УДК 004.8: 004.91

В. Б. Мокін[1]
Б. Ю. Варер[1]

ДОСЛІДЖЕННЯ ВПЛИВУ ПАРАМЕТРІВ ВЕЛИКОЇ МОВНОЇ МОДЕЛІ НА РІЗНОМАНІТНІСТЬ ЗГЕНЕРОВАНОГО ТЕКСТУ

[1] Вінницький національний технічний університет, Україна

Анотація

Робота присвячена дослідженню впливу параметрів генерування (temperature та top_p) на різноманітність тексту, створеного великими мовними моделями. Запропоновано методологію оцінювання різних аспектів різноманітності на основі шести метрик, об'єднаних в єдиний показник. Результати експериментального дослідження 30 різних промптів із 36 комбінаціями параметрів для моделі Meta Llama 3.2 3B-instruct показали, що оптимальні значення різноманітності досягаються при помірних і високих значеннях temperature та top_p. Дослідження має практичне значення для формування якісних синтетичних датасетів, що зменшує ризик перенавчання моделей і покращує їхню узагальнюючу здатність.

Ключові слова: великі мовні моделі, параметри генерування, різноманітність тексту, синтетичні дані, донавчання.

Abstract

The work is devoted to investigating the influence of generation parameters (temperature and top_p) on the diversity of text produced by large language models. A methodology is proposed for evaluating different aspects of diversity based on six metrics, combined into a single indicator. The results of an experimental study involving 30 different prompts and 36 parameter combinations for the Meta Llama 3.2 3B-instruct model showed that optimal diversity values are achieved with moderate and high settings of temperature and top_p. This research is of practical significance for creating high-quality synthetic datasets, which reduces the risk of model overfitting and improves their ability to generalize.

Keywords: large language models, generation parameters, text diversity, synthetic data, fine-tuning.

Вступ

Сучасні великі мовні моделі (англ. “Large Language Models”, скорочено - LLM), такі як OpenAI GPT, Anthropic Claude, Google Gemini, Meta LLaMA, DeepSeek продовжують демонструвати вражаючі здібності до оброблення природної мови, розв’язання складних задач та генерування якісних текстів, проте для максимальної ефективності у специфічних завданнях ці моделі часто потребують донавчання (англ. “fine-tuning”). Процес донавчання дозволяє адаптувати попередньо навчені моделі до заданої предметної області або до конкретних задач чи потреб, але це, у свою чергу, пов’язано з численними технічними викликами [1]. Зокрема, одним із таких викликів є вимога щодо максимальної різноманітності корпусу (набору текстів) тренувальних даних. Різноманітність тренувальних текстів є критично важливою для досягнення оптимальних результатів донавчання, оскільки вона забезпечує широке покриття можливих контекстів та патернів природномовного тексту, а також дозволяє уникнути перенавчання (англ. “overfitting”) та покращує здатність моделі до узагальнення. На практиці, для донавчання моделей часто використовують набори з синтетичних даних, згенерованих іншими LLM, оскільки такий підхід дає можливість швидко створити великі обсяги тренувальних даних без необхідності ручного збирання та розмічування, що суттєво скорочує тривалість підготовки та витрати на донавчання, а також дозволяє отримати вузькоспеціалізовані дані, які складно знайти у достатній кількості в природномовних текстах [2-5]. Щодо синтетичних даних питання різноманітності тексту є ще більш актуальним, оскільки синтетичні датасети часто страждають від однотипності та відсутності природномовної варіативності [6]. Саме тому дослідження впливу параметрів генерування на різноманітність тексту у синтетичних даних є дуже актуальним. У більшості сучасних LLM, серед найбільш поширених параметрів, що впливають на результат генерування, можна виділити такі параметри, як top_p та temperature. Ці параметри відіграють ключову роль в управлінні детермінованістю та стохастичністю генерації. Параметр temperature впливає на розподіл ймовірностей

при виборі наступного токена - збільшення цього параметра збільшує невизначеність при виборі наступного токена. Тим часом, параметр top_p (також відомий як “nucleous sampling”) обмежує вибір наступного токена лише найбільш ймовірними варіантами, сукупна ймовірність яких не перевищує встановлене порогове значення.

Результати дослідження

Методологія дослідження полягала в аналізі впливу ключових параметрів генерування на різноманітність згенерованого тексту на прикладі популярної моделі Meta “Llama 3.2 3B-instruct” [7]. Для цього було створено 30 різних промптів для LLM (з англ. “prompt” - підказка, яка є входом моделі), які в подальшому були використані для генерування. Для кожного промпту було згенеровано 36 варіантів результатів, кожен із різним набором параметрів temperature і top_p, обраним із 36 наперед заданих комбінацій - по 6 різних значень на кожен параметр. Ці значення були отримані з урахуванням типових рекомендацій щодо кожного параметру. Для забезпечення відтворюваності та об'єктивності порівняння всі генерації здійснювались із фіксованим початковим станом генератора випадкових чисел (seed).

Для обчислення різноманітності згенерованого тексту було застосовано 6 різних метрик, кожна з яких вимірює окремий лексико-семантичний аспект, який характеризує різні особливості різноманітності тексту:

- кількість унікальних слів (англ. “Unique Words Count” - UWC);
- середня довжина слова (англ. “Average Word Length” - AWL);
- частка значущих частин мови (англ. “Lexical POS Richness” - POS-R) - частка іменників, дієслів та прикметників серед усіх слів;
- відношення кількості змістовних слів до загальної довжини (англ. “Content Density Ratio” - CDR), де змістовними словами є ті, які не є так званими “стоп-словами” (англ. “a, an, the, is, of, in, on” тощо - як правило, використовується спеціальний список stopword з бібліотеки NLTK у Python);
- середня довжина речень у тексті (англ. “Average Sentence Length in Words” - ASL);
- індекс типовості (англ. “Genericness Score” - GS) - частка типових стоп-слів stopword.

Додатково було обчислено узагальнений індекс різноманітності “META”, що представляє собою нормалізоване середнє значення всіх шести метрик, враховуючи їхню спрямованість. На рис. 1 наведена теплова карта, на якій відображено усереднені результати шести показників різноманітності (UWC, AWL, POS-R, CDR, ASL, GS), об’єднані в метрику META, обчислену для 30 різних промптів. Із рисунка 1 видно, що деякі поєднання значень temperature і top_p (наприклад, temperature = 0,55 або 0,85 у поєднанні з top_p = 0,95) дають вищі значення META, а отже, забезпечують більш виражену різноманітність у згенерованих текстах. Водночас можна помітити, що за дуже низьких або дуже високих значень temperature 0,1 і 0,85 та різних top_p значення META часто виявляється дещо нижчим, хоча загальна різниця не є занадто великою. Цікавим результатом на рис. 1 є те, що top_p = 0,95 майже завжди дає кращу різноманітність, ніж за 0,99, ймовірно, у другому варіанті з’являється більше стоп-слів, але це ще потребує перевірки як в розрізі видів промптів, так і - складових критерію META.

На рис. 2 наведено графіки, які ілюструють те, як змінюється META для кожного з цих 30 промптів окремо, в залежності від конкретної пари значень параметрів temperature та top_p. Кожна точка на графіку відповідає одному згенерованому тексту із використанням відповідних параметрів та вихідного промпту. Позиція точки в межах одного графіку з фіксованим top_p визначається промптом та значенням META, а колір точки відображає значення temperature.

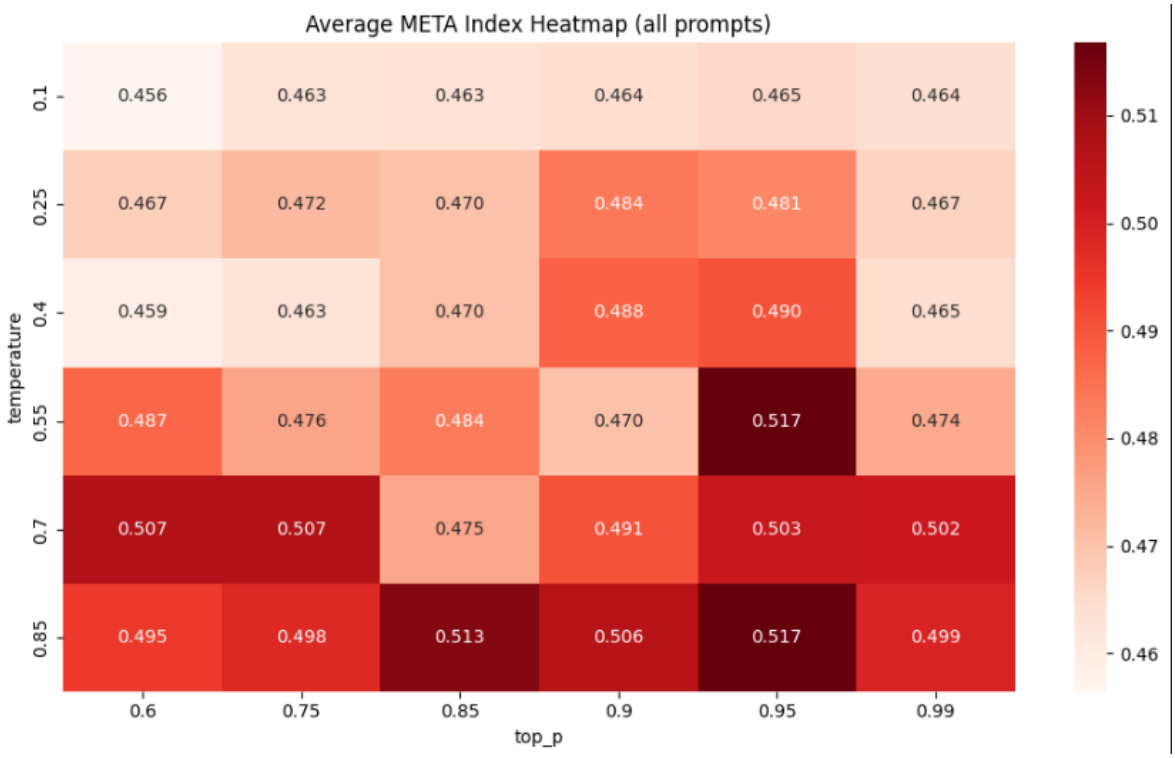


Рисунок 1. Теплова карта індексу META для різних значень temperature та top_p

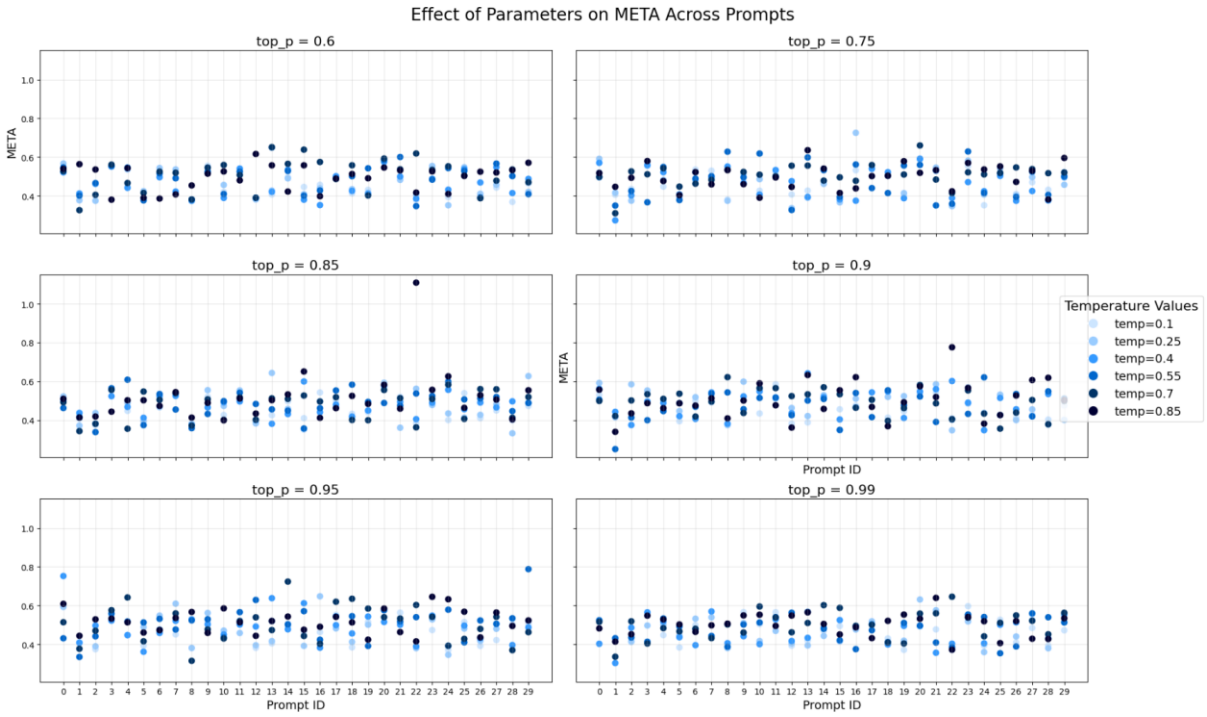


Рисунок 2. Вплив параметрів temperature та top_p на META в розрізі окремих промптів

На графіку на рис. 2 можна бачити, що, загалом, вищі значення temperature корелюють із вищим META (темніші точки частіше розташовані вище). Це логічно узгоджується з тим, що збільшення temperature, зазвичай, підвищує невизначеність у процесі генерування тексту. Втім, щодо окремих промптів іноді трапляються винятки, де занадто висока temperature може призвести до зниження різноманітності, що свідчить про важливість вибору параметрів під конкретну задачу. Також, цікаво, що на рис. 2, як і на рис. 1, варіант top_p = 0,99 для більшості промптів є гіршим за варіант 0,95.

Висновки

У дослідженні показано, що параметри temperature і top_p суттєво впливають на різноманітність згенерованого тексту, причому більші значення temperature, загалом, збільшують різноманітність, виміряну через усереднення шести загальних метрик різноманітності (UWC, AWL, POS-R, CDR, ASL, GS). При цьому, в окремих випадках, занадто високі значення temperature можуть, навпаки, зменшити різноманітність, що робить підбирання значень параметрів залежним від контексту конкретної задачі. Найбільші значення усередненої метрики різноманітності, здебільшого, спостерігалися за помірних або відносно високих значень top_p. Отже, для отримання максимально різноманітного тексту, найкращими параметрами для Llama 3.2-3B-instruct: temperature = 0.55 - 0.85 і top_p = 0.95, але варто провести більш глибоке дослідження як у розрізі видів промптів, так і - складових критерію META, а також випробувати інші види LLM.

Отримані результати мають важливе практичне значення для формування синтетичних датасетів (корпусів) для донавчання великих мовних моделей. Грамотне налаштування параметрів генерування дає змогу створити якісніший та більш різноманітний датасет, що зменшує ризик перенавчання моделей і покращує їхню узагальнюючу здатність.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. VM K., Warrier H., Gupta Y. и др. Fine tuning LLM for enterprise: Practical guidelines and recommendations // arXiv preprint arXiv:2404.10779. — 2024.
2. Kang A., Chen J.Y., Lee-Youngzie Z., Fu S. Synthetic data generation with LLM for improved depression prediction // arXiv preprint arXiv:2411.17672. — 2024.
3. Setlur A., Garg S., Geng X., Garg N., Smith V., Kumar A. RL on incorrect synthetic data scales the efficiency of LLM math reasoning by eight-fold // Advances in Neural Information Processing Systems. — 2024. — Т. 37. — С. 43000–43031.
4. Wei J., Huang D., Lu Y., Zhou D., Le Q.V. Simple synthetic data reduces sycophancy in large language models // arXiv preprint arXiv:2308.03958. — 2023.
5. Woolsey C.R., Bisht P., Rothman J., Leroy G. Utilizing large language models to generate synthetic data to increase the performance of BERT-based neural networks // AMIA Joint Summits on Translational Science Proceedings. — 2024. — 31 мая. — Т. 2024. — С. 429–438. — PMID: 38827067; PMCID: PMC11141799.
6. Bisbee J., Clinton J.D., Dorff C., Kenkel B., Larson J.M. Synthetic replacements for human survey data? The perils of large language models // Political Analysis. — 2024. — Т. 32, № 4. — С. 401–416. — DOI: 10.1017/pan.2024.5.
7. Meta. Llama 3.2 3B [Електронний ресурс]. — Режим доступу: https://huggingface.co/meta-llama/Llama-3.2-3B (дата звернення: 22.03.2025).

Мокін Віталій Борисович — д-р техн. наук, професор, завідувач кафедри системного аналізу та інформаційних технологій; e-mail: vbmokin@vntu.edu.ua;

Варер Борис Юхимович — аспірант кафедри системного аналізу та інформаційних технологій; e-mail: androbor17@gmail.com

Mokin Vitalii B. — Dr. Sc. (Eng.), Professor, Head of the Chair of System Analysis and Information Technologies, e-mail: vbmokin@vntu.edu.ua;

Varer Borys Yu. — Post-graduate Student of the Chair of System Analysis and Information Technologies, e-mail: androbor17@gmail.com