

КОНТЕКСТНО ЗАЛЕЖНА АДАПТАЦІЯ ВІДПОВІДЕЙ ГЕНЕРАТИВНИХ LLM

Березюк М.О.¹, Гуда А.Ю.²

¹ Український державний університет науки і технологій, аспірант, Україна

² Український державний університет науки і технологій,

доктор технічних наук, професор, Україна

Анотація. В умовах кризових ситуацій оперативність реагування набуває вирішального значення. Великі мовні моделі (LLM) здатні генерувати корисні рекомендації, однак їх стандартна поведінка часто не враховує специфічний контекст надзвичайних подій. У цій доповіді пропонується Adaptive Injectable Realignment Model (Adaptive IRM) як підхід для досягнення контекстно залежної генерації відповідей. Adaptive IRM – це невеликий нейронний модуль, який інтегрується у прямий прохід LLM та ін'єктує контекстні сигнали для корекції внутрішніх представлень моделі без зміни її основних ваг. Наша розробка орієнтована на кризовий контекст, з використанням N нейронів (у нашому випадку $N = 4$), які відповідають за стихійні лиха (землетрус, повінь, пожежу та ураган). Вихідні сигнали Adaptive IRM, які додаються на різних шарах трансформера, модифікує механізми уваги з метою підсилення інформації, релевантної конкретному контексту. У доповіді окреслено архітектуру IRM, описано запропоновані розширення, а також подано план використання датасету HumAID для навчання Adaptive IRM. Експериментальні результати поки відсутні; натомість, обговорено концепцію, мотивацію та перспективи впровадження запропонованої системи.

Ключові слова: Контекстно залежна генерація; Великі мовні моделі; Адаптивне вирівнювання; Модуляція нейронів; Кризова інформатика; Датасет HumAID; Втручання на рівні шарів; Архітектури трансформера; Адаптація поведінки моделі

Вступ

У сучасних умовах надзвичайних подій оперативне реагування є критично важливим [4]. Великі мовні моделі (LLM) мають потенціал надавати користувачам корисні рекомендації, наприклад, відповідати на запитання «що мені робити?», однак їх стандартна поведінка зазвичай не адаптується під конкретний контекст кризи. У ситуаціях надзвичайних подій контекст (наприклад, тип катастрофи) визначає, яка інформація є релевантною та які рекомендації найбільш доцільні. Для забезпечення такої адаптації необхідна система, здатна «впроваджувати» контекст у внутрішні представлення LLM.

Ця робота спрямована на розробку Adaptive IRM – адаптивного модуля інжекційного переналаштування, який інтегрується у прямий прохід LLM та впливає на її механізми уваги, ін'єктуючи контекстний сигнал. Сигнал представлено у вигляді вектора з N нейронів (планується $N = 4$), де кожен нейрон відповідає за певну категорію кризової події: землетрус, повінь, пожежа, ураган. Таким чином, система зможе генерувати відповіді, які адаптовані до конкретного кризового сценарію, забезпечуючи оперативність, точність та безпеку генерованих рекомендацій.

Adaptive IRM для контекстно залежної генерації

Injectable Realignment Model (IRM) – це сучасний підхід для корекції поведінки великої мовної моделі під час її генеративного процесу [1, 5]. IRM являє собою невелику прямопрохідну нейронну мережу, яку інтегрують паралельно з попередньо натренованою LLM. Під час прямого проходу LLM IRM отримує деякі внутрішні активації моделі і генерує вектор, який ін'єктується у певні шари трансформера через додавання у залишковий потік. Такий підхід дозволяє «переналаштувати» внутрішні представлення моделі без зміни основних ваг LLM [1, 2]. Оригінальний IRM був застосований для досягнення певного стилю генерації (наприклад, емоційної тональності) та показав, що невелика мережа може ефективно модифікувати поведінку LLM за допомогою малочисельних параметрів.

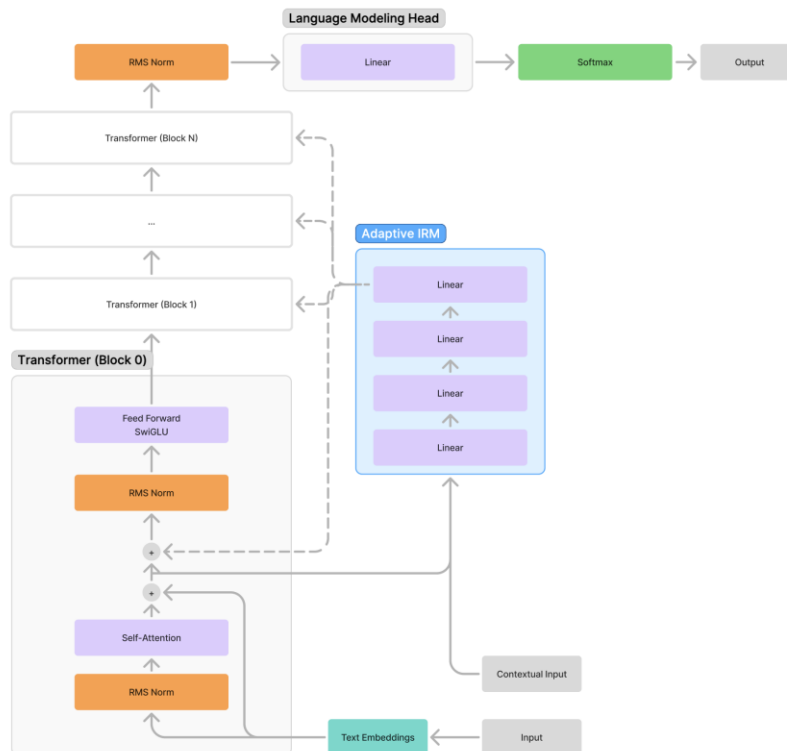


Рисунок 1 – Архітектура трансформера LLM з інтегрованим модулем Adaptive IRM. Adaptive IRM отримує на вході активації з ранніх шарів трансформера після механізму уваги та додатково — контекстний вектор. На виході модуль генерує корекційний сигнал, який додається до залишкових представлень у вибраних шарах трансформера. Таким чином Adaptive IRM впливає на увагу та генерацію LLM відповідно до заданого контексту.

Наша пропозиція полягає у розширенні концепції IRM для досягнення контекстної адаптації. Ми модифікуємо IRM так (Рисунок 1), щоб він приймав додатковий вхід – контекстний сигнал, представлений у вигляді N нейронів. Кожен з цих нейронів відповідає за конкретний кризовий контекст: землетрус, повінь, пожежа, ураган. Під час прямого проходу LLM, Adaptive IRM отримує як вихід з ранніх шарів LLM, так і контекстний вектор, який конкатенується з цим виходом. Таким чином, IRM генерує корекційний вектор, який ін'єктується на кількох шарах трансформера. Це дозволяє змінити розподіл уваги LLM, спрямовуючи її на релевантні для заданого контексту особливості. Наприклад, у випадку урагану відповідь може містити інформацію про евакуацію та повідомлення про погодні умови, а в умовах землетрусу – про безпечні зони та можливі післязсуви.

Датасет HumAID та стратегія навчання

Для навчання Adaptive IRM планується використання датасету HumAID – великої колекції кризових твітів з людською анотацією [3]. HumAID містить приблизно 77 000 твітів, зібраних під час 19 масштабних надзвичайних подій (2016–2019 рр.), охоплюючи різні типи катастроф. Кожен твіт анотовано за типом події та гуманітарними категоріями (наприклад, «пошук/термінова допомога», «підтримка», «пошкодження інфраструктури»).

Під час попередньої обробки даних ми будемо витягувати текст твітів та відповідні контекстні мітки (тип катастрофи: землетрус, повінь, пожежа, ураган). Контекстні мітки слугуватимуть вихідним сигналом для Adaptive IRM. Стандартна обробка тексту включатиме видалення URL, нормалізацію пробілів та очищення від нерелевантного контенту. Крім того, забезпечуватиметься збалансованість датасету по кожній категорії.

Під час навчання базова LLM буде заморожена, а лише параметри Adaptive IRM (включаючи контекстний вектор та ваги прямопрохідних шарів) будуть оптимізовані [1. 2]. Об'єктом оптимізації є така функція втрат, що заохочує модель до генерації відповідей, що точно відображають контекстну інформацію (наприклад, через завдання класифікації або генерації короткого звіту на основі заданого тексту). Таким чином, Adaptive IRM навчиться відображати сутність кожного типу кризи у своїх активаціях, що дозволить LLM генерувати більш адаптовані та релевантні відповіді під час надзвичайних ситуацій.

Висновки. У цій доповіді представлено концепцію та архітектурне рішення для контекстно залежної адаптації відповідей генеративних LLM за допомогою Adaptive IRM. Запропонована система дозволяє інтегрувати контекстний сигнал (у вигляді N нейронів, для нашої реалізації – 4, які відповідають за землетрус, повінь, пожежу, ураган) для корекції внутрішніх представлень LLM, що впливає на механізми уваги та генеративний процес.

Ми описали основні принципи оригінальної IRM, її впровадження та розширення для контекстної адаптації, а також подано стратегію використання датасету HumAID для навчання моделі. Хоч експериментальні

результати поки відсутні, запропонована концепція обіцяє значні переваги у застосуванні у кризових ситуаціях, де точність, релевантність та оперативність інформації мають критичне значення.

Подальші дослідження спрямовані на реалізацію та оцінку Adaptive IRM у лабораторних та польових умовах, а також на аналіз інтерпретованості отриманих активацій для розуміння того, як контекст впливає на процес генерації.

ЛІТЕРАТУРА

- 1.Smith B., Baker D., Chase C. та ін. The Mysterious Case of Neuron 1512: Injectable Realignment Architectures Reveal Internal Characteristics of Meta’s Llama 2 Model // arXiv [електронний препринт]. – 2024. – №: arXiv:2407.03621. – Режим доступу: <https://arxiv.org/abs/2407.03621> (дата звернення: [не вказано]).
- 2.Dathathri S. та ін. Plug and Play Language Models: A Simple Approach to Controlled Text Generation // Матеріали 8-ї Міжнародної конференції з питань Learning Representations (ICLR) [тезиси доповідей]. – 2020.
- 3.Alam F., Qazi U., Imran M., Ofli F. HumAID: Human-Annotated Disaster Incidents Data from Twitter with Deep Learning Benchmarks // Матеріали конференції ICWSM 2021 [тезиси доповідей]. – 2021.
- 4.Otal H. T., Canbaz M. A. LLM-Assisted Crisis Management: Building Advanced LLM Platforms for Effective Emergency Response and Public Collaboration // arXiv [електронний препринт]. – 2024. – №: arXiv:2402.10908. – Режим доступу: <https://arxiv.org/abs/2402.10908> (дата звернення: [не вказано]).
- 5.DRAGN Labs. Injectable Alignment Model – README [електронний ресурс]. – 2024. – GitHub Repository. – Режим доступу: <https://github.com/DRAGNLabs/InjectableAlignmentModel> (дата звернення: [не вказано]).

CONTEXT-AWARE ADAPTATION OF GENERATIVE LLM RESPONSES

Berezuk M.O., Guda A.I.

Abstract. *In crisis situations, the speed of response becomes critically important. Large Language Models (LLMs) are capable of generating useful recommendations; however, their default behavior often fails to account for the specific context of emergency events. This paper proposes the Adaptive Injectable Realignment Model (Adaptive IRM) as a method for achieving context-aware response generation. Adaptive IRM is a lightweight neural module that integrates into the LLM’s forward pass and injects contextual signals to adjust the model’s internal representations without modifying its original weights. Our approach is focused on disaster scenarios, using N neurons (in our case, $N = 4$) corresponding to natural hazards such as earthquakes, floods, fires, and hurricanes. The output signals from the Adaptive IRM, injected at various transformer layers, modulate*

attention mechanisms to emphasize information relevant to the given context. The paper outlines the IRM architecture, describes the proposed extensions, and presents a plan for using the HumAID dataset to train the Adaptive IRM. Experimental results are not yet available; instead, the concept, motivation, and implementation strategy of the proposed system are discussed.

Keywords: *Context-aware generation; Large Language Models; Adaptive alignment; Neuron modulation; Crisis informatics; HumAID dataset; Layerwise intervention; Transformer architectures; Model behavior adaptation*

REFERENCE:

1. Smith, B., Baker, D., Chase, C., et al. The Mysterious Case of Neuron 1512: Injectable Realignment Architectures Reveal Internal Characteristics of Meta's Llama 2 Model // arXiv [preprint]. – 2024. – ID: arXiv:2407.03621. – Available at: <https://arxiv.org/abs/2407.03621> (Accessed: [date not specified]).
2. Dathathri, S., et al. Plug and Play Language Models: A Simple Approach to Controlled Text Generation // Proceedings of the 8th International Conference on Learning Representations (ICLR). – 2020.
3. Alam, F., Qazi, U., Imran, M., Ofli, F. HumAID: Human-Annotated Disaster Incidents Data from Twitter with Deep Learning Benchmarks // Proceedings of the ICWSM 2021 Conference. – 2021.
4. Ota, H. T., Canbaz, M. A. LLM-Assisted Crisis Management: Building Advanced LLM Platforms for Effective Emergency Response and Public Collaboration // arXiv [preprint]. – 2024. – ID: arXiv:2402.10908. – Available at: <https://arxiv.org/abs/2402.10908> (Accessed: [date not specified]).
5. DRAGN Labs. Injectable Alignment Model – README [GitHub Repository, electronic resource]. – 2024. – Available at: <https://github.com/DRAGNLabs/InjectableAlignmentModel> (Accessed: [date not specified]).