Міжнародний конкурс студентських робіт зі штучного інтелекту 2026

II тур
Перший етап

Р Е Ц Е Н З І Я

на наукову роботу

назва Мультиагентна система штучного інтелекту для когнітивно-поведінкової самодопомоги з формальними гарантіями безпеки

шифр Платформа "Відчуваю"

представлену на Міжнародний конкурс студентських робіт зі штучного інтелекту

№ з/п	Характеристики та критерії оцінки рукопису наукової роботи	Рейтингова оцінка. Максимальна кількість балів (за 120-бальною шкалою)	Бали
1	Актуальність проблеми	10	9,8
2	Новизна та оригінальність ідей	15	10,8
3	Використані методи дослідження	15	10,3
4	Теоретичні наукові результати	10	6,5
5	Практична направленість результатів	20	16,0
6	Рівень використання наукової літератури та інших джерел інформації	5	4,5
7	Ступінь самостійності роботи	10	8,8
8	Якість оформлення	5	4,5
9	Наукові публікації	10	0
10	Недоліки роботи (пояснення зниження максимальних балів у пунктах 1-9): Тема має особливу суспільну вагу, а робота поєднує продуману архітектуру, увагу до безпеки й приватності та завершений програмний продукт. Твердження про першість мультиагентної реалізації КПТ-щоденника та графової організації безпеки потребують переконливішого зіставлення з найближчими аналогами. У списку джерел уже наведено дослідження мультиагентної оркестрації терапевтичних чатботів. Доцільно точніше визначити оригінальність конкретних механізмів, а не напряму загалом. Формальний інваріант передування безпеки встановлює порядок виконання вузлів графа, але не доводить правильності класифікації ризику чи змістовної безпечності відповіді. Автор коректно розмежовує ці властивості в експериментальному розділі, однак назва та загальні висновки потребують такого самого обмеження. Недостатньо узгоджено математичну модель і програмну реалізацію. Наведено оптимізаційний функціонал рефреймінгу та інтегральну функцію втрат для навчання й калібрування, але не описано відповідного навчального експерименту, значень коефіцієнтів і результатів оптимізації. Потребує пояснення спосіб визначення цільової інтерпретації. Віддаленість від вихідної думки у векторному просторі сама по собі не встановлює якості рефреймінгу. Експериментальні вибірки невеликі: по два приклади на кожен із 26 класів спотворень і по 20 повідомлень на клас безпеки. Не наведено достатнього опису походження сценаріїв, незалежної експертної розмітки та узгодженості експертів. Відсутні довірчі інтервали й порівняння з базовою одноагентною системою. Результат детекції оцінено за одноклассовим еталоном, хоча система повертає декілька класів. Припущення про змістову виправданість додаткових класів потребує незалежного експертного підтвердження. Відсутність хибних спрацювань на 20 прикладах без спотворень не є достатньою підставою для загального висновку про відсутність надмірного приписування спотворень.		

№ з/п	Характеристики та критерії оцінки рукопису наукової роботи	Рейтингова оцінка. Максимальна кількість балів (за 120-бальною шкалою)	Бали
10	Перевірка кризового класифікатора виявила один пропуск серед 20 кризових повідомлень. Це важливий відкрито наведений результат, але він обмежує висновки щодо надійності. Класифікатор аналізує лише останнє повідомлення користувача; не перевірено випадків, у яких зміст ризику розкривається тільки з попередніх реплік діалогу. Перевірку стійкості до впровадження інструкцій описано як структурне тестування складання промпту. Вона не замінює наскрізного випробування поведінки мовної моделі та її інструментів. Крім того, для заявлених 30 спроб необхідно пояснити кількість варіантів або повторів: п’ять форм атак і три канали самі по собі утворюють 15 комбінацій. Перевірка консолідації пам’яті охоплює лише два синтетичні набори. Коректність ідентифікаторів джерел підтверджує цілісність посилань, але не доводить збереження змісту фактів після злиття. Потрібно окремо оцінити втрату суттєвої інформації, появу непідтверджених узагальнень і поведінку на суперечливих записах. Практичну працездатність платформи продемонстровано, проте якість рефреймінгу, корисність для користувача та придатність до заявлених сценаріїв застосування окремо не перевірено. Висновки про використання в реабілітаційних установах доцільно подавати як перспективу подальшої фахової апробації. Оформлення загалом належне, але наявні редакційні неточності та неуніфіковані позначення. Зокрема, символ логічного слідування у формулі (2.9) використано для позначення посилання на текстовий фрагмент, хоча процедура перевірки такого зв’язку потребує окремого визначення. Відомостей про власні наукові публікації за результатами роботи не наведено. Тема когнітивно-поведінкової самодопомоги є актуальною, особливо в контексті зростання потреби в доступній психологічній підтримці. Але варто чіткіше обґрунтувати межі застосування системи та її відмінність від звичайних чат-ботів психологічної підтримки. У роботі бажано чіткіше показати, які компоненти є принципово новими, а які є адаптацією відомих підходів до мультиагентних систем і CBT. Архітектура ПЗ описана досить детально, водночас методологія експериментального оцінювання потребує більшої конкретики. Варто повніше описати вибірку користувачів або тестових діалогів, критерії оцінювання, базові системи для порівняння та процедуру перевірки відтворюваності результатів. Доцільно більш глибоко обґрунтувати коректність цих моделей, їхні припущення, обмеження та зв’язок із класичними положеннями когнітивно-поведінкової терапії. Платформа має високий практичний потенціал як інструмент первинної самодопомоги. Однак для реального застосування необхідні ширша валідація безпеки, аналіз помилкових або потенційно шкідливих відповідей, а також чіткі рекомендації щодо випадків, коли користувача слід скеровувати до фахівця. Перелік джерел охоплює основні напрями, але його доцільно розширити новітніми публікаціями щодо безпеки LLM, оцінювання мультиагентних систем, етичних аспектів цифрової психологічної підтримки. Слід чіткіше розмежувати власний внесок автора та використані готові LLM, бібліотеки, фреймворки, промптові шаблони й інструменти. Доцільно покращити наочність подання архітектури, результатів експериментів і прикладів діалогів. Наукові публікації в наведених матеріалах відсутні. У роботі описано програмну реалізацію мультиагентної платформи, включно з субагентами, модулем SafetyMiddleware та механізмом довгострокової пам’яті. Водночас доцільно детальніше описати умови розгортання, доступ до коду, залежності та можливість відтворення системи. Для практичного застосування важливі кількісні метрики безпеки виявлення ризикованих станів і порівняння з базовими діалоговими системами. Оптимізаційний функціонал і функція втрат слугують лише декларативним описом вимог до промпту, а не функцією навчання моделей. Замала навчальна вибірка. Низка теоретичних формул має декоративний характер. Немає документального підтвердження. Відсутні наукові публікації. Слід зазначити: формулювання «вперше» та «формальні гарантії» видаються надто сильними: структурна властивість графа обробки не означає, що небезпечна ситуація завжди буде розпізнана; значна частина формул розділу 2 описує бажані властивості, але в реалізації не застосовується, тож теоретична частина здебільшого має описовий характер; оцінювання проведено на невеликих наборах прикладів, підготовлених розробником; бракує незалежних експертів-психологів і порівняння з простішою архітектурою; для інструменту психологічної підтримки навіть поодинокі пропуски кризових повідомлень критичні, а вплив системи на користувачів поки не вивчався, тому до впровадження потрібна фахова апробація; апробації результатів у наукових виданнях немає.		

№ з/п	Характеристики та критерії оцінки рукопису наукової роботи	Рейтингова оцінка. Максимальна кількість балів (за 120-бальною шкалою)	Бали
10.1	Наявність розробленого програмного забезпечення, у т.ч. веб-сервісу, мобільного застосунку	10	8,8
10.2	Рівень отриманих практичних результатів наукової роботи	10	5,8
Сума балів			85,5

Загальний висновок рекомендується для захисту на ПНПК

(рекомендується, не рекомендується для захисту на науково-практичній конференції)

Рецензент *члени Конкурсної комісії напряму «Обробка природної мови»*

«06» листопада 2026 року