УДК 004.85

ПІДХОДИ ДО ПОЯСНЮВАНОСТІ РІШЕНЬ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ У ІНТЕЛЕКТУАЛЬНИХ СИСТЕМАХ

Матвєєв М.С.
e-mail: mykhailo.matvieiev@nure.ua
Науковий керівник – доц., канд. техн. наук Н.М. Сердюк
Харківський національний університет радіоелектроніки, каф. ІСТ
м. Харків, Україна

This work is devoted to the analysis of explainability approaches in Large Language Models used within intelligent information systems. The growing complexity of transformer-based architectures strengthens the “black-box” problem and increases the need for transparent and interpretable decision-making. The study examines post-hoc interpretation methods, attention-based analysis, and chain-of-thought prompting as practical tools for improving model transparency. Special attention is paid to emerging directions of mechanistic interpretability aimed at understanding internal representations of LLMs. The results emphasize the necessity of combining different interpretability strategies to enhance reliability, auditability, and user trust in LLM-based applications.

У сучасних умовах стрімкого розвитку інтелектуальних інформаційних систем зростає потреба не лише у підвищенні якості генерації та класифікації тексту, а й у забезпеченні прозорості прийняття рішень. Це особливо важливо для систем підтримки прийняття рішень, де користувачеві необхідно оцінити обґрунтованість результату, виявити потенційні помилки та сформувати довіру до алгоритмічного висновку. Базовою технологічною основою більшості Large Language Models є архітектура Transformer, у якій ключову роль відіграє механізм self-attention, що забезпечує моделювання залежностей у тексті та масштабованість навчання. Разом із тим, підвищення складності та розмірності трансформерних моделей посилює проблему “чорної скриньки”: модель демонструє високі метрики, але її внутрішня логіка залишається неочевидною для розробника і користувача, що ускладнює контроль якості, аудит і безпечне впровадження в прикладних системах [1].

У контексті LLM відповідні підходи доцільно розглядати як сукупність методів, що дозволяють аналізувати вплив вхідних даних, запиту та внутрішніх представлень моделі на кінцевий результат. Одним із практично значущих напрямів є застосування пост-хок методів, які не потребують модифікації моделі та формують інтерпретацію вже після отримання відповіді. Класичним прикладом є LIME, який будує локальну апроксимацію рішення поблизу конкретного прогнозу та дозволяє оцінити вагомість окремих токенів або ознак. Інший поширений підхід – SHAP, що спирається на значення Шеплі

та формує адитивну оцінку внеску ознак у результат, забезпечуючи узгодженість і порівнюваність аналізу між прикладами. Для інтелектуальних систем, що працюють із резюме та вакансіями, такі методи можуть використовуватись для визначення, які фрагменти тексту найбільше вплинули на оцінку відповідності вакансії, які компетенції були виявлені моделлю та які зміни можуть підвищити релевантність профілю [2].

Одним із напрямків наукових досліджень є вивчення ролі механізмів уваги в поясненні роботи моделей. Адже ваги уваги можуть бути використані як карта важливості окремих токенів у тексті. Проте дослідження показують, що увага не завжди може бути надійним поясненням роботи моделей. Іноді ваги уваги можуть бути змінені без суттєвого впливу на результат, тому потрібно бути обережним при тлумаченні результатів та проводити додаткові тести для підтвердження пояснення. Тому, якщо визначити пояснення правильно і застосувати строгі експериментальні процедури, то увага може бути корисним елементом у поясненні роботи моделей, особливо якщо використовувати комплексний підхід до інтерпретовності систем обробки природної мови. Практичним доповненням до пост-хок інтерпретації є методи, пов'язані з промптингом, зокрема chain-of-thought prompting, який стимулює модель формувати проміжні кроки мислення і тим самим підвищує прозорість процесу отримання відповіді в задачах, що потребують послідовних висновків. Важливо, однак, розрізняти зовнішню правдоподібність пояснення і його фактичну причинність: текстове пояснення може виглядати переконливо, але не завжди відображає внутрішні причини прийняття рішення, тому актуальним є використання різних підходів (пост-хок інтерпретацій, тестів вірності, аналізу стабільності) у реальних інтелектуальних системах.

Перспективним напрямом сучасних досліджень є механістичний аналіз великих мовних моделей. На відміну від підходів пост-хок, які базуються на інтерпретації результатів на основі вхідних даних, механістичний підхід спрямований на вивчення внутрішніх представлень моделі та структур її функціонування. Його метою є виявлення закономірностей формування латентних ознак і встановлення зв’язку між внутрішніми активаціями та згенерованим результатом.

У рамках цього напрямку особлива увага приділяється аналізу архітектури трансформера як базового компоненту великих мовних моделей. Результати досліджень показують, що застосування методів словникового навчання, розріджених автоенкодерів, факторизації активацій та аналізу внутрішніх представлень окремих шарів дозволяє ідентифікувати стабільні патерни, що формуються в процесі навчання та кодуються в параметрах моделі. Зокрема, виявляються латентні ознаки, пов’язані з синтаксичними структурами, семантичними категоріями або абстрактними концептами, що

беруть участь у генерації відповіді. Аналіз динаміки активацій та їх взаємодії між шарами відкриває можливості для дослідження механізмів узагальнення, локалізації функціональних підмодулів усередині мережі та оцінювання стабільності поведінки моделі за різних умов вхідних даних. Такий підхід дозволяє переходити від зовнішньої інтерпретації результату до структурного аналізу причин його формування.

У контексті застосування, зокрема в інтелектуальних системах автоматизованої генерації резюме, механістичний аналіз може бути використаний як інструмент системного контролю якості функціонування моделі. Використання цього підходу дозволяє виявляти систематичні упередження, діагностувати некоректні узагальнення, а також оцінювати стабільність результатів при зміні вхідних даних або формулюванні запиту. Аналіз внутрішніх представлень і активацій сприяє виявленню потенційних джерел помилок, локалізації нестабільних компонентів системи та підвищенню відтворюваності прийнятих рішень. У поєднанні з пост-хок підходами це сприяє формуванню пояснень на підставі не лише текстової інтерпретації результату, але й діагностичних показників внутрішнього функціонування моделі. Такий підхід підвищує надійність прикладних інтелектуальних сервісів і сприяє підвищенню довіри користувачів до автоматизованих рекомендацій.

Загалом, формалізація пояснюваності як функціональної вимоги до інтелектуальних систем узгоджується з підходом, що розвивається в рамках програм на кшталт DARPA XAI, де підкреслюється необхідність систем, здатних пояснювати раціональність, сильні та слабкі сторони, а також передбачувану поведінку [3]. Отже, поєднання пост-хок інтерпретації (LIME, SHAP), обережної роботи з attention-візуалізаціями, промпт-орієнтованих пояснень (chain-of-thought) та механістичних підходів формує науково обґрунтований базис для створення пояснюваних LLM-орієнтованих інтелектуальних систем і є перспективним напрямом подальших досліджень. Це особливо актуально для прикладних інтелектуальних систем, де прозорість рішень безпосередньо впливає на довіру користувача та безпечність впровадження.

Список використаних джерел:

1. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., Kaiser Ł., Polosukhin I. Attention Is All You Need / A. Vaswani // Advances in Neural Information Processing Systems (NeurIPS). – 2017. URL: https://arxiv.org/abs/1706.03762 (дата звернення: 2.03.2026).

2. Ribeiro M. T., Singh S., Guestrin C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier / M. T. Ribeiro // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2016. URL: https://arxiv.org/abs/1602.04938(дата звернення: 2.03.2026).

3. Gunning D., Aha D. DARPA’s Explainable Artificial Intelligence (XAI) Program / D. Gunning // AI Magazine. – 2019.