

Список використаних джерел

1. Andrukhiv D., Prydatko O. Інформаційні технології у менеджменті закладів освіти системи державної служби України з надзвичайних ситуацій. *Вісник Львівського державного університету безпеки життєдіяльності*. 2025. 31. С. 187–195. DOI: <https://doi.org/https://doi.org/10.32447/20784643.31.2025.19>
2. Ivanusa A., Tkachuk R., Maslova N., Yashchuk V., Tkachenko A. Управління інформаційною безпекою та кіберзахистом у закладах вищої освіти. *Вісник Львівського державного університету безпеки життєдіяльності*. 2025. 31. С. 101–116. DOI: <https://doi.org/https://doi.org/10.32447/20784643.31.2025.11>
3. Терлецька Т. С. Модель організації диференційованого навчання у електронному освітньому середовищі на базі LMS MOODLE. *Електронне наукове фахове видання «Відкрите освітнє е-середовище сучасного університету»*. 2025. 19. С. 177–190. DOI: <https://doi.org/10.28925/2414-0325.2025.1912>
4. Шумейко Н. «Використання систем управління навчанням в освітньому процесі: доцільність та переваги». *ОД*. Т. 45. № 2. червень 2024 р.

Мацкул Михайло Ігорович,
здобувач ступеня вищої освіти
магістра факультету радіофізики,
електроніки та комп'ютерних систем
Київського національного
університету імені Тараса Шевченка
Науковий керівник:
Шайда Валерій Віталійович,
магістрант факультету комп'ютерних
наук та кібернетики Київського
національного університету
імені Тараса Шевченка

КЕРІВНІ ПРИНЦИПИ ЮНЕСКО ЩОДО ВИКОРИСТАННЯ СИСТЕМ ШТУЧНОГО ІНТЕЛЕКТУ В СУДАХ І ТРИБУНАЛАХ: ТЕХНІЧНА АРХІТЕКТУРА ПРОТИДІЇ ГАЛЮЦІНАЦІЯМ ГЕНЕРАТИВНИХ МОДЕЛЕЙ

Штучний інтелект поступово інтегрується у системи правосуддя по всьому світу, відкриваючи можливості для покращення доступу до правосуддя, оптимізації управління справами та підтримки прийняття судових рішень, водночас несучи низку етичних, правозахисних і суто технічних викликів. Керівні принципи ЮНЕСКО щодо використання штучного інтелекту в судах і трибуналах [1] стали першою глобальною

етичною та операційною основою для забезпечення того, щоб ШІ служив справедливості, не підриваючи верховенство права. Серед п'ятнадцяти універсальних положень, сформульованих у Керівних принципах, ключове значення для технічної реалізації мають принципи інформаційної безпеки, точності та надійності, пояснюваності, можливості аудиту та відповідальності, оскільки саме вони визначають вимоги до архітектури програмних рішень, які можуть бути допущені до використання у сфері правосуддя.

Документ ЮНЕСКО окреслює, що саме повинно бути реалізовано, проте не відповідає на запитання, як саме це має бути реалізовано на рівні архітектури конкретної системи. Декларативно сформульовані принципи в умовах обмежених ресурсів і браку вузькоспеціалізованих експертів ризикують залишитись недосяжним орієнтиром, особливо зважаючи на стрімке проникнення публічних генеративних моделей у юридичну практику ще до того, як держава встигла сформулювати правила їх використання. Ситуація ускладнюється тим, що такі моделі демонструють критично високий рівень фактологічних помилок саме на правових завданнях.

Дослідження Stanford RegLab та Інституту людиноцентрованого ШІ Стенфордського університету [2] показало, що частота галюцинацій у відповідях найпопулярніших генеративних моделей на конкретні юридичні запити коливається від 58 % у GPT-4 до 88 % у Llama 2; стосовно центрального елементу будь-якого судового рішення – його holding (резолютивної частини) – нижня межа частоти помилок, усереднена за моделями, становить близько 63 %. Навіть спеціалізовані комерційні рішення на базі архітектури Retrieval-Augmented Generation (RAG), такі як Lexis+ AI та Westlaw AI-Assisted Research, які позиціонувалися виробниками як «hallucination-free», на практиці генерують недостовірні відповіді у 17–33 % запитів [3]. Найвідомішим практичним наслідком цієї проблеми став прецедент *Mata v. Avianca, Inc.* (S.D.N.Y., 2023), в якому суддя П. К. Кастель наклав солідарний штраф у розмірі 5 000 доларів США на двох адвокатів та їхню юридичну фірму за подання процесуального документа з шістьма повністю вигаданими ChatGPT судовими прецедентами та підробленими цитатами з них [4].

В українському контексті проблема набуває додаткової гостроти. Навчальні корпуси наявних публічних моделей містять мінімальний обсяг україномовних правових текстів, що погіршує точність розпізнавання норм національного законодавства; передавання тексту судового рішення або матеріалів кримінального провадження у промпт публічної моделі є потенційним порушенням принципу інформаційної безпеки через експорт чутливих даних на сервери іноземного оператора; нарешті, відсутність технічного механізму прив'язки кожного

твердження відповіді до конкретного джерела робить неможливою реалізацію принципу пояснюваності в тому вигляді, у якому його вимагають Керівні принципи ЮНЕСКО та Акт Європейського Союзу про штучний інтелект [5, 6], що класифікує системи для судової влади як високоризикові.

З метою практичної реалізації принципів № 5–8 та № 11 Керівних принципів пропонується архітектура довіреної системи підтримки прийняття судових рішень на основі поєднання технології Retrieval-Augmented Generation з механізмом верифікації цитування та обов'язковою відмовою моделі від відповіді у разі недостатньої впевненості. На відміну від класичної схеми «запит – відповідь» з публічною моделлю, така система фізично розгортається на серверах судової адміністрації або довіреного державного оператора (зокрема, в сертифікованій за КСЗІ хмарній інфраструктурі державного сектору), що знімає питання експорту даних і дозволяє контролювати весь життєвий цикл інформації.

Базовим компонентом системи є шар пошуку. Корпус документів формується з Єдиного державного реєстру судових рішень, нормативно-правових актів з офіційних джерел Верховної Ради України та постанов Кабінету Міністрів. Кожен документ розбивається на семантичні фрагменти (chunks) з контрольованим перекриттям, після чого для кожного фрагмента обчислюється векторне представлення (embedding) з використанням сучасних багатомовних моделей (зокрема, multilingual-E5-large або BGE-M3), які підтримують українську мову та використовуються як стартова точка з подальшою доменною адаптацією під юридичний корпус. Вектори зберігаються у спеціалізованій базі даних (Qdrant, Milvus або розширення pgvector до PostgreSQL) разом з метаданими: ідентифікатор документа, інстанція, дата, посилання на оригінал. Пошук здійснюється у гібридному режимі – одночасно за лексичною мірою BM25 та за косинусною подібністю векторів, після чого знайдені фрагменти впорядковуються cross-encoder моделлю (re-ranker), яка більш точно оцінює релевантність кожного фрагмента до запиту. Поєднання двох методів дозволяє знайти релевантні документи навіть у випадках, коли формулювання запиту лексично відрізняється від формулювань у тексті закону.

Шар генерації побудований на великій мовній моделі з відкритими ваговими коефіцієнтами (сімейство Llama, Mistral, Qwen або їх похідні), розгорнутій локально, що повністю виключає передачу даних зовнішнім провайдерам. Принциповою відмінністю від наявних рішень є робота моделі у режимі генерації з прив'язкою до контексту (grounded generation), тобто модель отримує запит користувача та відібрані шаром пошуку фрагменти і інструктується формувати відповідь виключно на їх основі, а структурований формат відповіді з

обов'язковими полями посилань забезпечується додатково через контрольоване декодування (constrained decoding) за схемою. Слід наголосити, що контрольоване декодування гарантує лише формальну коректність структури відповіді, але не достовірність вмісту полів цитування, саме тому критично необхідним є наступний етап постфактумної верифікації. Кожне твердження у згенерованій відповіді супроводжується посиланням на конкретний фрагмент із зазначенням ідентифікатора документа та зміщення у тексті. Після генерації окремий модуль постфактумної верифікації автоматично перевіряє кожну згенеровану цитату на буквальну присутність у вихідному фрагменті, а кожне твердження, що приписується джерелу, оцінюється моделлю природного логічного виводу (NLI) або спеціалізованою моделлю перевірки фактологічних тверджень (claim verification) на предмет того, чи дійсно воно впливає з процитованого тексту. Якщо для будь-якого твердження така перевірка не пройдена або якщо релевантність найкращих знайдених фрагментів за оцінкою re-ranker'a є нижчою за калібрований поріг, система повертає замість відповіді явне повідомлення про відмову (abstain) з поясненням причини. Саме механізм відмови є ключовим конструктивним рішенням, яке усуває природну схильність генеративних моделей вигадувати правдоподібні, але неіснуючі факти у разі браку інформації.

Запропонована архітектура дозволяє кількісно покращити показники, закладені у Керівних принципах ЮНЕСКО як вимоги. Локальне розгортання повністю виключає експорт даних і виконує принцип інформаційної безпеки (№ 5). Прив'язка кожного твердження до перевіреного першоджерела з ЄДРСР або офіційного НПА з постфактумною верифікацією покликана істотно знизити частоту галюцинацій, рівень якої у наявних рішеннях сягає 58–88 % для публічних моделей та 17–33 % для комерційних RAG-систем, що не передбачають жорсткої відмови у разі недостатньої впевненості. Кількісна оцінка ефективності запропонованої архітектури на україномовному юридичному корпусі є предметом окремого емпіричного дослідження; концептуально ж така архітектура реалізує принципи точності і надійності (№ 6) та пояснюваності (№ 7). Можливість простежити кожне твердження до конкретного фрагмента робить систему придатною до зовнішнього і внутрішнього аудиту (№ 8) на відміну від «чорних скриньок», що не повертають посилань, а явний режим відмови переносить відповідальність на людину лише тоді, коли система здатна обґрунтовано допомогти, що відповідає принципам № 11 та № 13.

Проте жодна архітектура не гарантує абсолютної відсутності помилок і що практична реалізація має низку нетривіальних складових. Якість пошуку залежить від того, наскільки використані багатомовні

моделі (multilingual-E5, BGE-M3) та cross-encoder re-ranker адаптовані до україномовного юридичного домену; для production-grade системи потрібне донавчання (contrastive fine-tuning) на парах «запит – релевантна норма» та «запит – релевантне судове рішення». Аналогічного донавчання потребує і модуль верифікації, оскільки готових NLI-моделей для україномовних юридичних текстів наразі не існує; як проміжне рішення допустиме використання потужної мовної моделі у режимі LLM-as-a-judge. Окремою задачею є темпоральне рішення: норми закону існують у кількох редакціях, тому індекс має зберігати їх версійність і враховувати дату подій справи. Корпус ЄДРСР частково деперсоналізований, проте процесуальні документи можуть містити персональні дані, тому перед індексацією потрібен окремий конвеєр деперсоналізації на базі NER-моделей. Поріг відмови має калібруватися на розміченій вибірці українських юридичних запитів. Попри ці обмеження, сама архітектурна логіка – локальне розгортання, обов'язкова прив'язка відповіді до джерела з постфактумною верифікацією та механізм відмови – є технічним переходом від декларативного до реального виконання Керівних принципів ЮНЕСКО і може слугувати основою для розроблення національного стандарту вимог до систем штучного інтелекту у судовій системі України.

Список використаних джерел

1. Guidelines for the use of AI systems in courts and tribunals. UNESCO, 2025. URL: <https://unesdoc.unesco.org/ark:/48223/pf0000396582>. DOI: <https://doi.org/10.58338/LIEY8089>.
2. Dahl M., Magesh V., Suzgun M., Ho D. E. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis*. 2024. Vol. 16, No. 1. P. 64–93. DOI: <https://doi.org/10.1093/jla/laae003>.
3. Magesh V., Surani F., Dahl M., Suzgun M., Manning C. D., Ho D. E. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. *Journal of Empirical Legal Studies*. 2025. Vol. 22, No. 2. P. 216–242. DOI: <https://doi.org/10.1111/jels.12413>.
4. Mata v. Avianca, Inc., 678 F. Supp. 3d 443 (S.D.N.Y. 2023). URL: <https://law.justia.com/cases/federal/district-courts/new-york/nysdce/1:2022cv01461/575368/54/>.
5. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L series, 12.07.2024. URL: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>.
6. Таран О. В., Гавловський В. Д. Правове регулювання штучного інтелекту в Європейському Союзі та Україні: основні підходи та права людини. *Інформація і право*. 2024. № 1 (48). С. 62–67. DOI: [https://doi.org/10.37750/2616-6798.2024.1\(48\).300706](https://doi.org/10.37750/2616-6798.2024.1(48).300706).