МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Національний технічний університет України «Київський політехнічний інститут ім. Ігоря Сікорського»

Навчально-науковий інститут прикладного системного аналізу

кафедра системного проектування

Звіт щотижневий

1-го тижня переддипломної практики зі

спеціальності 122 Комп'ютерні науки

Тема практики (індивідуального завдання): «Аналіз впливу методів активаційного інжинірингу на адаптацію особистісних характеристик великих мовних моделей»

Студентка 4-го курсу

групи ДА-11

Бєлікова Маргарита Євгенівна

Київ – 2025

Зміст

1. Вступ

У першому тижні роботи було зосереджено увагу на теоретичному осмисленні існуючих методів активаційного інжинірингу та аналізі архітектур великих мовних моделей (LLM) для управління особистісними рисами. В рамках цього етапу було проведено порівняльний аналіз двох основних архітектур: decoder-only (LLaMA 2, GPT-2, Mistral) та encoder-decoder (FLAN-T5). Також здійснено огляд досліджень, що стосуються використання активаційних напрямків для маніпуляцій з поведінкою мовних моделей, зокрема в контексті керування особистісними рисами. Робота була спрямована на визначення основних підходів, що використовуються в актуальних роботах, а також на виявлення прогалин в наявних знаннях, що стосуються перенесення активаційних напрямків між різними моделями.

2. Виконана робота

2.1 Порівняння архітектур decoder-only та encoder-decoder

Трансформери — це архітектури глибокого навчання, які були розроблені дослідниками Google і базуються на механізмі багатоголової уваги, запропонованому в 2017 році в статті "Attention Is All You Need". Текст перетворюється на числові представлення, звані токенами, і кожен токен перетворюється у вектор за допомогою таблиці вбудовування слів. На кожному шарі кожен токен контекстуалізується в межах контекстного вікна з іншими (немаскованими) токенами за допомогою паралельного механізму багатоголової уваги, що дозволяє підсилювати сигнал для ключових токенів і зменшувати менш важливі токени. Трансформери мають перевагу у відсутності рекурентних блоків, тому потребують менше часу на навчання, ніж попередні рекурентні нейронні архітектури (RNN), такі як довготривала короткочасна пам'ять (LSTM). Пізніші варіації були широко прийняті для навчання великих мовних моделей (LLM) на великих мовних наборах даних. Трансформери спочатку були розроблені як покращення попередніх архітектур для машинного перекладу, але з того часу знайшли багато застосувань. Вони використовуються у великомасштабній обробці природної мови, комп'ютерному зорі (візуальні трансформери), навчанні з підкріпленням, аудіо, мультимодальному навчанні, робототехніці та навіть у грі в шахи. Це також

призвело до розробки попередньо навчених систем, таких як генеративні попередньо навчені трансформери (GPT) та BERT (двонаправлені представлення кодувальника з трансформерів).

Архітектура decoder-only

Архітектура decoder-only складається виключно з декодерних блоків трансформера. Вона була популяризована моделями GPT-серії, такими як GPT-2, GPT-3 та GPT-4. Основна ідея полягає в автогенеративному підході, де модель навчається передбачати наступний токен на основі попередніх, використовуючи масковану самоувагу, як показано на рис. 1. Це дозволяє моделі генерувати текст послідовно, токен за токеном, враховуючи лише попередній контекст.

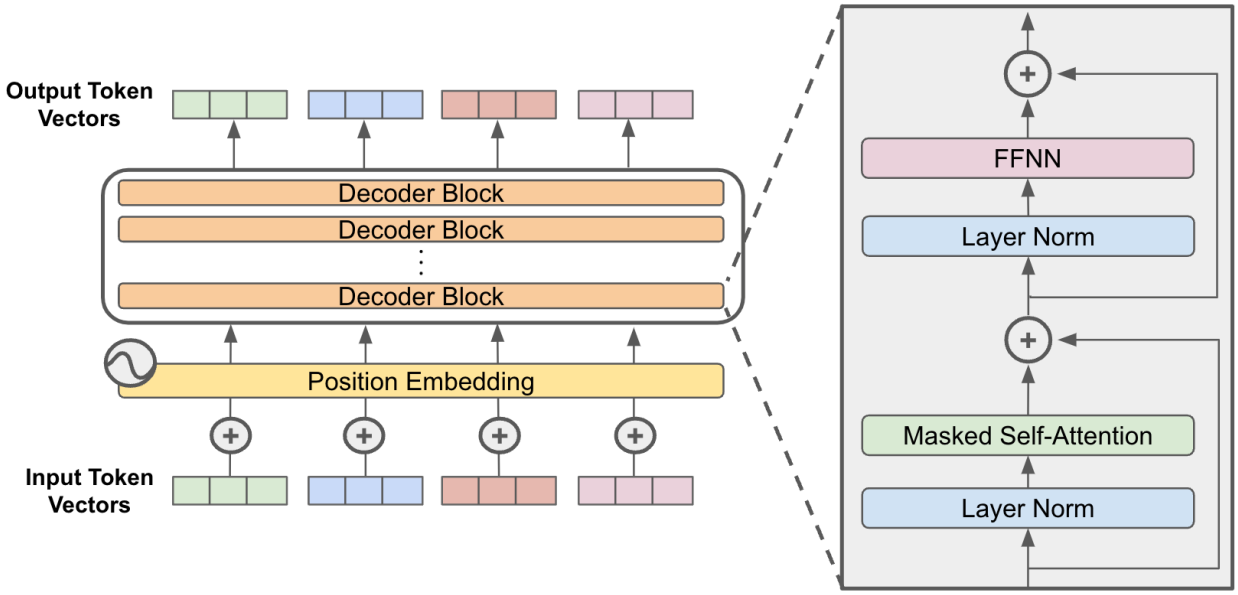

Рисунок 1. Схема архітектури decoder-only трансформера, де кожен токен обробляється послідовно з урахуванням лише попередніх токенів

Цей підхід ефективний для завдань, де необхідна генерація тексту без прямої залежності від повного вхідного контексту, таких як автозаповнення, генерація відповідей у чат-ботах та створення креативного контенту.

Архітектура encoder-decoder

Архітектура encoder-decoder, представлена в оригінальній статті "Attention Is All You Need", включає два основні компоненти: енкодер та декодер. Енкодер обробляє повний вхідний текст, створюючи контекстне

представлення, яке потім використовується декодером для генерації вихідного тексту, що показано на рис. 2, 3. Цей підхід дозволяє моделі враховувати як попередні, так і наступні токени вхідного тексту, що особливо корисно для завдань, де важливий повний контекст, таких як машинний переклад, узагальнення тексту та перефразування.

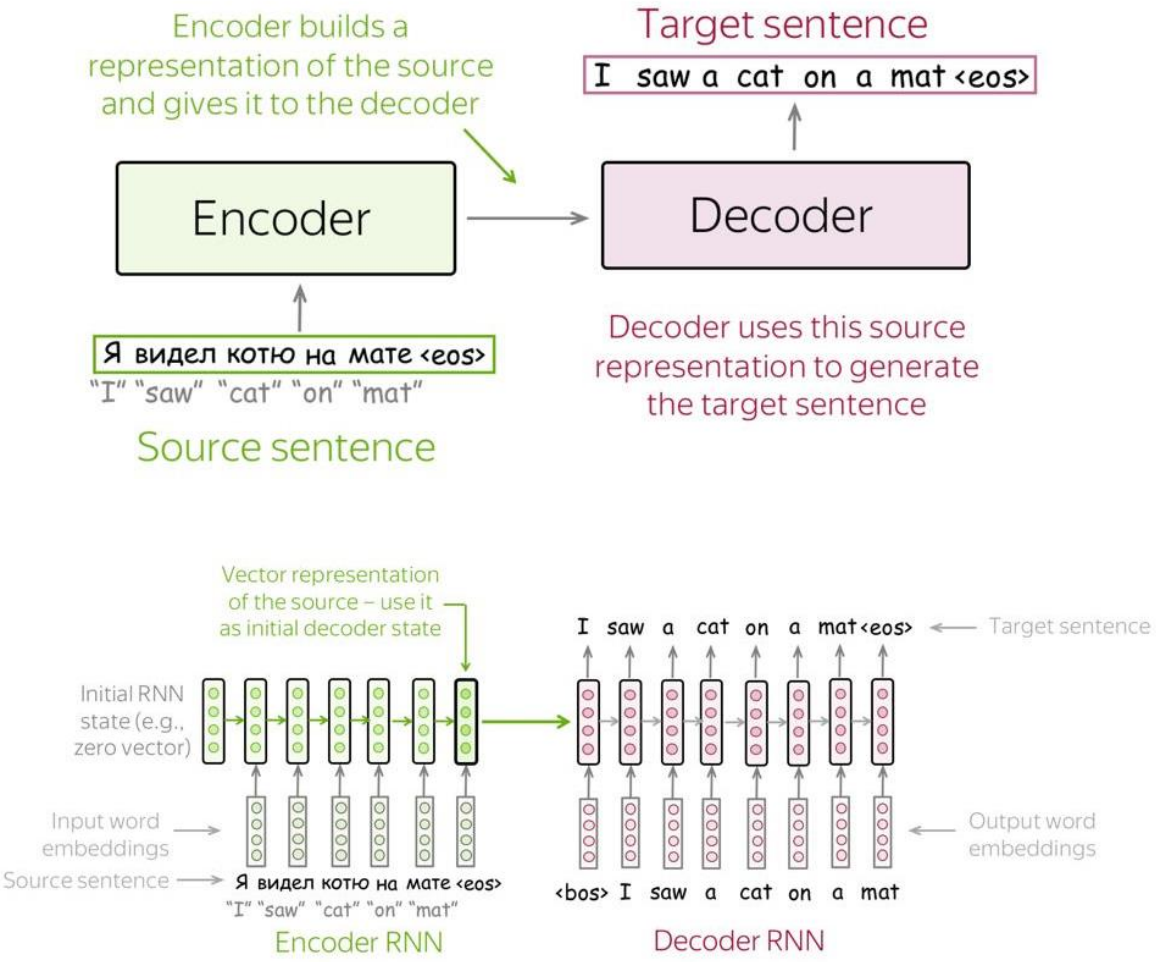

Рисунок 2. Схема архітектури encoder-decoder трансформера, де енкодер обробляє повний вхідний текст, а декодер генерує вихідний текст на основі контекстного представлення.

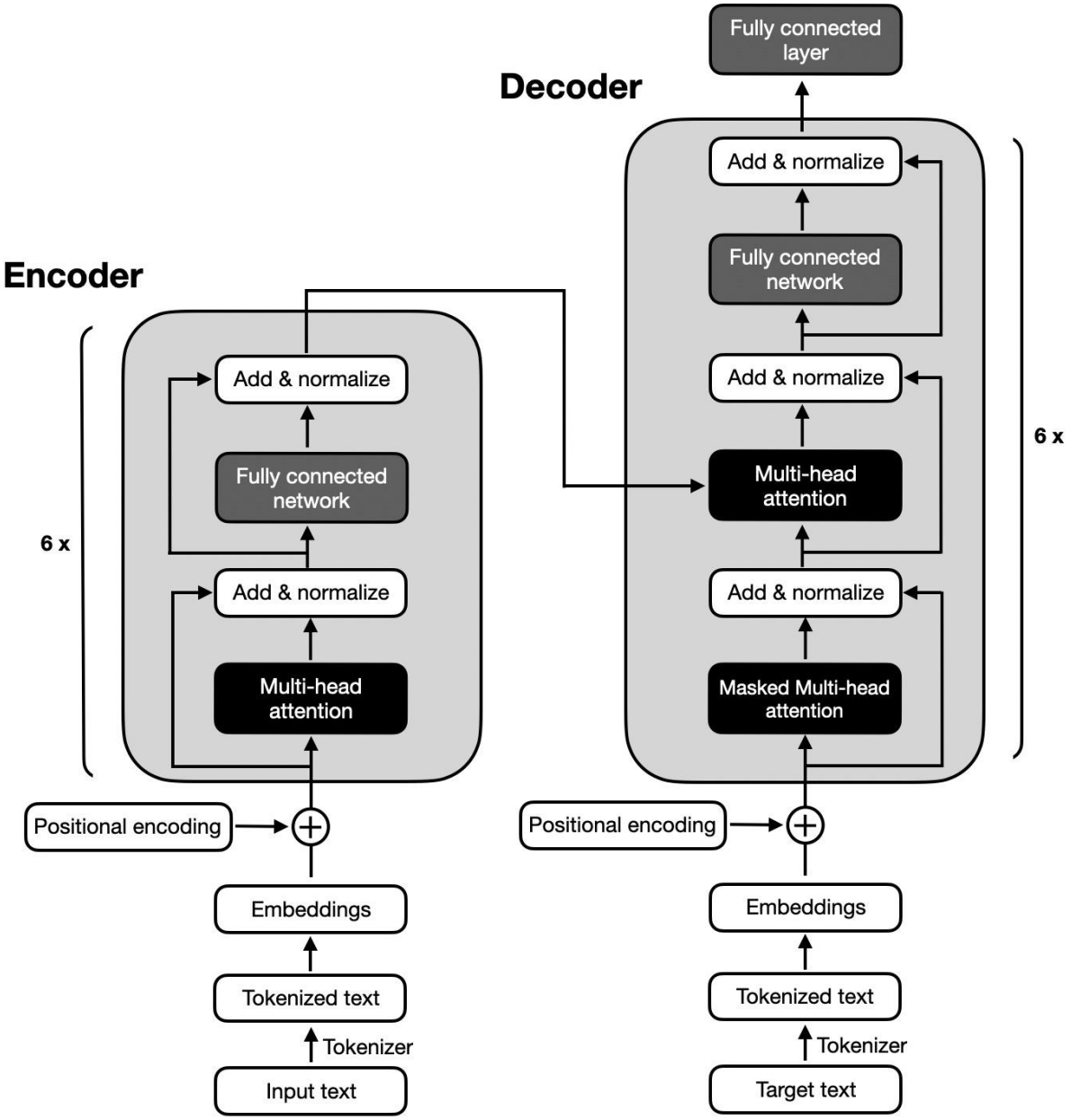

Рисунок 3. Компоненти архітектури encoder-decoder.

У моделях типу decoder-only, таких як GPT-2 чи LLaMA, інформація передається послідовно, і кожен токен залежить лише від попередніх. Це обмежує можливості для контекстуального впливу на активації, оскільки модель не має доступу до майбутніх токенів під час генерації. Натомість, в архітектурах encoder-decoder, таких як T5 або BART, енкодер обробляє весь вхідний текст, створюючи повне контекстне представлення, яке потім використовується декодером для генерації виходу. Це дозволяє більш гнучко маніпулювати активаціями, оскільки модель має доступ до повного контексту вхідного тексту. Таким чином, порівняння архітектур decoder-only та encoder-

decoder та підходів до впливу на їх активації є необхідним для розуміння того, як структурні особливості моделей впливають на їхню здатність до активаційного інжинірингу та моделювання особистісних рис. Це дозволяє обрати найбільш підходящу архітектуру для конкретних завдань, що вимагають контролю поведінки моделі через маніпуляції з її внутрішніми активаціями.

2.2 Аналіз методів управління особистісними рисами за допомогою активацій

Активаційний інжиніринг став потужним інструментом для модифікації поведінки великих мовних моделей (LLM), зокрема для цілеспрямованого впливу на їхні особистісні риси. Дослідження [1] показує, що внутрішні активації моделей містять інформацію, яка може бути зв’язана з психологічними характеристиками, такими як екстраверсія чи нейротизм. Наприклад, автори виявили, що вектор різниці між активаціями, отриманими під час генерації «екстравертних» та «інтровертних» відповідей, дозволяє систематично змінювати стиль комунікації моделі. Цей підхід ґрунтується на ідеї, що певні шари LLM відповідають за абстрактні концепти, які можна ізолювати та маніпулювати ними.

Один із найпоширеніших методів — Contrastive Activation Addition (CAA) — активно досліджується для decoder-only архітектур, таких як LLaMA чи GPT. Як продемонстровано в роботі [2], суть методу полягає в додаванні до активацій обраного шару (наприклад, 18-го) вектора, отриманого шляхом порівняння відповідей з протилежними рисами. Наприклад, для посилення креативності додається різниця між активаціями «креативних» і «шаблонних» відповідей. Це призводить до зсуву в поведінці моделі, який можна регулювати через коефіцієнт інтенсивності (α). Проте ефективність CAA обмежена для encoder-decoder моделей, таких як FLAN-T5, через розділення енкодера та декодера, що ускладнює визначення універсальних точок впливу.

Для архітектур encoder-decoder перспективним виявився підхід із застосуванням концепторів (Conceptors), описаний у дослідженні [3]. Концептори — це матриці, які фільтрують активації, посилюючи або пригнічуючи певні компоненти. Наприклад, вони дозволяють ізолювати

активації, пов’язані з «нейротизмом», і керувати їхньою інтенсивністю. Цей метод особливо корисний для моделей типу FLAN-T5, де активації енкодера відповідають за інтерпретацію контексту, а декодера — за генерацію тексту. Однак, як зазначають автори, концептори вимагають попереднього навчання на конкретній моделі, що ускладнює їхнє перенесення між архітектурами. Важливий внесок у розуміння активаційного інжинірингу вносить робота [4], де показано, що навіть складні явища (наприклад, «відмову» моделі відповідати) можна контролювати через окремі напрямки в активаційному просторі. Це підтверджує гіпотезу про те, що особистісні риси також можуть бути локалізовані в специфічних векторах. Проте, як підкреслюється в дослідженні [1], успіх таких маніпуляцій залежить від архітектури: у decoder-only моделях напрямки є більш стабільними, тоді як у encoder-decoder вони часто контекстно-залежні. Для моделей з двонаправленою обробкою (наприклад, BERT) активаційний інжиніринг ускладнюється через їхню здатність аналізувати весь текст одночасно. У таких випадках звертаються до методів на кшталт ортогоналізації ваг, які дозволяють виділити незалежні компоненти активацій. Наприклад, техніки, описані в роботі [5], використовують матричні перетворення для ізоляції векторів, відповідальних за конкретні риси. Це дозволяє уникнути побічних ефектів, але вимагає глибокого аналізу внутрішньої динаміки моделі.

Не дивлячись на перспективність цих методів, існують суттєві обмеження. Наприклад, маніпуляції з активаціями можуть порушувати когерентність відповідей або призводити до непередбачуваних змін у поведінці, особливо при роботі з encoder-decoder архітектурами. Крім того, як зазначається в [1], ефективність методів залежить від якості вихідних даних для ідентифікації активаційних напрямків: недостатньо репрезентативні запити можуть спричинити неточності.

2.3 Проблеми та обмеження існуючих досліджень

Існуючі дослідження методів управління особистісними рисами великих мовних моделей (LLM) через активації надають потужний інструментарій для модифікації поведінки моделей без потреби в перезапуску або перетренуванні. Однак, попри значний прогрес, ці підходи все ще мають низку суттєвих

проблем та обмежень, які можуть впливати на їх ефективність та універсальність.

Більшість методів, які активно використовуються в наукових дослідженнях, є специфічними для певних архітектур моделей, таких як GPT-2, GPT-3 та Llama. Техніки, орієнтовані на один тип трансформера або конкретну архітектуру, можуть не працювати з іншими моделями, що має важливе значення для їх застосування в реальних умовах. Наприклад, методи, розроблені для моделей на основі GPT, можуть не бути такими ж ефективними для encoder-decoder архітектур, таких як T5 або BART. Це обмежує застосування цих методів у широкому спектрі задач, де використовуються різні архітектури.

Одним з головних викликів є збереження загальних можливостей моделей при модифікації їх активацій. Хоча додавання керуючих векторів до активацій дозволяє досягати бажаних результатів, це може призвести до незначних змін в основних характеристиках моделі, таких як перплексія, точність або здатність до генерації коректного тексту на інших завданнях. Наприклад, в деяких випадках введення вектора керування може знижувати ефективність моделі на завданнях, які не пов'язані з управлінням особистісними рисами.

У більшості існуючих підходів контроль лише за однією конкретною рисою (наприклад, екстраверсія чи сентимент) є досить точним. Однак, керування кількома рисами одночасно залишається проблематичним через складність їх комбінування. Наприклад, спроби одночасно оптимізувати кілька параметрів, таких як баланс між агресивністю та співчуттям, часто призводять до погіршення результатів або до появи непередбачуваних взаємодій між різними рисами. Це вимагає створення більш складних і стабільних методів для комбінованого керування.

Ще однією важливою проблемою є відсутність єдиних стандартів для реплікації досліджень у цій сфері. Багато робіт проводяться на конкретних наборів даних та конкретних моделях, і через це результати можуть бути не відтворюваними або мати обмежене застосування в інших контекстах. Це також пов'язано з проблемою масштабованості: деякі методи, які показали хороші результати на невеликих моделях, можуть не працювати або виявляти

значні втрати при застосуванні до великих моделей, таких як GPT-4 або PaLM-2. Відсутність стандартизованих наборів для тестування таких методів у широкому контексті обмежує можливість проведення порівняльних досліджень і визначення, які методи є найбільш ефективними на практиці.

Хоча технології управління особистісними рисами через активації є потужними, вони також створюють потенційні етичні та безпекові ризики. Маніпуляції з особистісними рисами можуть бути використані для створення більш маніпулятивних або навіть шкідливих взаємодій між людиною та моделлю. Наприклад, в ситуаціях, де користувачі взаємодіють з моделями на основі конкретних психометричних профілів, може виникнути ризик маніпуляцій або несанкціонованого впливу на поведінку користувача. Більш того, деякі дослідження вказують на те, що управління такими рисами може підвищувати токсичність або неадекватну поведінку в залежності від налаштувань моделі, що вимагає ретельного контролю та безпеки при їх використанні.

Дослідження інтервенцій на рівні активацій можуть бути складними для інтерпретації та розуміння, що саме відбувається в процесі модифікації активацій. Часто використовуються методи, які вимагають значного технічного розуміння внутрішніх механізмів роботи трансформерів, що ускладнює роботу з ними для широкої аудиторії. Наприклад, методи, які використовують концептори або інші складні підходи до управління активаціями, можуть бути складними для практичного застосування без відповідних інструментів та знань про модель.

Крім того, важко визначити універсальний метод для різних завдань і типів застосування. Хоча дослідження показують, що деякі методи, такі як PAS (Personality Activation Search), дають відмінні результати на Llama-3, це не гарантує, що вони будуть такими ж ефективними для всіх моделей або завдань. Наприклад, для простих завдань генерації тексту може бути достатньо простих методів активаційного керування, тоді як для складних міждисциплінарних завдань (наприклад, медичні або юридичні консультації) можуть знадобитися більш складні або комбінаційні методи.

3. Висновки

Виконане дослідження методів управління особистісними рисами великих мовних моделей (LLM) через активації на архітектурах decoder-only та encoder-decoder показало значний потенціал цих методів у контексті модифікації поведінки моделей без необхідності в перетренуванні або зміни їх основної структури. Основна увага була зосереджена на порівнянні архітектур і розгляді існуючих методів активації, зокрема застосування методів на основі додавання контрастивних активацій (CAA), концепторів і пошуку активацій для управління особистісними рисами.

Аналіз існуючих досліджень дозволив виявити значні досягнення в галузі управління окремими рисами особистості, такими як екстраверсія чи нейротизм, однак під час дослідження було виявлено і ряд обмежень, зокрема обмежена універсальність методів для різних моделей, проблеми з комбінуванням кількох рис одночасно, а також складнощі з відтворюваністю та масштабованістю. Також важливим аспектом, що вимагає подальшого розвитку, є покращення безпеки та етичних стандартів при використанні таких технологій.

Поряд з цим, на основі отриманих результатів, можна стверджувати, що модифікація активацій у великих мовних моделях здатна забезпечити ефективне управління особистісними рисами, що відкриває нові можливості для створення адаптованих інтерфейсів та взаємодій у різноманітних додатках. Однак для подальшого розвитку цієї галузі необхідно здійснити додаткові дослідження, спрямовані на розробку універсальних та стабільних методів для комбінованого управління кількома рисами, підвищення відтворюваності результатів та інтеграцію цих методів у широке коло застосувань.

4. Джерела інформації

1. Allbert R. A., Wiles J. K., Grankovsky V. Identifying and Manipulating Personality Traits in LLMs Through Activation Engineering [Електронний ресурс] / R. A. Allbert, J. K. Wiles, V. Grankovsky. – 2024. – Режим доступу: https://doi.org/10.48550/arXiv.2412.10427. – Дата доступу: 17.04.2025.

2. Panickssery N., Gabrieli N., Schulz J., Tong M., Hubinger E., Turner A. M. Steering Llama 2 via Contrastive Activation Addition [Електронний ресурс] / N. Panickssery, N. Gabrieli, J. Schulz, M. Tong, E. Hubinger, A. M. Turner. – 2023. – Режим доступу: https://doi.org/10.48550/arXiv.2312.06681. – Дата доступу: 17.04.2025.
3. Postmus J., Abreu S. Steering Large Language Models using Conceptors: Improving Addition-Based Activation Engineering [Електронний ресурс] / J. Postmus, S. Abreu. – 2024. – Режим доступу: https://doi.org/10.48550/arXiv.2410.16314. – Дата доступу: 17.04.2025.
4. Arditi A., Obeso O., Syed A., Paleka D., Panickssery N., Gurnee W., Nanda N. Refusal in Language Models Is Mediated by a Single Direction [Електронний ресурс] / A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, N. Nanda. – 2024. – Режим доступу: https://doi.org/10.48550/arXiv.2406.11717. – Дата доступу: 17.04.2025.
5. Weng Y., He S., Liu K., Liu S., Zhao J. ControlLM: Crafting Diverse Personalities for Language Models [Електронний ресурс] / Y. Weng, S. He, K. Liu, S. Liu, J. Zhao. – 2024. – Режим доступу: https://doi.org/10.48550/arXiv.2402.10151. – Дата доступу: 17.04.2025.