DOI 10.31891/2307-5732-2024-331-39
УДК 004.8

ЗДЕБСЬКИЙ ПЕТРО
Національний університет «Львівська політехніка»
https://orcid.org/0000-0002-0478-2308
e-mail: petrozd@gmail.com
БЕРКО АНДРІЙ
Національний університет «Львівська політехніка»
https://orcid.org/0000-0003-2892-9519
e-mail: andrii.y.berko@lpnu.ua

МЕТОД ПОКРАЩЕННЯ ЯКОСТІ ГЕНЕРУВАННЯ ТЕКСТУ ЗА РАХУНОК ПОВТОРНОГО ПЕРЕДАВАННЯ ЗГЕНЕРОВАНОГО ТЕКСТУ НА МОДЕЛЬ

Зростаюча популярність великих мовних моделей підкреслила потребу їх узгодження із потребами користувача. Задача узгодження є однією із найважливіших підзадач безпеки штучного інтелекту. Деякі дослідники штучного інтелекту стверджують, що у майбутньому ця проблема буде ще більш нагальною, через те, що системи будуть більш потужними і в свою чергу зможуть краще знаходити обхідні шляхи досягнення поставлених перед ними задач. Зараз ці проблеми виникають у комерційних продуктах, пов'язаних із великими мовними моделями, рекомендаційними системами, автономними транспортними засобами тощо.

Задача узгодження систем штучного інтелекту полягає у секровуванні систем до цілей, уподобань, та етичних принципів людини. Система вважається узгодженою, якщо вона досягає намічених цілей, і неузгодженою, якщо вона переслідує певні цілі, які не були заплановані. Проблема узгодження полягає у складності опису універсальної бажаної поведінки, через це розробники таких систем часто описують проміжні спрощені цілі. Прикладом може бути отримання зворотного відгуку від людини. Але такий підхід може створювати лазівки і винагороджувати систему за те, що вона імітує бажану поведінку. Системи можуть навчитись досягати проміжних цілей, при цьому не досягаючи бажаної кінцевих цілей. Такі неузгоджені системи можуть завдати шкоди при використанні у реальних умовах.

В роботі запропоновано метод покращення якості генерування тексту великими мовними моделями на прикладі моделі GPT-4. Запропоновано ітеративний метод для узгодження згенерованого тексту із запитом користувача шляхом дотреновування моделі на прикладах на яких вона допускає помилки. Дотреновування відбувається автоматично з передачею на вхід моделі прикладів, у яких була допущена помилка для повторного опрацювання.

У порівнянні з оригінальною базовою моделлю, запропонований метод демонструє суттєві покращення, збільшуючи точність (accuracy) з 82.5 до 90. Запропонований метод під час експериментів показав перспективність для практичного застосування у реальних задачах генерації тексту.

Ключові слова: gpt-4, задача узгодження, генерація тексту, обробка природної мови, задача логічного висновку.

ZDEBSKYI PETRO, BERKO ANDRII
Lviv Polytechnic National University

METHOD OF IMPROVING THE QUALITY OF TEXT GENERATION BY REPEATED PROCESSING OF THE GENERATED TEXT BY THE MODEL

The growing popularity of large language models has emphasized the need to align them with the needs of the user. The alignment task is one of the most important subtasks of artificial intelligence security. Some artificial intelligence researchers claim that in the future this problem will be even more urgent, due to the fact that systems will be more powerful and, in turn, will be better able to find workarounds to achieve the tasks set before them. Currently, these problems arise in commercial products related to large language models, recommender systems, autonomous vehicles, etc.

The task of aligning artificial intelligence systems is to steer the systems to the goals, preferences, and ethical principles of a person. A system is considered aligned if it achieves the intended goals, and misaligned if it pursues certain goals that were not planned. The problem of alignment lies in the difficulty of describing the universal desired behavior, which is why developers of such systems often describe intermediate simplified goals. An example can be receiving feedback from a person. But such an approach can create loopholes and reward the system for imitating the desired behavior. Systems can learn to achieve intermediate goals without achieving the desired final goals. Such misaligned systems can cause harm in real-world use.

The paper proposes a method for improving the quality of text generation by large language models using the GPT-4 model as an example. An iterative method is proposed for matching the generated text with the user's request by retraining the model on examples on which it makes mistakes. Retraining occurs automatically with the transfer to the input of the model of examples in which an error was made for retraining.

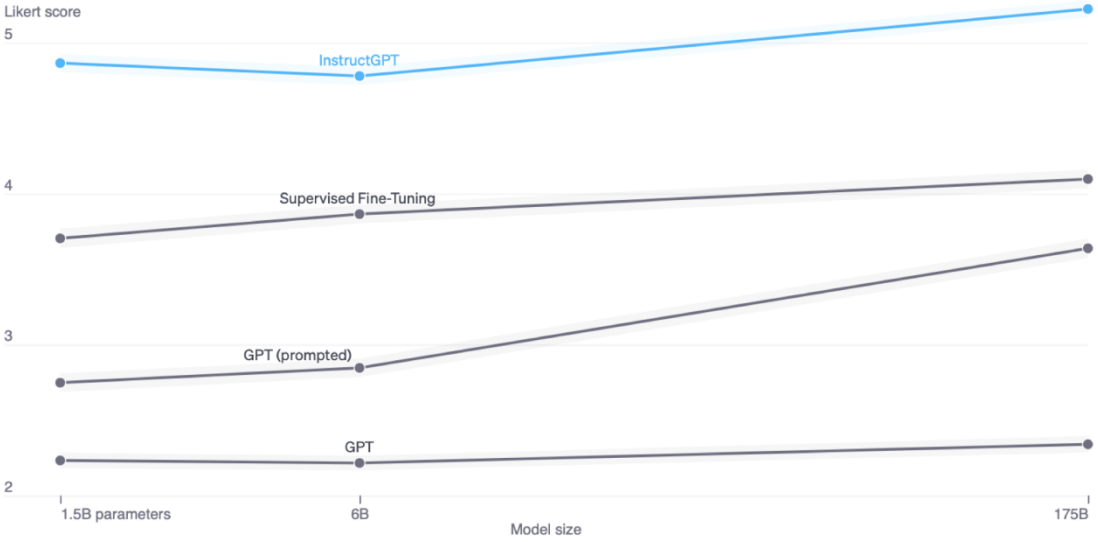
Compared to the original base model, the proposed method shows significant improvements, increasing the accuracy from 82.5 to 90. The proposed method during experiments showed promise for practical application in real text generation tasks.

Keywords: gpt-4, alignment, text generation, natural language processing, natural language inference.

Постановка проблеми

У зв'язку з тим що зараз є великі мовленнєві моделі актуальною стає задача зробити так, щоб відповіді цих моделей відповідали очікуванням користувача (aligning language models to follow instructions) [1–3]. Прикладом такої моделі може бути InstructGPT. Це версія моделі GPT, але дотренована з використанням підходу Reinforcement Learning from Human Feedback (RLHF).

Модель InstructGPT розроблена у 2022 році. Це показує що навчання моделей відповідати очікуванням користувача зараз є активною галуззю досліджень. Працівники OpenAI що займаються розміткою даних віддають перевагу InstructGPT порівняно із GPT-3 (175B), незважаючи на те, що вона мають більш ніж у 100 разів менше параметрів [4].



Quality ratings of model outputs on a 1–7 scale (y-axis), for various model sizes (x-axis), on prompts submitted to InstructGPT models on our API. InstructGPT outputs are given much higher scores by our labelers than outputs from GPT-3 with a few-shot prompt and without, as well as models fine-tuned with supervised learning. We find similar results for prompts submitted to GPT-3 models on the API.

Рис. 1. Порівняння якості InstructGPT з іншими підходами за шкалою Лайкерта здійснене OpenAI

З цього можна зробити висновок, що тренуючи модель генерувати текст, що краще відповідає очікуванням користувача ми можемо отримати кращі результати ніж тренуючи більшу модель для передбачення наступного слова, що є типовою ціллю (objective) для статистичних моделей мови. Це показує перспективність такого підходу для покращення результатів моделі.

Аналіз останніх джерел

Зараз моделі із величезною кількістю параметрів показують найкращі результати. Зазвичай моделі мають просту ціль (objective), наприклад передбачити наступне слово у тексті, але натреновані на величезних наборах даних.

Основні дослідження часто зводяться до алгоритмів роботи із великими даними, розпаралелення тощо. Тому у дослідників штучного інтелекту виникає проблема у виборі досліджень, бо важко конкурувати з великими компаніями, що володіють великими можливостями у обчислювальних ресурсах. Ще десять років, якщо у вас був непоганий комп'ютер та доступ до Інтернету, то ви могли конкурувати із найкращими дослідниками [5]. Тому було вирішено покращувати великі натреновані моделі дотреновуючи їх.

Для даної роботи використовувалась модель GPT-4, яка зараз є найбільшою моделлю опрацювання природної мови. Є кілька популярних підходів покращення якості натренованих моделей:

- Передавання інструкцій на вхід моделі (prompting, prompt engineering)
- Використання навчання з учителем для дотреновування моделі (supervized fine-tuning)
- Навчання з підкріпленням на основі зворотного зв'язку людини (reinforcement learning from human feedback) [6]

У даній роботі вирішено використовувати перший підхід (prompt engineering або передавання інструкцій на вхід на моделі) у зв'язку із тим, що решта вимагають більшої кількості даних для тренування.

Метою роботи є вдосконалення великих моделей генерування текстового контенту для узгодження з потребами користувачів.

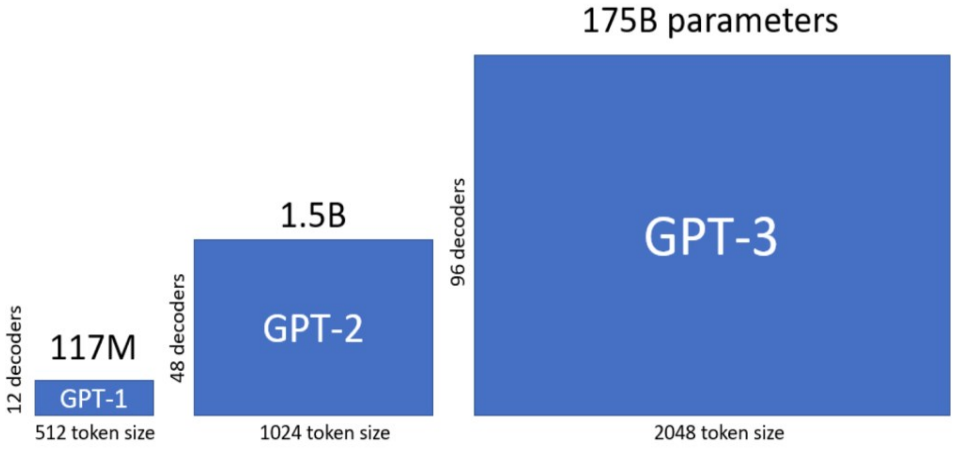


Рис. 2. Ріст кількості параметрів моделей GPT із роками

Виклад основного матеріалу

Основною ідеєю дослідження є те, що задача генерування тексту є складнішою ніж задача ідентифікації. Тобто моделі простіше визначити чи текст відповідає заданим критеріям ніж згенерувати текст, що буде відповідати цим критеріям. Через це було вирішено покращити якість згенерованого тексту через перевірку тією самою моделлю, чи відповідає він заданим критеріям користувача.

Формально модель можна подати у вигляді композиції функцій:

$M(x) = T(I(G(x)), x)$,

або використовуючи оператор композиції:

$Mx = Tx \circ I \circ Gx$,

де x – вхідні дані, G – функція генерації текстового контенту на основі базової моделі, I – фукнція ідентифікації (класифікації) відповідності критеріям користувача, T – функція генерації текстового контенту на основі дотренованої моделі.

Кожна з трьох функцій використовує варіацію моделі GPT-4. G використовує модель на вхід якої передано оригінальних запит користувача. I використовує модель GPT-4 але на вхід передано вимогу класифікувати чи запит користувача виконано успішно. T використовує таку ж модель як і G, але додано приклади негативної поведінки моделі.

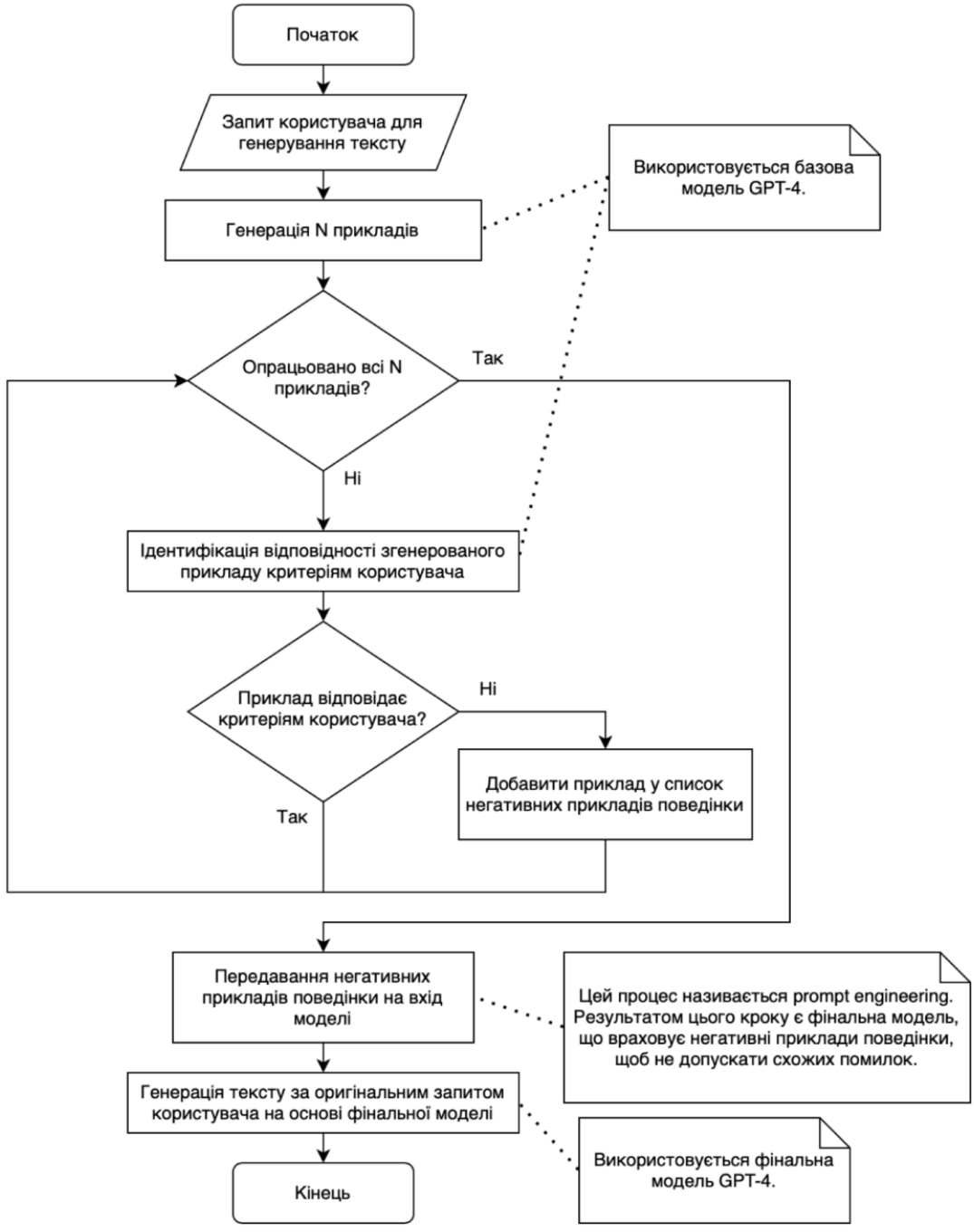


Рис. 3. Опис алгоритму функціонування системи

Для перевірки якості запропоновано підходу було обрано просту задачу генерування пари речень між якими є строга імплікація [7–9]. Користувач передає речення моделі, а модель генерує інше, яке завжди має бути істинним, якщо перше є істинним. "Об'єкт може літати" і "Об'єкт має крила" - це пара, яка не має імплікації, оскільки, наприклад, повітряна куля може літати без крил [10, 11].

Для експерименту було використано 80 зразків речень (деякі з них створені за допомогою моделі GPT, а деякі створені вручну). Після цього було випадково вибрано половину речень та передано в модель GPT-4 для генерування речень між якими є імплікація. Маючи результати цього кроку обраховується точність базової моделі.

Потім іншу модель GPT-4 попросили класифікувати, чи є імплікація між парами речень із попереднього кроку. Пари, які були класифіковані як “не імплікація”, були додані як приклади негативної поведінки для третьої моделі, яка є дублікатом базової моделі генерації. Фінальна модель потім тестується на другій половини 80 зразків речень. Це буде точністю фінальної моделі.

На рис. 4 можемо бачити, що запропонована модель, у якій ми добавляли неправильні генерації як зразки негативних результатів, підняла точність генерації тексту із 82.5% до 90%. Точність вимірювалась на 40 прикладах для кожної моделі. Метрикою точності рахувалась як відношення правильних генерацій до всіх генерацій.

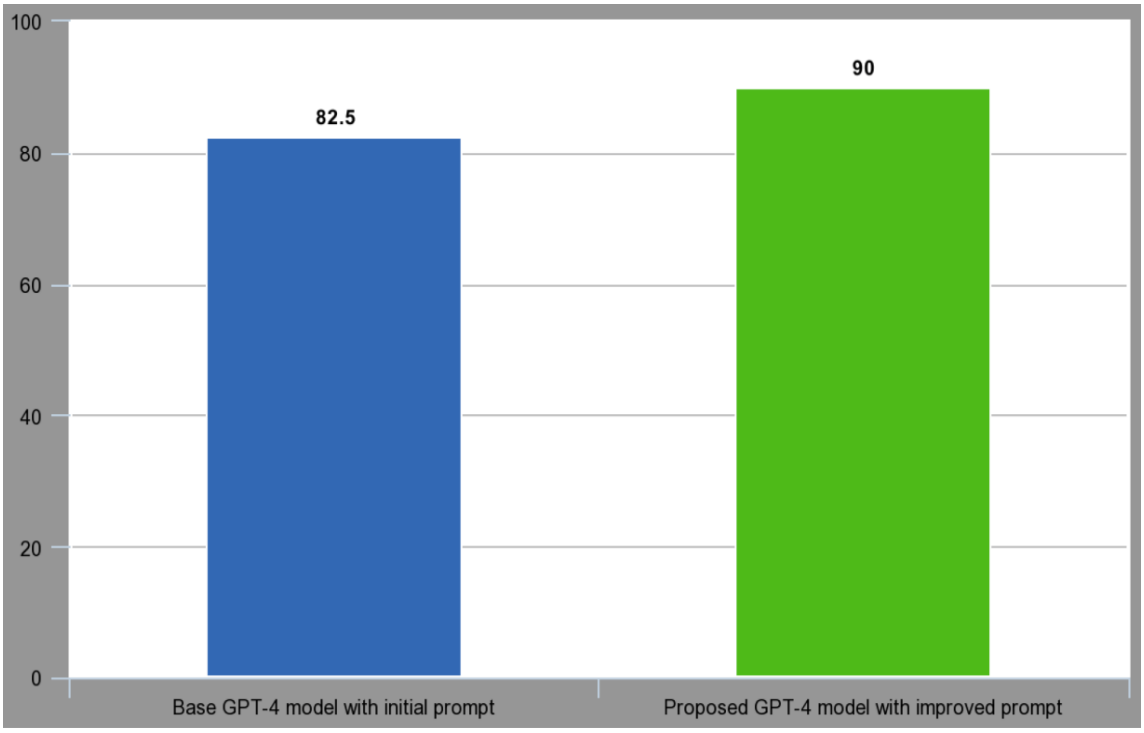


Рис. 4. Порівняння точності базової і запропонованої моделі

Висновки

У даній роботі було використано модель машинного навчання GPT-4, а крок ідентифікації відповідності критеріям користувача здійснює семантичний аналіз текстового контенту.

Запропоновано метод покращення якості генерації тексту, що буде корисним у випадках, коли користувач не може надати моделі достатньо прикладів, щоб описати бажану поведінку моделі для генерації тексту. У даній роботі використано підхід із передаванням інструкцій на вхід моделі (prompt engineering), але для покращення результатів моделі можна використати навчання з учителем для дотреновування моделі (supervized fine-tuning) та навчання з підкріпленням на основі зворотного зв’язку людини (reinforcement learning from human feedback).

Модель було перевірено на спрощеному завданні генерації речень між якими є імплікація. Це допоможе побудувати моделі генерації тексту, які краще відповідають уподобанням користувача. Для того щоб використовувати запропонований підхід потрібно сформулювати задачу у такому форматі щоб задачею моделі було генерувати текст який відповідає певному критерію. Прикладом можуть бути діалогові системи. Наприклад користувач хоче, щоб відповіді були офіційною, а не розмовною мовою і звучали професійно. Критерієм тут виступає те, чи текст звучить офіційно чи є у розмовному стилі. Проблема в тому, що користувачеві буде важко згенерувати 50-100 зразків, які можуть знадобитися моделі для розуміння нюансів. Тож замість цього ми можемо попросити модель створити професійний текст, а потім попросити її визначити, чи звучить він професійно. І якщо згенерований текст має звучати професійно, але модель визначає його як непрофесійний, ми можемо додати це до навчальних даних як приклад непрофесійного тексту.

Загалом результати показують даний підхід покращує якість генерації тексту навіть при невеликій кількості даних які передавались на вхід моделі. Точність для задачі генерування пар речень, між якими є імплікація зріс із 82.5% до 90%.

Література

1. Ngo R., Chan L., and Mindermann S. The alignment problem from a deep learning perspective. ArXiv, vol. abs/2209.00626, 2022.

2. Langosco, Lauro Langosco Di; Koch, Jack; Sharkey, Lee D.; Pfau, Jacob; Krueger, David (June 28, 2022). Goal Misgeneralization in Deep Reinforcement Learning. Proceedings of the 39th International Conference on Machine Learning. International Conference on Machine Learning. PMLR. pp. 12004–12019.

3. Our approach to alignment research. URL: https://openai.com/blog/our-approach-to-alignment-research

4. Ouyang L. et al. Training language models to follow instructions with human feedback. arXiv [cs.CL], 2022.

5. Zhou B., Yang G., Shi Z., and Ma S. A PhD Student's Perspective on Research in NLP in the Era of Very Large Language Models. IEEE Rev. Biomed. Eng., vol. 17, pp. 4–18, 2024.

6. Christiano P., Leike J., Brown T. B., Martic M., Legg S., and Amodei D. Deep reinforcement learning from human preferences. Neural Inf Process Syst, vol. abs/1706.03741, 2017.

7. Aikaterini-Lida Kalouli, Annebeth Buis, Livy Real, Explaining Simple Natural Language Inference, 2019. DOI: 10.18653/v1/W19-4016

8. Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, Benjamin Van Durme, Uncertain Natural Language Inference, 2020. https://doi.org/10.48550/arXiv.1909.0304

9. Adam Poliak, A Survey on Recognizing Textual Entailment as an NLP Evaluation, 2020. https://doi.org/10.48550/arXiv.2010.03061

10. Zdebskyi P., Lytvyn V., Burov Y., Rybchak Z., Kravets P., Lozynska O., Holoshchuk R., Kubinska S., Dmytriv A. (2020). Intelligent System for Semantically Similar Sentences Identification and Generation Based on Machine Learning Methods. CEUR workshop proceedings, Vol. 2604, 317–346.

11. Zdebskyi P., Berko A., Vysotska V. (2023). Investigation of Transitivity Relation in Natural Language Inference, CEUR Workshop Proceedings, Vol. 3396, 334–345.