https://doi.org/10.31891/2307-5732-2026-363-29
УДК 004.8:811(045)

РАДЗІХОВСЬКА ЛАРИСА
Вінницький торговельно-економічний інститут ДТЕУ
https://orcid.org/0000-0003-0185-8036
e-mail: larirad@ukr.net
ГУСАК ЛЮДМИЛА
Вінницький торговельно-економічний інститут ДТЕУ
https://orcid.org/0000-0002-0022-9644
e-mail: gusak-lyudmila@ukr.net
АБІЄВ АНДРІЙ
Вінницький торговельно-економічний інститут ДТЕУ
https://orcid.org/0009-0005-3114-7624
e-mail: andreyka.abiev@gmail.com

ВИКОРИСТАННЯ ЙМОВІРНІСНИХ ПРИНЦИПІВ ГЕНЕРАТИВНОГО ШТУЧНОГО ІНТЕЛЕКТУ У МОВНИХ МОДЕЛЯХ

Нині мовні моделі (зокрема, великі мовні моделі, наприклад, чат GPT), широко використовуються на практиці, оскільки задовільняють різні потреби користувачів: написання статей, спілкування з користувачами та інше. Мовні моделі − новітній напрямок штучного інтелекту, які дають можливість створення гнучких і адаптивних систем. Завдяки розвитку мовних моделей розробники, підприємці, користувачі можуть створювати власні застосунки на базі існуючих моделей, використовуючи ШІ для різних завдань. В основі роботи генеративного ШІ лежать ймовірнісні принципи. В статті здійснено аналіз особливостей застосування ймовірнісних принципів генеративного штучного інтелекту у мовних моделях. Зроблено висновок про те, що акцент на ймовірнісному підході, розуміння розподілу ймовірностей токенів є фундаментом для подальшого розвитку генеративного штучного інтелекту. При цьому поєднується математична точність із практичними методами підвищення якості й різноманітності згенерованого контенту.

Ключові слова: *ймовірнісні принципи, штучний інтелект, мовні моделі.*

RADZIKHOVSKA LARISA, HUSAK LYUDMILA, ABIEV ANDRIY
Vinnytsia Institute of Trade and Economics of the State University of Trade and Economics

USING PROBABILITY PRINCIPLES OF GENERATIVE ARTIFICIAL INTELLIGENCE IN LANGUAGE MODELS

Currently, language models (in particular, large language models, such as GPT chat) are widely used in practice, as they satisfy various user needs: writing articles, communicating with users, and so on. Language models are a new direction in artificial intelligence that enables the creation of flexible and adaptive systems. Thanks to the development of language models, developers, entrepreneurs, and users can create their own applications based on existing models, using AI for various tasks. The work of generative AI is based on probabilistic principles. The article analyzes the features of the application of probabilistic principles of generative artificial intelligence in language models. In modern generative artificial intelligence, language models learn to generate text by predicting the next words (tokens) based on a probability distribution. Understanding how the probability distribution of tokens is formed and how it can be manipulated (via softmax, temperature, sampling truncation) is key to tuning the behavior of a language model - from pragmatically accurate responses to creatively unexpected ones. Metrics such as entropy, KL-divergence, and perplexity allow us to analyze these distributions. They provide tools for quantifying the uncertainty and knowledge of a model, for comparing the model to a reference distribution or another model, and for tracking the progress of training. New approaches, such as diffusion models, demonstrate that it is possible to abandon step-by-step token prediction and still generate coherent text. However, in these models too, probabilities play a central role, guiding the process of adding and removing noise. Therefore, the emphasis on a probabilistic approach, understanding the probability distribution of tokens, is the foundation for the further development of generative artificial intelligence. This combines mathematical precision with practical methods for improving the quality and diversity of generated content.

Key words: *probabilistic principles, artificial intelligence, language models.*

Стаття надійшла до редакції / Received 18.01.2026
Прийнята до друку / Accepted 11.02.2026
Опубліковано / Published 26.03.2026

Постановка проблеми

Нині мовні моделі (зокрема, великі мовні моделі, наприклад, чат GPT), широко використовуються на практиці, оскільки задовільняють різні потреби користувачів: написання статей, спілкування з користувачами та багато ін.

Зазвичай це сучасні системи штучного інтелекту, призначені для обробки, розуміння та створення тексту.

Мовні моделі − новітній напрямок штучного інтелекту, які відкривають нову еру в обробці природної мови, надаючи можливість створення більш гнучких і адаптивних систем. З їх допомогою досягається високий рівень розуміння контексту, що збагачує досвід користувачів та розширює сфери застосування штучного інтелекту [1].

Розвиток мовних моделей дійшов до того, що користувачі можуть створювати власні застосунки на базі існуючих моделей, інтегруючи їх зі своїми сервісами. Це відкрило нові можливості для розробників, підприємців і навіть звичайних користувачів, які можуть використовувати ШІ для різних завдань.

У сучасному генеративному штучному інтелекті мовні моделі навчаються генерувати текст, прогнозуючи наступні слова (токени) на основі ймовірнісного розподілу.

Аналіз останніх досліджень

У працях І. Юрчак, А. Хіч, В. Оксентюк проведено історичний огляд розвитку мовних моделей, описано проблеми, пов'язані з їх використанням для дослідників та розробників [1]. Сімченко С.В., Лиходєєва Г.В, Левченко В.В., Морозова С.В., Демченко Н.Н. описали особливості використання великих мовних моделей в освітній, науковій та дослідницькій діяльності, вплив таких моделей на наукові відкриття [2]. Ворочек О.Г., Соловей І.В. розглянули принципи роботи мовних моделей штучного інтелекту та їхнє застосуванням для генерації публікацій у соціальних мережах [3] .

Метою даної роботи є аналіз особливостей застосування ймовірнісних принципів генеративного штучного інтелекту у мовних моделях.

Виклад основного матеріалу

У сучасному генеративному штучному інтелекті мовні моделі навчаються генерувати текст, прогнозуючи наступні слова (токени) на основі ймовірнісного розподілу.

Мовна модель після обробки вхідного тексту формує набір *logits* (ненормованих оцінок) для кожного можливого токена. Щоб отримати ймовірності токенів, ці logits перетворюються через функцію *softmax*, яка для і-го токена обчислює ймовірність як:

$$p_i = \frac{e^{x_i}}{\sum_j e^{x_j}} \quad (1)$$

де x_i - логіт і-го токена [4]. Цей результуючий розподіл ймовірностей є категоріальним: тобто модель призначає кожному токену з словника певну ймовірність, а сума всіх ймовірностей дорівнює 1. Такий категоріальний розподіл відображає невизначеність моделі щодо того, яке слово має йти далі.

Отримавши розподіл ймовірностей, найпростішим підходом було б завжди обирати токен з найбільшою ймовірністю («жадібний підхід»). Однак для мовних завдань постійний вибір лише найімовірнішого слова призводить до одноманітних, «нудних» текстів. Натомість, генеративні моделі зазвичай «семплюють» токени згідно з отриманим розподілом: тобто випадково вибирають наступне слово пропорційно до його ймовірності. Це додає варіативності − навіть якщо якесь слово має 50% ймовірності, інші слова з меншою, але суттєвою ймовірністю також мають шанс бути обраними. Таким чином генеруються різноманітні відповіді, а не завжди одна й та сама фраза.

Одним із ключових параметрів, що впливають на форму ймовірнісного розподілу, є температура (T). Температура використовується для перерозподілу ймовірностей: інтуїтивно, зниження T робить «піки» розподілу вищими (підвищує ймовірність найбільш вірогідних токенів) і зменшує ймовірності рідкісних токенів, а підвищення T, навпаки, вирівнює розподіл, збільшуючи відносну вагу менш імовірних токенів [35]. Реалізується це через поділ кожного логіту на T перед застосуванням softmax. Вища температура робить тексти більш креативними, вводячи різноманітність, але може знизити їх зв'язність; нижча температура робить вихід більш передбачуваним і однорідним, проте іноді надто простим [6]. Як окремий випадок, температура, наближена до 0, означає практично детермінований вибір найімовірнішого токена (модель фактично виконує argmax без випадковості).

Для керування генерацією часто застосовують методи обмеження простору вибору токенів. Популярний підхід − Top-k семплінг: модель бере лише топ-k найбільш ймовірних токенів і нормує ймовірності тільки серед них, відсіюючи решту [7]. Наприклад, якщо k = 50, то обиратиметься наступний токен лише з 50 найімовірніших кандидатів замість всього словника. Це знижує обчислювальне навантаження і відсікає малоймовірні продовження, де можуть бути галюцинації. Менше значення k робить текст більш передбачуваним, адже модель обмежена найбільш ймовірними словами, але може втратити рівень креативності [5]. Інша популярна методика − Top-p семплінг (нуклеус-семплінг): модель динамічно обирає поріг ймовірності p (наприклад 0.9) і включає до кандидатів найімовірніші слова, сумарна ймовірність яких досягає p. Таким чином розмір вибірки динамічно змінюється: для простих питань він може бути малим (якщо очевидна відповідь «yes/no», то вистачить двох варіантів), а для відкритих запитань – більшим. Нуклеус-метод забезпечує адаптивну рівновагу між різноманітністю і релевантністю відповіді. Параметри семплінгу (такі як температура, k та p) часто підбираються експериментально під задачу, щоб досягти балансу між творчістю тексту й його осмисленістю.

Ймовірнісний розподіл токенів, отриманий на виході моделі, можна аналізувати кількісно. Ентропія $H(P)$ є ключовою мірою невизначеності (або «інформаційної різноманітності») розподілу. Вищі значення ентропії відповідають більш рівномірному (невпевненому) розподілу, нижчі − більш «сконцентрованому» розподілу, де один або кілька токенів явно домінують за ймовірністю. Аналіз ентропії допомагає зрозуміти, наскільки модель «впевнена» у своєму прогнозі на кожному кроці. Якщо ентропія розподілу дуже висока (майже рівномірний розподіл по багатьох токенах), це означає, що модель вагається між багатьма варіантами − часто такий випадок свідчить про складність контексту або недостатність знань моделі. Навпаки, занадто низька ентропія (один токен має переважну ймовірність) може означати як впевненість моделі (наприклад, очевидне наступне слово у фразі), так і можливу упередженість або переіндукцію (явище, коли під час генерації тексту модель надміру дотримується певного шаблону або паттерну) моделі, якщо це не очікувано.

Для порівняння двох ймовірнісних розподілів використовується міра KL-дивергенції. Дивергенція Кульбака-Лейблера між розподілом $P = \{p_i\}$ і $Q = \{q_i\}$ (визначеними на тому самому просторі подій) задається як:

$$D_{KL}(P \parallel Q) = \sum_i \mathrm{p}_i \log \frac{\mathrm{p}_i}{\mathrm{q}_i} \quad (2)$$

і характеризує, наскільки розподіл Q неефективно описує дані, згенеровані згідно з P. Іншими словами, KL-дивергенція кількісно показує «відстань» (хоча формально це не симетрична відстань) між двома розподілами ймовірностей. Якщо $P = Q$, тоді $D_{KL} = 0$ (розподіли ідентичні, немає втрати інформації при заміні одного іншим). Якщо ж розподіли відрізняються, значення D_{KL} буде позитивним, причому більші значення відповідають більшому розходженню між ймовірностями. Важливо, що $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$, тобто порядок розгляду розподілів має значення (KL є асиметричною) [8].

У задачах мовного моделювання KL-дивергенція часто використовується для оцінки розбіжності між ймовірнісним розподілом, згенерованим моделлю, та певним «еталонним» розподілом. Такий цільовий розподіл може бути або емпіричним (на основі частот токенів у реальних даних), або − у випадку передбачення наступного слова − наближеним до імпульсного (one-hot) розподілу, де вся ймовірність зосереджена на одному правильному токені, а для решти вона дорівнює нулю. У такій ситуації KL-дивергенція від цього one-hot розподілу до ймовірнісного розподілу моделі зводиться до негативного логарифму ймовірності, яку модель надала правильному токену. Мінімізуючи середнє значення KL-дивергенції на тренувальному наборі, таким чином максимізується правдоподібність даних, що є ключовим критерієм у навчанні мовних моделей. Саме тому функція втрат у таких моделях зазвичай включає компонент крос-ентропії, яка безпосередньо пов'язана з KL-дивергенцією. У практиці KL також застосовується для регуляризації (наприклад, у варіаційних автоенкодерах) або для порівняння розподілів до й після донавчання моделі, або ж між двома різними моделями. Аналіз метрик дозволяє зрозуміти, наскільки розподіл моделі відхиляється від очікуваного, і в яких місцях модель робить найбільшу «помилку» у розподілі ймовірностей.

Для оцінки якості мовних моделей використовується метрика перплексії. Перплексія пов'язана з ентропією розподілу й по суті відображає середню кількість варіантів, між якими «вагається» модель при прогнозі наступного токена [9]. Формально перплексію можна визначити як експоненту від ентропії: $\mathrm{PPL} = 2^{H(P)}$ при логарифмі за основою 2 (або e^H при натуральному логарифмі). Наприклад, якщо ентропія вихідного розподілу моделі дорівнює 1 біт, то $\mathrm{PPL} = 2^1 = 2$, що інтерпретується як «модель в середньому вибирає з 2 рівноймовірних варіантів». Низька перплексія означає, що модель дуже впевнено передбачає наступне слово (фактично розподіл має низьку ентропію) − така модель показує кращу передбачуваність і якість. Висока перплексія, навпаки, вказує на велику невизначеність: модель розподіляє ймовірність між багатьма словами і не має чіткого прогнозу, що свідчить про гіршу прогностичну здатність [9]. У сфері обробки мови перплексія є стандартним показником: вона вимірюється на тестових корпусах для порівняння моделей. Сучасні великі мовні моделі (LLM) досягають перплексії на рівні десятків для складних корпусів. Зауважимо, що занадто низька перплексія (наближена до 1) на згенерованому тексті не завжди бажана − дослідження показують, що якщо модель генерує текст із перплексією, значно нижчою за притаманну людській мові, такий текст може стати повторюваним або нецікавим [7]. Тому метою є збалансувати перплексію: модель повинна достатньо впевнено передбачати слова, але не досягати тривіальної передбачуваності, щоб генерований текст залишався осмисленим і різноманітним.

Окрім авторегресивних мовних моделей, що генерують текст токен за токеном через категоріальні розподіли, розглянемо також дифузійні моделі. Це клас генеративних моделей, який реалізує зовсім інший підхід до породження даних. Дифузійні моделі працюють у два етапи: спочатку додають шум (стохастично «руйнують» структуру даних), а потім навчаються інверсного процесу − поступового денойзингу (процес видалення шуму з даних) для відновлення даних з шуму. У випадку зображень це означає перетворити чисте зображення на випадковий шум і назад, а в випадку тексту − поступово замінювати слова випадковими символами або спеціальними масками, а потім відновлювати речення. На відміну від традиційних GPT-подібних моделей, які прогнозують наступний токен на основі попередніх, дифузійні мовні моделі можуть відновлювати весь текст за декілька кроків, враховуючи контекст з обох боків (bidirectional modeling) і навіть генерувати всі токени паралельно [5]. Такий підхід дозволяє уникнути деяких обмежень авторегресивних методів − наприклад, «curse of irreversibility» (коли виправити помилку заднім числом неможливо, бо текст уже згенеровано). Дифузійні моделі вже продемонстрували конкурентні результати: зокрема, у 2025 році було представлено перші дифузійні мовні моделі, які за якістю тексту наблизились до авторегресивних GPT, водночас забезпечуючи вищу швидкість генерації (за рахунок паралелізму). Однак, ці моделі все ще потребують багато ітерацій для генерування виходу і складні у навчанні, але їхній розвиток відкриває нові можливості для керованого та більш гнучкого генерування тексту.

Висновки

Ймовірнісні принципи лежать в основі роботи генеративного ШІ. Розуміння того, як формується розподіл ймовірностей токенів і як ним можна керувати (через softmax, температуру, sampling truncation), є ключем до налаштування поведінки мовної моделі − від прагматично точних відповідей до творчо несподіваних. Метрики на кшталт ентропії, KL-дивергенції та перплексії дозволяють проаналізувати ці розподіли: вони дають інструменти для кількісного оцінювання невизначеності та знань моделі, для зіставлення

моделі з еталонним розподілом або іншою моделлю, для відстеження прогресу при навчанні. Нові підходи, такі як дифузійні моделі, демонструють, що можна відмовитися від покрокового передбачення токенів і все одно генерувати зв'язний текст, однак, і в цих моделях ймовірності відіграють центральну роль, керуючи процесом додавання та видалення шуму. Отже, акцент на ймовірнісному підході, розуміння розподілу ймовірностей токенів є фундаментом для подальшого розвитку генеративного штучного інтелекту, поєднуючи математичну точність із практичними методами підвищення якості й різноманітності згенерованого контенту.

Література

1. Юрчак І., Хіч А., Оксентюк В. Розуміння великих мовних моделей: майбутнє штучного інтелекту. Computer design systems. Theory and practice. Vol. 6, No. 2, 2024. С. 51-60.

2. Сімченко С.В., Лиходєєва Г.В, Левченко В.В., Морозова С.В., Демченко Н.Н. Використання великих мовних моделей в освітній, науковій та дослідницькій діяльності. URL: https://jai.in.ua/archive/2025/2025-1-5.pdf (Дата звернення 01.12.25).

3. Ворочек О.Г., Соловей І.В., Використання мовних моделей штучного інтелекту для генерації публікацій у соціальних мережах. Технічна інженерія. № 1 (93) 2024. С.128-133.

4. Chip H. Generation configurations: temperature, top-k, top-p, and test time compute. URL: https://huyenchip.com/2024/01/16/sampling.html (Дата звернення 01.12.25).

5. Diffusion languagem: the new paradigmhttp. URL: https://huggingface.co/blog/ProCreations/diffusion-language-model (Дата звернення 20.12.25).

6. Information theory in machine learning URL: https://www.geeksforgeeks.org/machine-learning/information-theory-in-machine-learning/ (Дата звернення 20.12.25).

7. The relationship between perplexity and entropy in NLP. URL: https://medium.com/data-science/the-relationship-between-perplexity-and-entropy-in-nlp-f81888775ccc (Дата звернення 20.12.25).

8. Дівергенція Кульбака-Лейблера. URL: https://surl.lu/ljpiop (Дата звернення 21.12.25).

9. Perplexity for LLM evaluation. URL: https://www.comet.com/site/blog/perplexity-for-llm-evaluation/ (Дата звернення 21.12.25).

References

1. Yurchak I., Khich A., Oksentiuk V. Rozuminnia velykykh movnykh modelei: maibutnie shtuchnoho intelektu. Somputer design systems. Theory and practice. Vol. 6, No. 2, 2024. S. 51-60.

2. Simchenko S.V., Lykhodieieva H.V, Levchenko V.V., Morozova S.V., Demchenko N.N. Vykorystannia velykykh movnykh modelei v osvitnii, naukovii ta doslidnytskii diialnosti. URL: https://jai.in.ua/archive/2025/2025-1-5.pdf (Data zvernennia 01.12.25).

3. Vorochek O.H., Solovei I.V., Vykorystannia movnykh modelei shtuchnoho intelektu dlia heneratsii publikatsii u sotsialnykh merezhakh. Tekhnichna inzheneriia. № 1 (93) 2024. S.128-133.

4. Chip H. Generation configurations: temperature, top-k, top-p, and test time compute. URL: https://huyenchip.com/2024/01/16/sampling.html (Data zvernennia 01.12.25).

5. Diffusion languagem: the new paradigmhttp. URL: https://huggingface.co/blog/ProCreations/diffusion-language-model (Data zvernennia 20.12.25).

6. Information theory in machine learning URL: https://www.geeksforgeeks.org/machine-learning/information-theory-in-machine-learning/ (Data zvernennia 20.12.25).

7. The relationship between perplexity and entropy in NLP. URL: https://medium.com/data-science/the-relationship-between-perplexity-and-entropy-in-nlp-f81888775ccc (Data zvernennia 20.12.25).

8. Diverhentsiia Kulbaka-Leiblera. URL: https://surl.lu/ljpiop (Data zvernennia 21.12.25).

9. Perplexity for LLM evaluation. URL: https://www.comet.com/site/blog/perplexity-for-llm-evaluation/ (Data zvernennia 21.12.25).