УДК 004.94:004.8

Тимошенко Д.О., здобувач,
Рудніченко М.Д., к.т.н., доцент
Національний університет «Одеська політехніка»

АНАЛІЗ СУЧАСНИХ МАЛИХ МОВНИХ МОДЕЛЕЙ ДЛЯ АВТОМАТИЗАЦІЇ ПРОЦЕСІВ АСИСТУВАННЯ В ОНЛАЙН-ЗУСТРІЧАХ

Вступ. Серед нових напрямів розвитку штучного інтелекту особливої уваги потребують малі мовні моделі (SLM), що поєднують здатність до глибокого семантичного розуміння та підвищену ефективність обчислень при значно меншій кількості параметрів, ніж великі трансформерні моделі. SLM демонструють потенціал для реального часу автоматизації процесів підсумовування, класифікації, витягнення інформації та інтерактивних діалогових відповідей в умовах обмежених обчислювальних ресурсів, що є критично важливим для інтеграції в сервіси онлайн-зустрічей і корпоративні комунікаційні платформи. Поточні дослідження зосереджують увагу на оптимальному балансі між продуктивністю й ефективністю моделей, структурних компонентах архітектур, адаптації до конкретних доменних задач і впровадженні в продуктивні системи [1].

Основна частина. Початкові дослідження малих мовних моделей концентруються навколо зменшення кількості параметрів при збереженні здатності до природномовних трансформацій. SLM, на відміну від великих мовних моделей з десятками і сотнями мільярдів параметрів, зазвичай мають кількість параметрів від декількох мільйонів до кількох мільярдів, що дозволяє їм ефективно працювати в режимі реального часу на пристроях із обмеженою обчислювальною потужністю або в кейсах, де важлива конфіденційність даних через локальне виконання обчислень.

Класичні архітектури малих мовних моделей часто базуються на трансформерних компонентах зі скороченою глибиною та шириною, що зумовлюється зменшенням кількості шарів уваги і параметрів моделі. Такі архітектурні оптимізації включають методи дистиляції знань, квантування параметрів та відсікання частин моделі, що незначною мірою впливають на якість вихідного контенту, але суттєво знижують розмір і час обчислення [2].

У контексті автоматизації асистування під час онлайн-зустрічей важливим аспектом є підтримка декількох мовних задач: транскрипція мови в текст, тематичне узагальнення обговорень, виділення ключових тез, класифікація сенсу висловлювань та формування релевантних

резюме після завершення події. Архітектурно система, що використовує SLM, може бути побудована як модуль конвеєра даних, де вхідні аудіо- чи відеопотоки спочатку трансформуються в текстовий формат за допомогою систем розпізнавання мовлення, після чого обробка природної мови здійснюється SLM, адаптованою до конкретної доменної області задач і заданих критеріїв якості відповіді чи резюме. У цьому контексті модель може бути навчена на попередньо зібраному корпусі даних, специфічних для онлайн-зустрічей та конференцій, із включенням прикладів структурованих діалогів, презентаційних команд і групових обговорень. Такі моделі можуть бути інтегровані у багатокомпонентну архітектуру, де SLM виступає одним із вузлів, що взаємодіє з компонентами контекстного менеджера й історії діалогу, з можливістю персоналізації відповідей відповідно до ролей учасників, часових обмежень і типу теми обговорення [3].

На відміну від великих моделей, SLM відзначаються значно нижчими вимогами до пам’яті й обчислювальної потужності, що робить їх привабливими для розробки локальних агентів у сервіси онлайн-зустрічей, де темп і масштаб інтерпретації мовної інформації критичні для ефективної підтримки учасників. Архітектурні рішення щодо інтеграції SLM включають розподілені схеми обчислень, де транскрипційний модуль працює автономно, а SLM відповідає за семантичне узагальнення та генерацію висновків у реальному часі..

Висновки. Аналіз сучасних досліджень в області SLM показує, що архітектурні рішення повинні враховувати специфіку домену, потребу в реальному часі та адаптивності до діалогових контекстів. Подальші дослідження можуть спрямовуватися на покращення архітектурних інтеграцій, оптимізацію продуктивності моделей в умовах обмежених ресурсів і тестування в реальних сценаріях онлайн-комунікації..

Список використаних джерел:

1. Pham T.M., Nguyen P.T., Yoon S., Lai V.D., Dernoncourt F., Bui T. SlimLM: An Efficient Small Language Model for On-Device Document Assistance. ArXiv. 2024. P. 1–11.

Xu B., Chen Y., Wen Z., Liu W., He B. Evaluating Small Language Models for News Summarization: Implications and Factors Influencing Performance. Proceedings of the 2025 NAACL-Long. 2025. Vol. 1. P. 4909–4922.

Corradini F. State of the Art and Future Directions of Small Language Models. Information. 2025. Vol. 9, No 7. P. 189–189.