isu-conference.com

ISU

INTERNATIONAL SCIENTIFIC UNITY

COLLECTION OF SCIENTIFIC PAPERS

ISSUE №43

3RD INTERNATIONAL SCIENTIFIC AND PRACTICAL CONFERENCE

THE FUTURE OF SCIENCE, TECHNOLOGY AND ECONOMY

OCTOBER 29-31, 2025
SOFIA, BULGARIA

UDC 001(08)

The Future of Science, Technology and Economy: Collection of Scientific Papers with Proceedings of the 3rd International Scientific and Practical Conference. International Scientific Unity. October 29-31, 2025. Sofia, Bulgaria. 470 p.

ISBN 979-8-89704-988-2 (series)
DOI 10.70286/ISU-29.10.2025

The conference is included in the Academic Research Index ReserchBib International catalog of scientific conferences.

The collection of scientific papers presents the materials of the participants of the 3rd International Scientific and Practical Conference "The Future of Science, Technology and Economy" (October 29-31, 2025. Sofia, Bulgaria).

ISBN 979-8-89704-988-2 (series)

ISU
INTERNATIONAL SCIENTIFIC UNITY

SECTION: INFORMATION TECHNOLOGY AND CYBERSECURITY

ПРИНЦИПИ ТА АРХІТЕКТУРИ LLM ДЛЯ ПАРСИНГУ ДАНИХ

Флегантов Леонід
кандидат фізико-математичних наук, доцент,
професор кафедри інформаційних систем та технологій
Левченко Юрій
здобувач вищої освіти магістерського рівня
Спеціальність «Інформаційні системи та технології»
Полтавський державний аграрний університет, Україна

У сучасну епоху інтелектуалізації обробки інформації великі мовні моделі (Large Language Models, LLM) стали одним із найважливіших досягнень у галузі штучного інтелекту. Здатність LLM до семантичного аналізу, синтаксичного розбору, контекстуального розуміння тексту та генерації узгоджених структур даних відкриває нові перспективи для автоматизації процесів парсингу даних, тобто структурованого вилучення інформації з неструктурованих джерел [1-3].

Традиційно парсинг даних реалізовувався через жорстко детерміновані алгоритми: регулярні вирази, граматики контекстно-вільних мов, або інструменти, побудовані на основі XPath, CSS-селекторів чи DOM-дерев. Проте ці методи мають суттєві обмеження у роботі з неструктурованими або напівструктурованими текстами – наприклад, із динамічними HTML-сторінками, текстами з помилками або неоднозначними структурами. На противагу їм LLM дозволяють виконувати семантичний парсинг, у якому провідну роль має інтерпретація контексту [8].

У цій роботі розглядаються основні принципи побудови LLM, їх архітектурні особливості, алгоритмічні підходи до парсингу, а також приклади сучасних LLM, зокрема таких, як GPT, BERT, T5, LLaMA та їх застосування у задачах структуризації даних.

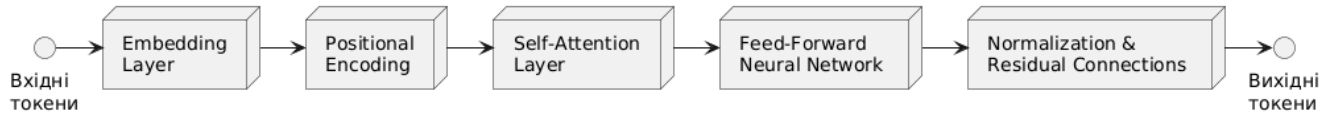

Рисунок 1. Основні компоненти архітектури трансформера.

Основною особливістю LLM є їх трансформерна архітектура (transformer architecture), запропонована в роботі [1]. Вона базується на механізмі self-attention (самоуваги), який дозволяє моделі аналізувати зв'язки між словами незалежно від їх позиції у реченні (рис. 1). Ця архітектура усуває обмеження послідовної обробки даних, характерної для RNN (Recurrent Neural Network) та

LSTM (Long Short-Term Memory), і забезпечує масштабованість моделей до мільярдів параметрів (табл. 1).

Таблиця 1 – Еволюція LLM

Модель	Тип	Архітектура	Кількість параметрів	Основна особливість
BERT (2019)	Bidirectional Encoder	Transformer Encoder	~340 млн	Двонаправлений контекст, Masked Language Modeling
GPT-3 (2020)	Autoregressive Decoder	Transformer Decoder	175 млрд	Одностороння генерація, навчання на великому корпусі
T5 (2020)	Seq2Seq Encoder–Decoder	Повний Transformer	11 млрд	Завдання «Text-to-Text», універсальність
LLaMA 3 (2024)	Autoregressive Decoder	Оптимізований Transformer	70 млрд	Енергоефективність, швидке донавчання

Представлений у моделі Embedding Layer (шар векторизації) перетворює слова або токени у числові вектори фіксованої розмірності. Positional Encoding (позиційне кодування) – додає інформацію про порядок слів у реченні, що необхідно для відтворення синтаксису. Self-Attention Layer (механізм самоуваги) – оцінює важливість кожного слова у контексті всього речення. Feed-Forward Neural Network (багатошарова нейронна мережа) – нелінійно перетворює результати уваги. Normalization & Residual Connections (нормалізація та залишкові зв'язки) – стабілізують навчання та пришвидшують збіжність.

Механізм Self-Attention дозволяє LLM формувати уявлення про контекст слова, беручи до уваги всі інші слова у реченні. Формально, він обчислює матрицю ваг, яка визначає вплив кожного токена на інші токени. Для кожного токена моделюються три вектори – Query (Q), Key (K), Value (V), а результат уваги визначається як:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

де d_k – розмірність векторів ключів [1].

Застосування LLM у парсингу даних базується на переході від синтаксичного аналізу до семантичного розуміння. Це означає, що сучасні LLM дозволяють не просто виділяти патерни у тексті, а й розпізнавати смислові зв'язки та структурувати дані, зокрема у форматах JSON, CSV, SQL або RDF [5].

Загалом, провідні сучасні LLM еволюціонували у напрямі підвищення контекстуальної точності, масштабованості та універсальності. Аналіз їх еволюції свідчить про поступовий перехід від двонаправлених до автогенеративних підходів у побудові великих мовних моделей.

Застосування LLM для парсингу даних спирається на принципи, до яких належать: контекстуальна інтерпретація – здатність моделі враховувати значення фрази у контексті всього документа; генеративний підхід – результати парсингу подаються як текстові відповіді, які можна безпосередньо перетворювати на структуровані дані; можливість few-shot і zero-shot навчання –

LLM не потребують повного донавчання на кожному новому типі даних; інтеграція з інструментами Data Engineering – LLM можуть взаємодіяти з API, базами даних, ETL-процесами тощо [9, 10].

Для задач парсингу даних важливо, як різні архітектурні підходи впливають на здатність LLM аналізувати текстову структуру, виявляти патерни і перетворювати неструктуровану інформацію у структуровану форму.

Модель BERT (Bidirectional Encoder Representations from Transformers), розроблена у Google AI [2], є двонаправленим енкодером трансформера, що дозволяє враховувати контекст слова як зліва, так і справа. Основна ідея BERT – Masked Language Modeling (MLM), тобто під час навчання певний відсоток токенів у реченні маскується, і модель навчається передбачати ці слова, використовуючи весь контекст. Це робить модель BERT ефективною для розуміння синтаксису та семантики, що є важливим у процесі семантичного парсингу.

Головні особливості BERT: архітектура моделі складається лише з енкодерів трансформера; використовується позиційне кодування для збереження порядку токенів; завдяки двонаправленості модель формує повний контекст для кожного слова; добре підходить для класифікації текстів, вилучення сутностей (NER) та синтаксичного аналізу, що може використовуватись для попереднього етапу парсингу даних. Зауважимо, що Embedding Layer в архітектурі BERT має три частини (Token, Segment, Position Embedding) (рис. 2).

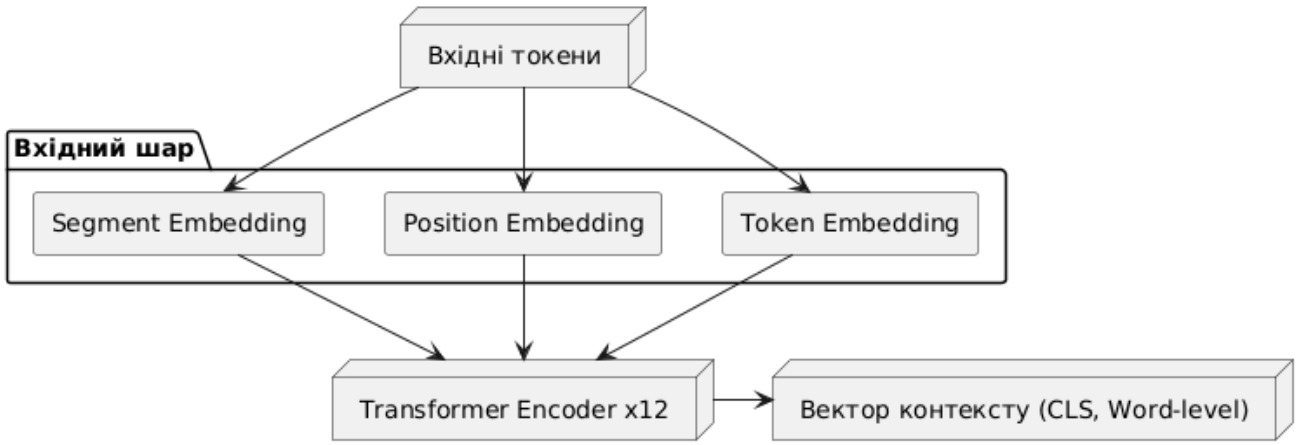

Рисунок 2. Архітектура моделі BERT.

Модель GPT (Generative Pre-trained Transformer), розроблена компанією OpenAI [3], грунтується на декодерній архітектурі трансформера. На відміну від BERT, GPT є авторегресивною моделлю, тобто прогнозує наступне слово, спираючись на попередні. Така архітектура набагато краще підходить для генеративних завдань, включно з автоматичним створенням парсерів, побудовою структурованих шаблонів, SQL-запитів, JSON-виходів тощо [9].

Головні особливості моделі GPT (рис. 3): архітектура складається лише з декодерів трансформера; навчання проводиться у каузальному режимі (Causal LM) – передбачення наступного токена за контекстом; модель підтримує few-shot / zero-shot навчання, що спрощує адаптацію під конкретні типи даних; у

новіших версіях (GPT-3.5, GPT-4, GPT-5) реалізовано механізми довгої пам'яті і розширене вікно контексту, що дозволяє аналізувати великі текстові документи.

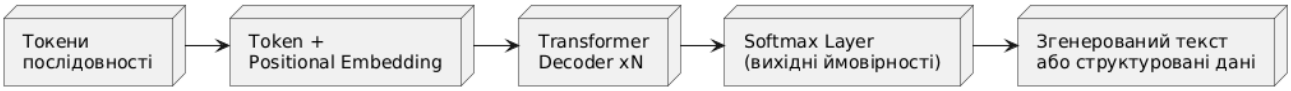

Рисунок 3. Архітектура моделі GPT.

Модель T5 (Text-to-Text Transfer Transformer), розроблена у Google Research [5], використовує підхід «text-to-text»: будь-яке завдання (класифікація, переклад, парсинг) формулюється як перетворення одного тексту в інший. Це робить T5 дуже зручною для семантичного парсингу, де завдання може бути описане як «перетвори текст у структуру даних».

Головні особливості моделі T5 (рис. 4): повна encoder-decoder архітектура; єдина уніфікована парадигма «input text → output text»; можливість виконання завдань типу «extract entities», «generate JSON», «summarize data»; у контексті Data Engineering може використовуватись для автоматичного створення парсерів для різних форматів (XML, HTML, API-відповідей).

Рисунок 4. Архітектура моделі T5 типу Encoder–Decoder.

Модель LLaMA (Large Language Model Meta AI) (рис. 5), розроблена Meta AI [6], оптимізована для ефективності та роботи на обмежених ресурсах. Архітектурно LLaMA є декодерною моделлю, подібною до GPT, але з низкою покращень: використання RoPE (Rotary Positional Embeddings) замість класичного позиційного кодування; оптимізована структура уваги для зменшення енергоспоживання; підтримка великого контекстного вікна (до 128 тис. токенів у LLaMA 3); підтримка інструкційного навчання (instruction tuning), що дозволяє ефективно виконувати завдання парсингу без додаткового донавчання [7].

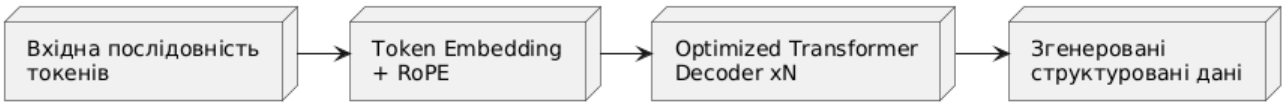

Рисунок 5. Архітектура моделі LLaMA.

Для ефективного застосування LLM у парсингу важливо враховувати не тільки їх архітектурні переваги, але й існуючі обмеження (табл. 2).

Таблиця 2 – Архітектурні переваги та обмеження LLM

Модель	Тип архітектури	Контекстна обробка	Придатність до парсингу	Основні переваги	Недоліки
BERT	Encoder-only	Двонаправлена	Висока точність розпізнавання сутностей	Добрий семантичний аналіз	Не генерує вихідний текст
GPT	Decoder-only	Одностороння	Генеративна побудова структур	Few-shot навчання, масштабованість	Обмежене розуміння повного контексту
T5	Encoder-Decoder	Повна	Семантичний та синтаксичний парсинг	Універсальність завдань	Високі вимоги до ресурсів
LLaMA	Decoder-only (оптимізована)	Одностороння	Гнучка генерація структур	Енергоефективність, адаптивність	Потребує великих даних для навчання

На підставі розгляду архітектурних принципів та еволюції великих мовних моделей (від BERT до LLaMA) було встановлено, що вони надають значні переваги для задач структурованого вилучення інформації порівняно з традиційними детермінованими методами. До основних переваг LLM належать: здатність обробляти неструктуровані текстові масиви без необхідності ручного налаштування правил парсингу; використання генеративного підходу, що дозволяє формувати структуровані вихідні дані; підвищення точності вилучення інформації завдяки контекстній інтерпретації тексту; та гнучка адаптація до нових форматів даних за допомогою методів few-shot / zero-shot навчання.

Серед розглянутих архітектур найбільш перспективними для автоматизованого парсингу виявились GPT і LLaMA завдяки їх здатності генерувати структуровані формати (JSON, XML, SQL). У той же час, моделі BERT і T5 є ефективними для попереднього семантичного аналізу, виділення сутностей і перетворення текстових даних на логічні структури, що формує базу для подальшої генерації структурованих результатів

Список використаних джерел

1. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. Attention is All You Need. Advances in Neural Information Processing Systems (NeurIPS), 2017. Vol. 30. P. 5998–6008. DOI: 10.48550/arXiv.1706.03762.
2. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT, 2019. P. 4171–4186. DOI: 10.48550/arXiv.1810.04805.
3. Brown T. B., Mann B., Ryder N., et al. Language Models are Few-Shot Learners. (GPT-3). Advances in Neural Information Processing Systems, 2020. Vol. 33. P. 1877–1901. DOI: 10.48550/arXiv.2005.14165.

4. OpenAI. GPT-4 Technical Report. OpenAI, 2023. 98 p. DOI: 10.48550/arXiv.2303.08774.
5. Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., Zhou Y., Li W., Liu P. J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (T5). Journal of Machine Learning Research, 2020. Vol. 21, No. 140. P. 1–67. DOI: 10.48550/arXiv.1910.10683.
6. Touvron H., Lavril T., Izacard G., et al. LLaMA: Open and Efficient Foundation Language Models. Meta AI Research, 2023. 44 p. DOI: 10.48550/arXiv.2302.13971.
7. Touvron H., MartinL., Stone, K., et al. LLaMA 2: Open Foundation and Fine-Tuned Chat Models. Meta AI Research Report, 2024. 56 p. DOI: 10.48550/arXiv.2307.09288.
8. Bommasani R., Hudson D. A., Adeli E., et al. On the Opportunities and Risks of Foundation Models. Stanford Center for Research on Foundation Models (CRFM), 2021. 227 p. DOI: 10.48550/arXiv.2108.07258.
9. OpenAI. Using GPT Models for Data Structuring and Parsing. OpenAI Developer Documentation, 2024. URL: https://platform.openai.com/docs/guides /structured-output (дата звернення: 20.10.2025).
10. Qintong Zhang, Victor Shea-Jay Huang, Bin Wang, Junyuan Zhang, Zhengren Wang, Hao Liang, Shawn Wang, Matthieu Lin, Wentao Zhang, Conghui He. Document Parsing Unveiled: Techniques, Challenges, and Prospects for Structured Information Extraction. (2024). URL: https://arxiv.org/html/2410.21169v1 (дата звернення: 21.10.2025).

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА КІБЕРБЕЗПЕКА

Нечипорук Владислав Вадимович
курсант 2 курсу
Тимошишин Роман Едуардович
курсант 2 курсу
Бундус Валерій Валерійович
курсант 2 курсу
Навчально-науковий інститут №4
Науковий керівник:
Лучик Василь Єфремович
д. е. н., професор
Кафедра протидії кіберзлочинності
Харківський національний університет внутрішніх справ, Україна

У сучасних умовах питання кібербезпеки є одним із найбільш актуальних у контексті як забезпечення безпеки окремих осіб та установ, так і забезпечення національної безпеки України. Під час повномасштабної війни, як свідчить практика, постійно зростає кількість кібератак та різноманітних кіберзагроз з боку ворога, спрямованих на те, щоб дестабілізувати загальну ситуацію в країні,

Collection of Scientific Papers
with Proceedings of the 3rd International Scientific and Practical Conference
«The Future of Science, Technology and Economy»
October 29-31, 2025
Sofia, Bulgaria

Organizing committee may not agree with the authors' point of view.
Authors are responsible for the correctness of the papers' text.

Contact details of the organizing committee:
Sole Proprietor Viktoriia Tsiundyk
E-mail: info@isu-conference.com
URL: https://isu-conference.com/

Certificate of the subject of the publishing business: ДК №7980 of 03.11.2023.