ERP/1: Істотність

Технічні вимоги

Локальна LLM-інфраструктура
для аналізу та оцінки істотності інформації

Зміст

Зміст

1 Умовні скорочення та визначення **3**

2 Загальні відомості **4**
2.1 Передумови 4
2.2 Питання, що вирішуються 4
2.3 Вимоги законодавства та міжнародних стандартів 4

3 Призначення та цілі впровадження **5**

4 Класифікація вимог **6**
4.1 Функціональні вимоги 6
4.1.1 Вимоги до локального RAG-контуру та векторного пошуку 6
4.1.2 Вимоги до автоматизованого знеособлення даних 6
4.1.3 Вимоги до локальної генерації текстів та архітектури інференсу 6
4.1.4 Вимоги до оцінки якості та метрик RAG-Triad 7
4.2 Нефункціональні вимоги 7
4.2.1 Апаратні вимоги до графічних прискорювачів 7
4.2.2 Вимоги до програмного стеку 7
4.2.3 Вимоги до продуктивності та паралельного інференсу 7
4.2.4 Вимоги до безпеки та захисту інформації (NIST SP 800-53) 8
4.2.5 Вимоги до відмовостійкості (CPU Fallback / Peer-to-Peer) 8

5 Додаток А. Специфікація сутностей ШІ-вузла (для ТЗ) **9**

6 Додаток Б. Технічні деталі RAG-контуру та оркестрації K8s **9**

Анотація

У цьому документі наведено детальні технічні вимоги до підсистеми ERP/1: Істотність, яка розробляється як сучасний інтелектуальний сервіс локального аналізу, семантичного пошуку та оцінки інформації на базі штучного інтелекту (ШІ). Платформа передбачає використання децентралізованих обчислювальних вузлів з споживчими графічними прискорювачами (GPU) для локального виконання квантованих великих мовних моделей (LLM) та семантичного RAG-пошуку. Вимоги визначають архітектуру апаратного забезпечення, програмний стек, вимоги безпеки (КСЗІ, LUKS, апаратний модуль безпеки, Cilium), комплаєнс із стандартами NIST SP 800-53 та NIST AI 600-1, а також метрики оцінки якості та відмовостійкості системи в периферійному контурі.

1. Умовні скорочення та визначення

Терміни та скорочення, що використовуються в цьому документі:

Термін / Скорочення	**Значення**
LLM	Large Language Model (Велика мовна модель)
GPU	Graphics Processing Unit (Графічний процесор / відеокарта)
RAG	Retrieval-Augmented Generation (Генерація з доповненням пошуку)
VRAM	Video Random Access Memory (Відеопам'ять)
K8s	Kubernetes (Платформа оркестрації контейнерів)
TCO	Total Cost of Ownership (Сукупна вартість володіння)
NER	Named Entity Recognition (Розпізнавання іменованих сутностей)
TPM	Trusted Platform Module (Апаратний модуль безпеки)
КСЗІ	Комплексна система захисту інформації
ЄСІКС	Єдина судова інформаційно-комунікаційна система

2. Загальні відомості

2.1. Передумови

Сучасний етап цифровізації державних та корпоративних систем вимагає впровадження інтелектуальних інструментів для аналізу великих масивів документів, автоматизованого реферування та підготовки проектів рішень. Водночас, специфіка обробки конфіденційної інформації та персональних даних накладає суворі обмеження на архітектуру обчислювальних засобів:

1. **Конфіденційність (КСЗІ)**: Відповідно до законодавства України, інформація з обмеженим доступом не може передаватися до сторонніх публічних хмарних сервісів (OpenAI, Anthropic, Google тощо). Обробка повинна відбуватися виключно в захищеному периметрі системи.
2. **Живучість**: Можливі блекаути, пошкодження каналів зв'язку та кібератаки вимагають повної автономності (Air-Gapped режим) ШІ-сервісів на місцях.
3. **Мінімізація затримок**: Локальний інференс безпосередньо в локальній мережі (LAN) усуває затримки передачі великих обсягів текстових документів.

Для вирішення цих завдань пропонується впровадити децентралізовані вузли обчислень у кожній установі на базі робочих станцій, інтегрованих у загальний контур ERP/1.

2.2. Питання, що вирішуються

Впровадження локальної ШІ-інфраструктури вирішує такі ключові завдання:

- Організація локального інференсу моделей надвисокої швидкості без ризику витоку інформації.
- Усунення галюцинацій мовної моделі через технологію локального семантичного RAG-пошуку.
- Автоматичне NER-знеособлення персональних даних перед передачею в мовну модель.
- Забезпечення централізованого управління великою кількістю периферійних вузлів.

2.3. Вимоги законодавства та міжнародних стандартів

Система повинна розроблятися та функціонувати відповідно до:

- Закону України «Про захист персональних даних»;
- Закону України «Про захист інформації в інформаційно-комунікаційних системах»;
- Постанови Кабінету Міністрів України № 205 від 21.02.2025 «Про затвердження Порядку використання засобів інформатизації»;
- Стандарту безпеки NIST SP 800-53 Rev. 5 (контролі SC-28, SC-8, AC-3, AU-2);
- Рекомендацій профілю безпеки Generative AI Profile (NIST AI 600-1).

3. Призначення та цілі впровадження

Основною метою впровадження модуля є надання користувачам (наприклад, суддям, помічникам, аудиторам) інтелектуального асистента для аналізу документів, виявлення істотних зв'язків та автоматичної підготовки проектів рішень без загрози витоку конфіденційної інформації та з дотриманням вимог щодо безпеки on-premises обчислень.

4. Класифікація вимог

4.1. Функціональні вимоги

4.1.1. Вимоги до локального RAG-контуру та векторного пошуку

Система повинна підтримувати локальний RAG (Retrieval-Augmented Generation) контур:

- Векторизація документів за допомогою багатомовної моделі отримання векторних представлень (embeddings).
- Швидкий локальний пошук найближчих векторів у HNSW-індексах за допомогою бібліотеки векторного пошуку.
- Зберігання метаданих та повних текстів рішень у швидкому локальному сховищі типу «ключ-значення».
- Асинхронна нічна синхронізація локальних баз із центральним об'єктним сховищем.

4.1.2. Вимоги до автоматизованого знеособлення даних

Перед відправкою тексту до моделі векторних представлень та збереженням у локальну базу даних, матеріали справи проходять обов'язковий етап локального знеособлення:

- NER-фільтрація: виявлення в тексті ПІБ фізичних осіб, адрес проживання, номерів телефонів, державних номерів автомобілів тощо.
- Маскування даних: заміна конфіденційних сутностей на узагальнені токени (`[ОСОБА_1]`, `[АДРЕСА_1]`).
- Локальне збереження таблиць відповідності в зашифрованій базі даних під захистом апаратного модуля безпеки.

4.1.3. Вимоги до локальної генерації текстів та архітектури інференсу

Система повинна забезпечувати високопродуктивну генерацію відповідей безпосередньо на периферійному вузлі установи на базі такої архітектури:

1. **Архітектура розподілу відеопам'яті (VRAM budget)**:
   - Статична область (ваги моделі): ваги квантованої великої мовної моделі повинні займати не більше 60–70% від загального обсягу доступної VRAM (до 9.5 ГБ для прискорювачів 12 ГБ VRAM та до 11 ГБ для 16 ГБ VRAM).
   - Динамічна область (контекст / KV Cache): виділення VRAM під збереження ключів та значень попередніх токенів (Key-Value Cache) з розрахунку забезпечення довжини контексту не менше 4000–8000 токенів.
   - Робоча пам'ять (Workspace / Scratch): резервування буферів для обчислення тензорів під час фаз prefill та decode (не менше 1–1.5 ГБ VRAM).
2. **Оптимізація та квантування**:
   - Застосування методів квантування ваг моделей без критичного зниження якості генерації (інференс моделей у 4-бітному та 8-бітному представленні).
   - Підтримка низькорівневих операторів уваги (attention kernels) з оптимізованим доступом до спільної пам'яті графічного процесора для мінімізації накладних витрат.

3. **Конвеєр обробки запитів (Execution Pipeline)**:
    - *Фаза попереднього обчислення (Prefill)*: паралельна обробка вхідного текстового промпту (контексту) з максимальною утилізацією обчислювальних ядер графічного прискорювача.
    - *Фаза авторегресійної генерації (Decode)*: послідовна генерація наступних токенів, яка обмежена пропускною здатністю пам'яті (memory-bandwidth bound) прискорювача.
4. **Багатокористувацьке планування (Concurrency Scheduling)**:
    - *Ітераційне пакетування (Continuous Batching)*: планування нових запитів на рівні окремих ітерацій генерації токенів, що дозволяє динамічно додавати нові сесії без очікування повного завершення попередніх.
    - *Сторінкове управління KV-кешем*: динамічний розподіл пам'яті для KV-кешу за логічними сторінками для усунення фрагментації VRAM.
5. **Режим автономності (Air-Gapped)**:
    - Локальне виконання інференсу без передачі промптів, проміжних ембедінгів або згенерованих відповідей через глобальну мережу Інтернет.

4.1.4. Вимоги до оцінки якості та метрик RAG-Triad

Для оцінки точності ШІ-сервісів система повинна вимірювати показники за трьома критеріями:

- **Context Relevance**: відповідність знайденого RAG-контексту запиту користувача.
- **Groundedness**: відсутність у генерованій відповіді тверджень, не підтверджених контекстом (боротьба з галюцинаціями).
- **Answer Relevance**: відповідність відповіді мовної моделі вихідному запиту користувача.

4.2. Нефункціональні вимоги

4.2.1. Апаратні вимоги до графічних прискорювачів

В якості периферійного апаратного прискорювача використовуються споживчі графічні прискорювачі (GPU) з об'ємом відеопам'яті (VRAM) 12 або 16 ГБ.

4.2.2. Вимоги до програмного стеку

Локальний сервер функціонує під управлінням Linux-дистрибутива корпоративного класу. Програмне забезпечення інференсу базується на:

- Фірмових драйверах графічного прискорювача та відповідному інструментарії розробки загального призначення на GPU;
- Оптимізованому інференс-сервері мовних моделей з низькорівневою оптимізацією під архітектуру використовуваного прискорювача;
- Бекенді на базі мов програмування з високим паралелізмом для взаємодії через драйвери портів або локальний API (HTTP/gRPC).

4.2.3. Вимоги до продуктивності та паралельного інференсу

Система повинна забезпечувати такі показники швидкості інференсу:

- Швидкість генерації для компактної мовної моделі розміром близько 8 млрд параметрів (при 4-бітному квантуванні) — не менше 70 токенів/сек.
- Швидкість генерації для компактної мовної моделі розміром близько 8 млрд параметрів (при 8-бітному квантуванні) — не менше 45 токенів/сек.
- Підтримка технології паралельного обслуговування запитів (continuous batching) для одночасної роботи від 5 до 8 активних користувачів на одну робочу станцію з графічним прискорювачем споживчого класу без відчутного просідання швидкості.

4.2.4. Вимоги до безпеки та захисту інформації (NIST SP 800-53)

Захист інформації забезпечується на базі таких контролів:

- **SC-28 (Protection of Information at Rest)**: шифрування дискових накопичувачів NVMe за допомогою стандартних засобів ОС та зберігання ключів у апаратному модулі безпеки.
- **SC-8 (Transmission Confidentiality and Integrity)**: шифрування LAN-трафіку за допомогою безпечного протоколу передачі (HTTPS/gRPC) з підтримкою TLS 1.3.
- **AC-3 (Access Enforcement)**: інтеграція з доменною службою каталогів організації та розмежування прав на базі атрибутивного доступу (ABAC).
- **AU-2 (Event Logging)**: незмінне ведення журналів аудиту всіх запитів та відповідей сервера ШІ.

4.2.5. Вимоги до відмовостійкості (CPU Fallback / Peer-to-Peer)

Для забезпечення безперервності процесів система повинна реалізовувати:

- **CPU Fallback**: автоматичний перехід інференсу на центральний процесор (з використанням векторних інструкцій) у разі виходу з ладу графічного прискорювача з падінням швидкості генерації до 8–12 токенів/сек.
- **Peer-to-Peer failover**: автоматичне перенаправлення запитів на сусідній периферійний вузол у межах регіону при повному збої локального сервера.

5. Додаток А. Специфікація сутностей ШІ-вузла (для ТЗ)

Для забезпечення роботи модуля «Істотність» у базі даних (Mnesia/KVS) ведуться такі реєстрові сутності:

1. **ai_node**: Дані про периферійний обчислювальний вузол (апаратні параметри, температура GPU, статус вузла).
2. **ai_model**: Параметри розгорнутих LLM та ембедінг моделей (шлях до файлу ваг, конфігурація квантування).
3. **ai_inference_task**: Журнал виконання запитів користувачів до ШІ (ідентифікатор користувача, промпт, час виконання).
4. **ai_anonymization_dictionary**: Зашифрована таблиця відповідностей для знеособлення персональних даних.

6. Додаток Б. Технічні деталі RAG-контуру та оркестрації K8s

Повна оркестрація 600+ периферійних вузлів здійснюється через централізований Kubernetes за методологією GitOps та інструментами автоматизації Ansible. Всі GGUF файли моделей та векторні індекси завантажуються в нічний час (cron Pull-механізм з 01:00 до 05:00) із центрального сховища централізованого об'єктного сховища. Оновлення контейнерів інференсу проводиться за схемою Blue-Green Deployment для виключення простоїв сервісу.