УДК 004.8

Б. Ю. Варер
В. Б. Мокін

ДОСЛІДЖЕННЯ ВПЛИВУ ВЗАЄМОДІЇ РІЗНИХ МОДЕЛЕЙ НА РОЗПОДІЛ ЙМОВІРНОСТЕЙ НАСТУПНОГО ТОКЕНА У ВЕЛИКИХ МОВНИХ МОДЕЛЯХ

Вінницький національний технічний університет

Анотація

Досліджено вплив зміни великої мовної моделі при фіксованому контексті на розподіл ймовірностей наступного токена, у порівнянні з впливом зміни контексту при фіксованій моделі. Проведено експериментальне порівняння факторів зміни моделі та зміни контексту з використанням моделей Meta LLaMA 3.2-3B та Microsoft Phi 4-mini на датасеті з 60 питань з різних предметних областей. За допомогою дивергенції Дженсена-Шеннона встановлено, що зміна моделі при фіксованому контексті призводить до змін у розподілі наступного токена (JSD 0.640-0.678), які є співставні за величиною зі зміною контексту при фіксованій моделі (JSD 0.638-0.721). Результати підтверджують важливість оптимального вибору моделей під час проєктування ефективних систем штучного інтелекту.

Ключові слова: великі мовні моделі, кооперація моделей, агентні системи, дивергенція Дженсена-Шеннона, розподіл ймовірностей, штучний інтелект

Abstract

The influence of changing a large language model with fixed context on the distribution of next token probabilities was investigated, compared to the influence of changing a context with a fixed model. An experimental comparison of model change and context change factors was conducted using Meta LLaMA 3.2-3B and Microsoft Phi-4-mini models on a dataset of 60 questions from various subject domains. Using Jensen-Shannon divergence, it was established that changing the model with fixed context leads to changes in the next token distribution (JSD 0.640-0.678) that are comparable in magnitude to changing context with a fixed model (JSD 0.638-0.721). The results confirm the importance of optimal model selection when designing effective artificial intelligence systems.

Keywords: large language models, model cooperation, agent systems, Jensen-Shannon divergence, probability distribution, artificial intelligence

Сучасні системи на базі штучного інтелекту дедалі частіше використовують композицію декількох великих мовних моделей (англ. “Large Language Model” - LLM) для розв’язання складних задач. Зокрема, цей підхід використовується як частина більш загального агентного підходу, який передбачає декомпозицію складної задачі на множину спеціалізованих підзадач та координацію їх розв’язання через систему взаємодіючих агентів. При цьому, агентний підхід може бути реалізований як через використання однієї базової моделі з різними інструкціями-промптами (англ. “prompt” - підказка), так і через композицію декількох різних моделей, кожна з яких оптимізована для конкретного типу задач [1].

Композиція декількох різних моделей дозволяє поєднувати сильні сторони різних архітектур та емерджентних особливостей кожної моделі. Наприклад, одна модель може бути оптимізована для логічного міркування, інша - дотренована для оброблення даних специфічної предметної області тощо [2-4]. Однак, при такому підході виникає фундаментальне питання: чи є вплив зміни моделі на результат генерування достатньо суттєвим порівняно з впливом зміни контексту (промпту)? Розуміння цього ефекту є критично важливим для проєктування ефективних систем кооперації великих мовних моделей, оскільки, в перспективі, дозволяє прогнозувати та контролювати характер їх взаємодії.

Більшість сучасних LLM побудовані з використанням архітектури трансформерів [5], що зумовлює авторегресійну природу генерування тексту. Передбачення в авторегресійній моделі здійснюється за наступною формулою:

$$P_M(x_{t+1} | x_{1:t}) = softmax(M(x_{1:t})),$$

де $x_{1:t}$ - послідовність токенів з першого елементу до елементу t; x_{t+1} - наступний токен, який передбачується моделлю; M - модель (наприклад, нейронна мережа на основі трансформерної архітектури), яка перетворює вхідну послідовність токенів на так звані логіти (англ. “logits”); softmax - функція нормалізації, що перетворює логіти на розподіл ймовірностей.

Для дослідження того, як змінюється розподіл наступного токену моделі, коли частина контексту генерується іншою моделлю, розглянемо дві моделі $M_1(X)$ і $M_2(X)$, на вхід яких подається одна і та ж послідовність токенів $X = x_{1:t}$ для генерування наступних N токенів. Після цього змоделюємо наступні сценарії для кожної з двох моделей:

1. Гомогенний контекст: модель прогнозує наступний токен після N токенів, згенерованих цією ж моделлю.
2. Гетерогенний контекст: модель прогнозує наступний токен після N токенів, згенерованих іншою моделлю.

В результаті буде отримано чотири розподіли ймовірностей для N+1 токена:

1. P^1_{M1} — розподіл, отриманий при використанні моделі M_1, коли вона продовжує власну послідовність (гомогенний контекст). Тобто, модель M_1 генерує N токенів після початкового контексту X, а потім використовує їх для передбачення наступного токена.
2. P^2_{M1} — розподіл, отриманий при використанні моделі M_1, коли вона продовжує послідовність, згенеровану моделлю M_2 (гетерогенний контекст). Таким чином, модель M_1 отримує змішаний вхід: початковий контекст X, а далі N токенів, які були згенеровані іншою моделлю.
3. P^1_{M2} — розподіл, отриманий при використанні моделі M_2, коли вона продовжує послідовність, згенеровану моделлю M_1.
4. P^2_{M2} — розподіл, отриманий при використанні моделі M_2, коли вона продовжує власну послідовність.

Формально це можна представити наступним чином:

$$P^1_{M_1}(x_{t+N+1} \mid X \oplus x^{(1)}_{t+1:t+N}) = \mathrm{softmax}(M_1(X \oplus x^{(1)}_{t+1:t+N}))$$

$$P^2_{M_1}(x_{t+N+1} \mid X \oplus x^{(2)}_{t+1:t+N}) = \mathrm{softmax}(M_1(X \oplus x^{(2)}_{t+1:t+N}))$$

$$P^1_{M_2}(x_{t+N+1} \mid X \oplus x^{(1)}_{t+1:t+N}) = \mathrm{softmax}(M_2(X \oplus x^{(1)}_{t+1:t+N}))$$

$$P^2_{M_2}(x_{t+N+1} \mid X \oplus x^{(2)}_{t+1:t+N}) = \mathrm{softmax}(M_2(X \oplus x^{(2)}_{t+1:t+N}))$$

де $x^{(i)}_{t+1:t+N}$ - послідовність N токенів, згенерованих моделлю M_i в авторегресійному процесі з початкового контексту $X = x_{1:t}$, тобто кожен токен $x^{(i)}_{t+j}$ визначається як:

$$x^{(i)}_{t+j} = argmax\ softmax(M_i(x^{(i)}_{1:t+j-1})), \quad j = 1, \ldots, N.$$

Метою дослідження є кількісне порівняння впливу двох факторів на формування розподілу наступного токена: модельного (зміна моделі при фіксованому контексті) та контекстного (зміна контексту при фіксованій моделі). Для кількісного оцінювання впливу було обчислено дивергенцію Дженсена-Шеннона між відповідними розподілами [6]: контекстний фактор - [JSD(P_{11}, P_{12}), JSD(P_{22}, P_{21})]; модельний фактор - [JSD(P_{11}, P_{21}), JSD(P_{12}, P_{22})], де:

$$\mathrm{JSD}(P, Q) = \frac{1}{2} \sum_i \left[P(i) \log_2 \left(\frac{2P(i)}{P(i) + Q(i)} \right) + Q(i) \log_2 \left(\frac{2Q(i)}{P(i) + Q(i)} \right) \right]$$

Для проведення такого експерименту були обрані моделі Meta LLaMA 3.2 3B [7] і Microsoft Phi 4-mini [8] та сформовано датасет з 60 питань, які належать до 6 різних предметних областей: загальні знання, пошук поради, програмування, математика, креативні завдання, повсякденні питання. Кожне питання було подано на вхід обом моделям, які незалежно згенерували по N = 10 токенів у жадібному (англ. “greedy”) режимі. В результаті, для кожного питання, було отримано пару контекстів. Далі, для кожної моделі і кожного питання був оцінений розподіл ймовірностей наступного токена для обох варіантів контексту і були отримані розподіли P^1_{M1}, P^1_{M2}, P^2_{M1}, P^2_{M2}.

Далі для відповідних пар розподілів було обчислено дивергенцію Дженсена-Шеннона. У таблиці 1 наведено середні значення JSD, отримані в результаті експерименту.

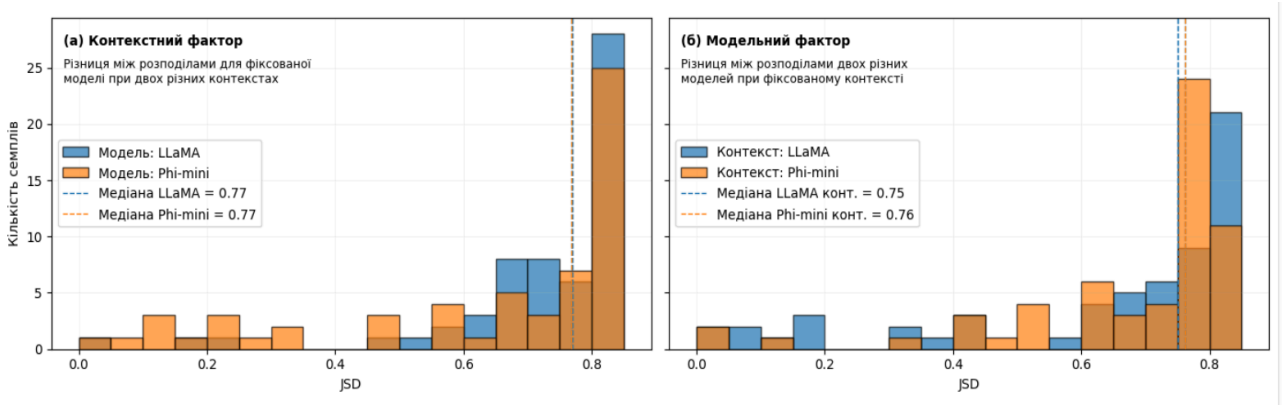
Таблиця 1. Середні значення JSD для контекстних та модельних факторів

Зміна контексту при фіксованій моделі	
Модель: LLaMA, Контекст: LLaMA ↔ Phi	**0.721 ± 0.161**
Модель: Phi, Контекст: Phi ↔ LLaMA	**0.638 ± 0.248**
Зміна моделі при фіксованому контексті	
Контекст: LLaMA, Модель: LLaMA ↔ Phi	**0.640 ± 0.248**
Контекст: Phi, Модель: Phi ↔ LLaMA	**0.678 ± 0.189**

Результати експерименту показують, що зміна LLM при фіксованому контексті впливає на розподіл наступного токена майже так само, як і зміна контексту при фіксованій моделі. Також слід зазначити різну чутливість моделей до зміни контексту: LLaMA демонструє вищу чутливість (JSD = 0.721) порівняно з Phi (JSD = 0.638) при меншій варіативності (σ = 0.161 проти σ = 0.248 у Phi). Це може свідчити про те, що LLaMA сильніше “реагує” на зміни у стилі та змісті попереднього тексту, але, водночас, демонструє більш стабільну та передбачувану поведінку. Натомість, Phi показує меншу загальну чутливість до контекстних змін, проте має більшу варіативність, що може вказувати на менш консистентне оброблення різних контекстів.

Аналіз модельного фактору також показує деяку асиметрію в поведінці моделей: зміна моделі при контексті, згенерованому Phi, призводить до більших змін у розподілі (JSD = 0.678) порівняно з контекстом, згенерованим LLaMA (JSD = 0.640), однак, при цьому, варіативність модельного фактору для контексту, згенерованому LLaMA, також вища. Це може свідчити про те, що хоча контекст Phi в середньому викликає більші зміни при переході між моделями, ці зміни є більш передбачуваними та консистентними. Натомість, контекст LLaMA, попри менший середній вплив, може містити елементи, які по-різному інтерпретуються різними моделями, що призводить до більшої варіативності результатів.

На рисунку 1 представлено розподіл значень дивергенції Дженсена-Шеннона, у разі зміни моделі для двох типів контексту (LLaMA та Phi), який показує переважання високих значень JSD, незалежно від походження контексту. Як можна побачити з гістограми, більшість значень JSD концентрується в діапазоні 0.6-0.8, що свідчить про стабільно високий рівень впливу зміни моделі на характеристики розподілу наступного токена.



а) б)

Рисунок 1. Гістограма значень JSD для контекстного (а) та модельного (б) факторів впливу на розподіл ймовірностей наступного токена

Проведене дослідження показало, що модельний фактор має суттєвий вплив на розподіл ймовірностей наступного токена, причому цей ефект є співставним за величиною з фактором зміни контексту, що підтверджує необхідність ретельного вибору моделей при проєктуванні систем штучного інтелекту. Співставність модельного та контекстного факторів вказує на те, що архітектурні особливості та специфіка навчання різних моделей відіграють не менш важливу роль у формуванні поведінки системи, ніж семантичний зміст вхідної інформації. Це спостереження є особливо

важливим для агентних систем, де різні моделі можуть використовуватись для отримання синергетичного ефекту, оскільки вибір конкретних моделей може суттєво впливати на загальну ефективність системи. Також, аналіз розподілу значень JSD показав, що модельний фактор проявляється доволі стабільно, незалежно від предметної області контексту, що дозволяє припустити, що виявлені закономірності можуть мати загальний характер. Обмеженням даного дослідження є використання лише двох моделей та фіксованої довжини контексту (N=10 токенів). Варто спробувати розширити аналіз, за рахунок використання більшої кількості моделей різних розмірів та архітектур.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Li X. та ін. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges // Vicinagearth. – 2024. – Т. 1, № 1. – С. 9. – DOI: 10.1007/s44336-024-00009-2. – ISSN 3005-060X.
2. Shen Y. та ін. HuggingGPT: solving AI tasks with ChatGPT and its friends in Hugging Face // Advances in Neural Information Processing Systems. – 2023. – Т. 36. – С. 38154–38180.
3. Hong S. та ін. METAGPT: meta programming for a multi-agent collaborative framework // Proceedings of the 12th International Conference on Learning Representations (ICLR 2024), Vienna (Austria), 7–11 May 2024. – 2024. – Текст : електронний ресурс. – URL: https://github.com/geekan/MetaGPT (дата звернення: 10.06.2025).
4. Jiang D., Ren X., Lin B. Y. LLM-Blender: ensembling large language models with pairwise comparison and generative fusion // Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), Toronto (Canada), 9–14 July 2023. – 2023. – С. 14165–14178. – DOI: 10.18653/v1/2023.acl-long.792. – URL: https://aclanthology.org/2023.acl-long.792 (дата зверн.: 10.06.2025).
5. Vaswani A. та ін. Attention is all you need // Advances in Neural Information Processing Systems. – 2017. – Т. 30. – С. 5998–6008.
6. Lin J. Divergence measures based on the Shannon entropy // IEEE Transactions on Information Theory. – 1991. – Т. 37, № 1. – С. 145–151. – DOI: 10.1109/18.61115. – ISSN 0018-9448.
7. Meta AI. LLaMA 3.2-3B-Instruct: multilingual large language model // Hugging Face : [Електронний ресурс]. – 2024. – URL: https://huggingface.co/meta-llama/Llama-3.2-3B-instruct (дата зверн.: 10.06.2025).
8. Microsoft Research. Phi-4-mini-instruct: small language model // Hugging Face : [Електронний ресурс]. – 2025. – URL: https://huggingface.co/microsoft/Phi-4-mini-instruct (дата зверн.: 10.06.2025).

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