УДК 004.89-044.332:378 **DOI: 10.33042/2522-1809-2024-6-187-12-17**

В.О. Дідок, М.П. Пан

Харківський національний університет міського господарства імені О.М. Бекетова, Україна

ОЦІНКА ЕФЕКТИВНОСТІ МЕТОДІВ АДАПТАЦІЇ МОВНОЇ МОДЕЛІ GPT-4O В ОСВІТНЬОМУ СЕРЕДОВИЩІ

Проблема адаптації великих мовних моделей, таких як GPT-4O, для освітнього середовища є актуальною. У дослідженні порівнюються п'ять методів адаптації GPT-4O: стандартна модель, Fine-Tuning, Prompt Engineering, Retrieval-Augmented Generation та агентні системи. Ефективність оцінюється за точністю, загальною точністю, F1-Score, кількістю токенів, час відповіді, частотою помилкових відповідей та фінальним балом.

Ключові слова: *GPT-4O, Fine-Tuning, Prompt Engineering, Retrieval-Augmented Generation, агенти, адаптація, освіта.*

Постановка проблеми

Проблема ефективної адаптації великих мовних моделей для різних прикладних завдань, зокрема в освітньому середовищі, стає все більш актуальною. Модель GPT-4O[1] має значний потенціал для покращення якості освітніх процесів завдяки своїй здатності генерувати відповіді на складні запити, що вимагають високого рівня розуміння текстової інформації. Однак ефективність її використання залежить від обраного методу адаптації.

На сьогодні існує кілька основних підходів до адаптації мовних моделей, включаючи Fine-Tuning (донавчання на специфічних даних), Prompt Engineering (інженерія підказок), Retrieval-Augmented Generation (RAG) та агентні системи. Кожен з цих методів має свої переваги та недоліки. Для того, щоб оцінити їхню ефективність, модель GPT-4O також буде порівнюватися зі стандартною моделлю без адаптації, яка слугуватиме контрольним прикладом.

Об'єктом дослідження є процес порівняння мовної моделі GPT-4O з різними методами адаптації для освітніх завдань. Ефективність кожного підходу оцінюватиметься за ключовими показниками, такими як точність (Precision), загальна точність (Accuracy), F1-Score, середня кількість токенів на відповідь (Average Tokens per Response) та частота "галюцинацій" (HallucinationRate), час відповіді. Також буде розглянуто фінальний бал (Final Score), який відображатиме комплексну оцінку якості відповідей.

Актуальність цього дослідження полягає у необхідності вибору найбільш оптимального методу адаптації великих мовних моделей для вирішення конкретних освітніх задач. Сучасні дослідження та практичні впровадження показують, що різні методи адаптації можуть мати різні рівні ефективності в залежності від контексту застосування

Аналіз останніх досліджень та публікацій

Адаптація великих мовних моделей, таких як GPT-4O, для освітніх та інших специфічних завдань є однією з ключових тем у сучасних дослідженнях штучного інтелекту. Основними підходами до адаптації є Fine-Tuning, Prompt Engineering, RAG та агентні системи. Кожен із цих методів має свої унікальні переваги та недоліки, що дозволяють використовувати їх у різних контекстах [1], [2], [4].

Fine-Tuning є ефективним методом, що дозволяє донавчати мовну модель на вузькоспеціалізованих даних, тим самим підвищуючи її продуктивність для специфічних завдань. Як показують дослідження, Fine-Tuning може значно покращити точність моделі, але вимагає значних обчислювальних ресурсів і часу [5]. Згідно з дослідженнями OpenAI, Fine-Tuning є оптимальним для застосування в завданнях, що потребують високої точності, таких як медичні або юридичні завдання [5].

Prompt Engineering, на відміну від Fine-Tuning, дозволяє оптимізувати результати моделі без необхідності донавчання. Цей підхід зосереджений на правильному формулюванні запитів, що дозволяє досягти значних результатів при мінімальних витратах ресурсів [2]. Однак, як показано в [3], цей підхід менш ефективний у складних сценаріях, що вимагають детального контексту або великих обсягів даних.

RAG — це підхід, який використовує зовнішні джерела даних для покращення релевантності відповідей моделей. RAG дозволяє мовним моделям, таким як GPT-4O, отримувати доступ до

актуальної інформації, що робить його корисним для завдань, де важлива фактична точність. Наприклад, у дослідженні [4] зазначається, що RAG знижує частоту так званих "галюцинацій" моделі та підвищує точність відповідей у навчальних середовищах, де зміст інформації може змінюватися.

Агентні системи, такі як P-Tuning, використовуються для адаптації моделей до контексту та складних багатоетапних завдань [6]. Цей підхід забезпечує високу продуктивність у сценаріях, що потребують динамічного планування або багаторівневих відповідей, що робить агентні системи ефективними для освітніх середовищ, де задачі можуть змінюватися залежно від прогресу користувача.

Дослідження підтверджують, що кожен із методів адаптації є оптимальним у своїх контекстах застосування [1], [2], [4]. Fine-Tuning є ідеальним для вузькоспеціалізованих завдань, де точність є головним критерієм успіху, тоді як Prompt Engineering та RAG підходять для швидкого та економічного налаштування моделей для загальних завдань [3], [4]. Агентні системи забезпечують динамічну та багатокрокову обробку завдань, що робить їх ефективними для використання в складних освітніх середовищах [6].

Мета статті

Метою даної роботи є порівняння ефективності стандартної моделі GPT-4O та різних методів її адаптації для освітнього середовища з метою виявлення найбільш продуктивного підходу.

Виклад основного матеріалу

Об'єкти та методи дослідження. Нехай дана велика мовна модель GPT-4O, яку необхідно адаптувати для освітнього процесу, використовуючи різні методи, такі як Fine-Tuning, Prompt Engineering, RAG та агенти. Основна задача полягає у порівнянні ефективності цих методів з базовою версією моделі за ключовими метриками: точністю, часом відповіді, частотою "галюцинацій", середньою кількістю токенів на відповідь, загальною точністю та фінальним балом.

Формально задача може бути представлена наступ-ним чином:

Нехай — стандартна версія моделі GPT-4O, а — адаптована модель за допомогою одного з методів. Потрібно знайти оптимальний метод (Fine-Tuning, Prompt Engineering, RAG, агенти), щоб виконувалася умова:

$$M(GPT_{adapt}, GPT_{std}, M_i) \longrightarrow opt,$$

де opt – оптимальний результат за наступними мет-риками:

- Точність: частка правильних відповідей серед тих, які модель визнала правильними.
- Загальна точність: частка правильних відповідей серед усіх запитів.
- F1-Score: гармонійне середнє між точністю і повнотою, що дозволяє врахувати баланс між помилковими позитивними і помилковими негативними відповідями.
- Середній час відповіді: середній час, необхідний для генерації відповіді в секундах.
- Середня кількість токенів: кількість токенів, використаних моделлю на одну відповідь.
- Частота "галюцинацій": частка відповідей, що містять нерелевантну або помилкову інформацію.
- Фінальний бал: інтегральна оцінка якості відповіді за кількома критеріями, такими як фактична правильність, релевантність, глибина відповіді, логічна зв'язність і ясність.

Таким чином, мета дослідження полягає в пошуку методу, який мінімізує частоту "галюцинацій" і час відповіді, одночасно максимізуючи точність, загальну точність, F1-Score, та фінальний бал.

Метод дослідження

Для оцінки ефективності різних методів адаптації моделі GPT-4O були використані п'ять підходів: стандартна версія моделі без додаткового навчання, Fine-Tuning, Prompt Engineering, RAG та агентні системи. Основна мета полягала в тому, щоб визначити, який із цих методів є найбільш ефективним для використання в освітніх середовищах, де точність і релевантність відповідей є критичними факторами успіху.

Для кожного з методів було встановлено низку ключових метрик для оцінки продуктивності моделі:

Fine-Tuning передбачає донавчання моделі на специфічних освітніх даних. Для цього використовувалася спеціальна вибірка з освітніх текстів, підручників, контрольних завдань, що дозволило оптимізувати модель для вирішення конкретних освітніх завдань. Параметри навчання включали 10 епох і коефіцієнт навчання (learning rate) 0.001, щоб уникнути перенавчання моделі. Fine-Tuning виявився найбільш ефективним у завданнях, де потрібна висока точність відповідей і глибоке розуміння контексту.

Prompt Engineering використовував підхід оптимізації запитів до моделі без необхідності до-навчання. Це дозволяло швидко отримувати релевантні відповіді на освітні запитання, але ефективність цього методу була обмежена складними запитами, де потрібно більше контексту.

Основний фокус тут був на правильному формулюванні запитів, що оптимізувало результати.

RAG дозволяв інтегрувати модель з зовнішніми джерелами даних, що надавало моделі можливість отримувати більш актуальну та точну інформацію для відповіді. Це особливо важливо для завдань, де потрібні точні фактичні відповіді на основі поточних або змінних даних.

Агентні системи використовувалися для динамічної обробки складних запитів і багатокрокових задач, таких як тестування або багатоетапне вирішення завдань у навчальному процесі. Цей підхід дозволив моделі GPT-4O ефективніше працювати з довготривалими завданнями, що потребували послідовних відповідей і аналізу.

Для кожного підходу ефективність оцінювалась за наступними метриками:

Точність (Precision) оцінювала частку правильних відповідей серед усіх позитивних передбачень (1):

$$P = \frac{T_p}{T_p + F_p} \quad (1),$$

де T_p — кількість правильних позитивних відповідей, а F_p — кількість хибних позитивних відповідей.

Загальна точність (Accuracy) оцінювала частку правильних відповідей серед усіх наданих відповідей, як позитивних, так і негативних (20:

$$A = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \quad (2),$$

де T_n — кількість правильних негативних відповідей, F_n — кількість хибних негативних відповідей.

F1-Score є гармонічним середнім між Precision і Recall, і ця метрика дозволяє врахувати як точність, так і повноту (3):

$$F1 = 2\frac{P * R}{P + R} \quad (3),$$

де P — Точність (Precision), а R — це повнота (Recall), яка вимірює, яку частку правильних відповідей модель змогла знайти серед усіх можливих позитивних результатів.

Середня кількість токенів на відповідь (Average Tokens per Response) використовувалася для оцінки того, наскільки модель здатна надавати лаконічні відповіді (4):

$$T_{avg} = \frac{T_{total}}{N_{total}} \quad (4),$$

де T_{total} — загальна кількість токенів, а N_{total} — кількість відповідей.

Ця метрика допомагала оцінити продуктивність і оптимізацію довжини відповідей.

Частота "галюцинацій" (Hallucination Rate) вимірювала частку відповідей, що містять нерелевантну або неправильну інформацію (5):

$$H = \frac{\mathrm{N_h}}{N_{total}} * 100\% \quad (5),$$

де $\mathrm{N_h}$ — кількість "галюцинацій".

Ця метрика була критичною для оцінки надійності моделі в освітніх завданнях, де важливо уникати дезінформації.

Фінальний бал (Final Score) розраховувався як середнє значення оцінок за кількома критеріями якості відповідей, такими як фактична правильність, контекстуальна релевантність, глибина відповіді, логічність та ясність (6):

$$S = \frac{\mathrm{C_f} + \mathrm{C_c} + \mathrm{C_d} + \mathrm{C_l} + \mathrm{C_q}}{5} \quad (6),$$

де:

$\mathrm{C_f}$ — фактична правильність,

$\mathrm{C_c}$ — контекстуальна релевантність,

$\mathrm{C_d}$ — глибина відповіді,

$\mathrm{C_l}$ — логічність,

$\mathrm{C_q}$ — ясність формулювань.

Таким чином, кожен із методів адаптації був оцінений за допомогою цих метрик, що дозволило отримати комплексну картину ефективності кожного підходу в освітньому середовищі.

Результати та обговорення

Експерименти були проведені з метою порівняння ефективності різних методів адаптації моделі GPT-4O для освітніх завдань. Було протестовано п'ять підходів: стандартна модель GPT-4O без донавчання, Fine-Tuning, Prompt Engineering, RAG та агентні системи.

Для Fine-Tuning модель було навчено на специфічних даних університетської платформи, включаючи питання студентів щодо використання освітніх інструментів і структур університету. Ці

дані забезпечили ефективні відповіді на освітні запити. Prompt Engineering передбачав додавання специфічних даних у запит, що збільшувало кількість споживаних токенів. RAG залучав зовнішні джерела для покращення відповідей, а агентні системи використовували динамічне планування для розв'язання складних завдань.

Модель GPT-4O було протестовано за кількома мет-риками: точність, загальна точність, F1-Score, середня кількість токенів на відповідь, частота "галюцинацій", фінальний бал та час відповіді. Для кожного методу було зафіксовано середні показники після кількох запусків.

Результати експериментів демонструють різницю між методами адаптації за всіма метриками. Середні значення для кожного методу на основі проведених тестів наведено в табл. 1.

Таблиця 1

Порівняльний аналіз ефективності методів адаптації моделі GPT-4O

Метод	**Точність (Precision)**	**Загальна точність (Accuracy)**	**F1-Score**	**Середня кількість токенів**	**Частота "галюцинацій" (%)**	**Фінальний бал (Final Score)**	**Час відповіді (сек)**
Стандартна модель	0.80	0.78	0.79	15	4	7.5	2.5
Fine-Tuning	0.92	0.91	0.91	18	2	9.1	1.8
Prompt Engineering	0.85	0.84	0.84	14	6	7.8	1.6
RAG	0.88	0.87	0.87	17	3	8.3	2.0
Агенти	0.87	0.86	0.85	16	5	8.0	2.2

Таблиця 1 демонструє, що Fine-Tuning забезпечує найвищу точність і найменшу частоту "галюцинацій". Prompt Engineering показав швидший час відповіді, але менш точні результати, оскільки додавання специфічних даних у запит збільшувало споживання токенів. RAG дозволив знизити частоту помилкових відповідей за рахунок використання зовнішніх джерел, однак мав довший час відповіді. Агентні системи виявилися корисними для складних багатокрокових завдань, але мали підвищену частоту "галюцинацій".

Результати експериментів чітко демонструють пере-ваги та недоліки різних методів адаптації моделі GPT-4O для освітнього середовища. Fine-Tuning виявився найефективнішим методом, показавши найвищу точність (92%) і найменшу частоту "галюцинацій" (2%), що підтверджує його придатність для вирішення специфічних освітніх завдань, де важлива точність відповіді та наявність адаптованих даних. Донавчання моделі на специфічних даних дозволило мінімізувати ресурси під час використання, оскільки модель вже мала всі необхідні знання для роботи з освітніми запитами.

Метод Prompt Engineering, хоча і забезпечує швидші відповіді (1.6 секунди), вимагає додаткових ресурсів через збільшену кількість споживаних токенів. Це робить його менш ефективним порівняно з Fine-Tuning, особливо в довготривалих процесах, коли запити можуть бути різноманітними. Крім того, частота "галюцинацій" у Prompt Engineering була значно вищою (6%), що знижує його корисність у критичних сценаріях, де важлива релевантність відповідей.

RAG показав хороші результати завдяки використанню зовнішніх джерел для генерації відповідей, що дозволило знизити частоту "галюцинацій" до 3%. Проте збільшений час відповіді (2 секунди) та ресурсоємність роблять цей метод менш оптимальним для завдань, де важлива швидкість обробки запитів. Цей підхід є корисним у випадках, коли модель повинна отримувати актуальні дані, однак він може бути менш ефективним у випадках, коли важлива оперативність відповіді.

Агентні системи забезпечили високу продуктивність для багаторівневих завдань, що вимагають динамічної адаптації та аналізу контексту. Проте частота "галюцинацій" у агентів була вищою (5%), а час відповіді довшим (2.2 секунди). Цей метод може бути ефективним для складних освітніх середовищ, де необхідно обробляти різноманітні й багатокрокові запити, однак у випадках, де важлива точність, його застосування може бути обмежене.

Fine-Tuning підтвердив свою ефективність не лише в точності, але й у зниженні витрат ресурсів під час тривалого використання, що робить його оптимальним для освітніх завдань. У свою чергу, Prompt Engineering та RAG можуть бути корисними у випадках, де потрібна швидка адаптація до нових даних або доступ до зовнішніх джерел інформації. Агентні системи забезпечують гнучкість у роботі з багаторівневими завданнями, але потребують оптимізації для зниження частоти помилкових відповідей.

Висновки

У даній роботі було проведено порівняльний аналіз п'яти різних методів адаптації моделі GPT-4O для освітнього середовища: Fine-Tuning, Prompt Engineering, RAG, агентні системи та базова модель без донавчання. Результати дослідження показали, що Fine-Tuning є найбільш ефективним методом, що забезпечує найвищу точність (92%) та найменшу частоту "галюцинацій" (2%). Цей підхід дозволив значно покращити точність відповідей завдяки навчанням на специфічних університетських даних. Prompt Engineering забезпечив швидкий час відповіді (1.6 секунди), але мав вищу частоту "галюцинацій" (6%) та був менш ефективним через збільшене споживання токенів під час кожного запиту. RAG виявився корисним для зниження частоти помилкових відповідей (3%) завдяки залученню зовнішніх джерел даних, але мав більший час відповіді (2 секунди). Агентні системи продемонстрували високу гнучкість у складних завданнях, однак їхня частота помилок (5%) була вищою за інші методи, а час відповіді довший (2.2 секунди).

Загалом, Fine-Tuning є оптимальним методом для завдань, де важливі точність і релевантність відповідей, особливо у сфері освіти. Prompt Engineering та RAG можуть бути використані у випадках, коли необхідна швидка адаптація до нових даних або доступ до зовнішніх джерел. Агентні системи найбільш підходять для багаторівневих завдань, які вимагають динамічної адаптації, проте їхня ефективність залежить від точності й швидкості відповіді.

Подальші дослідження можуть бути зосереджені на комбінованих підходах, які дозволять оптимізувати продуктивність моделі в освітніх середовищах, поєднуючи переваги різних методів адаптаці.

References

1. Investigating the Performance of Retrieval-Augmented Generation and Fine-Tuning. – [arXiv preprint] 2024. Available at: https://arxiv.org/abs/2403.09727

2. Prompt Engineering vs Fine-Tuning vs RAG. – [MyScale] 2024. Available at: https://myscale.com/blog/prompt-engineering-vs-finetuning-vs-rag/

3. Prompt Engineering or Fine Tuning: An Empirical Assessment of LLMs. – [arXiv preprint] 2023. Available at: https://arxiv.org/abs/2310.10508

4. Retrieval-Augmented Generation for Large Language Models: A Survey. – [arXiv preprint] 2023. Available at: https://arxiv.org/abs/2312.10997

5. Fine-Tuning GPT-4 Models. – [OpenAI] 2024. Available at: https://openai.com/index/gpt-4o-fine-tuning/

6. An Introduction to Large Language Models, Prompt Engineering, and P-Tuning. – [NVIDIA Developer Blog] 2023. Available at: https://developer.nvidia.com/blog/an-introduction-to-large-language-models-prompt-engineering-and-p-tuning

Рецензент: д-р техн. наук проф. В.Є. Плюгін, Харківський національний університет міського господарства імені О.М. Бекетова, Україна.

Автор: ДІДОК Володимир Олегович
здобувач вищої освіти 2-го курсу магістратури навчально-наукового інституту енергетичної, інформаційної та транспортної інфраструктури Харківський національний університет міського господарства імені О.М. Бекетова
E-mail – Volodymyr.Didok@kname.edu.ua
ID ORCID: https://orcid.org/0009-0001-0453-1808

Автор: ПАН Микола Павлович
кандидат технічних наук, доцент кафедри комп'ютерних наук та інформаційних технологій Харківський національний університет міського господарства імені О.М. Бекетова
E – mail – pan@kname.edu.ua
ID ORCID: https://orcid.org/0000-0001-7137-5174

EVALUATION OF ADAPTATION METHODS FOR THE GPT-4O LANGUAGE MODEL IN EDUCATIONAL ENVIRONMENTS

V. Didok, M. Pan
O. M. Beketov National University of Urban Economy in Kharkiv, Ukraine

The adaptation of large language models, such as GPT-4O, for educational settings is becoming increasingly relevant, given their potential to improve automated support in learning environments. This study focuses on evaluating five distinct adaptation methods: the standard model (without adaptation), Fine-Tuning, Prompt Engineering, Retrieval-Augmented Generation (RAG), and agent-based systems. These methods were compared to find the most effective for enhancing GPT-4O's performance in educational tasks. The analysis was conducted using key evaluation metrics such as precision, accuracy, F1-Score, average tokens per response, hallucination rate, and a final comprehensive score.

The experimental results indicate that Fine-Tuning, which involves additional training on domain-specific educational data, offers the most significant improvements in terms of accuracy and reduction of erroneous

responses (hallucinations). Fine-Tuning is especially effective in educational tasks requiring high accuracy and contextual understanding.

Retrieval-Augmented Generation showed promising results by leveraging external data to enhance accuracy and lower hallucinations, making it suitable for tasks needing up-to-date information. Prompt Engineering provided faster response times but had more inaccuracies due to reliance on optimal query formulation without retraining.

Agent-based systems excelled in handling complex tasks, though they showed a slight increase in hallucination rates due to their dynamic nature. The baseline performance of the standard GPT-4O model highlighted its limitations, like reduced accuracy and higher hallucination rates, especially in educational contexts.

These findings underscore that Fine-Tuning is the most effective adaptation method for educational tasks, offering substantial improvements in accuracy and reliability. Overall, this research highlights the necessity of selecting the appropriate adaptation method based on the specific requirements of educational tasks. The study contributes to the ongoing optimization of large language models for use in educational environments, ensuring that responses are both reliable and contextually relevant.

Key words*: GPT-4O, Fine-Tuning, Prompt Engineering, Retrieval-Augmented Generation, agents, adaptation, education.*