УДК 004.89
DOI:https://doi.org/10.17721/1812-5409.2025/2.30

Руслан ПРАВОСУД, асп.
ORCID ID: 0009-0008-9450-5515
e-mail: pronod9999@gmail.com
Київський національний університет імені Тараса Шевченка, Київ, Україна

Олександр МАРЧЕНКО, д-р фіз.-мат. наук, проф.
ORCID ID: 0000-0002-5408-5279
e-mail: omarchenko@univ.kiev.ua
Київський національний університет імені Тараса Шевченка, Київ, Україна

ПЕРЕНЕСЕННЯ СТИЛЮ ДІАЛОГУ ЗА ДОПОМОГОЮ ПАРАМЕТРИЧНО ЕФЕКТИВНОГО ДОНАВЧАННЯ

Розглянуто застосування сучасних методів донавчання, що характеризуються високою параметричною ефективністю, до задачі генерації діалогових реплік у стилі певного вигаданого або реального персонажа. Основна мета дослідження полягає в розробленні підходу, що допомагає адаптувати вже наявні мовні моделі до специфічного стилю спілкування або мовлення, властивого конкретній особистості, з використанням мінімальної кількості нових параметрів. Це дає змогу уникнути повного перенавчання великих моделей і зменшити обчислювальні витрати.

Однією із ключових цілей є можливість реалізації компактної, але функціональної мовної моделі, яку можна буде запускати на окремому комп'ютері без необхідності використання серверів чи потужних графічних процесорів. Такий підхід відкриває нові можливості для персоналізованих або розважальних застосувань, зокрема створення чат-ботів, які можуть взаємодіяти з користувачами в образі певного персонажа з книги, фільму чи історичної особи.

Для реалізації такого підходу запропоновано використання методу LoRA (Low-Rank Adaptation), що допомагає ефективно донавчати наявні трансформерні архітектури – зокрема модель GPT-2 – без повного перенавчання її всіх шарів. Застосування LoRA дає змогу додавати стилістичні особливості до мовлення моделі, не втрачаючи її базових можливостей генерувати логічні, зв'язні тексти.

Щоб перевірити, наскільки добре модель навчилася імітувати цільовий стиль мовлення, ми додатково трену-ємо модель BERT як класифікатор, що розрізняє тексти в бажаному стилі та звичайні, нейтральні репліки. У такий спосіб ми отримуємо об'єктивну метрику якості генерації, що надає можливість оцінити ефективність донавчання.

Набір даних, який використано для навчання, було згенеровано за допомогою моделі GPT4-mini, що дає змогу швидко створити велику кількість прикладів у потрібному стилі, забезпечуючи достатній обсяг даних для ефективного навчання та валідації результатів.

Ключові слова: ***обробка природних мов, параметрично ефективне донавчання, низькорангова адаптація, генерація тексту.***

Класифікація відповідно до AMS 2020: 68T50, 68T07.

Вступ

Упродовж останніх років великі мовні моделі (LLMs) продемонстрували значний прогрес у генерації природномовного тексту, що знаходить застосування у створенні діалогових систем, художніх історій і персоналізованого контенту. Попри це, налаштування таких моделей на специфічні задачі або стилі часто вимагає значних обчислювальних потужностей, що ускладнює їхнє використання на пристроях з обмеженими ресурсами. Отже, постає актуальне завдання – здійснити ефективне стилістичне донавчання без суттєвих витрат ресурсів і збереженням високої якості результатів.

Завдання генерації стилістично узгодженого тексту широко досліджується вченими в галузі обробки природної мови (NLP), й останнім часом у цій сфері досягнуто значного прогресу завдяки розвитку масштабних попередньо натренованих мовних моделей і методів донавчання.

Перенесення стилю в NLP спрямоване на трансформацію тексту так, щоб зберігався його зміст, але змінювалися стилістичні характеристики. Перші підходи до цієї задачі базувалися на системах із правилами та ручному доборі ознак для кодування стилістичних властивостей. З появою глибокого навчання почали застосовувати варіаційні автокодери (VAE) та генеративно-змагальні мережі (GAN) для перенесення стилю в текстах.

Сучасні методи дедалі частіше використовують попередньо навчені мовні моделі, такі як GPT-2 та GPT-3, для стилістичної генерації. Наприклад, у роботі (Keskar et al., 2019) було показано, як використання керувальних кодів у моделях на базі GPT дає змогу спрямовувати генерацію в заданому стилістичному напрямку. Однак такі методи зазвичай потребують значних обчислювальних ресурсів, що обмежує їхнє застосування в середовищах з обмеженими можливостями. Наша робота усуває це обмеження шляхом використання параметрично ефективного донавчання для досягнення стилістичної узгодженості з мінімальними витратами ресурсів.

Традиційне донавчання великих мовних моделей зазвичай передбачає оновлення всіх параметрів, що є надзвичайно ресурсоємним процесом. Сучасні підходи до параметрично ефективного донавчання, зокрема AdapterFusion (Pfeiffer et al., 2021) та LoRA (Hu et al., 2021), розв'язують цю проблему шляхом упровадження легких модулів з параметрами, орієнтованих на конкретні задачі. Ці методи допомагають адаптувати попередньо навчені моделі до нових завдань із мінімальними витратами ресурсів й обсягом пам'яті.

Особливої уваги заслуговує LoRA, що поєднує простоту реалізації з високою ефективністю в донавчанні масштабних моделей. Завдяки вбудовуванню матриць низького рангу в механізми уваги моделі LoRA дає змогу ефективно адаптувати поведінку без змін основних ваг моделі.

Створення якісних наборів даних часто є вузьким місцем у виконанні завдань стилістичного донавчання. Низка сучасних досліджень розглядає використання великих моделей, таких як GPT-3 або GPT-4, для генерації синтетичних

датасетів. Наприклад, у дослідженні (Schick, & Schütze, 2021) показано, що синтетичні дані можуть ефективно доповнювати навчання у випадках з обмеженими ресурсами. У нашій роботі ми використовуємо модель GPT4-mini для генерації різноманітного і якісного набору даних, що дпомагає зменшити потребу в ручній розмітці та забезпечити стилістичну узгодженість

Імітація мовлення конкретного персонажа відкриває нові можливості в галузях розваг, освіти й інтерфейсів "людина – комп'ютер". Наприклад, чат-боти з індивідуальним стилем, ігрові рольові сценарії чи генерація авторського тексту потребують здатності моделі генерувати стилістично відповідні та логічні репліки. Проте традиційні методи донавчання передбачають оновлення всієї моделі, що є надто ресурсозатратним.

Щоб усунути цю проблему, ми застосовуємо алгоритм Low-Rank Adaptation (LoRA) для адаптації попередньо навченої моделі GPT-2. Це дає можливість змінити поведінку моделі в бажаному напрямку, мінімізуючи кількість параметрів, які потрібно донавчати.

Для оцінки якості згенерованих відповідей ми використовуємо класифікатор на основі BERT, що навчається розпізнавати тексти в заданому стилі. Для створення навчального набору даних застосовуємо GPT4-mini, що дає змогу отримати різноманітні приклади високої якості.

Основні здобутки цієї роботи:

• *Ефективне донавчання з мінімальними ресурсами:* завдяки використанню LoRA ми зменшуємо обчислювальні витрати та забезпечуємо можливість запуску на менш потужних пристроях.

• *Практична придатність для сценаріїв, орієнтованих на персонажів:* запропонований підхід довів свою ефективність у створенні реплік, які відповідають стилю певної особи, що робить його корисним для історій, чат-ботів і синтетичних датасетів.

1. Методи

У процесі дослідження були використані методи аналізу, порівняння й експерименту.

2. Основні означення та результати

Модель для оцінки стилю

Підготовка датасету. Датасет для навчання моделі для оцінки стилю складається із двох частин: датасету філософських діалогів (Hypersniper, 2023) (близько 40 %), а також філософських питань, згенерованих GPT4-mini. Цей датасет містить приблизно 1000 філософських питань. Кожне питання згодом використовували як вхідний запит для GPT4-mini для генерації двох типів відповідей:

1. Відповідь у стилі філософа Сократа.
2. Відповідь з позиції звичайної людини.

Згенеровані пари були оброблені у такому форматі:

Людина: <питання>

Сократ: <відповідь>

Ці структуровані пари забезпечили навчальні дані для розрізнення стилістичних відмінностей між сократівськими та звичайними висловлюваннями.

Навчання моделі. Модель (Sanh, 2019) була донавчена для класифікації висловлювань як сократівських або звичайних шляхом присвоєння винагород. Позитивна винагорода очікувалася для відповідей у стилі Сократа, тоді як звичайним відповідям призначалася негативна винагорода.

Оцінювання моделі. Модель винагороди була оцінена за допомогою тестового датасету, що становив 20 % від початкового датасету. Результати показали:

- Середня винагорода понад *3.0* для відповідей у стилі Сократа.
- Середня винагорода менше *–3.0* для звичайних відповідей.
- Точність *99,5 %* у розрізненні сократівських і звичайних відповідей шляхом порівняння їхніх винагород.

Основна модель. У традиційному донавчанні всі або більшість параметрів моделі оновлюються. Натомість LoRA фіксує всі основні параметри моделі та додає невеликі додаткові низькорівневі (low-rank) матриці до певних компонентів трансформера – зазвичай до матриць запитів (Q), ключів (K), або значень (V) у механізмі самоуваги (self-attention). Замість оновлення великих оригінальних матриць, LoRA вставляє два додаткові шари: матрицю зниження розмірності. Під час донавчання оновлюються лише ці шари, а оригінальна модель залишається незмінною.

Попередньо натреновані моделі GPT2-large (Radford et al., 2019) (800 млн параметрів) та GPT2 (125 млн параметрів) були поліпшені за допомогою LoRA (Low-Rank Adaptation) шляхом додавання додаткових низькорівневих матриць для компонентів Q, K і V у шарах трансформера з рангом 32. Ці додаткові шари були єдиними, що підлягали донавчанню, і становили менше 1 % від загальної кількості параметрів моделі: 5,898,240 з 779,928,320 для GPT2-large та 1,179,648 з 125,620,225 для GPT2 відповідно.

Донавчання. Щоб краще відтворювати бажаний стиль мовлення, спершу ми провели стандартне донавчання моделей, мінімізуючи перехресну ентропійну втрату.

Як тренувальний датасет для цього підзавдання використовували діалоги з однієї репліки із Сократом у форматі:

Людина: <питання>

Сократ: <відповідь>

Після цього ми оцінили модель на тестовому датасеті, досягнувши середньої винагороди *3.123216* зі стандартним відхиленням *0.068859* для GPT2-Large та середньої винагороди *3.120361* зі стандартним відхиленням *0.074097* для GPT2, що досить суттєво відносно нагород, отриманих на моделі без донавчання (табл. 1).

Для оцінки зв'язності між собою за змістом питань та відповідей, а також відмінності між собою згенерованих відповідей, було використано BERTScore (Zhang et al., 2019). Результати наведено в табл. 2 – питання та відповіді мають відчутно більшу семантичну спорідненість, ніж пара довільних відповідей, що свідчить про різноматність генерованих реплік і зв'язність у межах поточного діалогу.

Таблиця 1

Порівняння моделей за стилістичною нагородою

Модель	Середнє значення метрики	Середнє відхилення метрики
GPT2-Large без донавчання	0.53929	0.69312
GPT2-Large з донавчанням	3.123216	0.068859
GPT2 без донавчання	0.48294	0.739292
GPT2 з донавчанням	3.120361	0.074097

Таблиця 2

Порівняння семантичної подібності відповідей

Модель	Середнє значення BERTScore	Середнє відхилення BERTScore
GPT2-Large питання vs відповідь	0.1896	0.1137
GPT2-Large питання vs інші питання	0.1432	0.0912
GPT2 питання vs відповідь	0.1753	0.1174
GPT2 питання vs інші питання	0.1487	0.0875

Приклад діалогу згенерованого GPT2 з донавчанням:

Person: What form of art resonates with you the most?

Socrates: Ah, my dear interlocutor, the art of expression has deep roots in Western philosophy, tradition, and human experience. To appreciate its nuances, I suggest you ponder what art embodies the highest ideals about what we seek out and strive for.

Person: I find that music resonates with me the most; its ability to evoke emotions and connect human experiences transcends words and speaks to the soul. It captures the essence of life's joys and sorrows in a way that feels both universal and deeply personal.

Socrates: Ah, indeed! Music is a mirror of the human experience. Tell me, what art has shaped your understanding of this realm? What music instills a sense of unity, unity in moments. What can you imagine to be the essence of music in the presence of the divine?

Дискусія і висновки

У дослідженні представлено ресурсоефективний підхід до навчання мовних моделей для генерації діалогів у стилі певного персонажа за допомогою параметроефективного налаштування. Запропонований метод використовує алгоритм Low-Rank Adaptation (LoRA) для налаштування моделей GPT-2 Large та GPT-2 за мінімальних обчислювальних витрат. Метрику оцінено з використанням класифікатора на основі DistilBERT для перевірки, що модель постійно генерує результати, узгоджені з бажаним стилем.

Представлена методологія може бути застосована в різних сферах, зокрема:

- чат-боти з особистістю персонажа;
- створення діалогових систем, що імітують історичних осіб, вигаданих персонажів або відомі особистості;
- генерація наративів, узгоджених з характером персонажів, для художніх творів;
- створення стилістично різноманітних наборів даних для навчання або оцінки моделей.

Напрями для подальших досліджень:

- розширення методу для підтримки кількох персонажів або стилів одночасно;
- використання навчання з підкріпленням для донавчання моделі;
- інтеграція зворотного зв'язку від людей або альтернативних механізмів оцінки для покращення відповідності;
- дослідження масштабованості підходу з використанням менш потужних попередньо навчених моделей, таких як дистильовані або квантизовані.

Отже, представлена робота демонструє, що параметроефективне налаштування є практичним рішенням для стилістичної генерації тексту, що сприяє ресурсозберігаючому впровадженню моделей обробки природної мови з особистісною орієнтацією. Отримані результати підкреслюють потенціал цього підходу для розвитку персоналізованого ШІ та генерації креативного контенту.

Внесок авторів: Олександр Марченко – концептуалізація; методика; Руслан Правосуд – програмне забезпечення, збір і перевірка емпіричних даних, емпіричне дослідження, аналіз джерел, підготовка огляду літератури та теоретичних основ дослідження.

Джерела фінансування. Публікацію частково профінансовано Київським національний університетом імені Тараса Шевченка.

Список використаних джерел

Hu, Edward J., Shen, Yelong, Wallis, Phillip, Allen-Zhu, Zeyuan, Yuanzhi, Li, Wang, Shean, Lu, Wang, & Chen, Weizhu. (2021). *Lora: Low-rank adaptation of large language models.* arXiv. https://doi.org/10.48550/arXiv.2106.09685

Hypersniper. *Philosophy dialog* [Dataset]. Hugging Face. https://huggingface.co/datasets/Hypersniper/philosophy_dialogue

Keskar, Nitish Shirish, McCann, Bryan, Varshney, Lav R., Caiming, Xiong, & Socher, Richard. (2019). *Ctrl: A conditional transformer language model for controllable generation.* arXiv. https://doi.org/10.48550/arXiv.1909.05858

Pfeiffer, Jonas, Kamath, Aishwarya, Rücklé, Andreas, Kyunghyun, Cho, & Gurevych, Iryna. (2020). *Adapterfusion: Non-destructive task composition for transfer learning.* arXiv. https://doi.org/10.48550/arXiv.2005.00247

Radford, Alec, Wu, Jeffrey, Child, Rewon, Luan, David, Amodei, Dario, & Sutskever, Ilya. (2019). *Language models are unsupervised multitask learners.* OpenAI blog. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Sanh, V. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.* arXiv. https://doi.org/10.48550/arXiv.1910.01108

Schick, Timo, & Schütze, Hinrich. (2020). *It's not just size that matters: Small language models are also few-shot learners.* arXiv. https://doi.org/10.48550/arXiv.2009.07118
Zhang, Tianyi, Kishore, Varsha, Wu, Felix, Weinberger, Kilian Q., Artzi Yoav. (2019). Bertscore: Evaluating text generation with bert. arXiv. https://doi.org/10.48550/arXiv.1904.09675

References

Hu, Edward J., Shen, Yelong, Wallis, Phillip, Allen-Zhu, Zeyuan, Yuanzhi, Li, Wang, Shean, Lu, Wang, & Chen, Weizhu. (2021). *Lora: Low-rank adaptation of large language models.* arXiv. https://doi.org/10.48550/arXiv.2106.09685
Hypersniper. *Philosophy dialog* [Dataset]. Hugging Face. https://huggingface.co/datasets/Hypersniper/philosophy_dialogue
Keskar, Nitish Shirish, McCann, Bryan, Varshney, Lav R., Caiming, Xiong, & Socher, Richard. (2019). *Ctrl: A conditional transformer language model for controllable generation.* arXiv. https://doi.org/10.48550/arXiv.1909.05858
Pfeiffer, Jonas, Kamath, Aishwarya, Rücklé, Andreas, Kyunghyun, Cho, & Gurevych, Iryna. (2020). *Adapterfusion: Non-destructive task composition for transfer learning.* arXiv. https://doi.org/10.48550/arXiv.2005.00247
Radford, Alec, Wu, Jeffrey, Child, Rewon, Luan, David, Amodei, Dario, & Sutskever, Ilya. (2019). *Language models are unsupervised multitask learners.* OpenAI blog. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
Sanh, V. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.* arXiv. https://doi.org/10.48550/arXiv.1910.01108
Schick, Timo, & Schütze, Hinrich. (2020). *It's not just size that matters: Small language models are also few-shot learners.* arXiv. https://doi.org/10.48550/arXiv.2009.07118
Zhang, Tianyi, Kishore, Varsha, Wu, Felix, Weinberger, Kilian Q., Artzi Yoav. (2019). Bertscore: Evaluating text generation with bert. arXiv. https://doi.org/10.48550/arXiv.1904.09675

Отримано редакцією журналу / Received: 25.05.25
Прорецензовано / Revised: 17.09.25
Схвалено до друку / Accepted: 10.10.25

Ruslan PRAVOSUD, PhD Student
ORCID ID: 0009-0008-9450-5515
e-mail: pronod9999@gmail.com
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

Oleksandr MARCHENKO, DSc (Phys. & Math.), Prof.
ORCID ID: 0000-0002-5408-5279
e-mail: omarchenko@univ.kiev.ua
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

DIALOG STYLE TRANSFER WITH PARAMETER-EFFICIENT FINE-TUNING

This work explores the application of modern fine-tuning methods, characterized by high parameter efficiency, to the task of generating dialogue responses in the style of a specific fictional or real-life character. The main goal of the study is to develop an approach that allows existing language models to be adapted to the specific communication or speech style of a particular personality using a minimal number of new parameters. This avoids the need for full retraining of large models and reduces computational costs.

One of the key objectives is to enable the implementation of a compact yet functional language model that can be run on a standalone computer without requiring servers or powerful graphics processors. Such an approach opens new possibilities for personalized or entertainment applications, such as creating chatbots capable of interacting with users in the persona of a character from a book, movie, or even a historical figure.

To implement this approach, we propose the use of the LoRA (Low-Rank Adaptation) method, which allows efficient fine-tuning of existing transformer architectures – in particular, the GPT-2 model – without retraining all of its layers. Using LoRA makes it possible to add stylistic features to the model's speech without compromising its core ability to generate logical and coherent texts.

To evaluate how well the model has learned to imitate the target speech style, we additionally train a BERT model as a classifier to distinguish between texts written in the desired style and ordinary, neutral responses. This provides an objective quality metric for generation and allows us to assess the effectiveness of the fine-tuning.

The dataset used for training was generated using the GPT-4 mini model, which enables the rapid creation of a large number of examples in the desired style.

K e y w o r d s : ***natural language processing, parameter-efficient fine-tuning, low-rank adaptation, text generation.***

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.