УДК 004.056.5
DOI https://doi.org/10.24144/2616-7700.2026.49(2).283-291

В. Б. Цап

Національний університет «Львівська політехніка»,
аспірант кафедри систем штучного інтелекту
vladyslav.b.tsap@lpnu.ua
ORCID: https://orcid.org/0000-0002-8062-0079

ПРОБЛЕМА МАСШТАБУВАННЯ КРИПТОГРАФІЧНИХ АЛГОРИТМІВ ДЛЯ ШТУЧНОГО ІНТЕЛЕКТУ

У статті розглянуто проблему масштабування криптографічних алгоритмів у системах штучного інтелекту, що працюють із чутливими даними. Метою дослідження є виявлення причин низької практичної придатності таких рішень і формування зрозумілих критеріїв їх оцінювання. У роботі проаналізовано повністю гомоморфне шифрування, безпечні багатосторонні обчислення, диференційну приватність і довірені виконувальні середовища. Запропоновано операційну модель масштабованості, яка враховує затримку відповіді, витрати пам'яті, обсяг передавання даних, кількість раундів взаємодії та складність окремих обчислювальних операцій. Також подано шкалу практичної придатності і протокол оцінювання, що дають змогу визначати межі застосовності криптографічних рішень для різних типів моделей і сценаріїв використання. Установлено, що основними причинами слабкої масштабованості є великі обчислювальні витрати, значне зростання пам'яттєвих потреб, високий мережевий трафік, багатораундовість протоколів і складність реалізації нелінійних операцій. Зроблено висновок, що для практичного розвитку приватного штучного інтелекту доцільним є проєктування гібридних архітектур і створення моделей, спочатку пристосованих до роботи в захищеному криптографічному середовищі.

Ключові слова: великі мовні моделі, криптографічні алгоритми, повністю гомоморфне шифрування, безпечні багатосторонні обчислення, приватність даних.

1. Вступ. Поширення великих мовних моделей (LLM) і систем машинного навчання у сферах, пов'язаних із медичними, фінансовими та персональними даними зробило приватність не додатковою властивістю, а базовою умовою практичного використання штучного інтелекту (ШІ). Тепер окрім точності до відповідей, очікуються також формальні гарантії, що дані користувача не будуть розкриті під час навчання або отримання передбачень.

Теоретичний фундамент для цього давно сформований: повністю гомоморфне шифрування (Fully Homomorphic Encryption, FHE), безпечні багатосторонні обчислення (Multi-Party Computation, MPC), диференційна приватність (Differential Privacy, DP), і довірені виконувальні середовища (Trusted Execution Environments, TEE) дозволяють обробляти дані без прямого розкриття їхнього вмісту. Проте для систем ШІ ключовим стає питання масштабованості: чи можна зберегти гарантії приватності за прийнятної затримки, пам'яттєвих витрат, мережевого трафіку та вартості інфраструктури.

Актуальність теми визначається тим, що розрив між теоретичною можливістю і сервісною придатністю в приватному ШІ досі залишається значним.

Метою статті є виявлення причин слабкої масштабованості криптографічних алгоритмів у задачах сучасного ШІ та формування критеріїв, за якими таку масштабованість можна оцінювати. Для досягнення цієї мети поставлено такі завдання: проаналізувати наявні підходи до приватних обчислень для ШІ,

визначити ключові обмеження, запропонувати інструмент оцінювання приватного ШІ та узагальнити інженерні напрями підвищення практичної придатності таких систем.

Наукова новизна полягає в тому, що проблему масштабування розглянуто як єдиний комплекс обмежень, у межах одного вимірюваного операційного контуру, який визначає межі практичного застосування криптографічних алгоритмів у ШІ.

2. Постановка завдання. Ключова проблема полягає у тому, що теоретична коректність операцій не гарантує операційної придатності. Для традиційного аналізу алгоритмів достатньо вказати, що метод має поліноміальну складність. Однак у криптографічному ШІ вирішальними часто стають не лише ступінь полінома, а й сталі множники, параметри безпеки, розмір шифртекстів, кількість раундів взаємодії, частота бутстрепінгу, тиск на пам'ять і мережу.

Висока ефективність трансформерів досягається завдяки поєднанню великих матричних множень із нелінійними функціями типу GELU, Softmax, нормалізації шару та механізму уваги. Саме ці елементи і є найгірше сумісними з базовими криптографічними примітивами. У Microsoft SEAL прямо зазначено, що гомоморфні обчислення природно підтримують додавання і множення, тоді як операції порівняння, сортування або інші складні нелінійні перетворення, як правило, не є практично здійсненними без спеціальних побудов.

Ширший спектр реалізацій показує, що це обмеження має різні інженерні форми. У PALISADE/OpenFHE практична придатність розширюється за рахунок бутстрепінгу, перемикання схем і спеціалізованої підтримки складніших функцій, але ці можливості підвищують вартість конфігурації та виконання. HElib надає гнучкий низькорівневий контроль над гомоморфним контуром, проте така гнучкість вимагає складнішого налаштування і ретельного керування параметрами. У Concrete прийнятніший режим досягається через квантування і свідоме обмеження архітектури моделі під FHE-виконання. У CrypTen, навпаки, вузьке місце часто переноситься з локальної арифметики у комунікацію, трафік і число раундів взаємодії. Тому тема дослідження має спиратися на сукупність реальних реалізацій.

Отже, постановка завдання полягає у з'ясуванні того, які саме технічні обмеження роблять криптографічно коректні схеми недостатньо масштабованими для сучасних моделей ШІ, які порогові режими непридатності виникають під час зростання масштабу і які інженерні напрями здатні зменшити цей розрив.

3. Огляд літератури. Однією з перших робіт, що практично продемонструвала можливість виконання нейромережі на зашифрованих даних, стала CryptoNets [1]. Для набору MNIST автори повідомили 99% точності, пропускну здатність близько 58,982 передбачень на годину та затримку 250 с для одного передбачення. Подальше зменшення затримки було продемонстровано в LoLa [2]. У цій роботі показано, що інше представлення даних, інше пакування шифртекстів і архітектурні спрощення можуть дати більше ніж десятикратне скорочення затримки: для LoLa-Small повідомляється 0,29 с на передбачення.

Межу масштабування добре ілюструє робота про великі DNN-моделі в HE-середовищі [3]. Її автори прямо вказують, що виконання моделі на зашифрованих даних часто супроводжується накладними витратами пам'яті та виконання у 100–10,000 разів, а навіть відносно компактні моделі можуть вимагати сотень

гігабайтів оперативної пам'яті.

Ще один критично важливий блок літератури стосується бутстрепінгу. Сучасний огляд Shen [4] прямо називає бутстрепінг водночас найбільш суттєвим і найбільш складним компонентом практичної FHE-реалізації. Це важливо для аналізу масштабованості, тому що в довгих або нелінійно насичених контурах саме бутстрепінг часто перетворює схему з формально універсальної на практично дорогу.

Для MPC-контурів системні обмеження добре видно на прикладі CrypTen [5]. Для класифікації тональності тексту CrypTen є на 3 порядки повільнішим за PyTorch, хоча в двосторонньому режимі з розміром пакета 32, виконання займає лише 0,03 с на зразок. Для Wav2Letter передавання даних на GPU при восьми сторонах забирає 63% часу виконання, при цьому комунікація сягає десятків гігабайтів на зразок, а число раундів порядку $2 \cdot 10^4$. Для класифікації зображень дві сторони можуть захищено обчислити ResNet-18 за 2,49 с, а ViT-B/16 — за 8,47 с, але комунікація зростає до сотень і тисяч гігабайтів на зразок, на одну сторону, а число раундів — до $4 \cdot 10^5$.

Для LLM надвеликого масштабу важливим орієнтиром є робота Zhang, Zheng і Bao [6], де для ChatGLM2-6B з LoRA-захистом приватної частини моделі повідомляється 1,61 с/токен. Принципове значення цього результату полягає не лише в самому числі, а в його архітектурному змісті: прийнятніша затримка досягається не повним наскрізним FHE для всіх параметрів моделі, а вибірковим захистом приватної частини.

Наведені джерела практично підтверджують тезу, що наразі криптографічні схеми не є достатньо оптимізовані для прямого використання у ШІ. Для формування подальших узагальнень використано корпус із джерел, де праці дають як експериментальні приклади захищеного виведення різного масштабу, так і порівняльні дані щодо реалізацій FHE. Із кожного джерела відбиралися лише явно наведені або безпосередньо обчислювані характеристики: затримка, пам'яттєві витрати, трафік, число раундів, роль бутстрепінгу та спосіб реалізації нелінійних операцій. На основі цього матеріалу можна сформувати метрично прив'язану сумісність найуживаніших компонентів нейромереж із криптографічними примітивами у табл. 1.

4. Основний результат. Основний результат проведеного аналізу полягає в тому, що слабка масштабованість криптографічних алгоритмів для ШІ визначається не одним окремим чинником, а сукупністю обчислювальних, пам'яттєвих і комунікаційних обмежень. Ці обмеження можна як відтворюваний інструмент оцінювання практичної придатності приватного ШІ. У межах цього інструмента система оцінюється за чотирма вимірюваними показниками: затримкою відповіді, пам'яттєвими накладними витратами, мережевим трафіком і числом раундів взаємодії. Наукова новизна такого підходу полягає в тому, що межа придатності визначається правилом найгіршої метрики: система придатна до цільового сценарію лише тоді, коли жоден із ключових показників не виходить за його межі.

Систематизацію ключових причин слабкої масштабованості криптографічних алгоритмів для ШІ наведено в табл. 2. Вона сформована як результат групування всіх зафіксованих обмежень у джерелах за домінувальною метрикою деградації. Якщо одна й та сама робота містила кілька незалежних вузьких

Таблиця 1.

Сумісність типових компонентів нейромереж із криптографічними примітивами

Компонент	**Вбудована підтримка у FHE/MPC**	**Причина накладних витрат**	**Типовий спосіб адаптації**	**Матричний наслідок**
Лінійні шари	Так	Велика кількість множень, ротацій, пакування	SIMD-пакування, оптимізація розміщення даних	Зростає затримка, але операція лишається базово придатною
ReLU	Ні	Порівняння і розгалуження	Поліномна апроксимація або захищене порівняння	Зростають глибина множень або число раундів
Sigmoid	Ні	Нелінійні функції	Низькостепенева апроксимація	Компроміс між точністю та затримкою
Softmax	Ні	Експонента, ділення, нормалізація	Апроксимація, вибірковий захист або винесення з приватного контуру	Зростають трафік, число раундів і ризик втрати точності
Нормалізація шару	Ні	Корені, ділення, дисперсія	Архітектурна заміна або спрощення	Зростають затримка і пам'яттєві витрати
Механізм уваги	Ні	Велика кількість матричних множень	Гібридний захист або спрощення механізму	Зростають затримка, трафік і пам'ять

місць, вони враховувалися окремо: наприклад, для CrypTen паралельно фіксувалися і затримка, і трафік, і раундовість, а для FHE-кейсів — і коефіцієнт розширення часу виконання, і пам'яттєвий надлишок, і роль бутстрепінгу. Тому табл. 2 відображає не окремі приклади, а повторювані класи обмежень, які відтворюються у різних архітектурах і реалізаціях.

Зазначений результат доцільно формалізувати моделлю операційної затримки криптографічного контуру:

$$T_{total} = C\, T_{plain} + T_{mem} + T_{comm} + T_{sync} + T_{boot},$$

де T_{total} — повний час сервісної відповіді в криптографічному контурі, C — криптографічний коефіцієнт розширення базової операції, T_{mem} — накладні витрати доступу до пам'яті та переміщення даних, T_{comm} — витрати на передавання даних, T_{sync} — витрати на синхронізацію між етапами протоколу, T_{boot} — час, пов'язаний із бутстрепінгом або іншими дорогими операціями оновлення криптографічного стану.

Пам'яттєвий надлишок визначається співвідношенням:

$$M_{ovh} = \frac{S_{cipher}}{S_{plain}},$$

Таблиця 2.

Основні чинники слабкої масштабованості криптографічних алгоритмів для штучного інтелекту

Група обмежень	**Конкретний прояв**	**Причина виникнення**	**Як вимірюється**	**Наслідок для систем ШІ**
Обчислювальні	Висока вартість базових операцій	Криптографічні перетворення значно дорожчі за операції у відкритому вигляді	Коефіцієнт розширення базової операції, повний час сервісної відповіді	Зростання часу відповіді
Пам'яттєві	Великі шифртексти та проміжні представлення	Роздування даних у зашифрованій формі	Пам'яттєвий надлишок, робочий набір у ГБ	Високі вимоги до пам'яті та кешу
Комунікаційні	Великий обсяг переданих даних	Протокольна взаємодія між сторонами	ГБ на запит або партію, витрати на передавання даних	Затримки і підвищені вимоги до мережі
Протокольні	Багатораундовість, синхронізація	Структура безпечного протоколу	Число раундів, чутливість до RTT, витрати на синхронізацію між етапами протоколу	Час відповіді починає визначатися мережею
Алгоритмічні	Погана сумісність із нелінійностями	Базові схеми природно підтримують насамперед додавання і множення	Потреба в апроксимації, зростання глибини множень, втрата точності	Приріст вартості або погіршення якості
Криптографічні	Накопичення шуму	Багатошарові множення	Частота бутстрепінгу, час бутстрепінгу	Довгі контури стають дорогими
Системні	Обмежений ефект локального прискорення	Вузьке місце зміщується в іншу підсистему	Частка прискорюваного часу	Непропорційний приріст від апаратного прискорення

де S_{cipher} — розмір зашифрованого подання, а S_{plain} — розмір відповідних відкритих даних.

Для багатосторонніх обчислень комунікаційну складову відображає така оцінка:

$$T_{MPC} \approx \sum_{i=1}^{R} (\frac{V_i}{B} + L_i) + T_{loc},$$

де R — кількість раундів, V_i — обсяг переданих даних у і-му раунді, B — ефективна пропускна здатність каналу, L_i — мережева затримка і-го раунду, а T_{loc} — локальні обчислення сторін.

Практичне застосування цих формул вимагає явного підбору параметрів. Значення C доцільно брати з опублікованих порівняльних вимірювань [7] або

профілювання [8], M_{ovh} — безпосередньо обчислювати як відношення розміру шифртексту до розміру відкритого подання, T_{comm} — оцінювати через обсяг передавання і пропускну здатність каналу, T_{sync} — через число раундів і середню мережеву затримку, T_{boot} — за публікаціями або документацією конкретної реалізації. Відтворення оцінки передбачає послідовність із чотирьох кроків: спочатку для конкретної моделі та криптографічної схеми фіксуються T_{plain} і розмір відкритого подання, далі з літератури або профілювання підставляються C, M_{ovh}, T_{boot} і мережеві параметри, після цього обчислюються T_{total}, M_{ovh} та, за потреби, T_{MPC}, на завершальному етапі отримані значення зіставляються з порогами операційної придатності. Щоб довести формалізацію до рівня робочого інструмента, достатньо показати, як вона працює на числових прикладах.

У межах цієї статті операційні пороги слід задавати не довільно, а від опорних значень, які вже присутні в розглянутому корпусі джерел. Для затримки такими опорними точками є 0,29 с у LoLa-Small [2], 1,61 с/токен у ChatGLM2-6B з приватним LoRA-контуром [6], 2,49 с для ResNet-18 і 8,47 с для ViT-B/16 у CrypTen [5], а також 250 с у CryptoNets [1]. Саме тому межа до 1 с відділяє реально інтерактивний режим від повільніших, інтервал 1-10 с охоплює зафіксовані в літературі ще практично придатні сценарії захищеного виведення, а значення понад 100 с спирається на вже опублікований непридатний випадок 250 с [1], проміжок 10-100 с є логарифмічною перехідною зоною між останнім придатним спостереженням порядку 10^1 і непридатним спостереженням порядку 10^2. Для пам'яті нижньою емпіричною межею критичності є порядок 10^2 ГБ на вузол, оскільки саме про сотні гігабайтів для відносно компактних моделей прямо повідомляє одне з джерел [3]. Для трафіку опорними є «десятки гігабайтів» для Wav2Letter і до $2 \cdot 10^3$ ГБ для ViT-B/16 у CrypTen [5], тому межі 10 ГБ і 10^2 ГБ відбивають перехід від напруженого до явно непридатного режиму. Для MPC-раундовості опорними точками є порядок $2 \cdot 10^4$ і $4 \cdot 10^5$ раундів у тих самих експериментах [5], тому інтервали до 10^4, 10^4-10^5 і понад 10^5 є не довільними, а прив'язаними до спостережуваних порядків величин. Отже, використані тут межі мають статус відтворюваних інтервальних оцінок, безпосередньо виведених із наведених публікацій.

Базова валідація запропонованої шкали узгоджується з реальними випадками з літератури саме тому, що її пороги побудовано з цих випадків, а не накладено на них постфактум. LoLa-Small із затримкою 0,29 с [2] фіксує нижню область реально інтерактивного захищеного виведення. ChatGLM2-6B з приватним LoRA-контуром і затримкою 1,61 с на токен [6], а також CrypTen для ResNet-18 (2,49 с) і ViT-B/16 (8,47 с) [5] задають інтервал практично придатних, але вже не миттєвих сценаріїв. CryptoNets із затримкою 250 с на передбачення [1] показує емпірично зафіксований вихід у сервісно непридатний режим. Трафік і раундовість у CrypTen — від десятків гігабайтів і $2 \cdot 10^4$ раундів до $2 \cdot 10^3$ ГБ і $4 \cdot 10^5$ раундів [5] — додатково показують, що саме мережеві характеристики, а не лише затримка, можуть першими зробити систему непридатною. Отже, правило найгіршої метрики працює на реальних прикладах, а формалізація дає змогу класифікувати систему ще до повного розгортання.

Нехай для FHE-випадку незашифроване виконання моделі займає $T_{plain} = 0,05$ с, криптографічний коефіцієнт розширення становить $C = 100$, а додаткові витрати дорівнюють $T_{mem} = 0,3$ с, $T_{comm} = 0$, $T_{sync} = 0$, $T_{boot} = 0$.

Тоді

$$T_{total} = 100 \cdot 0,05 + 0,3 = 5,3 \text{ с}.$$

Отримане значення потрапляє у відкладено-інтерактивний режим. Якщо ж для тієї самої моделі взяти верхню межу, тобто $C = 10000$, то

$$T_{total} = 10\ 000 \cdot 0,05 + 0,3 = 500,3 \text{ с},$$

що вже означає сервісну непридатність. Отже, формула показує не просто те, що FHE повільно працює, а те, при яких значеннях C система виходить за межі цільового сценарію.

Для оцінки пам'яттєвого надлишку можна взяти простий приклад пакета активацій розміром $S_{plain} = 4$ МБ. Якщо після шифрування він займає $S_{cipher} = 400$ МБ, тобто відповідає нижній межі 100× накладних витрат [3], тоді

$$M_{ovh} = \frac{400}{4} = 100.$$

Це означає збільшення обсягу в 100 разів. Для партії з 32 зразків лише цей фрагмент вимагатиме вже 12,8 ГБ пам'яті, тобто навіть нижня межа емпірично зафіксованого накладного коефіцієнта швидко виводить систему з комфортного серверного режиму, а при переході до «сотень гігабайтів» [3] — і за межі практичної придатності.

Для MPC-режиму ефект раундовості й мережі видно з формули T_{MPC}. Нехай $R = 20000$, $T_{loc} = 0,5$ с, а ефективний внесок одного раунду становить лише 0,2 мс. Тоді

$$T_{MPC} \approx 20\ 000 \cdot 0,0002 + 0,5 = 4,5 \text{ с}.$$

У локальній мережі це ще прийнятний результат. Але якщо ефективний штраф на раунд зростає до 10 мс, маємо

$$T_{MPC} \approx 20\ 000 \cdot 0,01 + 0,5 = 200,5 \text{ с},$$

тобто перехід у практично непридатний режим лише через мережеву складову. Саме так формула пояснює, чому одна й та сама схема може виглядати прийнятною в лабораторії й непридатною у розподіленому сервісі.

Поширене припущення, що низьку швидкодію можна повністю компенсувати апаратним прискоренням, для криптографічного ШІ працює лише частково. Межу цієї стратегії кількісно виражає співвідношення:

$$S_{tot} = \frac{1}{(1-p) + \frac{p}{s}},$$

де p визначає частку часу, що справді піддається апаратному прискоренню, а s — коефіцієнт прискорення.

Для криптографічних систем частка p описується співвідношенням:

$$p = \frac{T_{comp}}{T_{comp} + T_{mem} + T_{comm} + T_{sync}},$$

де T_{comp} — час виконання криптографічних і арифметичних перетворень у межах обчислювального контуру.

Отже, альтернативні підходи перерозподіляють компроміс між приватністю, точністю і продуктивністю між різними типами витрат. Найбільш реалістичним напрямом лишається гібридна архітектура і співпроєктування моделі з криптографічним контуром, коли найчутливіші фрагменти захищаються найсильнішими засобами, а решта контуру оптимізується під конкретний сценарій використання.

5. Висновки та перспективи подальших досліджень. Проведений аналіз показує, що проблема масштабування криптографічних алгоритмів для ШІ є насамперед проблемою операційної придатності, а не лише математичної коректності: навіть коли схема формально забезпечує правильне виконання обчислень над захищеними даними, вона може виявитися непридатною для інтерактивного сервісу через сукупність затримки, пам'яттєвих витрат, мережевого трафіку, кількості раундів взаємодії і складних нелінійностей.

Головний висновок полягає в тому, що найближчим реалістичним шляхом розвитку приватного ШІ є побудова гібридних архітектур, де сила гарантій захисту диференціюється за чутливістю фрагментів системи. Для FHE це означає фокус на вузьких критичних контурах, зменшення частоти дорогих нелінійних операцій і криптографічно узгоджене проєктування, для MPC — мінімізацію раундів, перетворення часток секрету та вузьких місць комунікації, для DP — балансування бюджету приватності, обчислювальної вартості і корисності, для TEE — жорсткіший аналіз моделі довіри.

Перспективи подальших досліджень пов'язані з трьома напрямами: розробленням наборів порівняльних тестів саме для проміжного масштабу приватного ШІ, проєктуванням моделей, які від початку оптимізуються під криптографічний контур, побудовою порівняльних методик, у яких FHE, MPC, DP і TEE оцінюються в єдиному операційному просторі «ризик — затримка — пам'ять — трафік — вартість». Саме така зміна оптики є необхідною умовою переходу від концептуально коректних прототипів до сервісів, придатних для застосування.

Конфлікт інтересів

Автор заявляє, що немає конфлікту інтересів щодо даного дослідження, включаючи фінансовий, особистий, авторський або будь-який інший, який міг би вплинути на дослідження, а також на результати, представлені в даній статті.

Фінансування

Дослідження було проведено без фінансової підтримки.

Доступність даних

Усі дані доступні в цифровій або графічній формі в основному тексті рукопису.

Використання штучного інтелекту

Автор підтверджує, що при створенні даної роботи він не використовував технології штучного інтелекту.

Список використаної літератури

1. Gilad-Bachrach, R., Dowlin, N., Laine, K., Lauter, K., Naehrig, M., & Wernsing, J. (2016). CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy. *Proceedings of Machine Learning Research,* 48. https://proceedings.mlr.press/v48/gilad-bachrach16.pdf
2. Brutzkus, A., Elisha, O., & Gilad-Bachrach, R. (2019). Low Latency Privacy Preserving Inference. *Proceedings of the 36th International Conference on Machine Learning, PMLR 97,* 812–821. https://proceedings.mlr.press/v97/brutzkus19a.html
3. Lloret-Talavera, G., Jordà, M., Servat, H., Boemer, F., Chauhan, C., Tomishima, S., Shah, N. N., & Peña, A. J. (2022). Enabling Homomorphically Encrypted Inference for Large DNN Models. *IEEE Transactions on Computers,* 71(5), 1145–1155. https://doi.org/10.1109/TC.2021.3076123
4. Shen, H., Xu, Q., Yu, B., Yang, Y., & He, W. (2025). Bootstrapping in approximate fully homomorphic encryption: a research survey. *Cybersecurity,* 8, 87. https://doi.org/10.1186/s42400-025-00384-3
5. Knott, B., Venkataraman, S., Hannun, A., Sengupta, S., Ibrahim, M., & van der Maaten, L. (2021). CrypTen: Secure Multi-Party Computation Meets Machine Learning. *Advances in Neural Information Processing Systems,* 34, 4961–4973. https://proceedings.neurips.cc/paper_files/paper/2021/file/2754518221cfbc8d25c13a06a4cb8421-Paper.pdf
6. Zhang, R., Zheng, Z., & Bao, W. (2025). Practical Secure Inference Algorithm for Fine-tuned Large Language Model Based on Fully Homomorphic Encryption. *arXiv:2501.01672.* https://doi.org/10.48550/arXiv.2501.01672
7. Zhu, H., Suzuki, T., & Yamana, H. (2023). Performance Comparison of Homomorphic Encrypted Convolutional Neural Network Inference Among HElib, Microsoft SEAL and OpenFHE. *IEEE Asia-Pacific Conference on Computer Science and Data Engineering.* https://doi.org/10.1109/CSDE59766.2023.10487709
8. Takeshita, J., Koirala, N., McKechney, C., & Jung, T. (2025). HEProfiler: an in-depth profiler of approximate homomorphic encryption libraries. *Journal of Cryptographic Engineering,* 15(2). https://doi.org/10.1007/s13389-025-00377-5

Tsap V. B. The problem of scaling cryptographic algorithms for artificial intelligence.

The article examines the problem of scaling cryptographic algorithms in artificial intelligence systems that process sensitive data. The study aims to identify the factors limiting the practical applicability of such solutions and to formulate evaluation criteria. It analyzes fully homomorphic encryption, secure multi-party computation, differential privacy, and trusted execution environments. The paper proposes an operational model of scalability that accounts for response latency, memory overhead, data transfer volume, interaction rounds, and computational complexity. It also introduces a scale of practical applicability and an evaluation protocol for determining the applicability limits of cryptographic solutions across different model types and usage scenarios. The study shows that weak scalability is mainly caused by high computational costs, increased memory requirements, network traffic, protocol round complexity, and nonlinear operations. It concludes that privacy-preserving artificial intelligence requires hybrid architectures and models initially adapted to protected cryptographic environments.

Keywords: large language models, cryptographic algorithms, fully homomorphic encryption, secure multi-party computation, data privacy.

Отримано: 08.03.2026 Прийнято: 26.03.2026 Опубліковано: 30.04.2026