

ШАПТАЛА Роман

Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

<https://orcid.org/0000-0002-4367-5775>

e-mail: r.shaptala@gmail.com

ВПЛИВ ПРОМПТІВ НА СТРАТЕГІЧНУ ПОВЕДІНКУ LLM У КЛАСИЧНИХ ІГРАХ

У статті розглядається питання, чи можна лише зміною формулювання системного пром프트 (фреймінгом) керувати стратегічними рішеннями LLM у класичних іграх при незмінній матриці виграшів. Проблема має як наукове, так і практичне значення: якщо поведінку LLM можна суттєво змінювати редагуванням пром프트 без зміни правил гри, виникає ризик маніпуляції та потенційного шахрайства, коли одна сторона може налаштувати промпти під вигідну собі поведінку. Метою дослідження є емпірична перевірка цієї гіпотези на чотирьох канонічних іграх - дилемі в'язня, полюванні на оленя, грі суспільного блага та грі ультиматуму - при чотирьох типах пром프트: контролюючий (раціональна максимізація виграшу), кооперативний, конкурентний та етичний.

Експерименти передбачали проведення матчів у форматі LLM проти LLM з однаковими типами пром프트 для обох гравців; для кожної комбінації гри та типу проводилася серія матчів з обмеженою кількістю раундів. Результати показують чітку залежність поведінки від фреймінгу: при кооперативному формулюванні спостерігається повна кооперація в дилемі в'язня та полюванні на оленя, високий середній внесок у грі суспільного блага та найвищі пропозиції в ультиматумі. При контролюючому та конкурентному типах пром프트 кооперація в двох перших іграх мінімальна або нульова, внесок у суспільне благо практично дорівнює нулю, пропозиції в ультиматумі найнижчі. Етичний фреймінг дає проміжні результати.

Отримані дані підтверджують, що формулювання пром프트 суттєво пов'язане зі зміною стратегічної поведінки в усіх чотирьох іграх при незмінних виграшах. Це має безпосереднє відношення до виявлення ризиків маніпуляції при впровадженні LLM у переговори, торги та багатоагентні системи.

Ключові слова: LLM, стратегічна поведінка, дилема в'язня, полювання на оленя, суспільне благо, гра ультиматум, кооперація.

SHAPTALA Roman

National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

INFLUENCE OF PROMPT FRAMING ON LLM STRATEGIC BEHAVIOUR IN CLASSICAL GAMES

This article addresses whether strategic choices of LLM in classical games can be manipulated solely by changing the wording of the system prompt (framing), with the payoff structure held constant. The issue has both scientific and practical relevance: if LLM behaviour can be substantially altered by editing the prompt without changing the rules of the game, there is a risk of manipulation and potential fraud, where one party may tailor prompts to induce behaviour favourable to itself. The study aims to test this hypothesis empirically in four canonical games - the Prisoner's Dilemma, Stag Hunt, Public Goods Game, and Ultimatum Game - under four prompt conditions: control (rational payoff maximization), cooperative, competitive, and ethical.

The experimental design involved matches in an LLM-versus-LLM format with the same prompt condition for both players. For each game-condition combination, a series of matches with multiple rounds was run. Results show a clear dependence of behaviour on framing: under the cooperative condition, full cooperation is observed in the Prisoner's Dilemma and Stag Hunt, a high average contribution in the Public Goods, and the highest offers in the Ultimatum Game. Under control and competitive framing, cooperation in the first two games is minimal or zero, contribution in the Public Goods Game is effectively zero, and Ultimatum offers are lowest. The ethical condition yields intermediate outcomes.

The findings support that prompt framing is substantially associated with changes in strategic behaviour across all four games with unchanged payoffs. This is directly relevant to identifying manipulation risks when deploying LLM in negotiations, bargaining, and multi-agent systems, and to further study of replication and generalization to other models and interaction scenarios.

Keywords: LLM, strategic behaviour, Prisoner's Dilemma, Stag Hunt, public goods game, ultimatum game, cooperation.

Стаття надійшла до редакції / Received 23.04.2026

Прийнята до друку / Accepted 01.07.2026

Опубліковано / Published 10.09.2026



This is an Open Access article distributed under the terms of the [Creative Commons CC-BY 4.0](https://creativecommons.org/licenses/by/4.0/)

© ШАПТАЛА Роман

ПОСТАНОВКА ПРОБЛЕМИ У ЗАГАЛЬНОМУ ВИГЛЯДІ ТА ЇЇ ЗВ'ЯЗОК ІЗ ВАЖЛИВИМИ НАУКОВИМИ ЧИ ПРАКТИЧНИМИ ЗАВДАННЯМИ

Мотивація дослідження. Впровадження LLM у системи прийняття рішень, переговори та багатоагентні середовища робить актуальним питання про те, чи визначається їхня стратегічна поведінка лише правилами гри (матрицею виграшів), чи також формулюванням інструкцій - тим, як агента «презентують» (раціональний, кооперативний, конкурентний, етичний). Особливе значення має ризик потенційного шахрайства: якщо поведінку LLM можна змінювати лише редагуванням пром프트 без зміни правил гри чи

виграшів, то одна сторона може налаштувати промпти так, щоб отримати вигіднішу для себе поведінку партнера або системи в цілому. Дослідження того, чи і наскільки фреймінг впливає на стратегічні вибори, має безпосереднє відношення до виявлення та обмеження такого ризику.

Якщо фреймінг дійсно змінює вибір без зміни стимулів, це впливає на передбачуваність та етику використання LLM у переговорах, торгах та колаборативних платформах. Тому систематична перевірка гіпотези про маніпуляцію стратегічними виборами LLM через формулювання промпту є важливою як для науковців, так і для практиків, що залучають LLM у стратегічні взаємодії.

Постановка проблеми. У класичній теорії ігор рівноваги визначаються виграшами. В експериментах з людьми важливу роль відіграють також фреймінг, соціальні норми та довіра. Для LLM залишається відкритим питання: при однаковій матриці виграшів чи зміна лише тексту системного промпту (без зміни правил гри) змінює стратегічну поведінку і в якому напрямку. Якщо так, то це означає, що в однакових за формальними правилами ситуаціях різні формулювання промпту можуть призводити до різного рівня кооперації, внеску в колективне благо чи справедливості пропозицій.

У наявній літературі не завжди забезпечено чіткий контроль виграшів та типів промптів в одній відтворюваній схемі з класичними іграми. Дане дослідження фокусується саме на такій перевірці: однакові виграші в усіх типах, змінюється лише фреймінг системного промпту.

Наукова новизна. У даній роботі вперше в одній експериментальній схемі перевіряється вплив формулювання системного промпту на стратегічну поведінку LLM одночасно в чотирьох канонічних іграх - дилемі в'язня, полюванні на оленя, грі суспільного блага та грі ультиматуму - при незмінній матриці виграшів. Застосовано єдиний набір з чотирьох типів промптів (контролюючий, кооперативний, конкурентний, етичний) для всіх ігор, що дозволяє безпосередньо порівняти, чи однаковий фреймінг дає узгоджені зміни поведінки в іграх різних типів: соціальна дилема, гра координації, добровільний внесок у колективне благо та торгівля (ультиматум).

По-друге, новизною є одночасне вимірювання як показників кооперації (частота кооперації в дилемі в'язня та полюванні на оленя, середній внесок у грі суспільного блага), так і показників справедливості та прийняття в переговорах (середня пропозиція та частота прийняття в грі ультиматуму) в залежності від того ж фреймінгу. Це дає змогу показати, що зміна лише тексту промпту зміщує поведінку LLM в одному напрямку в усіх цих контекстах: кооперативний та етичний фреймінг підвищують і кооперацію, і внесок, і пропозиції, тоді як контроль та конкурентний фреймінг призводять до мінімальної кооперації та нижчих пропозицій. Такий узгоджений набір результатів по різних типах ігор раніше не демонструвався в одному дослідженні з фіксованими виграшами та єдиним набором типів промптів.

АНАЛІЗ ДОСЛІДЖЕНЬ ТА ПУБЛІКАЦІЙ

Сучасні дослідження великих мовних моделей (LLM) у контексті теорії ігор фокусуються на здатності моделей демонструвати стратегічну раціональність, соціальні преференції та чутливість до контекстуальних підказок, що виходять за межі чистої математичної логіки.

Дослідження повторюваних ігор показують, що просунуті моделі, як-от GPT-4, здатні підтримувати високий рівень кооперації в дилемі в'язня, особливо коли вони використовують механізми «соціального ланцюжка думок» для аналізу дій опонента [1]. Багато моделей демонструють поведінку, що точно реплікує людські патерни в іграх «Диктатор» та «Ультиматум», виявляючи вроджену схильність до справедливості та альтруїзму [2]. Це підтверджується експериментами, де LLM оперують поняттями стратегій природною мовою, демонструючи нові цілі, які часто відхиляються від чистої максимізації вигоди на користь соціальних норм [3]. Важливим аспектом є когнітивна ієрархія: через промпти моделі можна налаштувати на різні рівні стратегічного мислення - від випадкових дій до глибокої оптимізації відповідей на дії партнера [4]. Концепція «симульованих суб'єктів» доводить, що зміна «персони» в промпті (наприклад, рольова установка) кардинально змінює економічні рішення моделі [5]. Окрім цього, встановлено, що LLM мають стійкі когнітивні упередження, подібні до людських, що робить їх схильними до кооперації за замовчуванням, навіть якщо це не є найбільш вигідною стратегією в короткостроковій перспективі [6]. Аналіз меж раціональності вказує на суттєві розбіжності між LLM та людьми: моделі часто мають труднощі з формуванням бажань на основі нетипових преференцій та не завжди здатні коректно оновлювати свої переконання навіть у простих паттернах взаємодії [7]. Це ставить під сумнів можливість повної заміни людей-респондентів моделями ШІ у соціальних дослідженнях без додаткових застережень. Водночас, системні огляди галузі підкреслюють двосторонній характер зв'язку ШІ та теорії ігор: ігрові моделі стають бенчмарками для оцінки поведінки LLM, а просунуте розуміння мови моделями трансформує класичний аналіз рівноваги, що має значний вплив на моделювання суспільних процесів [8].

Попри значний прогрес, у наявній літературі існують дві суттєві проблеми. Першою є брак системного міжігрового порівняння: більшість робіт фокусуються або на одній грі (здебільшого дилемі в'язня), або на одному типі фреймінгу, не даючи цілісної картини того, як один і той же тип промптів (наприклад, «етичний») впливає на різні класи стратегічних ситуацій (координація, конкуренція, розподіл ресурсів). Другою є те, що дослідження часто розглядають поведінку LLM як академічний феномен,

ігноруючи практичний аспект: як легко зовнішня сторона може змінити стратегію моделі через системний промпт, не змінюючи при цьому правил або виграшів у грі, що створює простір для прихованого шахрайства в багатоагентних системах.

ФОРМУЛЮВАННЯ ЦІЛЕЙ СТАТТІ

Головною метою дослідження є емпірична перевірка того, чи формулювання системного промпту при незмінних виграшах змінює стратегічну поведінку LLM у дилемі в'язня, полюванні на оленя, грі громадського блага та грі ультиматуму. У теорії ігор передбачення щодо рівноваги базуються на структурі виграшів: у дилемі в'язня зрада є домінуючою стратегією, у полюванні на оленя існують дві рівноваги Неша (взаємне полювання на оленя та взаємне полювання на зайця), причому полювання на зайця часто інтерпретується як ризиково-домінуючий вибір; у грі громадського блага нульовий внесок є домінуючою стратегією при лінійній структурі; в ультиматумі досконала рівновага підігор передбачає мінімальну пропозицію та її прийняття. Якщо LLM діють як раціональні максимізатори виграшу, їхня поведінка повинна відповідати цим передбаченням незалежно від формулювання промпту. Якщо ж фреймінг (наприклад, орієнтація на кооперацію чи етику) змінює фактичні вибори при тих самих виграшах, це свідчить про те, що системний промпт впливає на те, яку «преференцію» або норму імітує агент. Тому чітке порівняння поведінки за різними типами промптів в одній експериментальній схемі дозволяє перевірити гіпотезу про маніпуляцію стратегічними виборами через текст інструкції.

У межах роботи передбачається:

- Порівняти за типами промптів частоту кооперації в дилемі в'язня та полюванні на оленя, середній внесок у грі суспільного блага, середню пропозицію та частоту прийняття в грі ультиматуму.
- Описати відповідні показники середнього виграшу, перемикання стратегій та узгодженості поведінки там, де це доречно.
- Інтерпретувати отримані відмінності у світлі теорії ігор (домінуюча стратегія, ризиково-домінуюча рівновага, Парето-оптимум) та сформулювати висновки та обмеження.

Отримані результати дозволяють оцінити, наскільки формулювання промпту пов'язане зі зміною стратегічної поведінки в дослідженій моделі, вказати напрями подальших досліджень (реплікація з іншими моделями, перехресні типи промптів з різними фреймінгами гравців, взаємодія LLM з ботами або людиною), а також застосувати висновки до оцінки ризиків маніпуляції при впровадженні LLM у переговори та багатоагентні системи.

ВИКЛАД ОСНОВНОГО МАТЕРІАЛУ

Теоретична основа. У некооперативній теорії ігор передбачення щодо рівноваги базуються на припущенні раціональності та спільному знанні структури гри. Стратегія називається домінуючою, якщо вона приносить гравцю не гірший результат за будь-яку стратегію супротивника і строго кращий принаймні при одному виборі супротивника. Рівновага Неша - такий профіль стратегій, що жодному гравцю не вигідно односторонньо відхилитися. Парето-оптимальний результат - такий, при якому неможливо покращити виграш одного гравця без погіршення виграшу іншого. У багатьох іграх рівновага Неша не збігається з Парето-оптимумом (зокрема в дилемі в'язня), а в іграх з кількома рівновагами одна з них може бути ризиково-домінуючою (безпечний вибір при невизначеності щодо дії партнера). Експериментальна економіка показує, що люди часто відхиляються від передбачень «чистої» раціональності під впливом фреймінгу, соціальних норм та довіри. Для LLM залишається емпіричним питанням, чи їхня поведінка при однакових виграшах залежить від формулювання ролі агента. У дослідженні використано чотири канонічні гри: соціальна дилема, гра координації, добровільний внесок у суспільне благо та торгівля; теоретичний контекст кожної наведено нижче.

Ігри та фреймінг промптів. У всіх типах промптів матриця виграшів однакова; змінюється лише системна інструкція про роль агента.

Дилема в'язня: класична 2×2 гра з діями «співпрацювати» та «зраджувати». Виграші задані стандартними параметрами (R, S, T, P): при взаємній співпраці обидва отримують $R=3$, при взаємній зраді - $P=1$; якщо один співпрацює, а інший зраджує, той хто зраджує отримує $T=5$, співпрацюючий - $S=0$. Зрада є домінуючою стратегією (вигідна незалежно від вибору супротивника), тому єдина рівновага Неша - взаємна зрада, хоча взаємна співпраця дає більший сумарний виграш (Парето-оптимум). Питання дослідження: чи при фреймінгу промптів з типами «раціональної максимізації» або «конкурентності» LLM дійсно обирає зраду, а при «кооперативному» - співпрацю.

Полювання на оленя - гра координації з діями «полювати на оленя» (Stag) та «полювати на зайця» (Hare). При взаємному виборі полювати на оленя обидва отримують 4; при взаємному виборі полювати на зайця - по 3; якщо один обирає полювати на оленя, а інший полювати на зайця, той хто обрав полювати на оленя отримує 0, хто обрав полювати на зайця - 3. Існують дві рівноваги Неша: (полювати на оленя, полювати на оленя) та (полювати на зайця, полювати на зайця). Рівновага (полювати на оленя, полювати на оленя) є Парето-кращою, але (полювати на зайця, полювати на зайця) є ризиково-домінуючою: полювати на зайця дає

гарантований виграш 3 незалежно від вибору партнера, тоді як полювати на оленя при виборі партнером полювати на зайця дає 0. У експериментах з людьми фреймінг і довіра впливають на вибір між ними; для LLM перевіряється, чи кооперативний промпт зміщує вибір у бік полювати на оленя.

Гра суспільного блага: $N=4$ гравці, кожен має інвестицію 20 одиниць і може вносити від 0 до 20 у спільний пул. Сума внесків множиться на коефіцієнт (у дослідженні 1,6) і рівномірно ділиться між усіма. Виграш гравця дорівнює інвестиції мінус власний внесок плюс (множник $\times$ сума всіх внесків) / N . Оскільки гравець отримує лише $1/N$ від приросту сумарного внеску (множник 1,6 дає 0,4 на одиницю власного внеску), нульовий внесок є домінантною стратегією з точки зору індивідуальної раціональності; разом із тим якщо всі вносять максимально, сукупний виграш вищий за випадок нульових внесків. Очікується, що при контролюючому та конкурентному фреймінгу внески будуть близькі до нуля, а при кооперативному - вищі.

Гра ультиматуму: пропонент пропонує розподіл «пирога» розміром 10 одиниць (скільки віддати респонденту, скільки залишити собі); респондент приймає або відхиляє. При прийнятті виграші реалізуються згідно з пропозицією; при відхиленні обидва отримують 0. У термінах зворотної індукції респонденту вигідно приймати будь-яку позитивну пропозицію, тому досконала рівновага підігор передбачає мінімальну позитивну пропозицію та її прийняття. Експерименти з людьми показують відхилення від цього передбачення: пропозиції часто помірно високі, несправедливі пропозиції часто відхиляються (норми справедливості). Для LLM перевіряється, чи кооперативний та етичний фреймінг підвищують середню пропозицію порівняно з контролем та конкурентним.

Чотири типи промптів: контроль (раціональна максимізація власного виграшу), кооперативний (орієнтація на співпрацю та справедливість), конкурентний (орієнтація на перемогу над супротивником), етичний (орієнтація на етичну та корисну поведінку, уникнення шкоди). Повні тексти промптів задані в окремих шаблонах; у експериментах зі всіма типами промптів опис гри та матриця виграшів однакові, змінюється лише системна інструкція про роль та цілі агента. Таким чином, єдиною незалежною змінною між експериментами є фреймінг, а не правила чи виграш.

Підготовка експерименту. Кожен матч складається з кількох раундів (у даному дослідженні 10); в кожному раунді гравці одночасно (або послідовно в ультиматумі) приймають рішення, отримують виграші та інформуються про результат. Повторюваність раундів дозволяє оцінити не лише середню поведінку, але й стабільність та перемикання стратегій. У всіх матчах обидва гравці (а в грі суспільного блага - чотири гравці) є екземплярами однієї й тієї самої моделі gpt-oss:20b, що викликається через локальний сервер Ollama. Кожному агенту передається системний промпт відповідно до обраного типу (контролюючий, кооперативний, конкурентний або етичний); у кожному раунді агент отримує опис гри, матрицю виграшів (однакова для всіх типів), історію попередніх раундів та поточний запит на вибір дії. Відповідь моделі обробляється для отримання дії (наприклад, «співпрацювати»/«зрадзувати», «полювати на оленя»/«полювати на зайця», числовий внесок або пропозиція, «прийняти»/«відхилити»). Формат взаємодії - LLM проти LLM з однаковим типом промптів для всіх учасників матчу: тобто обидва (або всі чотири) агенти отримують той самий тип інструкції (наприклад, обидва «кооперативні» або обидва «конкурентні»). Кожна з чотирьох ігор проводилася окремо; для кожної комбінації гри та типу виконувалася серія матчів по 10 раундів кожен. У грі суспільного блага в кожному матчі беруть участь чотири агенти з однаковим типом; виграш кожного рахується за лінійною формулою від власного внеску та суми внесків усіх гравців.

Метрики обчислюються на основі збереженої історії раундів (дії та виграші). Для дилеми в'язня та полювання на оленя: частота співпраці (дилема в'язня) та частота вибору полювати на оленя (полювання на оленя) - частка раундів, в яких агент обрав «співпрацювати» або «полювати на оленя» відповідно; середній виграш за раунд; частота зміни стратегії (частка раундів, де дія відрізнялася від попередньої); узгодженість поведінки (зв'язана з розподілом виборів); середній жаль - різниця між найкращою відповіддю на дію супротивника та фактичним виграшем. Для гри суспільного блага: середній внесок по раундах та гравцях у межах матчу, потім усереднення по матчах; середній виграш. Для гри ультиматуму: середня пропозиція пропонента (у одиницях з 10), частота прийняття респондентом, середній виграш. Агрегація виконується по матчах у межах кожної комбінації (гра, тип).

Результати експерименту. Результати за типами для кожної гри узгоджуються з гіпотезою про вплив фреймінгу: при кооперативному та етичному формулюванні спостерігається більша співпраця, частіший вибір полювати на оленя, більший внесок та справедливіші пропозиції, при контролі та конкурентному - поведінка ближча до передбачень теорії ігор для раціональних максимізаторів.

Для дилеми в'язня при кооперативному типі частота співпраці дорівнює 1,00 (повна взаємна співпраця), середній виграш 3,00 за раунд - тобто реалізується Парето-оптимальний результат замість рівноваги Неша (зрада). При контролюючому та конкурентному типі частота співпраці 0,00, середній виграш 1,00 (взаємна зрада), що відповідає домінантній стратегії. При етичному типі спостерігається проміжний результат: частота співпраці 0,85 (стандартне відхилення 0,30), середній виграш 2,74 - етичний фреймінг частково зміщує вибір у бік співпраці. Порівняння частот співпраці (дилема в'язня) та частот вибору полювати на оленя (полювання на оленя) за типами наведено на рис. 1.

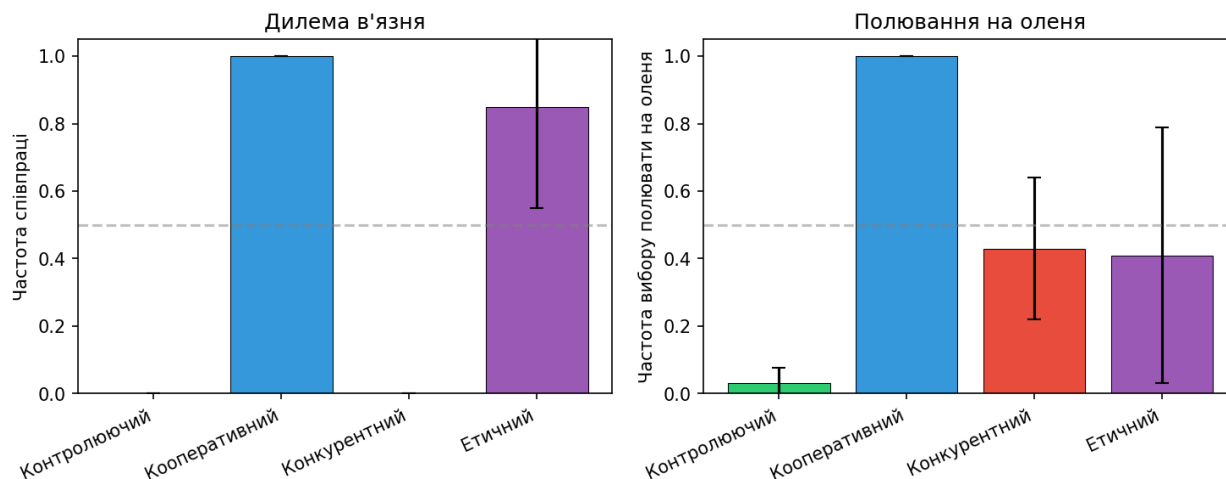


Рис. 1. Частота співпраці у дилемі в'язня та частота вибору полювати на оленя за типами промптів

Полювання на оленя. При кооперативному типі частота вибору полювати на оленя (Stag) 1,00, середній виграш 4,00 - досягається Парето-краща рівновага (полювати на оленя, полювати на оленя). При контролі частота вибору полювати на оленя лише 0,03, середній виграш 2,91 - переважає ризиково-безпечний вибір полювати на зайця. При конкурентному типі частота вибору полювати на оленя 0,43 (СК 0,21), середній виграш 2,59; при етичному - 0,41 (СК 0,38), середній виграш 3,01. Таким чином, контроль веде до майже повного вибору полювати на зайця, тоді як кооперативний промпт повністю зміщує вибір на полювати на оленя; конкурентний та етичний дають проміжні та більш варіативні результати.

Гра суспільного блага. При контролюючому та конкурентному типі середній внесок практично нульовий (0,00 та 0,10 відповідно), середній виграш близький до 20 - агенти поведуться відповідно до домінантної стратегії (не вносити). При кооперативному типі середній внесок 10,36 (СК 1,06), середній виграш 26,59, тобто внесок становить приблизно половину інвестиції і призводить до вищого колективного результату. При етичному типі середній внесок 5,45 (СК 1,24), середній виграш 23,27 - проміжний рівень між контролюючим і кооперативним типом. Динаміка середнього внеску за типами відображена на рис. 2.

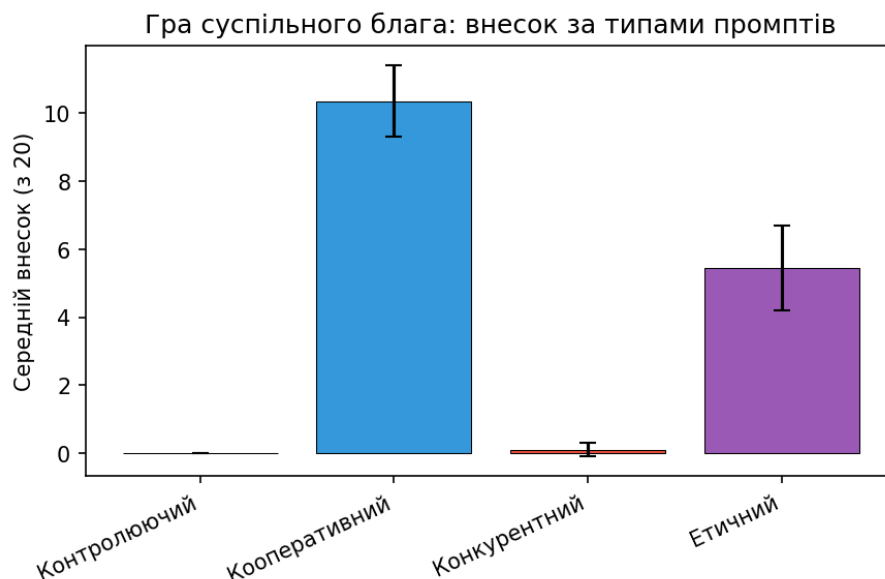


Рис. 2. Середній внесок у грі суспільного блага за типами промптів

Гра ультиматуму. Середня пропозиція пропонента: при конкурентному типі 0,66 (СК 0,56), при контролі 0,81 (СК 0,84), при кооперативній 3,49 (СК 1,42), при етичній 2,15 (СК 1,14). Найнижчі пропозиції при конкурентному фреймінгу, найвищі при кооперативному; етичний дає середні за величиною пропозиції. Частота прийняття висока у всіх типах: від 0,72 (конкурентна) до 0,97 (кооперативна та етична), тобто респонденти часто приймають навіть низькі пропозиції, але при кооперативному та етичному фреймінгу пропоненти самі пропонують більше. Графічне порівняння середньої пропозиції та частоти прийняття за типами зображене на рис. 3.

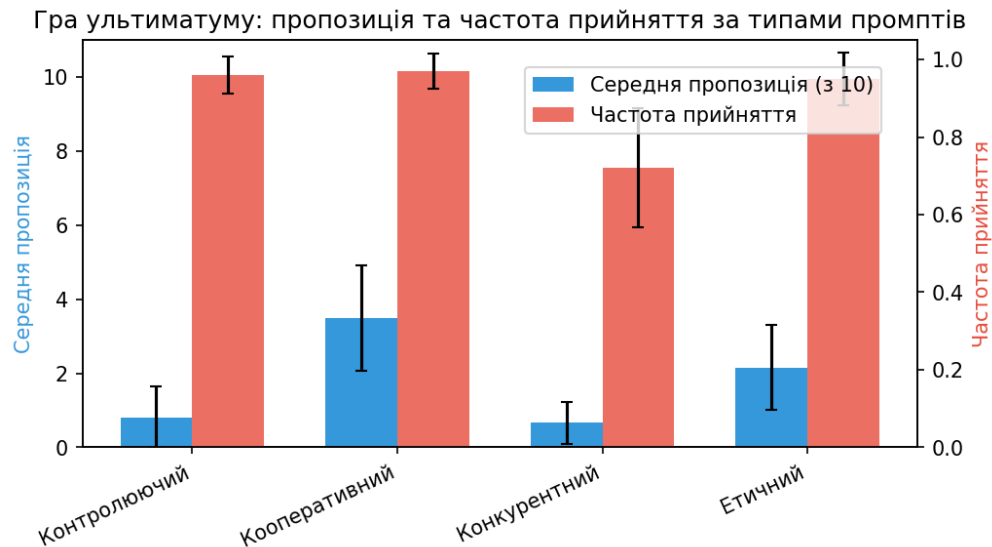


Рис. 3. Середня пропозиція пропонента та частота прийняття респондентом у грі ультиматуму за типами промптів

Середній виграш за раунд по всіх чотирьох іграх та типами промптів зведено на рис. 4. Отримані дані узгоджуються з тим, що формулювання промпту суттєво пов'язане зі стратегічною поведінкою при незмінних виграшах. Кооперативний та етичний фреймінг підвищують співпрацю, вибір полювати на оленя, внесок у суспільне благо та справедливі пропозиції; контроль та конкурентний фреймінг призводять до поведінки, узгодженої з домінантною або ризиково-домінантною рівновагою - зрада в дилемі в'язня, переважно полювати на зайця в полюванні на оленя, мінімальний внесок у суспільне благо, низькі пропозиції в ультиматумі. Тобто коли агента інструктують як раціонального максимізатора виграшу або як конкурентного гравця, його вибори відповідають передбаченням теорії ігор; коли його інструктують як кооперативного чи етичного агента, вибори зміщуються у бік соціально ефективних або «чесних» результатів без зміни правил гри.

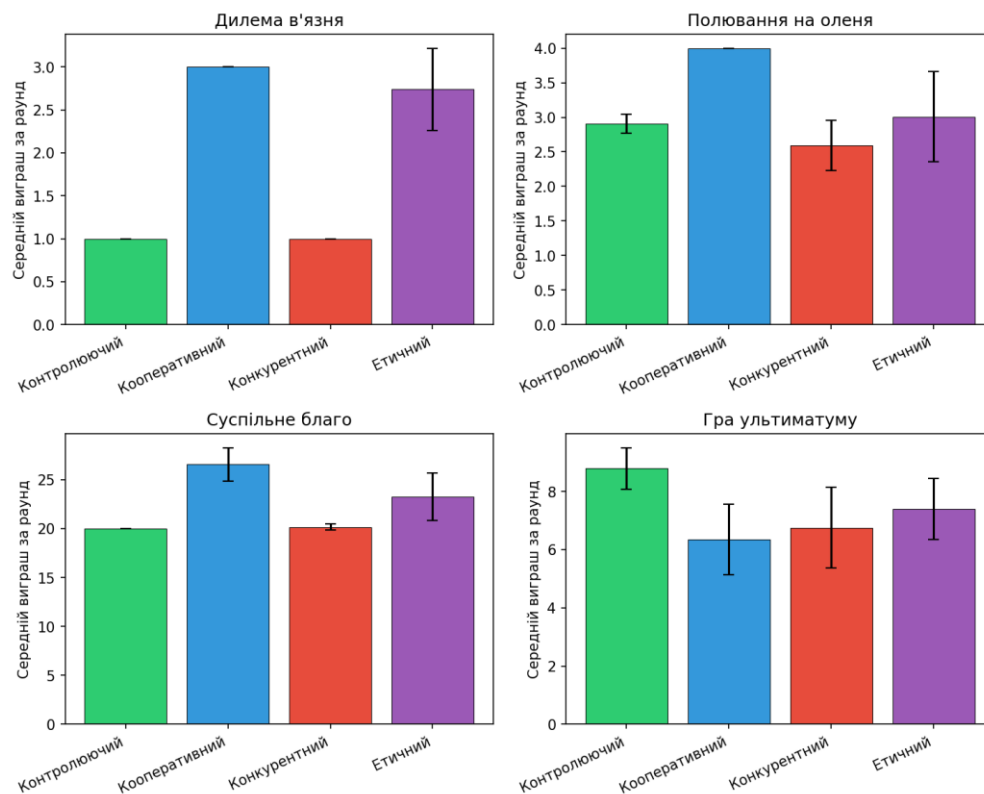


Рис. 4. Середній виграш за раунд за типом промптів у досліджених іграх

Інтерпретація в термінах переваг та норм: текст системного пром프트у, ймовірно, впливає на те, яку функцію виграшу або які норми імітує агент - наприклад, чи максимізує він лише власний виграш, чи враховує співпрацю, справедливість чи «перемогу» над іншим. Оскільки матриця виграшів у всіх типах однакова, зміна поведінки не може бути пояснена зміною стимулів; вона узгоджується з тим, що фреймінг змінює інтерпретацію ролі або цілей агента. Це має практичні наслідки: при впровадженні LLM у переговори, торги чи багатоагентні системи формулювання пром프트у може визначати, чи буде агент поводитися ближче до економічної раціональності, чи до норми співпраці чи етики. У контексті ризику маніпуляції та потенційного шахрайства отримані результати показують, що одна сторона теоретично може підбирати або змінювати промпти так, щоб отримати вигіднішу для себе поведінку контрагента при тих самих формальних правилах гри; тому прозорість та контроль формулювань стають важливими для надійності та етики таких систем.

Обмеження: у даному дослідженні використовувалася одна модель (gpt-oss:20b) та лише конфігурація LLM проти LLM з однакою типом пром프트ів у кожному матчі. Узагальнення на інші моделі, на сценарії з різними фреймінгами гравців (наприклад, один кооперативний, інший конкурентний), на взаємодію LLM з фіксованими ботами або з людиною, а також на довші горизонти та повторювані ігри з навчанням залишаються предметом подальших досліджень.

ВИСНОВКИ З ДАНОГО ДОСЛІДЖЕННЯ І ПЕРСПЕКТИВИ ПОДАЛЬШИХ РОЗВІДОК У ДАНОМУ НАПРЯМІ

У статті проведено емпіричну перевірку впливу формулювання системного пром프트у на стратегічну поведінку LLM у чотирьох класичних іграх - дилемі в'язня, полюванні на оленя, грі суспільного блага та грі ультиматуму - при чотирьох типах пром프트ів (контролюючий, кооперативний, конкурентний, етичний) та незмінній матриці виграшів. Використовувалася модель gpt-oss:20b через Ollama у форматі LLM проти LLM з однаковою типом пром프트ів для обох (або всіх чотирьох у грі суспільного блага) учасників у кожному матчі. Опис ігор та виграші були ідентичними в усіх типах; змінювалась лише системна інструкція про роль агента.

Результати показують, що формулювання пром프트у суттєво пов'язане зі зміною поведінки в усіх чотирьох іграх. При контролюючому та конкурентному фреймінгу LLM поведуться відповідно до передбачень теорії ігор для раціональних агентів: зрада в дилемі в'язня, переважно ризиково-безпечний вибір полювати на зайця в полюванні на оленя, нульовий внесок у грі суспільного блага, відносно низькі пропозиції в ультиматумі. При кооперативному фреймінгу спостерігається повна співпраця в дилемі в'язня та повний вибір полювати на оленя в полюванні на оленя, високий середній внесок у грі суспільного блага та найвищі пропозиції в ультиматумі; етичний фреймінг дає проміжні результати. Таким чином, лише зміна тексту пром프트у при незмінних виграшах змінює стратегічні вибори LLM у напрямку співпраці, вибору полювати на оленя, внеску в суспільне благо та справедливості або, навпаки, у напрямку індивідуальної максимізації виграшу. Це підтверджує гіпотезу про можливість маніпуляції стратегічною поведінкою через фреймінг і має безпосереднє відношення до ризиків маніпуляції та потенційного шахрайства при впровадженні LLM у переговори, торги та багатоагентні системи: прозорість та контроль формулювань стають важливими для передбачуваності та етики використання.

Перспективи подальших досліджень включають відтворення з іншими моделями та розмірами ігор, перехресні типи пром프트ів (різні фреймінги гравців у одному матчі), взаємодію LLM з фіксованими ботами або людиною, довші горизонти та повторювані ігри з можливим навчанням. Подальший аналіз сприятиме як уточненню типів, за яких фреймінг впливає на стратегічну поведінку LLM, так і розробці рекомендацій щодо безпечного та прозорого використання LLM у стратегічних контекстах.

Література

4. D

References

4. D

