DOI: 10.34185/1991-7848.itmm.2026.01.088

ПРОТОКОЛ ДИНАМІЧНОЇ АДАПТАЦІЇ ІНФОРМАЦІЙНОЇ ЩІЛЬНОСТІ ПОВІДОМЛЕНЬ У МУЛЬТИАГЕНТНИХ СИСТЕМАХ НА ОСНОВІ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

Пономаренко П. А.[1] [ORCID], Божуха Л. М.[2] [ORCID]

[1]Дніпровський національний університет імені Олеся Гончара, аспірант, України

[2]Дніпровський національний університет імені Олеся Гончара,

д.т.н., професор, України

Анотація*. Зростання складності задач, що розв'язуються мультиагентними системами на основі великих мовних моделей, висуває підвищені вимоги до ефективності міжагентної комунікації. Існуючі підходи до стиснення контексту не враховують поточний стан агента-отримувача, що призводить до деградації якості виходів, зростання затримки та квадратичного збільшення токен-трафіку при масштабуванні системи. Ця робота пропонує протокол адаптивної міжагентної комунікації Receiver-Load-Aware Compression Protocol (RLACP), в якому рівень стиснення повідомлень динамічно визначається поточним когнітивним навантаженням агента-отримувача, що формалізується як композитна метрика з чотирьох компонентів: кількості активних задач, семантичної невизначеності виходів, затримки відповіді та рівня заповненості контексту. Залежно від значення цієї метрики агент-компресор застосовує один із чотирьох режимів стиснення — від збереження повного тексту до повної відмови від природної мови на користь структурованих пар ключ-значення.*

Ключові слова*: мультиагентні системи, великі мовні моделі, адаптивне стиснення, міжагентна комунікація, управління навантаженням контексту.*

Вступ

Мультиагентні системи на основі великих мовних моделей набувають широкого застосування у задачах складного планування, програмної інженерії та наукових досліджень. Проте зі зростанням кількості агентів та складності їхніх взаємодій виникає фундаментальна проблема: агенти обмінюються повідомленнями без урахування поточного стану агента-отримувача. Наслідком цього є накопичення когнітивного навантаження, деградація якості виконання задач та непередбачувана затримка системи. Зокрема, при наближенні контексту агента до максимальної довжини спостерігаються

ефекти «lost-in-the-middle» [1] та розсіювання уваги, що призводить до часткового або помилкового врахування вхідних повідомлень. Існуючі підходи до стиснення контексту, зокрема LLMLingua [2] та AutoCompressor [3], є статичними одноагентними рішеннями і не враховують стан отримувача в реальному часі, а також поточну затримку у відповіді.

Проведено аналіз архітектури Receiver-Load-Aware Compression Protocol (RLACP), що представляє собою протокол адаптивної міжагентної комунікації для мультиагентних систем на основі великих мовних моделей (LLM). Запропонований підхід динамічно регулює рівень стиснення повідомлень між агентами залежно від поточного когнітивного навантаження агента-отримувача, що дозволяє суттєво знизити деградацію якості виконання задач при зростанні завантаженості контексту.

Основний матеріал

Запропонована архітектура складається з трьох взаємодіючих компонентів: агентів-учасників із моніторами стану, шини статусів та агента-компресора. Кожен агент оснащений монітором стану, що безперервно обчислює композитну метрику когнітивного навантаження (1).

$$C_j = \alpha \cdot \tilde{T}_j + \beta \cdot H(Y_j) + \gamma \cdot \lambda_j + \delta \cdot F_j \quad (1)$$

де $\tilde{T}_j$ — нормалізована кількість активних задач агента j; $H(Y_j)$ — узагальнена міра семантичної невизначеності останніх виходів агента j (конкретна реалізація обирається залежно від типу задач агента та є предметом подальшої емпіричної валідації); λ_j — нормалізована затримка відповіді; F_j — заповненість контексту; α, β, γ, δ — вагові коефіцієнти.

Кожен агент регулярно публікує поточну метрику C_j у централізовану шину статусів за push-моделлю. Агент-компресор перехоплює вихідні повідомлення від інших агентів та застосовує функцію стиснення $M(C_j)$, де τ_1, τ_2, τ_3 — гіперпараметри розподілу режимів табл. 1.

Таблиця 1

Розподіл моделей стиснення повідомлень

Значення $M(C_j)$	**Назва режиму**	**Умова**
Mode 0	No Compression	$C_j < \tau_1$
Mode 1	Semantic Pruning	$\tau_1 \leq C_j < \tau_2$
Mode 2	Thematic Grouping	$\tau_2 \leq C_j < \tau_3$
Mode 3	Key-Value Extract	$C_j > \tau_3$

Використані режими мають своє призначення та відповідні властивості:

1. No Compression — стандартний режим без стиснення.

2. Semantic Pruning — видалення прикметників, вставних слів та «ввічливості»; залишаються лише іменники та дієслова [4].

3. Thematic Grouping — метод «центроїдного» стиснення, де замість послідовного викладу фактів агент-відправник групує інформацію за смисловими кластерами [5].

4. Key-Value Extract — передача лише пар *key: value* [6].

Ключові властивості протоколу. Архітектура RLACP характеризується такими принциповими властивостями:

1. **Асиметричність комунікації** — рівень стиснення є функцією стану отримувача, а не лише семантики повідомлення. Одне й те саме повідомлення може мати різний рівень стиснення в різні моменти часу.
2. **Неінвазивність** — протокол не змінює внутрішню логіку агентів і підключається до існуючих фреймворків (AutoGen, LangGraph, CrewAI) як middleware-прошарок.
3. **Graceful degradation** — при зростанні навантаження система поступово підвищує рівень стиснення, зберігаючи функціональність за рахунок деталізації, аналогічно механізму TCP congestion control.
4. **Субквадратичне масштабування** — при зростанні кількості агентів, токен-трафік системи зростає субквадратично завдяки автоматичному підвищенню стиснення при збільшенні навантаження.

Метрики оцінювання. Для кількісного підтвердження ефективності архітектури пропонується оцінювати три групи метрик: Task Completion Rate та Answer Faithfulness Score для вимірювання деградації якості при зростанні заповненості контексту; System Throughput для оцінки ефективності протоколу управління потоком; Total Token Traffic та Compression Efficiency Ratio для

аналізу масштабованості системи при збільшенні кількості агентів. Експериментальна валідація планується на сценаріях як рівномірного, так і нерівномірного розподілу навантаження між агентами.

Висновки

Запропонований протокол RLACP вирішує фундаментальну прогалину сучасних мультиагентних LLM-систем — відсутність механізму адаптації комунікації до поточного стану отримувача. На відміну від існуючих підходів, архітектура використовує активну трансляцію стану отримувача для динамічного регулювання рівня стиснення, що дозволяє одночасно підвищити якість виконання задач при перевантаженні контексту, знизити загальний токен-трафік системи та забезпечити стійке масштабування при зростанні кількості агентів. Неінвазивність архітектури робить можливим її інтеграцію до існуючих мультиагентних фреймворків без модифікації логіки агентів.

ЛІТЕРАТУРА / REFERENCE

1. Liu N. F. et al. Lost in the Middle: How Language Models Use Long Contexts. URL: https://arxiv.org/abs/2307.03172
2. Jiang H. et al. LLMLingua: Compressing Prompts for Accelerated Inference of Large Language Models. URL: https://arxiv.org/abs/2310.05736
3. Chevalier A. et al. Adapting Language Models to Compress Contexts. URL: https://arxiv.org/abs/2305.14788
4. Filippova, K., & Strube, M. (2008). Dependency Tree Based Sentence Compression. Proceedings of the Fifth International Natural Language Generation Conference (INLG'08), pp. 25–32. Association for Computational Linguistics. URL: https://aclanthology.org/W08-1105/
5. Radev, D., Jing, H., Stys, M., & Tam, D. (2004). Centroid-Based Summarization of Multiple Documents: Sentence Extraction, Utility-Based Evaluation, and User Studies. Proceedings of the NAACL Workshop on Automatic Summarization. URL: https://dl.acm.org/doi/10.3115/1117575.1117578
6. LLM-TKIE: Large Language Model Driven Transferable Key Information Extraction Mechanism. (2025). Scientific Reports. URL: https://www.nature.com/articles/s41598-025-15627-z

PROTOCOL FOR DYNAMIC ADAPTATION OF MESSAGE INFORMATION DENSITY IN MULTI-AGENT SYSTEMS BASED ON LARGE LANGUAGE MODELS

Pavlo Ponomarenko, Liliia Bozhukha

Abstract. *The increasing complexity of tasks solved by multi-agent systems based on large language models imposes higher demands on the efficiency of inter-agent communication. Existing approaches to context compression do not take into account the current state of the receiving agent, which leads to output quality degradation, increased latency, and quadratic growth of token traffic when scaling the system. This work proposes the Receiver-Load-Aware Compression Protocol (RLACP), an adaptive inter-agent communication protocol in which the level of message compression is dynamically determined by the current cognitive load of the receiving agent. This load is formalized as a composite metric consisting of four components: number of active tasks, semantic uncertainty of outputs, response latency, and context saturation level. Depending on the value of this metric, the compressing agent applies one of four compression modes — ranging from preserving full text to complete abandonment of natural language in favor of structured key-value pairs.*

Keywords: *multi-agent systems, large language models, adaptive compression, inter-agent communication, context load management.*