

Ілля Карпинець, Ігор Цимбалюк*

¹ студент, Національного лісотехнічного університету України, Львів, Україна. karpynets.i@nltu.lviv.ua

² асистент кафедри ІПЗ, Національного лісотехнічного університету України, Львів, Україна. <https://orcid.org/0009-0008-8489-473X>.
ihortsymbaliuk@nltu.edu.ua

Децентралізована система виконання великих мовних моделей штучного інтелекту у гетерогенному обчислювальному середовищі

***Анотація.** В роботі досліджується застосування сучасних моделей ШІ у середовищі, де користувачі з потужними пристроями отримують винагороду за надання ресурсів тим, хто потребує доступу до моделей без достатньої потужності. Поширення моделей підвищує попит на ресурси, централізовані дата-центри створюють проблеми централізації, втрати приватності, доступу компаній до даних і бар'єрів доступності. Децентралізована система розподілених вузлів виконує запити, вимагаючи маршрутизації завдань до вузлів та захисту приватності без прив'язки транзакцій до ідентифікаторів.*

***Ключові слова** – маршрутизація запитів, аноніміфікація даних, розподілені обчислення, децентралізовані фінанси.*

Маршрутизація запитів до вузлів з потрібними ресурсами виступає ключовим елементом системи, забезпечуючи швидкий пошук підходящих вузлів для виконання запиту з урахуванням вимог, таких як вибір моделі з певною архітектурою, розміром чи спеціалізацією [1], бюджет обчислень та обмежень приватності, наприклад, заборони відправки даних на незнайомі сервери. У гетерогенній мережі вузлів з різними обчислювальними можливостями ця маршрутизація мінімізує латентність і вартість, гарантуючи дотримання умов якості, швидкості та стабільності. Реєстр можливостей, як децентралізований індекс з локальним кешем і розподіленим реєстром, містить метадані вузлів, включаючи доступні моделі, обсяг VRAM, пропускну здатність, рівень довіри та поточне навантаження [3], слугуючи основою для швидкого виявлення ресурсів. Оцінювач сумісності, у формі багатовимірної функції ранжування, обчислює оцінку для пари запит-вузол, інтегруючи фактори латентності та витрат, що оптимізує

вибір у системі. Мультимодельна маршрутизація дозволяє розділяти запити щоб перемикалися на компактніші суміжні моделі, якщо потрібна відсутня в одному вузлі, розширюючи доступність системи. Контрактні угоди, як короткі криптографічно підписані SLA, фіксують очікувані параметри між замовником і вузлом, включаючи час виконання та відшкодування при невиконанні, забезпечуючи надійність взаємодії. Верифікація результатів через статистичні метрики, контрольні приклади, голосування кількох вузлів чи випробувальні запити з відомими відповідями захищає від помилок, тоді як управління неполадками шляхом автоматичної повторної маршрутизації при відмові вузла та захисту від некоректних результатів підвищує стійкість системи. Аноніміфікація запитів і транзакцій доповнює маршрутизацію, приховуючи метадані, такі як ініціатор запиту, обсяг ресурсів, час і частоту, для запобігання профілюванню та виявленню зв'язків між користувачами й моделями, застосовуючи схеми на кшталт кільцевих підписів, для маскування відправника, одержувача та суми в транзакціях ресурсів. Кільцеві підписи для транзакцій ресурсів ускладнюють відстеження підписувача угод SLA, адаптуючи метадані запиту в зашифрованому вигляді, тоді як приватні адреси чи одноразові ідентифікатори генеруються замовником для повернення результатів чи підтвердження оплати, забезпечуючи анонімність на рівні ідентифікації. Розподілені змішані мережі чи покровові ретранслятори приховують мережевий шлях запитів, знижуючи ризики кореляційного аналізу, а диференційна приватність з локальною обфускацією додає шум у метадані чи з'єднує запити для зменшення точності профілювання [2]. Змішування платежів інтегрується для подальшого маскування фінансових потоків, доповнюючи криптографічні інструменти. Архітектура системи (рис. 1) інтегрує ці елементи, починаючи з клієнта, який формує запит з параметрами моделі, затримки та приватності, створює одноразовий ідентифікатор і шифрує відповідь для безпечної передачі.

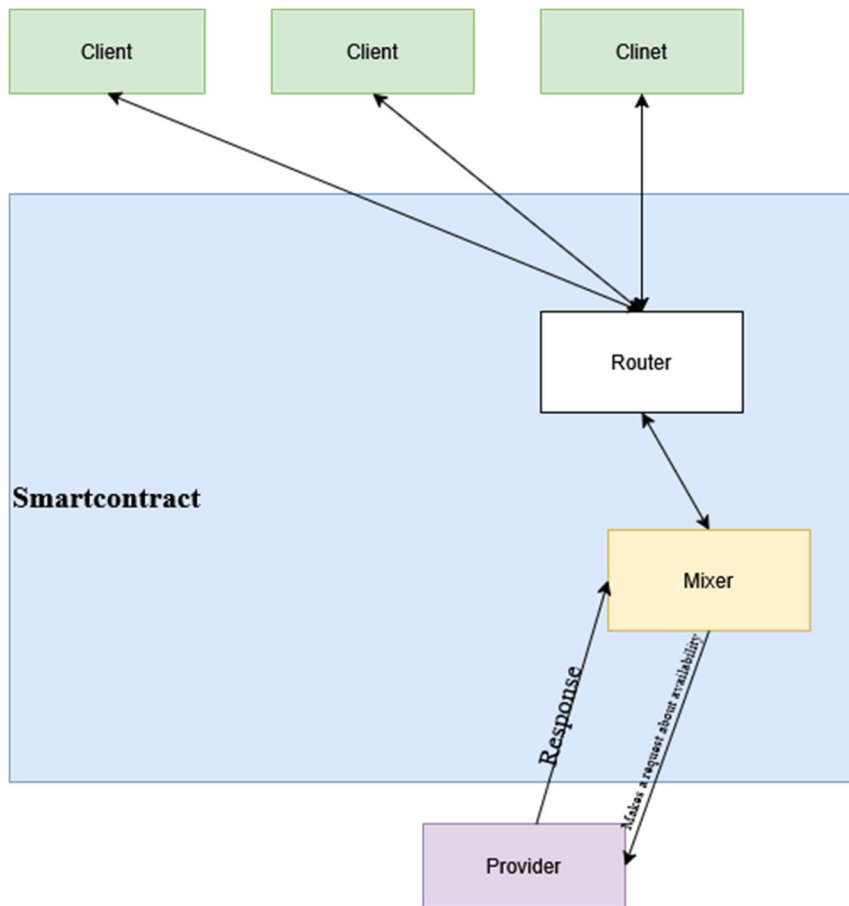


Рисунок 1. Архітектура системи

Клієнтська програма об'єднує запити, проводить локальну оптимізацію через кеш чи агрегацію, контактуючи з реєстром можливостей як DHT чи легковажим реєстром з підписаними метаданими вузлів для початкової оцінки. Модуль маршрутизації оцінює сумісність, підписує SLA та визначає стратегію поділу чи паралельності, спрямовуючи завдання до виконавчих вузлів, які виконують обчислення, підписують результат кільцевим підписом чи через ідентифікатор і повертають зашифрований результат. Мережевий рівень приватності через змішану мережу забезпечує анонімну передачу пакетів між клієнтом і вузлом, формуючи єдину децентралізовану платформу.

Висновки. Комбінація інтелектуальної маршрутизації та криптографічних механізмів анонімізації створює основу для практичної децентралізованої платформи виконання великих мовних моделей у гетерогенних середовищах, потенційно зменшуючи бар'єри доступу, підвищуючи стійкість інфраструктури та захищаючи приватність користувачів, хоча вимагає подальших досліді-

джень у верифікації результатів, витратах на розподілене збереження моделей та регуляторній сумісності.

Список використаних літературних джерел

- [1] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, Dario Amodei. Scaling Laws for Neural Language Models. URL: <https://doi.org/10.48550/arXiv.2001.08361>
- [2] Francesco Luigi De Faveri, Guglielmo Faggioli, Nicola Ferro. Words Blending Boxes. Obfuscating Queries in Information Retrieval using Differential Privacy. URL: <https://doi.org/10.48550/arXiv.2405.09306>
- [3] Keyu Zhang, Andrew Martin. RaceTEE: A Practical Privacy-Preserving Off-Chain Smart Contract Execution Architecture. URL: <https://doi.org/10.48550/arXiv.2503.09317>

Decentralized system for executing large-scale artificial intelligence language models in a heterogeneous computing environment

Ilya Karpynets, Igor Tsymbalyuk

The paper explores the application of modern AI models in an environment where users with powerful devices are rewarded for providing resources to those who need access to models without sufficient power. The proliferation of models increases the demand for resources, centralized data centres create problems of centralization, loss of privacy, access of companies to data and accessibility barriers. A decentralized system of distributed nodes executes requests, requiring routing of tasks to nodes and privacy protection without tying transactions to identifiers.