DOI: https://doi.org/10.26642/ten-2025-2(96)-136-146
УДК 623.61

С.О. Дупелич, к.т.н., доц.
В.В. Дзюбенко, ад'юнкт
Житомирський військовий інститут імені С.П. Корольова

Аналіз можливостей великих та малих мовних моделей (LLM/SLM) для оптимізації процесів прийняття рішень у системі радіозв'язку

Сучасна система радіозв'язку функціонує в умовах екстремальної складності, що характеризуються динамічно мінливою обстановкою, високою щільністю інформаційних потоків та активною протидією противника. Ефективність управління в таких умовах визначається здатністю не лише обробляти значні масиви даних, але й приймати оптимальні рішення в режимі реального часу. Ключовими критеріями успішності є своєчасність реакції, достовірність аналізу та прихованість дій. Інтенсивний розвиток великих мовних моделей (англ. – Large Language Models (LLM)) та малих мовних моделей (англ. – Small Language Models (SLM)), що демонструють унікальні здібності до семантичної інтерпретації, логічного мислення та генерації рекомендацій, відкриває принципово нові горизонти для автоматизації когнітивних завдань та створення інтелектуальних систем підтримки прийняття рішень (СППР) нового покоління. Ця робота є першим комплексним узагальненням, що системно розглядає застосування мовних моделей штучного інтелекту саме в цьому специфічному та критично важливому контексті.

У роботі проведено компаративний аналіз ключових архітектур мовних моделей, що домінують на ринку, висвітлюючи дихотомію між пропрієтарними системами (сімейства GPT від OpenAI, Gemini від Google, Claude від Anthropic) та моделями з відкритим вихідним кодом (Llama від Meta, Mistral від Mistral AI, Phi-3 від Microsoft). Обґрунтовано, що для завдань, які вимагають гнучкості, можливості локального розгортання та швидкого прототипування в умовах високої прихованості, особливий інтерес становлять компактні мовні моделі (SLM). Зокрема, детально розглянуто архітектуру Microsoft Phi-3-mini, чия висока продуктивність при малому розмірі (3,8 мільярда параметрів) досягається завдяки інноваційному підходу до навчання на високоякісних синтетичних даних, що ставить якість навчального масиву вище за його обсяг.

Ключові слова: *великі мовні моделі; малі мовні моделі; LLM; SLM; Phi-3-mini; системи підтримки прийняття рішень; радіозв'язок.*

Актуальність теми. Актуальність дослідження визначається критичними умовами, в яких функціонує сучасна система радіозв'язку. Вона стикається з екстремальною складністю, що включає динамічну зміну обстановки, перевантажені інформаційні потоки та активну протидію з боку противника, що своєю чергою висуває жорсткі вимоги до оперативності. Ефективність управління в такому середовищі залежить не стільки від здатності обробляти великі масиви даних, скільки від спроможності приймати оптимальні рішення в реальному часі. Ключовими факторами успіху стають своєчасність реакції, достовірність аналізу та прихованість дій. Стрімкий розвиток великих (LLM) та малих (SLM) мовних моделей, що демонструють унікальні здібності у семантичній інтерпретації, логічному мисленні та формуванні рекомендацій, відкриває нові перспективи для автоматизації когнітивних процесів. Це створює технологічне підґрунтя для розробки інтелектуальних систем підтримки прийняття рішень (СППР) нового покоління. Відтак адаптація та інтеграція цих потужних інструментів ШІ для вирішення специфічних і критично важливих завдань радіозв'язку є важливою науково-прикладною проблемою. Ця робота є першим комплексним узагальненням, що системно розглядає цей потенціал, аналізуючи застосування мовних моделей саме в цьому специфічному контексті.

Аналіз останніх досліджень та публікацій, на які спираються автори. Дослідження ґрунтується на фундаментальних працях, що описують ключові архітектури сучасного штучного інтелекту. Зокрема, основоположною є робота A.Vaswani та ін. [11, 24, 31], що ввела архітектуру Transformer. Задачі глибокого розуміння мови та контексту, що стали базовими для багатьох систем, викладені у праці, що описує архітектуру BERT [20]. Значний пласт публікацій присвячений аналізу можливостей великих пропрієтарних моделей. Сюди належать технічні звіти та описи, що деталізують сімейства моделей GPT від OpenAI [1–3], мультимодальну архітектуру Gemini від Google [4–6] та орієнтовані на безпеку моделі Claude від Anthropic [7]. Також аналізуються специфічні корпоративні платформи, призначені для бізнес-завдань, зокрема розробки компанії «Cohere» [8, 9]. Паралельно досліджуються ключові моделі з відкритим вихідним кодом. Детально розглядаються роботи, присвячені сімейству Llama від Meta [12–14], що стало каталізатором для open-source спільноти, а також обчислювально-ефективні архітектури Mistral та Mixtral [15, 16] і потужна модель Falcon від TII [17–19]. Особлива увага в статті приділяється концепції та перевагам малих мовних моделей (SLM), що спирається на технічні звіти Microsoft, присвячені

сімейству Phi-3 [21–26]. Теоретичною основою для їх високої ефективності є інноваційний підхід до навчання на високоякісних синтетичних даних, принципи якого викладені у праці «TinyStories» [27].

Метою статті є проведення комплексного аналізу можливостей великих та малих мовних моделей для оптимізації процесів прийняття рішень у системі радіозв'язку.

Викладення основного матеріалу. Сучасна індустрія штучного інтелекту характеризується інтенсифікацією досліджень у галузі LLM. Ці моделі є класом архітектур глибинного навчання, тренованих на великомасштабних текстових корпусах та інших масивах даних. Їх функціональність охоплює широкий спектр завдань, зокрема, генерацію, семантичну інтерпретацію, переклад та автоматизоване реферування інформації людською мовою.

Поточний ринок LLM демонструє виражену дихотомію, яка полягає у конкурентній взаємодії між двома ключовими парадигмами розробки. З одного боку, це пропрієтарні системи, що створюються та контролюються великими технологічними корпораціями. З іншого – моделі з відкритим вихідним кодом (open-source), які розвиваються та підтримуються глобальною спільнотою розробників. Кожна з цих парадигм має свої унікальні переваги та недоліки, що визначають доцільність їх використання для конкретних завдань та бізнес-моделей.

1. Пропрієтарні моделі закритого типу

Ця категорія LLM, розробка та контроль над якими здійснюються великими технологічними корпораціями та приватними комерційними організаціями, зазвичай демонструє найвищі показники продуктивності. Їхня перевага є прямим наслідком колосальних інвестицій в обчислювальні ресурси та передову дослідницьку діяльність.

Доступ до функціональності цих систем, як правило, є закритим і надається на комерційній основі через програмні інтерфейси додатків (API) або шляхом інтеграції в комерційні продукти. Таким чином, користувачі взаємодіють з моделлю як із сервісом, не маючи доступу до її вихідного коду чи архітектурних деталей.

Сімейство GPT (OpenAI). Сімейство моделей Generative Pre-trained Transformer (GPT), розроблене дослідницькою лабораторією OpenAI, відіграє центральну роль у прискоренні розвитку та популяризації технологій генеративного штучного інтелекту. Ці архітектури демонструють високу продуктивність у вирішенні комплексних завдань, пов'язаних з обробкою та генерацією людської мови.

Сучасні ітерації, зокрема GPT-4 та GPT-4o, фактично стали ринковими стандартами, демонструючи найвищу універсальність та передові мультимодальні можливості. Їхня еволюція є яскравим підтвердженням ефективності законів масштабування (scaling laws), згідно з якими послідовне збільшення кількості параметрів та обсягу навчальних даних веде до якісного зростання когнітивних здібностей моделі.

Зокрема, ітерації, такі як GPT-3.5 та GPT-4, слугували фундаментальною технологією для створення та функціонування широко відомого діалогового агента ChatGPT. Поточна флагманська модель, GPT-4o, є значним кроком уперед, оскільки реалізована як нативна мультимодальна архітектура. Це дозволяє їй здійснювати інтегровану та одночасну обробку гетерогенних потоків даних, що включають текст, аудіо та візуальну інформацію, що суттєво розширює спектр її потенційних застосувань [1–3].

Архітектура GPT є визначальною для сучасної епохи генеративного ШІ, виконуючи роль технологічного драйвера та ринкового стандарту. Її еволюція демонструє ефективність законів масштабування (scaling laws), де збільшення кількості параметрів та обсягу навчальних даних призводить до якісного зростання когнітивних можливостей. Ключовим внеском GPT є не лише досягнення найвищих показників у широкому спектрі завдань, але й успішна комерціалізація та популяризація LLM через продукти на кшталт ChatGPT. Модель GPT-4o закріплює лідерство, інтегруючи нативну мультимодальність як новий стандарт для флагманських універсальних моделей.

Архітектура Gemini (Google). Серія моделей Gemini є мультимодальною архітектурою та стратегічною розробкою Google, спроєктованою для інтегрованої обробки різнорідних типів даних, таких як текст, програмний код, зображення та відео. Еволюційно цим розробкам передували проєкти LaMDA та PaLM 2, а сама лінійка глибоко інтегрована у власну екосистему продуктів Google.

Дане сімейство включає градацію моделей за потужністю для вирішення різних завдань:

1. **Gemini Ultra** – флагманська модель для найбільш складних обчислювальних завдань;
2. **Gemini Pro** – універсальне рішення, інтегроване, зокрема, у чат-бот Gemini;
3. **Gemini Nano** – компактна версія, оптимізована для виконання на периферійних пристроях.

Ключовою особливістю сучасної версії Gemini 1.5 Pro є величезне контекстне вікно, що сягає одного мільйона токенів. Це дозволяє моделі ефективно аналізувати великомасштабні масиви даних, враховуючи довгі документи, програмний код або тривалі відеозаписи, в межах одного запиту [4–6].

Gemini 1.5 Pro поточна версія, що вирізняється значно розширеним контекстним вікном (до одного мільйона токенів), що дозволяє аналізувати великомасштабні масиви інформації.

Gemini 1.0 (Ultra, Pro, Nano): перше покоління архітектури, представлене у трьох конфігураціях, що масштабуються для різних обчислювальних платформ та завдань.

Gemini є стратегічною відповіддю Google, що має на меті не просто конкурувати, а й технологічно перевершити наявні архітектури. Фундаментальним диференціатором є її нативно мультимодальна основа, що дозволяє проводити інтегровану обробку гетерогенних даних без потреби у поєднанні окремих компонентів. Розширене контекстне вікно в моделі Gemini 1.5 Pro свідчить про фокус на застосунках, що вимагають аналізу великомасштабних масивів інформації. Таким чином, Gemini позиціонується як глибоко інтегрована в екосистему Google, мультимодальна за своєю суттю архітектура для наступного покоління ШІ-додатків.

Моделі Claude (Anthropic). Моделі Claude, розроблені компанією «Anthropic», позиціонуються на ринку з особливим акцентом на підвищенні стандартів безпеки та етичності функціонування систем штучного інтелекту. Цей фокус на безпеці та етичній узгодженості є ключовим диференціатором, що робить дані моделі оптимальним вибором для корпоративного використання та роботи з конфіденційними даними.

Сучасне сімейство Claude 3 (зокрема версії Haiku, Sonnet та Opus) характеризується розширеним контекстним вікном. Ця технічна особливість уможливлює глибокий аналіз документів значного обсягу, таких як наукові праці, фінансова звітність або великі фрагменти програмного коду, в межах одного запиту [7]. Claude 3 (Opus, Sonnet, Haiku): сімейство враховує градацію моделей за співвідношенням продуктивності та обчислювальної складності. На момент випуску версія Opus демонструвала конкурентоспроможні результати на низці стандартних тестів (бенчмарків) порівняно з GPT-4.

Розробки Claude від Anthropic репрезентують парадигму, в якій ключовим пріоритетом є безпека, керованість та етична узгодженість моделі. Їхній головний науковий внесок полягає у практичній імплементації методів, як-от «Конституційний ШІ» (Constitutional AI), для мінімізації небажаної та шкідливої поведінки. Моделі Claude демонструють, що висока продуктивність, особливо в задачах, що вимагають аналізу об'ємних текстових корпусів (завдяки великому контекстному вікну), може бути досягнута паралельно з підвищенням рівня безпеки, що робить їх оптимальним рішенням для корпоративних клієнтів з високими вимогами до надійності.

Платформа Cohere. Компанія «Cohere», заснована вихідцями з Google, спеціалізується на розробці LLM, призначених виключно для корпоративного використання (Enterprise AI). Їхні розробки позиціонуються як інструменти для вирішення прикладних бізнес-задач, таких як імплементація систем семантичного пошуку, створення клієнтоорієнтованих чат-ботів та автоматизація узагальнення текстових даних. Ключова відмінність Cohere від конкурентів полягає у фокусі на наданні бізнесу інструментів, що гарантують конфіденційність даних, глибоку кастомізацію та підвищення достовірності відповідей. Для зменшення галюцинацій моделі активно використовують технологію Retrieval-Augmented Generation (RAG), що дозволяє підключати їх до приватних баз даних компанії. Cohere пропонує набір моделей, що функціонують як єдина інтегрована система для побудови розумних та безпечних корпоративних додатків [8–11].

Command R+ – це флагманська генеративна модель, оптимізована для діалогових застосувань та робочих процесів, що використовують RAG. Вона добре підходить для створення чат-ботів, помічників та систем, що відповідають на питання на основі документів.

Embed v3 – це провідна модель для створення текстових вбудувань (embeddings). Вбудування – це перетворення тексту (слова, речення, документи) у числовий вектор, що відображає його семантичне значення. Ця технологія є основою для семантичного пошуку, який розуміє сенс запиту, а не лише ключові слова. Модель є багатомовною та підтримує понад 100 мов.

Rerank v3 – це унікальна модель переранжування. Її завдання – покращити якість вже наявних результатів пошуку. Наприклад, ви можете отримати 100 документів за допомогою традиційного пошуку (за ключовими словами) або векторного пошуку, а модель Rerank проаналізує їх відносно запиту і розставить у порядку максимальної релевантності, виводячи на перші місця найточніші відповіді.

Стратегія Cohere характеризується чіткою орієнтацією на корпоративний сегмент (Enterprise) та вирішення прикладних бізнес-завдань. Їхній основний внесок – це фокус на практичній імплементації технології RAG як основного методу для підвищення достовірності та релевантності відповідей шляхом їх «заземлення» на приватних даних клієнта. Пропонуючи спеціалізований інструментарій (моделі Embed, Rerank), Cohere позиціонується не як розробник універсального «чату», а як постачальник фундаментальних компонентів для побудови надійних пошукових та аналітичних систем корпоративного рівня.

2. Моделі з відкритим вихідним кодом (Open-Source)

Парадигма open-source відіграє фундаментальну роль у демократизації доступу до передових технологій штучного інтелекту. Моделі цієї категорії розповсюджуються на умовах ліцензій, що дозволяють їх вільне використання, модифікацію та подальше розповсюдження, в тому числі для комерційних цілей.

Наявність таких моделей дозволяє дослідницьким групам та комерційним стартапам проводити незалежні розробки з меншими початковими капіталовкладеннями. Таким чином, ця парадигма є важливим фактором, що стимулює інноваційну діяльність та конкуренцію у глобальній дослідницькій спільноті.

Сімейство Llama (Meta). Лінійка моделей Llama від компанії «Meta» значно вплинула на розвиток сегмента open-source, надавши спільноті потужні та відкриті архітектури. Релізи Llama 2 та особливо Llama 3 стали поворотним моментом для всієї системи ШІ.

Ці моделі вперше продемонстрували рівень продуктивності, що можна порівняти з провідними пропрієтарними аналогами, що стимулювало нову хвилю інновацій. На сьогодні архітектури Llama слугують фундаментальною базою для тисяч похідних дослідницьких проєктів та комерційних імплементацій по всьому світі [12–14].

Llama 3: остання версія з покращеними можливостями у сферах генерації коду та виконання логічних операцій.

Llama 2: популярна ітерація, яка стала фундаментальною основою для численних дослідницьких та комерційних проєктів.

Серія моделей Llama відіграла визначальну роль у демократизації доступу до високопродуктивних великих мовних моделей, ставши каталізатором для всієї екосистеми open-source. Її випуск значно зменшив технологічний розрив між пропрієтарними системами та відкритими розробками. Стратегічне значення Llama полягає у створенні фундаментальної основи для численних похідних проєктів, академічних досліджень та комерційних стартапів, що сприяє децентралізації інновацій у галузі штучного інтелекту.

Моделі Mistral AI. Розробки європейської компанії Mistral AI здобули широке визнання завдяки високій обчислювальній ефективності та інноваційним архітектурним рішенням. Ключовою технологією, що забезпечує переваги їхніх моделей, є архітектура «Mixture-of-Experts» (MoE), яка дозволяє значно оптимізувати використання обчислювальних ресурсів під час роботи.

Зокрема, модель Mixtral 8x7B, що базується на MoE, та більш компактна Mistral 7B демонструють одне з найкращих на ринку співвідношень продуктивності до апаратних вимог, що робить їх популярним вибором для розгортання в середовищах з обмеженими ресурсами [15–16].

Mixtral 8x7B: модель, що використовує архітектуру «Mixture of Experts» (MoE). Цей підхід дозволяє активувати лише релевантні підмножини параметрів для обробки вхідних даних, що оптимізує використання ресурсів.

Mistral 7B: компактна модель, що демонструє високу продуктивність у своєму класі параметрів.

Моделі Mistral AI є яскравим прикладом пріоритету на обчислювальній ефективності без суттєвої втрати продуктивності. Їхній ключовий внесок – це популяризація та успішне застосування архітектури MoE у моделі Mixtral. Цей підхід, що базується на розрідженій активації параметрів, довів свою спроможність досягати результатів, що порівнюються із значно більшими монолітними моделями, при менших ресурсних витратах. Таким чином, Mistral демонструє вектор розвитку у бік створення більш економічних та доступних для розгортання LLM.

Модель Falcon (TII). Розроблена Інститутом технологічних інновацій (OAE), модель Falcon-180B тривалий час займала провідні позиції в незалежних рейтингах продуктивності серед LLM з відкритим кодом. Моделі Falcon стали важливим етапом у розвитку великих мовних моделей з відкритим кодом, оскільки на момент свого виходу вони продемонстрували найвищу продуктивність серед усіх доступних open-source альтернатив, кинувши виклик пропрієтарним системам.

На відміну від багатьох інших моделей, що базуються на архітектурі Llama, Falcon використовує власну, оптимізовану архітектуру. Однією з ключових інновацій є використання багатозапитової уваги (Multi-Query Attention). У традиційній архітектурі кожен «нейрон» (голова уваги) має власні унікальні параметри для запитів, ключів та значень. У Falcon кілька голів уваги можуть спільно використовувати ті самі параметри для ключів і значень, що значно зменшує обсяг пам’яті, необхідний для роботи моделі, і прискорює процес генерації тексту. Одним із головних секретів успіху Falcon є якість даних, на яких він навчався. Спеціально для цього проєкту команда Technology Innovation Institute створила набір даних RefinedWeb, що складається з кількох трильйонів токенів. Це ретельно відфільтрований та дедуплікований контент з інтернету, з якого було видалено значну частину «шуму», повторів та низькоякісних текстів, що дозволило моделі ефективніше засвоювати знання.

Falcon 180B – це найпотужніша модель у сімействі, що містить 180 мільярдів параметрів і навчалася на 3,5 трильйонах токенів. На момент виходу у 2023 році вона очолювала рейтинг Hugging Face Open LLM Leaderboard, випереджаючи такі моделі, як Llama 2 70B. Завдяки своїм розмірам та якості тренування, вона демонструє виняткові здібності у логічному мисленні, генерації коду та відповідях на складні питання. Спочатку моделі Falcon (наприклад, 40B) були випущені під ліцензією Apache 2.0, яка є дуже гнучкою і дозволяє комерційне використання. Потужніша версія, Falcon 180B, має власну ліцензію «Falcon-180B TII License», яка також допускає комерційне використання, але з певними обмеженнями, тому користувачам необхідно ознайомитися з її умовами. Окрім флагманської моделі на 180 мільярдів параметрів, існують і менші, більш ефективні версії, призначені для різних завдань та апаратних можливостей [17–19].

Falcon-40B: потужна модель, що тривалий час була лідером серед моделей свого розміру.

Falcon-7B та Falcon-1.8B: компактніші версії, які можна запускати на менш потужному обладнанні, включно з високопродуктивними споживчими пристроями.

Модель Falcon, зокрема її версія на 180 мільярдів параметрів, стала важливим етапом розвитку open-source LLM, встановивши на момент свого випуску новий стандарт продуктивності. Її головний внесок полягає у демонстрації критичної значущості якості та масштабу навчальних даних, для чого було створено спеціалізований корпус RefinedWeb. Falcon довів, що дослідницькі центри за межами традиційних технологічних хабів здатні створювати передові фундаментальні моделі, що стимулює глобальну конкуренцію та інновації у відкритій спільноті.

Окрім зазначених вище, існує низка інших важливих моделей. BigScience Large Open-science Open-access Multilingual Language Model (BLOOM) є результатом масштабного міжнародного дослідницького проєкту. Bidirectional Encoder Representations from Transformers (BERT) від Google, хоч і є енкодерною, а не генеративною моделлю, здійснив революцію в задачах розуміння природної мови (NLU) і заклав архітектурні принципи для багатьох наступних розробок. Архітектура BERT є фундаментальною, хоча й не належить до класу генеративних моделей. Її ключовим внеском став механізм двонаправленого кодування, який забезпечив прорив у задачах розуміння природної мови завдяки аналізу контексту слова з обох боків. BERT став концептуальною основою для значної кількості подальших архітектурних розробок [20].

Архітектура BERT являє собою фундаментальний прорив, що спричинив парадигмальний зсув у задачах розуміння природної мови. Її ключовий внесок полягає у впровадженні механізму двонаправленого кодування, який дозволив моделям формувати глибокі контекстуалізовані представлення слів, враховуючи як попередній, так і наступний контекст одночасно. Хоча BERT не є генеративною моделлю (на відміну від GPT), вона заклала концептуальну основу для парадигми «попереднє навчання та подальше доналаштування» (pre-training and fine-tuning), яка стала промисловим стандартом. Таким чином, спадщина BERT полягає у створенні архітектурного фундаменту для всіх сучасних завдань, що вимагають глибокого семантичного аналізу тексту.

Модель BLOOM є унікальним явищем, чия головна цінність полягає не стільки в абсолютній продуктивності, скільки в методології її створення. Це результат масштабного міжнародного дослідницького проєкту, що об’єднав сотні вчених і продемонстрував модель колективної та прозорої розробки ШІ. Ключовим внеском BLOOM є безпрецедентний рівень відкритості (доступні дані, код та журнали тренування), що сприяє відтворюваності результатів та демократизації досліджень. Окрім цього, її нативна багатомовність (46 мов) робить її критично важливим інструментом для наукових спільнот, що працюють з мовами, недостатньо представленими в інших великих моделях.

Microsoft Phi-3-mini. Представляє клас компактних моделей (SLM), де акцент зроблено не на розмірі, а на якості навчальних даних.

Microsoft Phi-3 – це сімейство компактних мовних моделей (SLM), що представляє значний крок у напрямі оптимізації співвідношення продуктивності та обчислювальних ресурсів. На відміну від великих мовних моделей (LLM), що налічують десятки і сотні мільярдів параметрів, стратегія розробки Phi-3 зосереджена на досягненні максимальних когнітивних здібностей у межах суворо обмеженого розміру [21–23].

Фундаментальною інновацією, що лежить в основі високої продуктивності Phi-3, є якість навчальних даних. Замість того, щоб тренувати модель на необроблених масивах даних з інтернету, дослідники Microsoft використали підхід «TinyStories» та згенерували високоякісний, синтетичний навчальний масив, що нагадує дитячі підручники. Цей масив даних був ретельно структурований для навчання моделі фундаментальним концепціям логіки, причинно-наслідкових зв’язків та загальних знань. Такий підхід довів, що якість та структура навчальних даних може компенсувати їхній значно менший обсяг.

Флагманська модель сімейства, Phi-3-mini, маючи лише 3,8 мільярда параметрів, демонструє результати в стандартних бенчмарках (MMLU, HellaSwag), які можна порівняти з моделями значно більшого розміру, такими як Mixtral 8x7B або GPT-3.5. Це робить її однією з найефективніших моделей з відкритим кодом за співвідношенням продуктивності на один параметр. Завдяки своєму компактному розміру, Phi-3-mini є надзвичайно ефективною для розгортання. Вона може працювати: на споживчих GPU **(**модель легко запускається на одній відеокарті, як-от NVIDIA RTX 4090/3090); на центральних процесорах (CPU) квантовані (оптимізовані) версії моделі можуть виконуватися на сучасних CPU без потреби у дискретній графіці); на мобільних пристроях (модель оптимізована для роботи на пристроях з обмеженими ресурсами, що відкриває шлях до створення складних ШІ-додатків, які працюють локально на телефонах чи в системах Інтернету речей (IoT)).

Модель Phi-3 не є прямим конкурентом для GPT-4 або Claude 3 у задачах максимальної складності. Її стратегічне значення полягає у демократизації та розширенні сфери застосування ШІ. Вона робить можливим створення потужних, автономних та конфіденційних ШІ-рішень, які працюють на периферії (Edge AI), а не в централізованій хмарі. Це критично важливо для додатків, що вимагають низької затримки, роботи в офлайн-режимі та гарантованої приватності даних. Порівняльний аналіз розглянутих мовних моделей наведено у таблиці 1.

Таблиця 1

Порівняльний аналіз LLM та SLM

№ з/п	Модель / Сімейство	Розробник	Переваги	Недоліки	Оптимальні сценарії використання
1	2	3	4	5	6
1.	GPT (4, 4o)	OpenAI	Найвища загальна продуктивність та універсальність. Провідні творчі та мультимодальні можливості. Величезна екосистема та легка інтеграція через API	Висока вартість для масштабних завдань. Залежність від хмарної інфраструктури OpenAI	Створення передових комерційних продуктів, швидке прототипування, складні аналітичні та креативні завдання, мультимодальні додатки
2.	Gemini	Google	Нативна мультимодальність (з нуля). Дуже велике контекстне вікно (1.5 Pro). Глибока інтеграція з екосистемою Google	Дещо менша екосистема розробників порівняно з OpenAI. Продуктивність може бути нерівномірною	Аналіз великих масивів даних (відео, код, довгі документи), корпоративні рішення в екосистемі Google, мультимодальний пошук
3.	Claude 3	Anthropic	Акцент на безпеці та етичності. Найкраща продуктивність в аналізі довгих текстів. Керована та передбачувана поведінка	Може бути надто «обережною» і відмовлятися від відповідей. Менш гнучка у суто творчих, «неформальних» завданнях	Юриспруденція, медицина, фінанси, робота з конфіденційними даними, де безпека та аналіз документів є критичними
4.	Llama 3	Meta	Найкраща продуктивність серед open-source моделей. Гнучка ліцензія для комерційного використання. Величезна спільнота та тисячі похідних моделей	Високі вимоги до ресурсів для розгортання. Потребує доналаштування (fine-tuning) для конкретних завдань	Створення кастомних комерційних продуктів, академічні дослідження, стартапи, що потребують повного контролю над моделлю
5.	Mistral / Mixtral	Mistral AI	Найкраще співвідношення продуктивності та вартості. Інноваційна та ефективна архітектура (MoE). Висока швидкість генерації	Менші моделі (7B) поступаються у складних логічних завданнях. Молодша екосистема порівняно з Llama	Додатки реального часу, середовища з обмеженими ресурсами, економічно ефективні чат-боти та аналітичні системи
6.	Cohere	Cohere	Спеціалізація на корпоративних задачах (RAG). Високий рівень конфіденційності даних. Потужні інструменти для семантичного пошуку	Не призначена для загальних творчих діалогів	Побудова корпоративних баз знань, внутрішніх пошукових систем, аналітичних інструментів, де точність даних є критичною
7.	Falcon	TII	Висока продуктивність (особливо 180B). Навчена на дуже якісних даних (RefinedWeb). Гнучке ліцензування	Надвисокі вимоги до обладнання для версії 180B. Менш активна спільнота, ніж у Llama	Великі дослідницькі проєкти, державні ініціативи, де потрібна максимальна потужність від open-source моделі
8.	BERT	Google	Фундаментальна для NLU (розуміння мови). Висока точність у класифікації та аналізі тексту. Велика кількість готових моделей для різних мов	Не є генеративною моделлю (не пише тексти). Архітектурно застаріла для сучасних діалогових систем	Аналіз тональності відгуків, класифікація документів, розуміння пошукових запитів, видобування іменованих сутностей (NER)

Закінчення табл. 1

1	2	3	4	5	6
9.	BLOOM	BigScience	Справжня багатомовність (46 мов). Повна прозорість та відкритість проєкту. Чудовий інструмент для наукових досліджень	Поступається у продуктивності новішим моделям. Дуже високі апаратні вимоги	Академічні дослідження у багатомовному середовищі, розвиток технологій для мов з обмеженими ресурсами, лінгвістичний аналіз
10.	Phi-3-mini	Microsoft	Найкраща продуктивність у своєму класі (small models). Дуже низькі вимоги до ресурсів (може працювати на CPU). Ідеальна для експериментів та швидкого навчання	Поступається великим моделям у глибині знань. Може робити помилки у складних, нішевих темах	Академічні дослідження, навчання та експерименти, розгортання на мобільних пристроях, локальні ШІ-асистенти

За результатами проведеного аналізу можливостей відомих LLM та SLM встановлено, що для використання в системі радіозв'язку важливо використовувати різні моделі для вирішення різних типів завдань. Разом із цим, для збільшення оперативності прийняття рішень на етапі планування застосування сил і засобів радіозв'язку акцент слід зосередити саме на SLM, що можуть бути локально розгорнуті на відносно невеликих продуктивних потужностях.

Крім цього, зазначене завдання потребує можливості розгортання у мобільних засобах за умов забезпечення високої продуктивності та забезпечення fine-tuning. Зазначені умови можуть бути реалізовані в сімействі Phi-3-mini компанії «Microsoft», оскільки для завдань, що вимагають експериментів, швидкого прототипування та функціонування в ізольованому середовищі, компактні моделі, є більш раціональним вибором. Успіх цієї моделі ґрунтується на унікальній філософії розробки та інноваційному підході до процесу навчання.

На відміну від більшості LLM, які навчаються за принципом «випити океан», обробляючи трильйони токенів необроблених даних з інтернету, підхід Microsoft до створення Phi-3 базується на гіпотезі, що якість навчальних даних є важливішою за їх кількість. Ця фундаментальна інновація змінює парадигму навчання, фокусуючись на ефективності засвоєння інформації.

Для навчання Phi-3 був створений великий масив високоякісних «синтетичних» даних. За своєю структурою та змістом цей масив нагадує ретельно написані підручники та енциклопедії. Такий підхід дозволив моделі ефективно засвоїти фундаментальні концепції логіки, міркувань та причинно-наслідкових зв'язків. Замість простого поглинання сирої інформації, модель навчалася на структурованих знаннях, що забезпечило значно вищу ефективність тренувального процесу.

В основі математичного апарату моделі Phi-3-mini лежить архітектура трансформер (Transformer). Її функціонування полягає у послідовності матричних операцій, які перетворюють вхідний текст, представлений у числовій формі (токени), на ймовірнісний розподіл. Цей розподіл своєю чергою використовується для вибору наступного слова в послідовності, що дозволяє генерувати зв'язний та логічний текст [24, 25].

Характеристики моделі Phi-3-mini

Розмір: 3,8 мільярда параметрів. Це значно менше, ніж у флагманських моделей, що робить її набагато легшою та швидшою.

Контекстне вікно: модель існує у двох версіях: стандартній з 4,000 токенів та розширеній до 128,000 токенів. Велике контекстне вікно дозволяє моделі аналізувати об'ємні документи або «пам'ятати» дуже довгі діалоги.

Ліцензія: MIT License. Це дуже гнучка ліцензія, яка дозволяє вільне використання, модифікацію та розгортання моделі навіть у комерційних продуктах.

Сильні сторони

Якість міркувань. Головна перевага Phi-3-mini – це її надзвичайно висока для такого компактного розміру здатність до логічного мислення та аналізу. Вона може розв'язувати задачі, які раніше були під силу лише набагато більшим моделям.

Ефективність. Модель оптимізована для роботи на апаратному забезпеченні середнього класу, включно зі споживчими GPU (наприклад, RTX 3090/4090), сучасними центральними процесорами (CPU) і навіть мобільними пристроями.

Узагальнення та RAG. Завдяки гарним здібностям до міркування та великому контекстному вікну, модель чудово підходить для узагальнення документів та використання в системах RAG (Retrieval-Augmented Generation).

Слабкі місця

Обмеженість нішевих знань. Оскільки модель не навчалася на всьому інтернеті, вона може не знати дуже специфічних, рідкісних фактів або деталей про останні події. Великі моделі, як-от GPT-4, мають ширшу, хоча й менш структуровану, базу загальних знань.

Одномовність. Модель в першу чергу оптимізована для англійської мови. Хоча вона має певні можливості в інших мовах, її продуктивність буде поступатися спеціалізованим багатомовним моделям [26, 27].

1) Токенізація: текст розбивається на менші одиниці – токени (слова, частини слів). Наприклад, «рекомендація» – [«рекомен», «дація»].

2) Ембединги: кожен токен перетворюється на вектор фіксованої довжини (d_{mode}), який представляє його семантичне значення. Це робиться за допомогою матриці ембедингів W_e. До цього вектора також додається позиційний вектор, який кодує положення токена в реченні.

Математично, для послідовності токенів $x = (x_1, \dots \dots, x_n)$, вхідна матриця *X* отримується так [28]:

$$X = Embedding(x) + Positional Encoding(x). \quad (1)$$

Кожне слово – це об'єкт на карті. Формула дає кожному об'єкту дві координати:

Embedding (x) – його семантична координата;

Positional Encoding (x) – його позиційна координата.

Сума + об'єднує ці дві координати в одну, даючи моделі повну інформацію.

Ця формула готує текст для обробки нейромережею, поєднуючи два ключові аспекти: значення слів та їхній порядок у реченні. Вона є першим і найважливішим кроком на вході до архітектури Трансформер [29–31].

У результаті виходить фінальна матриця *X*, де кожен рядок (вектор) одночасно кодує і «що це за слово», і «де воно знаходиться» у реченні. Саме ця збагачена інформація і подається на перший блок уваги (Attention) Трансформера, дозволяючи йому розуміти контекст та граматику.

З метою реалізації елементів підсистеми оцінювання обстановки в районі виконання завдань з використанням SLM типу Phi-3-mini на першому етапі створюється вхідний вектор стану середовища S_t.

3) Ядро моделі: Блоки Трансформера (The Core: Transformer Blocks)

Вхідна матриця *X* послідовно проходить через декілька однакових блоків Трансформера. Кожен блок складається з двох основних частин: механізму багатоголової уваги та нейронної мережі прямого поширення.

А. Механізм Багатоголової Уваги (Multi-Head Attention)

Це найважливіша частина, яка дозволяє моделі зважувати важливість різних токенів у послідовності для розуміння контексту. Вона потребує такі дії.

Для кожного токена (рядка в матриці *X*) створюються три вектори: Запит (Query, *Q*), Ключ (Key, *K*) та Значення (Value, *V*). Це робиться шляхом множення матриці *X* на три окремі матриці ваг W_Q, W_K, W_V, які навчаються:

$$Q = XW_Q, K = XW_K, V = XW_V. \quad (2)$$

Обчислення оцінок уваги (Attention Scores). Модель обчислює, наскільки кожен токен «цікавий» для кожного іншого токена. Це робиться шляхом скалярного добутку матриці Запитів *Q* на транспоновану матрицю Ключів K^T. Результат масштабується, щоб уникнути занадто великих значень.

$$Scores = \frac{QK^T}{\sqrt{d_k}}, \quad (3)$$

де d_k – розмірність векторів Ключів.

Перетворення оцінок у ваги (Softmax). Оцінки перетворюються на ймовірності (ваги уваги) за допомогою функції Softmax. Це гарантує, що сума ваг для кожного токена дорівнює 1.

$$A = soft\max(Scores). \quad (4)$$

Зважування значень. Матриця ваг уваги *A* множиться на матрицю Значень *V*. Токени з вищими вагами більше впливатимуть на кінцевий результат.

$$Attention(Q, K, V) = A \times V. \quad (5)$$

У «багатоголовій» увазі цей процес виконується паралельно декілька разів (h голів) з різними матрицями ваг, що дозволяє моделі фокусуватися на різних аспектах контексту одночасно.

Б. Нейронна мережа прямого поширення (Feed-Forward Network)

Після механізму уваги результат проходить через просту двошарову нейронну мережу, яка застосовується до кожного токена окремо. Це додає моделі нелінійності.

$$FFN(x) = \mathrm{Re}\,LU(xW_1 + b_1)W_2 + b_2, \quad (6)$$

де $W_1, W_2,\ b_1, b_2$ – матриці ваг та зсувів, що навчаються.

ReLU (Rectified Linear Unit) – це функція, яка застосовується до кожного числа в отриманому векторі. Вона працює дуже просто: якщо число від'ємне, вона замінює його на 0; якщо додатне – залишає як є. Цей крок вводить у модель нелінійність. Без неї вся мережа, незалежно від кількості шарів, поводилася б як один простий лінійний перетворювач і не змогла б вивчати складні залежності в даних.

4) Вихід моделі: Прогнозування (Output: Prediction)

Після проходження через усі блоки Трансформера, кінцева матриця векторів Y проходить через лінійний шар та функцію Softmax, щоб спрогнозувати ймовірність кожного можливого токена у словнику для наступної позиції.

Лінійне перетворення: Logits = Y·Wfinal

Ймовірності: P = softmax(Logits)

Модель вибирає токен з найвищою ймовірністю (або семплує з розподілу ймовірностей) як свою відповідь.

Для системи радіозв'язку, де мовна модель є ядром СППР, її математичний апарат потрібно розглядати не лише як генератор тексту, а як функцію, інтегровану в більший контур управління. Наприклад, на етапі планування застосування сил і засобів радіозв'язку слід провести аналіз обстановки в районі виконання завдання.

Робота Phi-3-mini основана на архітектурі Transformer, що передбачає певну послідовність матричних операцій які перетворюють послідовність чисел (токенів) на іншу послідовність чисел (ймовірності наступного токена), а саме: для цього необхідно формалізувати вхідні дані. «Нечіткий опис обстановки» у момент часу *t* представляється як текстовий рядок C_t. Цей рядок перетворюється на вхідний вектор стану S_t за допомогою математичного апарату моделі:

$$S_t = TransformerEncoder\,(Embedding\,(C_t) + Positional\ Embedding\,(C_t)), \quad (7)$$

де *TransformerEncoder* – це повний стек блоків Трансформера моделі Phi-3. Результат S_t – це матриця контекстуалізованих векторів, що є багатим числовим представленням поточної обстановки.

На другому етапі проводиться генерація простору можливих дій ($\boldsymbol{A_t}$). Модель повинна згенерувати не одну відповідь, а набір можливих рекомендацій (дій). Для цього використовується метод семплування з вихідного розподілу ймовірностей моделі.

Нехай $P(a \mid S_t)$ – це ймовірність згенерувати послідовність токенів a (рекомендацію) на основі стануS_t. Генеруємо k можливих дій:

$$A_t = \{a_1, a_2, \ldots\ldots a_k\}, \quad (8)$$

де $a_i \sim P(a \mid S_t)$.

Кожна дія a_i – це текстова рекомендація, наприклад, «Збільшити потужність передавача на 10 %» або «Перейти на частоту 450 МГц».

Третій етап передбачає реалізацію функції оцінки корисності (Utility Function, $\boldsymbol{U}$).

Це ядро системи прийняття рішень. Кожна згенерована дія $\boldsymbol{a_i}$ повинна бути оцінена з точки зору її корисності для системи. Функція корисності $\boldsymbol{U}$ – це математичне вираження цілей. Вона приймає дію $\boldsymbol{a_i}$ та поточний стан $\boldsymbol{S_t}$ і повертає числову оцінку.

$$U(a_i, S_t) \to R. \quad (9)$$

Функція U може бути складною і враховувати багато факторів:

Ефективність зв'язку ($\boldsymbol{E}$): пропускна здатність, відношення сигнал/шум.

Безпека (Sec): ймовірність перехоплення, стійкість до завад.

Витрати ресурсів (Cost): енергоспоживання, використання частотного спектра.

Спрощена функція корисності може виглядати так:

$$U(a_i, S_t) = w_1 \cdot E(a_i) + w_2 \cdot Sek(a_i) - w_3 \cdot Cost(a_i), \quad (10)$$

де w_1, w_2, w_3 – вагові коефіцієнти, що визначають пріоритети.

Ця формула є універсальною математичною конструкцією для багатокритеріальної оптимізації, її можна знайти будь-якій галузі, де потрібно прийняти раціональне рішення, збалансувавши кілька позитивних та негативних факторів.

На четвертому етапі здійснюється вибір оптимальної дії ($\boldsymbol{a_t^*}$).

Система приймає рішення, вибираючи дію, яка максимізує функцію корисності.

$$a_t^* = \arg\max_{a_i \in A_t} U(a_i, S_t). \quad (11)$$

Обрана дія a_t^* є фінальною рекомендацією системи, яку потрібно виконати для переконфігурації системи радіозв'язку.

На завершальному етапі проводиться адаптація та навчання.

Це етап, на якому методика з генетичним алгоритмом (ГА) застосовується для вдосконалення самої системи прийняття рішень.

Популяція: «Особинами» у ГА є не ваги моделі, а параметри самої СППР. Це можуть бути:

1. Вагові коефіцієнти функції корисності ($w_1, w_2, w_3, \dots \dots \dots$);
2. Параметри генерації рекомендацій (наприклад, температура семплування);
3. Правила або мета-інструкції (meta-prompts), що подаються в модель разом із S_t.

Фітнес-функція (F): продуктивність системи оцінюється на наборі симуляцій. Для кожної «особини» (набору параметрів) система проганяється через низку сценаріїв зміни обстановки. Фітнес-функція оцінює, наскільки добре система досягала своїх цілей.

$$F(особина) = \sum_{t=1}^{N} R(S_{t+1}), \tag{12}$$

де $R(S_{t+1})$ – це винагорода, отримана в результаті переходу системи в новий стан S_{t+1} після застосування дії a_t^*.

Еволюція: ГА використовує оператори схрещування та мутації для створення нових поколінь параметрів СППР, які з часом максимізують фітнес-функцію F.

Таким чином, математичний апарат методики – це замкнений контур, де Phi-3-mini є генератором гіпотез (можливих дій), а зовнішній оптимізаційний шар (ГА) налаштовує критерії вибору найкращої дії для досягнення глобальних цілей системи.

Разом з цим вибір між великими та малими мовними моделями не є взаємовиключним. Для завдань, що вимагають інноваційних підходів, найбільш обґрунтованою є двофазна стратегія.

Фаза 1: апробація гіпотез та прототипування на базі Phi-3-mini. На цьому етапі використання компактної моделі дозволяє з мінімальними обчислювальними та часовими витратами перевірити життєздатність концепції СППР, відпрацювати архітектуру та функцію корисності, провести навчання операторів. Це швидкий, дешевий та безпечний спосіб довести принципову можливість вирішення задачі.

Фаза 2: масштабування до промислового рівня на базі LLM. Після успішного завершення першої фази відпрацьована методологія може бути перенесена на велику модель (наприклад, Llama 3 70B). Цей крок має на меті досягнення максимальної точності, надійності та глибини аналізу, що є необхідним для систем критичного призначення.

Таким чином, моделі Phi-3-mini та Llama 3 формують синергетичний тандем, де одна є гнучким інструментом для досліджень, а інша – потужною платформою для фінального рішення.

Висновки. Проведений аналіз сучасних мовних моделей демонструє, що вибір між великими (LLM) та компактними (SLM) архітектурами не є взаємовиключним, а являє собою стратегічне рішення щодо етапності розробки складних систем. Для завдань, що вимагають інноваційних підходів та інтенсивних експериментів, найбільш обґрунтованою є двофазна стратегія впровадження.

На першому етапі – апробації та прототипування – використання компактних моделей, наприклад, Phi-3-mini, є оптимальним рішенням. Завдяки низьким вимогам до ресурсів, високій ефективності та можливості локального розгортання вони дозволяють з мінімальними витратами перевірити життєздатність концепцій, відпрацювати архітектуру та довести принципову можливість вирішення поставленої задачі.

Після успішної валідації гіпотез проєкт переводять на другий етап – масштабування до промислового рівня. На цій стадії відпрацьована методологія застосовується до великих моделей, наприклад, Llama 3 70B, для досягнення максимальної точності, надійності та глибини аналізу, що є критично важливим для систем спеціального призначення.

Таким чином, замість протиставлення компактні та великі мовні моделі формують синергетичний тандем. SLM є гнучким інструментом для досліджень, тоді як LLM слугують потужною платформою для фінального, високоякісного рішення. Запропонована методологія інтеграції дозволяє раціонально підійти до розробки СППР радіозв'язку нового покоління та кардинально підвищити ефективність і оперативність планування радіозв'язку.

References:

1. OpenAI, *GPT-4o*, [Online], available at: https://openai.com/index/hello-gpt-4o/
2. *GPT-4*, [Online], available at: https://arxiv.org/abs/2303.08774
3. «Language Models are Few-Shot Learners», *GPT-3*, [Online], available at: https://arxiv.org/abs/2005.14165
4. *Gemini 1.5 Pro Google*, [Online], available at: https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/
5. *Gemini 1.5*, [Online], available at: https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf
6. *Gemini 1.0*, [Online], available at: https://arxiv.org/abs/2312.11805
7. *Anthropic*, [Online], available at: https://www.anthropic.com/news/claude-3-family
8. Cohere, [Online], available at: https://txt.cohere.com/blog/
9. *Command R+*, [Online], available at: https://txt.cohere.com/command-r-plus-rag-optimized-model/
10. [Online], available at: https://docs.cohere.com/
11. Vaswani, A., Gomez, A. et al., «Attention Is All You Need», [Online], available at: https://arxiv.org/abs/1706.03762

12. *Llama 3 Meta AI*, [Online], available at: https://ai.meta.com/blog/meta-llama-3/
13. *Llama 2*, [Online], available at: https://arxiv.org/abs/2307.09288
14. *Llama Meta AI*, [Online], available at: https://ai.meta.com/llama/
15. «Mixtral в блозі компанії», [Online], available at: https://mistral.ai/news/mixtral-of-experts/
16. *Mixtral 8x7B*, [Online], available at: https://arxiv.org/abs/2401.04088
17. «Detalnyi opys modeli, yii perevah ta rezultativ u benchmarkakh», *Falcon 180B*, [Online], available at: https://www.tii.ae/news/uaes-tii-releases-falcon-180b-advanced-open-source-ai-model
18. Hugging Face, [Online], available at: https://huggingface.co/tiiuae/falcon-180B
19. *RefinedWeb*, [Online], available at: https://huggingface.co/datasets/tiiuae/falcon-refinedweb
20. «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding», [Online], available at: https://arxiv.org/abs/1810.04805
21. *Microsoft Phi-3*, [Online], available at: https://azure.microsoft.com/en-us/blog/introducing-phi-3-a-new-family-of-open-ai-models-from-microsoft/
22. Hugging Face, [Online], available at: https://huggingface.co/microsoft/Phi-3-mini-4k-instruct
23. «TinyStories: How Small Can Language Models Be and Still Speak Coherent English?», [Online], available at: [Online], available at: https://arxiv.org/abs/2305.07759
24. Vaswani, A. et al. (2017), «Attention Is All You Needhttps», [Online], available at: https://arxiv.org/abs/1706.03762
25. Microsoft (2024), «Phi-3 Technical Report», [Online], available at: https://arxiv.org/abs/2404.14219
26. Hugging Face, [Online], available at: https://huggingface.co/microsoft/Phi-3-mini-4k-instruct
27. «TinyStories: How Small Can Language Models Be and Still Speak Coherent English?», [Online], available at: https://arxiv.org/abs/2305.07759
28. Vaswani A. et al. (2017), «Attention Is All You Need», [Online], available at: https://arxiv.org/abs/1706.03762
29. Microsoft (2024), *Phi-3 Technical Report*, [Online], available at: https://arxiv.org/abs/2404.14219
30. Alammar, J., «The Illustrated Transformer», [Online], available at: http://jalammar.github.io/illustrated-transformer/
31. Vaswani, A., Shazeer, N., Parmar, N. et al. (2017), «Attention is All you Need Part of Advances in Neural Information Processing Systems 30», *NIPS 2017*.

Дупелич Сергій Олексійович – кандидат технічних наук, доцент Житомирського військового інституту ім. С.П. Корольова.

https://orcid.org/0000-0002-9438-5778.

Наукові інтереси:

– інформаційні технології радіо- та супутникового зв'язку.

Дзюбенко Володимир Васильович – ад'юнкт штатний науково-організаційного відділу Житомирського військового інституту ім. С.П. Корольова.

https://orcid.org/0009-0002-4548-8903.

Наукові інтереси:

– інформаційні технології радіо- та супутникового зв'язку.

Dupelich S.O., Dzyubenko V.V.

Analysis of the capabilities of large and small language models (LLM/SLM) for optimizing decision-making processes in a radio communication system

The modern radio communication system operates in conditions of extreme complexity, characterized by a dynamically changing environment, high density of information flows, and active enemy countermeasures. The effectiveness of management in such conditions is determined by the ability not only to process large data sets, but also to make optimal decisions in real-time. The key success criteria are the timeliness of the reaction, the reliability of the analysis, and the concealment of actions. The intensive development of large language models (LLM) and small language models (SLM), which demonstrate unique abilities for semantic interpretation, logical thinking, and the generation of recommendations, opens up fundamentally new horizons for the automation of cognitive tasks and the creation of intelligent decision support systems (DSS) of a new generation. This work is the first comprehensive generalization that systematically considers the application of artificial intelligence language models in this specific and critically important context.

The paper provides a comparative analysis of the key architectures of language models dominating the market, highlighting the dichotomy between proprietary systems (GPT families from OpenAI, Gemini from Google, and Claude from Anthropic) and open-source models (Llama from Meta, Mistral from Mistral AI, and Phi-3 from Microsoft). It is argued that for tasks that require flexibility, local deployment, and rapid prototyping in conditions of high secrecy, compact language models (SLM) are of particular interest. In particular, the Microsoft Phi-3-mini architecture is considered in detail, whose high performance at a small size (3,8 billion parameters) is achieved thanks to an innovative approach to training on high-quality synthetic data, which puts the quality of the training set above its volume.

Keywords: large language models; small language models; LLM; SLM; Phi-3-mini; decision support systems; radio communication.

Стаття надійшла до редакції 29.09.2025.