ISSN 2522-9818 (print)
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2024. No. 2 (28)*

УДК 004.8 DOI: https://doi.org/10.30837/2522-9818.2024.2.048

В. Волоховський

АНАЛІЗ МЕТОДІВ ТРЕНУВАННЯ ВУЗЬКОСПРЯМОВАНИХ МОВНИХ МОДЕЛЕЙ У СФЕРІ ГЕНЕРАЦІЇ ДОГОВОРІВ

Предметом дослідження є моделі та методи машинного навчання для генерації договорів в умовах обмежених ресурсів і способи порівняння та оцінювання їх ефективності. **Мета роботи** – аналіз підходів до розроблення вузькоспрямованих великих мовних моделей та визначення оптимального методу створення незалежних спеціалізованих систем, що дають змогу генерувати договори різними мовами в різних правових системах. У статті розв'язуються такі **завдання**: визначення наявних компаній та рішень, виявлення підходів до створення текстів природною мовою, аналіз способів оцінювання та порівняння таких систем, виявлення обмежень і недоліків сучасних рішень і підходів, пошук оптимального методу розроблення систем за умови обмежених ресурсів. **Досягнуті результати**: досліджено підходи до генерації текстів природною мовою та їх особливості; визначено архітектуру "Трансформер" як сучасний стандарт у сфері генерації текстової інформації; розглянуто види моделей на основі зазначеної архітектури; проаналізовано джерела даних для їх тренування; розглянуто методи адаптації моделей у вузькоспрямованих галузях; виявлено способи порівняння та оцінювання ефективності виконання різних завдань мовними моделями; виявлено недоліки наявних спеціалізованих мовних моделей і неповноту наборів метрик оцінювання завдання генерації договорів. Унаслідок аналітичного експерименту було визначено, що метод пошуково-доповненої генерації є найбільш оптимальним для розв'язання поставленого завдання в заданих умовах. Проведений експеримент та його результати можуть бути використані як основа для подальших досліджень у сфері розроблення вузькоспрямованих мовних моделей за умови обмежених ресурсів. **Висновки**. У статті проаналізовано методи генерації текстової інформації природною мовою за допомогою сучасних підходів машинного навчання. Виокремлено їх переваги й недоліки для невеликих компаній та наукових установ, які мають обмежені матеріальні та людські ресурси. Як приклад у роботі розглянуто спеціалізовану юридичну галузь і проблему генерації договорів та визначено найбільш оптимальний метод її розв'язання.

Ключові слова: велика мовна модель; генерація природної мови; договір; юридичний документ.

Вступ

Договори відіграють важливу роль у повсякденному житті людей та в ефективному функціонуванні компаній. Їх правильне формулювання та оформлення відповідно до чинного законодавства та інтересів зацікавлених сторін потребує вузькоспрямованих знань у галузі права. Складання угод вручну та внесення змін потребує чимало часу навіть для досвідчених юристів і може призвести до помилок під час копіювання його частин з інших документів та до низького рівня повторного використання цього договору в інших ситуаціях. Витрачаючи час на ці завдання, фахівці приділяють менше уваги клієнтам та розумінню їхніх потреб, що призводить до погіршення якості наданих послуг і неефективного використання часу й навичок.

Перелічимо проблеми, з якими найчастіше стикаються юристи та компанії, що працюють з договорами [1]:

– значна кількість часу та зусиль на складання та розуміння договорів;

– залежність неюридичних відділів компанії, таких як відділ кадрів і продажів, від команди юристів для укладання угод;

– неузгодженість між документами всередині компанії через різноманітність їх видів і формулювань.

Автоматизація процесу генерації та аналізу договорів за допомогою сучасних технологій може подолати ці проблеми. Методи машинного навчання, зокрема великі мовні моделі, показали себе найкраще у виконанні завдань генерації природної мови. Однак наявні комерційні рішення зазвичай перебувають у приватній власності компаній, а розроблення систем на основі мовних моделей потребує значних обсягів вузькоспрямованих тренувальних даних і чималих обчислювальних ресурсів, що обмежує потенціал використання цього методу невеликими компаніями, науковими інститутами та окремими дослідниками.

Аналіз останніх досліджень і публікацій

У сфері розуміння, оброблення та генерації природної мови в останні роки проведено чимало

Сучасний стан наукових досліджень та технологій в промисловості. 2024. №2 (28)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

досліджень. Автори роботи [2] запропонували архітектуру "Трансформер", яка замінила наявні рекурентні нейронні мережі, а в праці [3] її вдосконалено з метою отримання кращих результатів у галузі генерації тексту. Базові моделі, створені внаслідок багатьох досліджень, використовуються в різних галузях [4, 5]. Розроблення та вивчення вузькоспрямованих моделей показали переваги адаптації моделей до певної галузі порівняно із загальними моделями [6, 7]. Автори студій [8, 9] описують, як сучасні великі мовні моделі, що мають сотні мільярдів параметрів, без додаткового тренування можуть виконувати нові завдання в різних спеціалізованих галузях, використовуючи тільки інформацію та інструкції, які отримують з контексту. У роботі [10] сформовано великий корпус юридичних документів різними мовами для різних юрисдикцій, що спрощує для інших дослідників доступ до вузькоспрямованої інформації окресленої галузі. Автори праці [11] пропонують додати нову метрику оцінювання виконаних завдань з оброблення природної мови. Розроблені набори метрик дають змогу комплексно оцінювати знання моделей, їх здатність виконувати різні завдання та дотримуватися правил [12, 13]. У фахових дослідженнях сформовано аналогічні набори метрик для оцінювання ефективності моделей в обробленні та розумінні юридичних документів [11, 14, 15].

Визначення не розв'язаних раніше частин загальної проблеми. Мета роботи й завдання

Оброблення та генерація текстів природною мовою є нелегким завданням. Наявні підходи та моделі складні в розробленні, потребують значних обсягів обчислювальних, матеріальних і людських ресурсів. Ці фактори зумовили обмежене використання таких систем.

Поява архітектури "Трансформер", що є більш ефективною за попередні види нейронних мереж, відкрила нові можливості її застосування в галузях науки, виробництва та бізнесу. Попри це тренування й розроблення таких моделей стали складнішими, відповідно потребують ще більше ресурсів, які мають тільки корпорації або великі дослідницькі установи. Створення чималої кількості базових мовних моделей, що є у відкритому доступі, дали змогу багатьом невеликим компаніям і науковим інститутам досліджувати та розробляти нові системи.

Оскільки базові моделі натреновані на загальнодоступній інформації, їх можливості та навички в спеціалізованих галузях є обмеженими. Достатня кількість робіт присвячена дослідженню способів створення та адаптації моделей до вузькоспрямованих галузей. Інші праці присвячені розробленню моделей для певної царини та їх адаптації до вузького набору завдань і знань.

Автори студій у сфері оброблення юридичних документів і договорів приділяють увагу використанню тільки одного підходу – безпосередньому тренуванню параметрів моделі. Такі дослідження, як зазначалось раніше, потребують значних ресурсів і часу. У них зазвичай беруть участь багато науковців із різних університетів і компаній. Тому для невеликих комерційних і наукових організацій можливості використання новітніх технологій та підходів усе ще залишаються досить обмеженими. З огляду на окреслену проблему визначимо мету дослідження.

Метою роботи є аналіз підходів до розроблення вузькоспрямованих великих мовних моделей та визначення оптимального методу створення незалежних спеціалізованих систем, що уможливлюють генерацію договорів різними мовами в різних правових системах.

Сформулюємо *завдання*, що необхідно виконати для досягнення поставленої мети:

– визначення наявних компаній та рішень у цій сфері;

– виявлення підходів до створення текстів природною мовою;

– аналіз способів оцінювання та порівняння таких систем;

– виявлення обмежень і недоліків наявних рішень та підходів;

– пошук оптимального методу розроблення систем за умови обмежених ресурсів.

Перелічені завдання спрямовані на глибокий аналіз методів генерації текстових даних у вузькоспрямованих галузях та виявлення способів вирішення окресленої проблеми на прикладі генерації договорів у юридичній сфері.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2024. No. 2 (28)

Матеріали та методи

Огляд ринку

Перелічимо наявні компанії на ринку автоматизації юридичних документів (*legal document automation*):

– *Juro* – використовує велику мовну модель *GPT (Open AI)* для створення угод і уможливлює взаємодію із системою за допомогою *AI*-чату, підтримує англійську мову [16];

– *Luminance* – застосовує власну модель *Legal Inference Transformation Engine*, уможливлює автоматичне ведення переговорів щодо змісту договору (*Autopilot*) та підтримує понад 80 мов [17];

– *Icertis* – упроваджує систему *Icertis ExploreAI Service*, що поєднує можливості великих мовних моделей *Open AI*, власні *AI*-системи та *Icertis Data Lake* для аналізу та генерації документів, має підтримку декількох мов;

– *Oneflow* – використовує *GPT (Open AI)* і надає велику кількість готових шаблонів договорів, підтримує 10 європейських мов.

Серед основних функцій сервісів, що застосовують методи машинного навчання, можна виокремити такі:

– автоматичне складання договорів різних типів;

– аналіз і резюмування;

– підтримка правил та обмежень під час складання документів.

Для генерації договорів компанії найчастіше використовують нейронні мережі, а саме великі мовні моделі (*LLM*), розроблені спеціально для автоматизації документів, як у разі *Luminance*, або моделі загального призначення – *GPT* із додатковими модифікаціями для роботи в цій сфері.

З відкритих джерел встановлено, що модель *Legal Inference Transformation Engine* була натренована на понад 150 мільйонах перевірених юридичних документах. Більшість інших систем основані на моделі *OpenAI GPT*, тому ключові розбіжності між ними полягають у додатковому налаштуванні з використанням інформації в юридичній галузі. Однак щодо цих систем не вдалось отримати більш детальної інформації про підходи та дані, що використовуються для розроблення, оскільки вони є закритими та належать до інтелектуальної власності компаній.

Отже, бачимо, що сфера автоматизації юридичних документів активно адаптує та використовує сучасні підходи до оброблення природної мови, надаючи нові можливості формулювання угод, значно спрощуючи та прискорюючи цей процес. Автоматизація зазначеного процесу зменшує необхідну кількість юридичного персоналу компанії, даючи змогу неюридичним відділам складати угоди.

Аналіз методів генерації природної мови

Розглянемо підходи до розв’язання проблеми генерації природної мови (*NLG*). Найбільш ранні методи моделювання та генерації тексту використовували модульні архітектури, побудовані на наборі модулів, об’єднаних послідовно в системі. На зміну цим методам прийшли підходи, основані на плануванні, що визначали послідовність одного або декількох кроків для досягнення конкретної мети. Мета розбивалася на менші завдання, які виконувалися за допомогою дій, що мали певний набір умов та ефектів, які впливали на кінцевий результат. Наступним етапом розвитку були стохастичні підходи, що активно використовували набори даних для виявлення статистичних залежностей у природній мові для подальшої генерації тексту.

Нейронні мережі були найбільш популярними та ефективними підходами для розв’язання проблем моделювання тексту й машинного перекладу. Розглянемо деякі з найбільш використаних моделей:

– рекурентні нейронні мережі (*RNN*);

– мережі з довгою короткочасною пам’яттю (*LSTM*);

– вентильні рекурентні вузли (*GRU*);

– варіаційні автокодувальники (*VAE*);

– згорткові нейронні мережі (*CNN*);

– генеративні змагальні мережі (*GAN*) [18].

Попри широке використання окреслених підходів, вони мають певні недоліки, які обмежують їх ефективність для виконання завдання генерації текстів.

Рекурентні нейронні мережі мають тенденцію зазнавати проблеми зникнення градієнта на довгих послідовностях, що може обмежувати їх здатність до генерації довгих текстів зі складною структурою. Їх можна використовувати для генерації коротких фрагментів тексту або в завданнях, де контекст не потребує глибокого аналізу.

LSTM та *GRU* краще працюють із довгими залежностями в тексті, порівняно зі звичайними *RNN*, завдяки своїм механізмам керування пам’яттю та можуть бути ефективнішими для генерації більш складних текстів. Проте через свою рекурентну природу навчання моделей відбувається послідовно, а можливості паралелізації є обмеженими.

Сучасний стан наукових досліджень та технологій в промисловості. 2024. №2 (28)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

Варіаційні автокодувальники використовуються для генерації нових зразків зазвичай із деякою змінністю відповідно до вхідної інформації та застосовуються для створення різноманітності у тексті. Основною проблемою цього підходу є згортання розходження Кульбака – Лейблера, що призводить до генерації вихідних даних незалежно від вхідних.

Хоча згорткові нейронні мережі зазвичай використовуються для оброблення зображень, вони можуть бути адаптовані для роботи з текстом та бути ефективними для його генерації з огляду на локальні шаблони та структури. Вони широко не використовувалися через проблеми з вибором архітектури та оптимального значення гіперпараметрів.

GAN-моделям властива здатність генерувати реалістичні дані, зокрема природну мову, і вони можуть бути застосовані для створення тексту з високою якістю та натуралізмом. Проте навчання таких моделей є більш складним через недиференційну природу дискретних символів, а також модель може генерувати поверхневі, повторювані та недалекоглядні відповіді.

Сучасним підходом до генерації тексту є архітектура "Трансформер", що має структуру "кодувальник-декодувальник" [2]. Основним принципом запропонованого підходу є механізм самоуваги, що визначає важливість частин вхідної послідовності токенів до інших слів у цій послідовності. Зазначений підхід, на відміну від попередніх, дає змогу:

– виконувати обчислення паралельно, зменшуючи необхідний час тренування;

– обробляти послідовності тексту більшої довжини без втрати контексту.

З огляду на отримані метрики (*benchmarks*) ця архітектура демонструє кращі результати виконання певних завдань оброблення природної мови, наприклад, машинного перекладу [2].

Для генерації тексту використовується варіація оригінальної архітектури, що має тільки декодувальник та генерує тестову послідовність на основі початкового вхідного тексту *(prompt*).

Ці характеристики зумовили стрімкий розвиток моделей на основі архітектури "Трансформер". Нові можливості генерації природної мови викликали значний інтерес різних сфер бізнесу, що сприяло впровадженню систем на основі мовних моделей у багатьох індустріях за останні декілька років.

Аналіз джерел даних

Для навчання моделей використовують великі за обсягом набори даних, що можуть містити терабайти корпусів тексту.

Зміст тренувальних даних *загального* призначення не обмежується однією галуззю, що робить їх більш придатними для навчання загальних моделей. Ці дані можна поділити на декілька основних класів [19]:

– текст вебсторінок, що отримується за допомогою сканування великої кількості вебсторінок в інтернеті та визначається чималим обсягом, динамічністю змісту, наявністю різних мов та багатьох тем, високим рівнем неперевіреної інформації (*Common Crawl, C4, mC4*);

– книги, яким властива висока якість змісту, граматична та лексична точність, значна довжина тексту, наявність складних мовних зворотів, термінів і фразеологізмів (*Anna's Archive, BookCorpusOpen*);

– академічні матеріали, які мають високий рівень професіоналізму та знань, що зумовлює виняткову якість їх робіт (*arXiv, PubMed Central*);

– програмний код, що містить приклади використання мов програмування для розв’язання різних завдань (*BIG-QUERY* та *phi-1*);

– дані соціальних медіа, що охоплюють створені користувачами дописи, коментарі та діалоги й визначаються потенційною присутністю шкідливої інформації, такої як упередження, дискримінація та насильство (*Pushshift Reddit* та *OpenWebText*);

– дані енциклопедій, які є в онлайн-енциклопедіях або інших базах знань і яким властивий певний рівень надійності інформації (*Wikipedia*).

Вузькоспрямовані моделі потребують навчальної інформації, яка містить знання, особливі для певної галузі.

У юриспруденції *Pile of Law* та *MultiLegalPile* є найбільшими наборами даних [10]. *Pile of Law* містить близько 256 ГБ правової та адміністративної інформації. Для її формування використано 35 різних джерел, зокрема юридичні документи, судові висновки, публікації державних установ, контракти, статути, нормативні акти, журнали справ тощо. *MultiLegalPile* об’єднує 689 ГБ юридичних документів 24 мовами з 17 різних юрисдикцій. Він містить набір *Pile of Law*, а також декілька натренованих моделей на основі *RoBERTa* та *Longformer*. Ці набори даних можна використовувати для тренування як одномовних, так і багатомовних моделей і адаптувати їх під законодавство різних країн.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2024. No. 2 (28)

Базові моделі

Розроблення мовних моделей потребує значних ресурсів, часу та навичок. Більшість компаній та дослідницьких установ не мають достатньо матеріальних і людських ресурсів для тренування нових моделей. А час, необхідний для досягнення бажаних результатів, може виявитися занадто тривалим.

Для тренування *LLaMA* – базової моделі з відкритим доступом, розробленої компанією *Meta*, з 65 мільярдами параметрів – використано 2048 *Nvidia* A100 *GPU*, кожен з яких мав 80 ГБ пам'яті *RAM* [20]. Тренування на наборі даних обсягом 1,4 мільярда токенів тривало приблизно 21 день. Вартість застосування таких ресурсів за оцінками може становити приблизно 2,4 мільйона доларів [21].

Щоб зробити великі мовні моделі доступнішими та відкрити перспективи ширших досліджень навіть для окремих фахівців, корпорації та великі компанії, що спеціалізуються на мовних моделях, публікують у відкритий доступ базові, попередньо натреновані моделі (*foundational models*):

– *Google: LaMDA, Chinchilla, Gemma* [22, 23, 24];

– *Meta: LLaMA 2* [4];

– *Mistral: Mixtral* [5].

Зазвичай перелічені моделі містять загальні знання та можуть виконувати різні завдання, проте їх використання для вузькоспрямованих завдань та в специфічних сферах є обмеженим. Подальше налаштування може дати змогу моделі набувати нових знань та виконувати нові завдання, потребуючи значно менше ресурсів для навчання, порівняно з розробленням нової моделі.

Вузькоспрямовані комерційні моделі

Розроблення вузькоспрямованих моделей у певних галузях показало їх перевагу над моделями загального призначення.

Компанія *Microsoft* розробила продукт *Microsoft Sales Copilot*, що допомагає менеджерам з продажу збільшувати ефективність роботи та створювати персоналізовані пропозиції клієнтам, використовуючи *GPT*-моделі від *OpenAI* з додатковим налаштуванням у сфері продажу [25]. Система дає змогу створювати персоналізовані листи на основі інформації про клієнта, деталей угоди та попередньої комунікації із замовниками. Також можна аналізувати наради й зустрічі, додаючи виділення ключових слів, теми розмов, конкурентів, ключові метрики оцінювання ефективності та запропоновані завдання.

Компанія *Luminance* спільно з дослідниками з Кембриджського університету розробили модель *Legal Pre-Trained Transformer*, що була натренована на понад 150 мільйонах перевірених юридичних документів [26]. Ця модель дає змогу генерувати документи, зокрема контракти й договори, аналізувати та резюмувати їх зміст (*summarization*). Чат-бот на основі зазначеної моделі може вести переговори щодо змісту контракту, перевіряти компанії на відповідність вимогам та обмеженням, виявляти зміни в нових версіях [17].

Основним недоліком розглянутих моделей є те, що вони належать компаніям і доступ до них є закритим. Це обмежує можливості їх використання, покращення та незалежного оцінювання іншими дослідниками. Тому постає питання розроблення власних спеціалізованих систем.

Методи тренування та налаштування вузькоспрямованих моделей

Для розроблення вузькоспрямованих мовних моделей упроваджують різні підходи, що відрізняються за кількістю необхідних ресурсів і часу, ефективністю та обсягом тренувальних даних.

Тренування нової вузькоспрямованої моделі, наприклад *Legal Pre-Trained Transformer*, є найбільш складним підходом. Цей процес схожий до того, як розробляються базові моделі або моделі загального призначення.

Для налаштування параметрів моделі на основі архітектури "Трансформер" застосовується механізм самоуваги [2]:

$$Attention(Q,K,V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V\ , \qquad (1)$$

де d – розмірність моделі;

n – кількість запитів;

m – кількість пар "ключ – значення";

$Q \in \mathbb{R}^{n\times d_k}$ – матриця запитів;

$K \in \mathbb{R}^{m\times d_k}$ – матриця ключів;

$V \in \mathbb{R}^{m\times d_v}$ – матриця значень;

$\sqrt{d_k}$ – коефіцієнт масштабування.

Механізм багатоголової самоуваги паралельно обчислює функції самоуваги (1) для отримання

Сучасний стан наукових досліджень та технологій в промисловості. 2024. №2 (28)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

інформації з декількох підпросторів подання на різних позиціях:

$$MultiHead\left(Q,K,V\right)=Concat\left(head_1,\ldots,head_h\right)W^o, \quad (2)$$

$$head_i=Attention\left(QW_i^Q,KW_i^K,VW_i^V\right),$$

де h – кількість голів самоуваги;

W^o – фінальна матриця ваг, отримана від усіх голів самоуваги;

W_i^Q – матриця ваг запитів;

W_i^K – матриця ваг ключів;

W_i^V – матриця ваг значень.

Кожен шар моделі додатково містить нейронну мережу прямого зв'язку:

$$FFN\left(x\right)=\mathrm{Re}LU\left(xW_1+b_1\right)W_2+b_2\ , \quad (3)$$

де x – вхідний вектор;

$W_1\in\mathbb{R}^{d\times d_m}$ і $W_2\in\mathbb{R}^{d_m\times d}$ – матриці ваг лінійних трансформацій.

Основною розбіжністю тренування вузько-спрямованих моделей від моделей загального призначення є використання властивих для певної галузі даних, що навчають модель нових знань і завдань. Перевагами цього методу є висока ефективність моделі, повний контроль над тренувальною інформацією та можливість адаптації до різних завдань і потреб. Контроль над навчальними даними також дає змогу визначити мови, які підтримуватимуться системою. Як і в ситуації тренування базових моделей, недоліками окресленого підходу є необхідність у значних обсягах тренувальної інформації, а також обчислювальних ресурсах та часі, що зумовлюють вартість розроблення моделі. Збір вузькоспрямованих тренувальних наборів може викликати труднощі, оскільки дані можуть бути недоступні для широкого використання, належати до інтелектуальної власності інших компаній або їх якість та обсяг можуть бути недостатніми для навчання великих систем.

У процесі роботи з формування набору даних *MultiLegalPile* було натреновано нові вузькоспрямовані моделі в галузі права *Legal-XLM-R* за допомогою зібраної інформації. Вони демонструють значно вищі показники під час оцінювання на наборах метрик *LEXTREME* та *LexGLUE*, якщо порівнювати з моделями загального застосування (*DeBERTa* та *RoBERTa*), а також мають переваги щодо інших вузькоспрямованих моделей у цій галузі (*Legal-BERT* та *CaseLaw-BERT*) [10].

Тонке налаштування (*fine-tuning*) великих мовних моделей для конкретних галузей передбачає адаптацію попередньо навченої моделі для кращого розуміння та генерації тексту, що відповідає цьому домену. Розглянемо основні підходи тонкого налаштування.

Під час *повного тонкого налаштування* (*full-fine tuning*) попередньо навчена мовна модель налаштовується для роботи з інформацією, властивою для певної галузі [23]. Цей підхід передбачає оновлення всіх параметрів моделі під час тренування, що дає змогу їй отримати знання в окресленій сфері та вивчити нові закономірності. Проте налаштування вимагає великого обсягу інформації для ефективного виявлення та розуміння особливостей цільової галузі. Хоча цей підхід може бути ефективним, він також потребує значних обчислювальних ресурсів та часу на тренування, насамперед для великих моделей, що мають мільярди параметрів. Нещодавні дослідження у сфері оптимізації тренувального процесу показали, що за допомогою нових методів, зокрема *LOMO (LOw-Memory Optimization)*, можна виконувати налаштування моделей навіть із десятками мільярдів параметрів лише на декількох *GPU*-процесорах [27].

Для розв'язання проблем попереднього підходу розроблено *параметро-ефективні методи* тонкого налаштування (*PEFT*) [28]. Вони передбачають упровадження різних методів глибокого навчання для зменшення кількості параметрів, які треба навчити, і водночас зберігають схожий рівень ефективності повного тонкого налаштування. *PEFT*-методи оновлюють лише незначну кількість додаткових параметрів або підмножину попередньо навчених, зберігаючи наявні знання моделі та адаптуючи їх до нового завдання, що зменшує ризик катастрофічного забування. Застосування повного тонкого налаштування на специфічних тренувальних даних, обсяг яких зазвичай набагато менший, ніж набір даних, що використовувався для тренування базової моделі, може призвести до перенавчання. *PEFT*-методи дають змогу розв'язати цю проблему способом вибіркового оновлення попередньо навчених параметрів або їх повного "заморожування".

Виокремлюють декілька основних груп параметро-ефективних методів тонкого налаштування.

Підхід *адитивного* тонкого налаштування (*Additive Fine-tuning*) додає нові параметри під час налаштування моделі для нового завдання, поділяється на методи на основі адаптерів (*adapter-based*), м'якого налаштування на основі підказок (*soft prompt-based*) тощо, наприклад *AttentionFusion* і *AdapterFusion*. Метод

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2024. No. 2 (28)

послідовного адаптера додає мережі адаптерів після шарів самоуваги та нейронної мережі прямого зв'язку [28]. Кожен адаптер є модулем нижчого рангу, що містить низхідну проєкцію, нелінійну функцію активації та висхідну проєкцію, а також залишкового з'єднання. Для вхідної інформації h результат обчислюється таким чином:

$$h = h + \left(\mathrm{Re}\,LU\left(hW_{down}\right)\right)W_{up}\ , \qquad (4)$$

де h – вхідний вектор;

$W_{down} \in \mathbb{R}^{d\times r}$ – низхідна проєкція;

$W_{up} \in \mathbb{R}^{r\times d}$ – висхідна проєкція;

$\mathrm{Re}\,LU\left(x\right) = \max\left(0, x\right)$ – нелінійна функція активації.

Підхід *часткового* тонкого налаштування (*Partial Fine-tuning*) зменшує кількість налаштовуваних параметрів способом обрання критичних попередньо навчених ваг і відкидання неважливих, поділяється на методи оновлення зміщення (*Bias Update*), маскування попередньо натренованих ваг (*Pretrained Weight Masking*) та дельта-маскування ваг (*Delta Weight Masking*). Метод порогового маскування використовує визначене значення порогу τ для побудови матриці двійкових масок M з метою вибору попередньо навчених ваг W шарів самоуваги та нейронної мережі прямого зв'язку з допомогою множення елементів матриці:

$$W' = W \odot M\ , \qquad (5)$$
$$M = \mathbb{I}_{s_{i,j}>\tau},$$

де $\odot$ – добуток Адамара;

τ – визначений поріг;

$W \in \mathbb{R}^{d\times k}$ – ваги моделі;

$S \in \mathbb{R}^{d\times k}$ – матриця, що ініціалізується випадковими рівномірно розподіленими дійсними числами; якщо елемент матриці перевищує поріг τ, відповідній позиції в матриці двійкових масок присвоюється значення 1, інакше – 0.

Підхід *перепараметризованого* тонкого налаштування (*Reparameterized Fine-tuning*) застосовує перетворення низького рангу, щоб зменшити кількість параметрів для навчання. Він поділений на методи розкладання низького рангу (*Low-rank Decomposition*), наприклад *Intrinsic SAID*, і методи на основі *LoRA* (*Delta-LoRA*). Метод *LoRA* (*Low-Rank Adaptation*) додає дві матриці низького рангу, що оновлюються в процесі тренування [29]. Низхідна та висхідна матриці проєкцій використовуються паралельно з матрицями запитів Q, ключів K і значень V у шарі самоуваги моделі. Під час навчання оновлюються тільки матриці W_{down} та W_{up}. Нові ваги обчислюються таким чином:

$$\Delta W = W_{down}W_{up}\ , \qquad (6)$$

де $W_{down} \in \mathbb{R}^{d\times r}$ – низхідна проєкція;

$W_{up} \in \mathbb{R}^{r\times k}$ – висхідна проєкція;

$r \ll \left\{k, d\right\}$ – обрана розмірність.

Під час генерації результату моделлю вагові коефіцієнти ΔW об'єднуються з вихідною матрицею ваг

$$h = W_0 x + \Delta W x\ , \qquad (7)$$

де x – вхідний вектор;

$W_0 \in \mathbb{R}^{d\times k}$ – ваги моделі;

$\Delta W \in \mathbb{R}^{d\times k}$ – додатково натреновані ваги за допомогою методу *LoRA*.

Підхід *гібридного* тонкого налаштування (*Hybrid Fine-tuning*) поєднує різні підходи *PEFT* для посилення їх переваг і зменшення недоліків. Розрізняють такі підходи до комбінації методів: ручні, що вимагають складного дизайну (*MixAnd-Match Adapter, Compacter*), та автоматичні, які використовують структурний пошук (*AutoPEFT*). Метод *Compacter* розроблено на основі підходів адаптера, розкладання низького рангу та параметризованого гіперкомплексного множення (*parameterized hypercomplex multiplication*) [30]. Він має схожу до адаптерів структуру, однак замінює низхідну та висхідну проєкції шаром параметризованого гіперкомплексного множення низького рангу (*low-rank parameterized hypercomplex multiplication*). Значення ваг обчислюються таким чином:

$$W = \sum_{i=1}^{n} A_i \otimes B_i\ , \qquad (8)$$

де $A_i \in \mathbb{R}^{n\times n}$ – загальна матриця ваг для всіх шарів адаптера;

$B_i \in \mathbb{R}^{\frac{k}{n}\times\frac{d}{n}}$ – матриця ваг, притаманна для окремих шарів адаптера;

$\otimes$ – добуток Кронекера.

Підхід *уніфікованого* тонкого налаштування (*Unified Fine-tuning*) є єдиною структурою для додавання різноманітних методів налаштування в єдину архітектуру (*AdaMix* та *ProPETL*) і зазвичай використовує один *PEFT*-метод, а не комбінацію різних. Метод *AdaMix* містить M модулів адаптації, що долучаються в кожен шар моделі:

Сучасний стан наукових досліджень та технологій в промисловості. 2024. №2 (28)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

$A_{ij} : i \in [1, L]1, L, j \in [1, M]$ – j-й модуль адаптації в i-му шарі [31]. Під час тренування на кожному кроці випадковим способом обирається пара матриць проєкцій для i-го шару:

$$A_i = \{W_{ij}^{up}, W_{ik}^{down}\}, \quad (9)$$

$$B_i = \{W_{ij'}^{up}, W_{ik'}^{down}\},$$

де W_{ij}^{down} – низхідна проєкція i-го шару j-го модуля адаптації;

W_{ij}^{up} – висхідна проєкція i-го шару j-го модуля адаптації.

Отже, вся вхідна інформація обробляється тим самим набором модулів. Використовуючи матриці, наведені у (9), відбувається така трансформація:

$$h = h + f\left(h \cdot W^{down}\right) W^{up} \quad (10)$$

де h – вхідний вектор;

W^{down} – низхідна проєкція;

W^{up} – висхідна проєкція;

$f(x)$ – функція активації.

Серед основних переваг параметро-ефективних методів можна виокремити необхідність у значно меншому обсязі вузькоспрямованих тренувальних даних і обчислювальних ресурсів, оскільки цей метод тренує тільки незначну частину параметрів. Недоліками підходу є менша ефективність моделі та необхідність додаткового тренування параметрів для виконання різних завдань, властивих для обраної сфери. Набір мов, з якими модель може взаємодіяти, обмежується тими, які підтримує базова модель. Додавання нових мов потребує значних обсягів інформації, щоб навчити модель різних аспектів та особливостей її використання. Зауважимо, що це майже неможливо за допомогою цього методу через незначну кількість ваг, які оновлюються.

Метод *пошуково-доповненої генерації* (*Retrieval-Augmented Generation*) ефективно застосовується в галузях, що потребують доступу до актуальних і точних знань, які постійно оновлюються та змінюються [19]. Він оснований на використанні зовнішніх джерел, що містять перевірену та точну інформацію. Під час генерації відповіді система шукає та отримує додаткові релевантні дані із цих джерел. На підставі запиту, контексту й додаткової інформації модель дає більш точні та аргументовані відповіді, основані на достовірних фактах.

RAG-модель описується таким чином: на основі вхідної послідовності x система отримує текстові документи z і використовує їх як додатковий контекст під час генерації цільової послідовності y. Основними елементами системи є пошуковий компонент і генератор. Пошуковий компонент $p_\eta(z|x)$ знаходить k найбільш релевантних фрагментів тексту на основі x. *Dense Passage Retriever* використовує щільний кодувальник, який перетворює фрагменти тексту в d-вимірні вектори дійсних чисел та створює індекс для всіх фрагментів тексту, що застосовуються для пошуку [32]. Компонент генератора $p_\theta(y_i|x, z, y_{1:i-1})$, який є великою мовною моделлю, генерує токени тестової послідовності на основі контексту попередніх $i-1$ токенів $y_{1:i-1}$, вхідних даних x та фрагментів z, отриманих від пошукового компонента.

RAG-модель на основі токенів дозволяє компоненту генератора обирати вміст із кількох документів під час генерації відповіді. Найбільш релевантні k-документи отримуються за допомогою пошукового компонента. Наступний токен обчислюється таким чином:

$$p(y|x) \approx \prod_i^N \sum_{z \in top-k(p(\cdot|x))} p_\eta(z|x) p_\theta(y_i|x, z, y_{1:i-1}), \quad (11)$$

де $top-k\left(p(\cdot|x)\right)$ – k-документи з набору z найвищої попередньої ймовірності $p_\eta(z|x)$.

Можна виокремити три групи підходів, що використовуються для реалізації зазначеного методу [33]:

– наївний (*naive*) – найпростіший підхід, оснований на індексації даних, отриманні їх зі сховища та генерації відповіді; можуть траплятися галюцинації, а отримана інформація може бути нерелевантною та повторюваною;

– розширений (*advanced*) – зосереджується на покращенні якості отриманої інформації способом додаткового оброблення перед пошуком і після нього; використовує індексування за допомогою підходів ковзного вікна, дрібної сегментації та додавання метаданих;

– модульний (*modular*) – містить різноманітні стратегії для вдосконалення компонентів способом додавання нових модулів: пошукових, злиття (*fusion*), пам'яті, маршрутизації, передбачення та адаптації до завдання.

Метод пошуково-доповненої генерації має низку переваг. Він значно дешевший для впровадження

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2024. No. 2 (28)

та використання, оскільки застосовує наявні моделі як основу та не потребує додаткового тренування. Завдяки отриманню даних із баз знань система має доступ до інформації, яка з'явилась після тренування моделі та не була додана до її тренувального набору. Звернення до перевірених та достовірних джерел інформації зменшує рівень галюцинацій.

Основними викликами під час його використання можна назвати отримання якісної інформації з баз знань, їх оцінювання та впорядкування за актуальністю. Метод не передбачає тренування та зміни параметрів моделі, тому загальні знання та завдання, які модель може виконувати, залишаються незмінними. Так само підтримка різних мов обумовлюється обраною базовою моделлю.

Метод *навчання з контексту* (*in-context learning*) дає змогу мовним моделям виконувати нові завдання на основі інструкцій природної мови та кількох прикладів, що демонструють виконання нового завдання та надаються моделі через вхідний текст (*prompt*). Зазначений метод оснований на можливостях великих мовних моделей розвивати широкий набір навичок і здібностей під час навчання, а потім використовувати їх під час генерації відповідей, швидко адаптуватися до нового завдання та його розпізнавання.

У межах цього підходу можна виокремити три категорії інструкцій, що допомагають моделі навчитися нового завдання:

– навчання на основі декількох прикладів (*few-shot learning*) – модель отримує інструкції, що описують завдання, та демонстрації його виконання (зазвичай від 10 до 100);

– навчання на основі одного прикладу (*one-shot learning*) – передбачає інструкції та лише одну демонстрацію;

– навчання без прикладів (*zero-shot learning*) – надає тільки інструкції за допомогою природної мови, проте не наводить жодних прикладів.

Дослідження на прикладі *GPT*-3 та інших моделей меншого розміру виявили таке: що більша модель, то кращі результати можна отримати, застосовуючи навчання з контексту [8]. Додаткові експерименти з використання цього методу та *GPT*-4 моделі в галузі медицини продемонстрували можливості досягнення кращих результатів, якщо порівнювати з вузькоспрямованою моделлю *Med-PaLM* 2, яка пройшла тонке налаштування на декількох спеціалізованих наборах даних [7, 9]. Під час цього дослідження розроблено підхід *Medprompt*, що формує вхідний текст, використовуючи декілька різних технік: динамічний вибір декількох прикладів (*few-shot*), автоматично створений ланцюжок думок (*chain of thought*) та ансамблевий метод вибору (*choice shuffling ensemble*). Його було застосовано до інших галузей знань, зокрема до права. Ефективність оцінювалася за допомогою набору метрик *MMLU (Massive Multitask Language Understanding)* [27]. Експеримент показав, що без додаткового налаштування *GPT*-4 отримав оцінку 68,3. Унаслідок додаткового застосування підходу *Medprompt* досягнуто значно вищий результат – 72,9 бала.

Основною перевагою зазначеного методу є відсутність налаштування параметрів, що унеможливлює витрати на обчислювальні ресурси, необхідні для тренування та зміни ваг. Також він потребує значно меншого обсягу спеціалізованої інформації порівняно з тонким налаштуванням. Складання прикладів та інструкцій є інтуїтивним процесом, схожим на те, як люди взаємодіють між собою під час опису та визначення завдань. Це дає змогу експертам певної галузі, не маючи знань про машинне навчання та мовні моделі, впроваджувати метод для виконання нових вузькоспрямованих завдань.

Недоліками навчання з контексту є обмеження розміру вхідного тексту, що лімітує кількість прикладів і розмір інструкцій, які можна навести. Ефективність цього методу зазвичай є меншою порівняно з тренуванням нової моделі та тонким налаштуванням. Набір підтримуваних мов зазвичай не може бути розширеним.

Аналіз ефективності моделей

Для оцінювання ефективності та надійності роботи мовних моделей використовують різні метрики. Найпростішими з них є *ROUGE* та *BERTScore*.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) вимірює повноту відповіді та застосовується для оцінювання якості резюмування та генерації тексту. Існує декілька варіантів цієї метрики, зокрема *ROUGE-N* (вимірює перекриття n-грам), *ROUGE-L* (вимірює найдовшу спільну підпослідовність) та *ROUGE-W* (вимірює зважену найдовшу спільну підпослідовність).

BERTScore обчислює схожість між згенерованим та еталонним текстом [34]. Обидві послідовності подаються за допомогою векторів вбудовувань (*embeddings*). Схожість між ними обчислюється за допомогою косинуса подібності (*cosine similarity*).

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2024. №2 (28) ISSN 2524-2296 (online)

Зі зростанням необхідності оцінювати більші та складніші моделі розроблено набори тестів (*benchmarks*), що оцінюють їх рівень знань і можливості виконання завдань.

SuperGLUE – фреймворк для оцінювання здібностей моделей розуміти тексти природною мовою [13]. Він містить різні групи завдань англійською мовою: відповіді на запитання з одним чи декількома правильними варіантами, класифікація тексту та визначення причинно-наслідкових зв'язків.

MMLU (Massive Multitask Language Understanding) – набір тестів загального призначення, розроблений для перевірки знань, набутих на етапі навчання моделі із застосуванням методу навчання з контексту [27]. Він охоплює 57 різних предметів, як загальних (математика та історія), так і більш фахових (право та медицина). Складність завдань коливається від початкового до професійного рівня, що перевіряють як знання про світ, так і здатність вирішувати проблеми.

Для юриспруденції також розроблено фахові набори тестів.

LexGLUE (Legal General Language Understanding Evaluation) – набір тестів для юридичної галузі, що містить сім різних наборів даних англійською мовою та перевіряє здібності моделей у виконанні завдань класифікації тексту та відповіді на запитання з декількома варіантами [14].

LegalBench – набір тестів, що містить 162 метрики [11]. Вони охоплюють шість типів юридичних завдань, які по-різному оцінюють модель: виявлення проблем, згадування правил, їх застосування, прийняття рішення на основі правил, тлумачення тексту, розуміння риторики.

LEXTREME – інший фреймворк, що передбачає 11 наборів даних, які охоплюють 24 мови [15]. Він містить три групи завдань: класифікація з визначення одного та декількох класів і розпізнавання іменованих об'єктів у юридичних документах.

Можна помітити, що більшість загальних і спеціалізованих наборів метрик зосереджуються на аналітичних здібностях моделі та її розумінні природної мови: класифікація та маркування тексту, резюмування, прогнозування висновків судових справ, розпізнавання іменованих сутностей. Однак вони майже не приділяють увагу завданню генерації нового контенту та його оцінюванню.

Пошук оптимального підходу

Визначимо критерії порівняння різних підходів тренування великих мовних моделей для використання у сфері генерації договорів.

Обсяг обчислювальних ресурсів визначає кількість серверів, що оптимізовані для машинного навчання та мають відповідні *GPU*-процесори, необхідні для навчання моделі за прийнятний час. Критерій є категоріальним, оскільки абсолютне порівняння не буде ефективним для оцінювання через значну розбіжність для різних методів. Використаємо такі категорії: великий – кількість обчислювальних серверів перевищує 10, малий – менше ніж 10 серверів, відсутній – не потребує додаткових обчислювальних ресурсів для тренування.

Обсяг тренувальних даних визначає розмір вузькоспрямованого тренувального набору, що дасть змогу моделі набути знань і навичок у новій сфері й ефективно виконувати поставлені завдання. Критерій є категоріальним, оскільки абсолютне порівняння не буде ефективним для оцінювання через значну варіацію необхідних обсягів для різних методів. Використаємо такі категорії: великий – кількість тренувальних даних понад 10 ГБ, середній – від 10 МБ до 10 ГБ, малий – менше ніж 10 МБ. Варто зауважити, що для методу навчання з контексту обсяг інформації обмежується розміром вхідного тексту, який модель може обробити [35]. Аналогічно й метод пошуково-доповненої генерації може застосовувати обмежений обсяг інформації для кожного запиту. Проте загальний набір даних, що використовується для пошуку найбільш релевантної інформації, може бути значно більшим.

Час тренування визначає обсяг необхідних часових ресурсів для налаштування моделі з використанням спеціалізованих даних. Критерій є категоріальним, оскільки абсолютне порівняння не буде ефективним для оцінювання через значну розбіжність значень для різних методів. Використаємо такі категорії: великий – тренування потребує понад 24 год, середній – від 1 до 24 год, малий – менше ніж 10 хв. Варто зауважити, що критерій має одноразовий ефект на тренування моделі за допомогою методів, що змінюють ваги системи. Для методів пошуково-доповненої генерації, оскільки вони не модифікують параметри, значення цього критерію визначає час, необхідний для пошуку й оброблення релевантної інформації, і впливає на кожен запит до моделі в процесі генерації відповіді. У разі використання

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2024. No. 2 (28)

методу навчання з контексту тренувальні дані є частиною запиту, тому навчання моделі за допомогою цих даних є частиною генерації відповіді.

Контроль тренувальних даних визначає можливість значно впливати на інформацію, що застосовується для навчання моделі. Критерій є категоріальним. Використаємо такі категорії: повний – розробники самостійно визначають тип даних, їх походження та обсяги, частковий – можливість визначати тільки частину тренувальної інформації, зазвичай вузькоспрямовані набори даних, відсутній – неможливість значно вплинути на інформацію, яку модель застосовує під час генерації. Зауважимо, що для методів пошуково-доповненої генерації та навчання з контексту тренувальною є інформація, що модель отримує через вхідний текст. Хоча ці дані впливають на процес генерації відповіді, вони не запам'ятовуються моделлю та не будуть використані для наступних запитів.

Підтримка великих документів визначає можливість обробляти значну за обсягом інформацію на десятки сторінок. Критерій має два значення: так чи ні. Зауважимо, що методи тренування та тонкого налаштування обробляють такі документи під час налаштування параметрів. За умови впровадження інших методів розмір документа обмежується розміром вхідного тексту моделі. У разі застосування методу пошуково-доповненої генерації можна зменшити вплив цього фактора завдяки використанню тільки найбільш релевантних і значущих частин документа.

Підтримка нових мов визначає можливість упровадження моделі в різних країнах та юрисдикціях. Критерій має два значення: так – існує можливість навчити систему нової мови та ефективно її використовувати; ні – набір мов обмежується базовою моделлю, а впровадження вузькоспрямованої інформації іншими мовами не матиме значного ефекту.

Обсяг обчислювальних ресурсів для актуалізації знань визначає кількість серверів, необхідних для оновлення даних моделі у разі зміни законодавства та правил. Критерій є категоріальним, оскільки абсолютне порівняння не буде ефективним для оцінювання через значну розбіжність для різних методів. Використаємо такі категорії: великий – кількість обчислювальних серверів перевищує 10, малий – менше ніж 10 серверів, відсутній – не потребує додаткових обчислювальних ресурсів. У разі застосування підходів тренування нової моделі, повного та параметро-ефективного тонкого налаштування модель набуває нових знань унаслідок повторного процесу налаштування ваг. За умови використання інших методів модель набуває актуальних знань під час генерації кожної відповіді та не потребує додаткового тренування.

Рівень підтримки уваги до бізнес-правил визначає здатність моделі дотримуватися заданих правил і обмежень. Критерій є категоріальним. Використаємо такі категорії: високий – можливість підтримки багатьох правил різної складності, середній – базові нескладні правила, низький – базові нескладні правила в обмеженому обсязі. У разі застосування методів тренування та тонкого налаштування модель під час тренування навчається виконувати завдання аналізу та уваги до правил. Підхід навчання з контексту може використовувати шаблони вхідного тексту, що містять обмеження, а здатність системи дотримуватися їх визначається можливостями базової моделі. Аналогічно й підхід пошуково-доповненої генерації, крім отримання інформації з джерел на запит користувача, може визначити релевантні правила та додати їх до контексту запиту.

Значення критеріїв для кожного методу налаштування моделі наведено в табл. 1.

Таблиця 1. *Критерії оцінювання методів тренування вузькоспрямованих мовних моделей*

Критерій	Тренування моделі	Повне тонке налаштування	Параметро-ефективне тонке налаштування	Пошуково-доповнена генерація	Навчання з контексту
Обсяг обчислювальних ресурсів	Великий	Великий	Малий	Відсутній	Відсутній
Обсяг тренувальних даних	Великий	Великий	Середній	Середній	Малий
Час тренування	Тривалий	Тривалий	Нетривалий	Відсутній	Відсутній
Контроль тренувальних даних	Повний	Частковий	Частковий	Відсутній	Відсутній
Підтримка великих документів	Так	Так	Так	Так	Ні
Підтримка нових мов	Так	Так	Ні	Ні	Ні
Обсяг обчислювальних ресурсів для оновлення знань	Великий	Великий	Малий	Відсутній	Відсутній
Рівень підтримки уваги до бізнес-правил	Високий	Високий	Високий	Середній	Низький

Сучасний стан наукових досліджень та технологій в промисловості. 2024. №2 (28)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

Використовуючи розглянуті вище критерії, визначимо найбільш оптимальний підхід за допомогою методу адитивного згортання з ваговими коефіцієнтами.

Спочатку перетворимо категоріальні критерії до числових значень:

– обсяг обчислювальних ресурсів: великий – 2, малий – 1, відсутній – 0;

– обсяг тренувальних даних: великий – 2, середній – 1, малий – 0;

– час тренування: тривалий – 2, нетривалий – 1, відсутній – 0;

– контроль тренувальних даних: повний – 2, частковий – 1, відсутній – 0;

– підтримка великих документів: так – 1, ні – 0;

– підтримка нових мов: так – 1, ні – 0;

– обсяг обчислювальних ресурсів для оновлення знань: великий – 2, малий – 1, відсутній – 0;

– рівень підтримки уваги до бізнес-правил: високий – 2, середній – 1, низький – 0.

З метою спрощення подання інформації використаємо абревіатури для методів налаштування моделей: тренування моделі – ТМ, повне тонке налаштування – ПТН, параметро-ефективне тонке налаштування – ПЕТН, пошуково-доповнена генерація – ПДГ, навчання з контексту – НК.

Далі виконаємо нормування кожного з них на проміжку [0, 1], використовуючи мінімальне й максимальне значення, та подамо результати в табл. 2.

Таблиця 2. *Нормовані критерії*

Критерій	ТМ	ПТН	ПЕТН	ПДГ	НК
Обсяг обчислювальних ресурсів	1	1	0,5	0	0
Обсяг тренувальних даних	1	1	0,5	0,5	0
Час тренування	1	1	0,5	0	0
Контроль тренувальних даних	1	0,5	0,5	0	0
Підтримка великих документів	1	1	1	1	0
Підтримка нових мов	1	1	0	0	0
Обсяг обчислювальних ресурсів для оновлення знань	1	1	0,5	0	0
Рівень підтримки уваги до бізнес-правил	1	1	1	0,5	0

Оскільки наведені критерії потребують як мінімізації, так і максимізації, перетворимо їх таким чином, щоб більше значення мало більшу корисність відповідно до поставленого завдання. Перетворенню підлягають такі критерії: обсяг обчислювальних ресурсів, обсяг тренувальних даних, час тренування, обсяг обчислювальних ресурсів для оновлення знань. Значення оновлених критеріїв подані в табл. 3.

Таблиця 3. *Критерії максимізації корисності*

Критерій	ТМ	ПТН	ПЕТН	ПДГ	НК
Обсяг обчислювальних ресурсів	0	0	0,5	1	1
Обсяг тренувальних даних	0	0	0,5	0,5	1
Час тренування	0	0	0,5	1	1
Контроль тренувальних даних	1	0,5	0,5	0	0
Підтримка великих документів	1	1	1	1	0
Підтримка нових мов	1	1	0	0	0
Обсяг обчислювальних ресурсів для оновлення знань	0	0	0,5	1	1
Рівень підтримки уваги до бізнес-правил	1	1	1	0,5	0

Визначимо вагові коефіцієнти для кожного критерію. Було проведено експертне дослідження серед фахівців в Україні та за кордоном, які спеціалізуються на великих мовних моделях. За допомогою опитування було визначено такі вагові коефіцієнти:

– обсяг обчислювальних ресурсів: 0,15;

– обсяг тренувальних даних: 0,15;

– час тренування: 0,15;

– контроль тренувальних даних: 0,08;

– підтримка великих документів: 0,15;

– підтримка нових мов: 0,07;

– обсяг обчислювальних ресурсів для оновлення знань: 0,1;

– рівень підтримки уваги до бізнес-правил: 0,15.

Результати досліджень

Наведемо результати виконання завдання з оптимізації за допомогою адитивного згортання з ваговими коефіцієнтами та визначимо максимально корисний метод тренування вузькоспрямованих моделей (табл. 4).

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2024. No. 2 (28)

Таблиця 4. *Корисність методів тренування вузькоспрямованих мовних моделей*

Метод	Корисність
ТМ	0,45
ПТН	0,41
ПЕТН	0,615
ПДГ	**0,7**
НК	0,55

З огляду на досягнуті результати аналізу можемо зробити висновок, що метод пошуково-доповненої генерації є найбільш оптимальним за заданих умов. Методи ТМ та ПТН, що потребують тренування всіх параметрів моделі, виявилися менш корисними через значний обсяг необхідних ресурсів для їх налаштування. Метод ПЕТН є другим за корисністю та використовує значно менше ресурсів для налаштування, якщо порівнювати з ТМ та ПТН, проте все ж таки потребує певного тренування параметрів. Метод НК є менш корисним за ПЕТН, але для виконання деяких завдань простота його використання та швидкість налаштування можуть бути вагомими факторами. Хоча ПДГ має певні недоліки, зокрема залежність від можливостей базової моделі та ефективності алгоритму пошуку релевантної інформації, вони компенсуються відсутністю необхідності тренування моделі та простотою актуалізації знань.

Висновки

У процесі дослідження проаналізовано підходи до розроблення вузькоспрямованих великих мовних моделей, виявлено їх переваги, недоліки та обмеження, визначено найбільш оптимальний метод створення незалежних спеціалізованих систем, що дають змогу генерувати договори різними мовами в різних правових системах.

Аналіз попередніх досліджень та ринку виявив, що більшість моделей у відкритому доступі, натренованих для роботи в юридичній галузі, мають архітектуру кодувальника, яка не є ефективною для завдань генерації тексту. А наявні моделі з архітектурою декодувальника є або закритими, або подані моделями загального призначення, що потребують додаткової адаптації в обраній галузі.

Щоб порівняти підходи між собою, було сформовано набір критеріїв і наведено значення для кожного з методів налаштування моделей. Для визначення найбільш оптимального та корисного підходу впроваджено метод лінійного адитивного згортання з ваговими коефіцієнтами.

Унаслідок аналітичного експерименту виявлено, що метод пошуково-доповненої генерації є найбільш оптимальним за заданих умов, хоча програє більш складним підходам у гнучкості налаштування. Значно менший обсяг тренувальних ресурсів і наявний набір можливостей дають змогу ефективно адаптувати цей підхід для вузькоспрямованих галузей. Водночас метод параметро-ефективного тонкого налаштування за наявності додаткових часових і обчислювальних ресурсів може бути так само ефективним для адаптації в юридичній галузі.

Додатково визначено, що більшість спеціалізованих наборів метрик зосереджуються на аналітичних завданнях класифікації, резюмування та розуміння тексту, однак майже не приділяють уваги оцінюванню якості генерації нового контенту.

Перспективи подальшого розвитку

У подальших дослідженнях плануємо приділити увагу тренуванню моделей на основі деяких розглянутих методів і порівняти їх ефективність для виконання завдань аналізу, розуміння та генерації договорів у юридичній галузі. Також плануємо дослідити можливості поєднання декількох методів у межах однієї системи таким чином, щоб підсилити переваги кожного підходу та позбавитися недоліків для підвищення ефективності системи у виконанні поставлених завдань.

Зважаючи на те, що наявні набори метрик не мають повноцінних можливостей для оцінювання ефективності генерації юридичних документів, цей напрям також потребує подальших досліджень і розвитку.

Подяка

Автор висловлює подяку Збройним силам України за можливість написати повноцінну роботу під час повномасштабного вторгнення Російської Федерації на територію України. Також дякує науковому керівникові О. С. Назарову за підтримку та допомогу під час написання роботи.

Сучасний стан наукових досліджень та технологій в промисловості. 2024. №2 (28)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

Список літератури

1. Generative AI for Legal Contracts. Nasdaq. URL: https://www.nasdaq.com/articles/generative-ai-for-legal-contracts (дата звернення: 27.05.2024).
2. Vaswani A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. Attention is all you need. Advances in neural information processing systems. *31st Conference on Neural Information Processing Systems*. 2017. 30. DOI: https://doi.org/10.48550/arXiv.1706.03762
3. Devlin J., Chang M.W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. 2018. DOI: https://doi.org/10.48550/arXiv.1810.04805
4. Touvron H. та ін. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*. 2023. DOI: https://doi.org/10.48550/arXiv.2307.09288
5. Jiang A. Q. та ін. Mixtral of experts. *arXiv preprint arXiv:2401.04088*. 2024. DOI: https://doi.org/10.48550/arXiv.2401.04088
6. Wu S. та ін. BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564*. 2023. DOI: https://doi.org/10.48550/arXiv.2303.17564
7. Singhal K. та ін. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*. 2023. DOI: https://doi.org/10.48550/arXiv.2305.09617
8. Brown T. та ін. Language models are few-shot learners. *Advances in neural information processing systems*. 2020. № 33. P. 1877–1901. DOI: https://doi.org/10.48550/arXiv.2005.14165
9. Nori H. та ін. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv preprint arXiv:2311.16452*. 2023. DOI: https://doi.org/10.48550/arXiv.2311.16452
10. Niklaus J., та ін. Multilegalpile: A 689gb multilingual legal corpus. *arXiv preprint arXiv:2306.02069*. 2023. DOI: https://doi.org/10.48550/arXiv.2306.02069
11. Guha N. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*. 2024. № 36. DOI: https://doi.org/10.48550/arXiv.2308.11462
12. Hendrycks D. та ін. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*. 2020. DOI: https://doi.org/10.48550/arXiv.2009.03300
13. Wang A., та ін. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*. 2019. № 32. DOI: https://doi.org/10.48550/arXiv.1905.00537
14. Chalkidis I., та ін. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. 2022. С. 4310–4330. DOI: https://aclanthology.org/2022.acl-long.297
15. Niklaus J., та ін. LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain. *Findings of the Association for Computational Linguistics: EMNLP 2023*. 2023. С. 3016–3054. DOI: https://aclanthology.org/2023.findings-emnlp.200
16. Mabey R. Unveiling our legal AI Assistant. Juro. URL: https://juro.com/blog/legal-ai-assistant (дата звернення: 10.03.2024).
17. Browne R. An AI just negotiated a contract for the first time ever and no human was involved. CNBC. URL: https://www.cnbc.com/2023/11/07/ai-negotiates-legal-contract-without-humans-involved-for-first-time.html (дата звернення: 10.03.2024).
18. Ian G. та ін. Generative adversarial nets. *Advances in neural information processing systems*. 2014. № 27. DOI: https://doi.org/10.48550/arXiv.1406.2661
19. Lewis P. та ін. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*. 2020. № 33. P. 9459–9474. DOI: https://doi.org/10.48550/arXiv.2005.11401
20. Touvron H., та ін. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*. 2023. DOI: https://doi.org/10.48550/arXiv.2302.13971
21. Vanian J., Leswing K. ChatGPT and generative AI are booming, but the costs can be extraordinary. CNBC. URL: https://www.cnbc.com/2023/03/13/chatgpt-and-generative-ai-are-booming-but-at-a-very-expensive-price.html (дата звернення: 27.05.2024).
22. Thoppilan R. та ін. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*. 2022. DOI: https://doi.org/10.48550/arXiv.2201.08239
23. Hoffmann J. та ін. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*. 2022. DOI: https://doi.org/10.48550/arXiv.2203.15556
24. Mesnard T. та ін. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*. 2024. DOI: https://doi.org/10.48550/arXiv.2403.08295

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2024. No. 2 (28)*

25. Microsoft Copilot for Sales. Microsoft. URL: https://www.microsoft.com/en-us/ai/microsoft-sales-copilot (дата звернення: 27.05.2024).
26. Luminance's Legal Pre-Trained Transformer. Luminance. URL: https://www.luminance.com/technology.html (дата звернення: 27.05.2024).
27. Lv K. та ін. Full parameter fine-tuning for large language models with limited resources. *arXiv preprint arXiv:2306.09782*. 2023. DOI: https://doi.org/10.48550/arXiv.2306.09782
28. Xu L. та ін. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint arXiv:2312.12148*. 2023. DOI: https://doi.org/10.48550/arXiv.2312.12148
29. Hu Edward J., та ін. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*. 2021. DOI: https://arxiv.org/abs/2106.09685
30. Karimi Mahabadi, R., Henderson, J., Ruder, S. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*. 2021. № 34. P. 1022–1035. DOI: https://doi.org/10.48550/arXiv.2106.04647
31. Wang Y., та ін. AdaMix: Mixture-of-Adaptations for parameter-efficient model tuning. *arXiv preprint arXiv:2205.12410*. 2022. DOI: https://doi.org/10.48550/arXiv.2205.12410
32. Karpukhin V., та ін. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*. 2020. DOI: https://doi.org/10.48550/arXiv.2004.04906
33. Gao Y. та ін. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*. 2023. DOI: https://doi.org/10.48550/arXiv.2312.10997
34. Zhang T., та ін. BERTScore: Evaluating Text Generation with BERT. *International Conference on Learning Representations*. 2020. DOI: https://doi.org/10.48550/arXiv.1904.09675
35. GPT-4o. OpenAI. URL: https://platform.openai.com/docs/models/gpt-4o (дата звернення: 27.05.2024).

References

1. "Generative AI for Legal Contracts", available at: https://www.nasdaq.com/articles/generative-ai-for-legal-contracts (last accessed 27.05.2024).
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017), "Attention is all you need", *Advances in neural information processing systems*, № 30. DOI: https://doi.org/10.48550/arXiv.1706.03762
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2018), "Bert: Pre-training of deep bidirectional transformers for language understanding", *arXiv preprint arXiv:1810.04805*. DOI: https://doi.org/10.48550/arXiv.1810.04805
4. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D. (2023), "Llama 2: Open foundation and fine-tuned chat models", *arXiv preprint arXiv:2307.09288*. DOI: https://doi.org/10.48550/arXiv.2307.09288
5. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D.D.L., Hanna, E.B., Bressand, F., Lengyel, G. (2024), "Mixtral of experts", *arXiv preprint arXiv:2401.04088*. DOI: https://doi.org/10.48550/arXiv.2401.04088
6. Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., Mann, G. (2023), "Bloomberggpt: A large language model for finance", *arXiv preprint arXiv:2303.17564*. DOI: https://doi.org/10.48550/arXiv.2303.17564
7. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M. (2023), "Towards expert-level medical question answering with large language models", *arXiv preprint arXiv:2305.09617*. DOI: https://doi.org/10.48550/arXiv.2305.09617
8. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S. (2020), "Language models are few-shot learners", *Advances in neural information processing systems*, № 33, P. 1877–1901. DOI: https://doi.org/10.48550/arXiv.2005.14165
9. Nori, H., Lee, Y.T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., Luo, R. (2023), "Can generalist foundation models outcompete special-purpose tuning? case study in medicine", *arXiv preprint arXiv:2311.16452*. DOI: https://doi.org/10.48550/arXiv.2311.16452
10. Niklaus, J., Matoshi, V., Stürmer, M., Chalkidis, I., Ho, D.E. (2023), "Multilegalpile: A 689gb multilingual legal corpus", *arXiv preprint arXiv:2306.02069*. DOI: https://doi.org/10.48550/arXiv.2306.02069
11. Guha, N., Nyarko, J., Ho, D., Ré, C., Chilton, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D., Zambrano, D., Talisman, D. (2024), "Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models", *Advances in Neural Information Processing Systems*, № 36. DOI: https://doi.org/10.48550/arXiv.1904.09675

Сучасний стан наукових досліджень та технологій в промисловості. 2024. №2 (28)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

12. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J. (2020), "Measuring massive multitask language understanding", *arXiv preprint arXiv:2009.03300*. DOI: https://doi.org/10.48550/arXiv.2009.03300
13. Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S. (2019), "Superglue: A stickier benchmark for general-purpose language understanding systems", *Advances in neural information processing systems*, № 32. DOI: https://doi.org/10.48550/arXiv.1905.00537
14. Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D., Aletras N. (2022), "LexGLUE: A Benchmark Dataset for Legal Language Understanding in English", *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, P. 4310–4330. DOI: https://aclanthology.org/2022.acl-long.297
15. Niklaus, J., Matoshi, V., Rani, P., Galassi, A., Stürmer, M., Chalkidis I. (2023), "LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain", *Findings of the Association for Computational Linguistics: EMNLP 2023*, P. 3016–3054. DOI: https://aclanthology.org/2023.findings-emnlp.200
16. Mabey, R. "Unveiling our legal AI Assistant", available at: https://juro.com/blog/legal-ai-assistant (last accessed 27.05.2024).
17. Browne, R. "An AI just negotiated a contract for the first time ever – and no human was involved", available at: https://www.cnbc.com/2023/11/07/ai-negotiates-legal-contract-without-humans-involved-for-first-time.html (last accessed 27.05.2024).
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017), "Attention is all you need", *Advances in neural information processing systems*, № 30. DOI: https://doi.org/10.48550/arXiv.1706.03762
19. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.T., Rocktäschel, T., Riedel, S. (2020), "Retrieval-augmented generation for knowledge-intensive nlp tasks", *Advances in Neural Information Processing Systems*, № 33, P. 9459–9474. DOI: https://doi.org/10.48550/arXiv.2005.11401
20. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A. (2023), "Llama: Open and efficient foundation language models", *arXiv preprint arXiv:2302.13971*. DOI: https://doi.org/10.48550/arXiv.2302.13971
21. Vanian, J., Leswing, K. "ChatGPT and generative AI are booming, but the costs can be extraordinary", available at: https://www.cnbc.com/2023/03/13/chatgpt-and-generative-ai-are-booming-but-at-a-very-expensive-price.html (last accessed 27.05.2024).
22. Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y. (2022), "Lamda: Language models for dialog applications", *arXiv preprint arXiv:2201.08239*. DOI: https://doi.org/10.48550/arXiv.2201.08239
23. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.D.L., Hendricks, L.A., Welbl, J., Clark, A., Hennigan, T. (2022), "Training compute-optimal large language models", *arXiv preprint arXiv:2203.15556*. DOI: https://doi.org/10.48550/arXiv.2203.15556
24. Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M.S., Love, J., Tafti, P. (2024), "Gemma: Open models based on gemini research and technology", *arXiv preprint arXiv:2403.08295*. DOI: https://doi.org/10.48550/arXiv.2307.09288
25. "Microsoft Copilot for Sales", available at: https://www.microsoft.com/en-us/ai/microsoft-sales-copilot (last accessed 27.05.2024).
26. "Luminance's Legal Pre-Trained Transformer, available at: https://www.luminance.com/technology.html (last accessed 27.05.2024).
27. Lv, K., Yang, Y., Liu, T., Gao, Q., Guo, Q., Qiu, X. (2023), "Full parameter fine-tuning for large language models with limited resources", *arXiv preprint arXiv:2306.09782*. DOI: https://doi.org/10.48550/arXiv.2306.09782
28. Xu, L., Xie, H., Qin, S.Z.J., Tao, X., Wang, F.L. (2023), "Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment", *arXiv preprint arXiv:2312.12148*. DOI: https://doi.org/10.48550/arXiv.2312.12148
29. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W. (2021), "Lora: Low-rank adaptation of large language models", *arXiv preprint arXiv:2106.09685*. DOI: https://arxiv.org/abs/2106.09685
30. Karimi Mahabadi, R., Henderson, J., Ruder, S. (2021), "Compacter: Efficient low-rank hypercomplex adapter layers", *Advances in Neural Information Processing Systems*, № 34, P. 1022–1035. DOI: https://doi.org/10.48550/arXiv.2106.04647
31. Wang, Y., Agarwal, S., Mukherjee, S., Liu, X., Gao, J., Awadallah, A.H., Gao, J. (2022), "AdaMix: Mixture-of-Adaptations for parameter-efficient model tuning", *arXiv preprint arXiv:2205.12410*. DOI: https://doi.org/10.48550/arXiv.2205.12410
32. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.T. (2020), "Dense passage retrieval for open-domain question answering", *arXiv preprint arXiv:2004.04906*. DOI: https://doi.org/10.48550/arXiv.2004.04906
33. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H. (2023), "Retrieval-augmented generation for large language models: A survey", *arXiv preprint arXiv:2312.10997*. DOI: https://doi.org/10.48550/arXiv.2312.10997

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2024. No. 2 (28)*

34. Zhang, T., Kishore, V., Wu, F., Weinberger, K., Artzi Y. (2020), "BERTScore: Evaluating Text Generation with BERT", *International Conference on Learning Representations*. DOI: https://doi.org/10.48550/arXiv.1904.09675
35. "GPT-4o", available at: https://platform.openai.com/docs/models/gpt-4o (last accessed 27.05.2024).

Надійшла (Received) 29.05.2024

Відомості про авторів / About the Authors

Волоховський Віталій Євгенович – Харківський національний університет радіоелектроніки, аспірант кафедри програмної інженерії, Харків, Україна; e-mail: vitalii.volokhovskyi@nure.ua; ORCID ID: https://orcid.org/0009-0006-5682-1889

Volokhovskyi Vitalii – Kharkiv National University of Radio Electronics, PhD student at the Department of Software Engineering, Kharkiv, Ukraine.

ANALYSIS OF METHODS FOR TRAINING DOMAIN-SPECIFIC LANGUAGE MODELS IN THE AREA OF LEGAL CONTRACTS GENERATION

The **subject** of the research is machine learning models and methods for generating legal contracts with limited resources and performance evaluation benchmarks. The **goal** of the work is to analyse approaches of domain-specific Large Language Models development and to find the optimal method of creating independent specialized systems that can generate contracts in different languages and legal systems. The article addresses the following **tasks**: identification of existing companies and solutions in this area, exploring approaches to create texts in natural language, analysis of evaluation and comparison methods of such systems, inspecting limitations and shortcomings of existing solutions and approaches, finding the optimal method of developing systems with limited resources. The following **results** were obtained: approaches of natural language generation and their features were investigated; the "Transformer" architecture was defined as a modern standard in the field of text information generation; different model types which are based on this architecture were considered; data sources for training were analysed; methods of adapting models in specialized areas were considered; model evaluating benchmarks for various tasks were reviewed; shortcomings of the existing specialized language models and the incompleteness of existing benchmarks for contract generation task evaluation were revealed. As a result of the analytical experiment, it was determined that the Retrieval-Augmented Generation method is the most optimal for solving the given task under the given conditions. The conducted experiment and its results can be used as a basis for further research of domain-specific language models development with limited resources. **Conclusions**: the article provides an overview of natural language generation methods using modern machine learning techniques, considers their advantages and disadvantages for small companies and scientific institutions that have limited resources. The work examinates a specialized legal domain and the problem of contract generation and determines the most optimal method to solve it.

Keywords: large language model; natural language generation; contract; legal document.

Бібліографічні описи / Bibliographic descriptions

Волоховський В. Є. Аналіз методів тренування вузькоспрямованих мовних моделей у сфері генерації договорів. *Сучасний стан наукових досліджень та технологій в промисловості*. 2024. № 2 (28). С. 48–64. DOI: https://doi.org/10.30837/2522-9818.2024.2.048

Volokhovskyi, V. (2024), "Analysis of methods for training domain-specific language models in the area of legal contracts generation", *Innovative Technologies and Scientific Solutions for Industries*, No. 2 (28), P. 48–64. DOI: https://doi.org/10.30837/2522-9818.2024.2.048