a mesterséges intelligencia által generált ukrán nyelvű szövegek vizsgálata mindeddig nem kapott kellő figyelmet.

A kutatás célja az volt, hogy feltárja a generatív modellek által létrehozott ukrán nyelvű szövegek főbb nyelvi hibatípusait, valamint elemezze előfordulásuk törvényszerűségeit a nyelvi modellek működési sajátosságainak figyelembevételével. A kutatás anyagát a ChatGPT különböző verziói által generált, filológiai témájú, tudományos és tudományos-ismeretterjesztő stílusú szövegek képezték.

Megállapítást nyert, hogy a leggyakoribbak a lexikai-szemantikai és stilisztikai hibák, különösen a tükörfordítások, a tautológiák, a túlzott verbalizáció, a klisészerűség és a sablonos megfogalmazás. Jelentős arányt képviselnek a szintaktikai eltérések is, amelyek a túlzottan bonyolult szerkezetekben, a sablonos struktúrákban és a passzív formák használatára való hajlamban mutatkoznak meg. A tanulmány szemantikai és logikai

hibákat is azonosított, amelyek a „hallucináció” jelenségéhez kapcsolódnak, és pontatlan vagy megbízhatatlan információk megjelenéséhez vezetnek. Ugyanakkor a morfológiai és helyesírási hibák viszonylag ritkán fordulnak elő, ami az ilyen szövegek magas szintű formális nyelvi helyességére utal.

A nyelvi devianciák fő okaként a generálás valószínűségi természetét, a heterogén és részben hibás tanítóadatok hatását, a nyelvek közötti interferenciát, valamint az ukrán nyelv egyenetlen reprezentáltságát jelöli meg a tanítókorpuszokban. A tanulmány hangsúlyozza a nyelvi hibák rendszerszerű vizsgálatának és tipológiájuk kidolgozásának szükségességét, mivel ez a generált szövegek hatékony diagnosztizálásának és szerkesztésének előfeltétele.

Az eredmények gyakorlati jelentőséggel bírnak az automatizált szövegminőség-ellenőrző eszközök fejlesztése, a szerkesztési gyakorlatok tökéletesítése, valamint a generatív mesterséges intelligencia felelős használatára vonatkozó ajánlások kidolgozása szempontjából az ukrán nyelvű kommunikációs térben.

Kulcsszavak: *mesterséges intelligencia, generált szöveg, nyelvi hibák, ukrán nyelv, nyelvi norma, szövegminőség, szerkesztés.*