

УДК 621.431.74

DOI: 10.31653/1819-3293-2026-1-31-13-31

ARTICLE HISTORY

Received 20.02.2026

Accepted 24.03.2026

Горб Сергій Іванович
Національний університет "Одеська морська академія", Одеса,
Україна
gs@onma.edu.ua

Ефективність мовних моделей штучного інтелекту для оцінювання операційних ризиків на судах

Gorb Sergii
National University "Odessa Maritime Academy", Odessa, Ukraine
gs@onma.edu.ua

The Effectiveness of Artificial Intelligence Language Models in Assessing Operational Risks on Ships

Резюме. Незважаючи на зростання інтересу до застосування штучного інтелекту в судновій практиці, судноплавні компанії висловлюють сумніви щодо надійності цієї технології під час оцінювання операційних ризиків. У статті наведено порівняльне оцінювання ефективності відомих мовних моделей штучного інтелекту для підтримки оцінювання операційних ризиків на судах. Запропоновано систему критеріїв, що дає змогу оцінити властивості моделей, які безпосередньо впливають на практичну надійність їх застосування в судновій діяльності, а також визначити моделі, що забезпечують найвищу якість аналізу. Установлено, що найбільш доцільним є використання штучного інтелекту для первинної ідентифікації ризиків, формування переліку запобіжних заходів і прискорення документування результатів. Водночас ефективність такого підходу залежить від правильного вибору моделі та обов'язкової експертної верифікації отриманих висновків.

Summary. Despite the growing interest in applying artificial intelligence in shipboard practice, shipping companies express doubts about the

reliability of this technology in operational risk assessment. The article presents a comparative evaluation of the performance of well-known artificial intelligence language models in supporting operational risk assessment on board ships. A system of criteria is proposed to assess the properties of the models that directly affect the practical reliability of their application in ship operations, as well as to identify the models that provide the highest quality of analysis. It was found that the most appropriate use of artificial intelligence is for the initial identification of risks, the development of a list of preventive measures, and the acceleration of result documentation. At the same time, the effectiveness of this approach depends on the correct choice of model and the mandatory expert verification of the conclusions obtained.

Міжнародний кодекс з управління безпечною експлуатацією суден та запобіганням забруднення (ISM Code) вимагає, щоб власники суден або інші суб'єкти, які взяли на себе відповідальність від судновласника за експлуатацію судна, визначили своєю метою "оцінювати всі виявлені ризики для своїх суден, персоналу та навколишнього середовища та встановлювати відповідні заходи захисту". Ця вимога набула чинності з 1 липня 2010 року відповідно до резолюції Міжнародної морської організації (ІМО) MSC.273(85).

У національному законодавстві аналогічні положення містяться у пунктах 1.1 Додатків 1 та 2 до *Положення про систему управління безпекою судноплавства на морському та річковому транспорті*, що затверджено наказом Міністерства транспорту України від 20.11.2003 № 904. У цьому документі також визначено, що метою діяльності судноплавної компанії є "забезпечення захисту від усіх ризиків, що виявлені". Термін впровадження цієї вимоги – 1 липня 2004 року, тобто на 6 років раніше, ніж аналогічної вимоги ISM Code.

У Керівництві [1] зазначається, що ISM Code не визначає будь-якого конкретного підходу до управління ризиками і судноплавна компанія повинна самостійно обрати методи, що відповідають її організаційній структурі, типу суден та характеру господарської діяльності. Обрані методи можуть бути більш або менш формалізованими, але вони мають бути систематичними, забезпечувати повноту та ефективність оцінювання і реагування на ризики, а також бути належним чином задокументованими для підтвердження обґрунтованості рішень.

Відповідно до п. 6.4 ДСТУ ISO 31000:2018 *Менеджмент ризиків. Принципи та настанови*, оцінювання ризику включає три послідовні етапи:

1. *Ідентифікацію ризику* — виявлення, розпізнавання та опис ризиків, які можуть вплинути на досягнення цілей організації;

2. *Аналіз ризику* — з'ясування природи та характеристик ризику, включно з оцінкою рівня ризику, невизначеностей, джерел ризику, можливих наслідків, ймовірностей та ефективності наявних засобів управління ризиком;

3. *Порівняльну оцінку ризику* — співставлення отриманих результатів з установленими критеріями для визначення необхідності додаткових дій.

Результати оцінювання ризику формують основу для подальшої обробки ризику, яка передбачає вибір і реалізацію відповідних варіантів реагування. Можливі варіанти реагування охоплюють:

уникнення ризику шляхом відмови від діяльності, яка його спричиняє;

прийняття або навіть посилення ризику з метою використання сприятливої можливості;

усунення джерела ризику;

зміну ймовірності реалізації ризику;

зміну потенційних наслідків ризику;

розподіл ризику між сторонами (наприклад через аутсорсинг чи страхування);

свідоме утримання ризику на основі обґрунтованого рішення.

Методи аналізу ризиків детально викладено у ДСТУ EN IEC 31010:2022.

Практика показує, що нормативні процедури оцінювання ризиків та їх реалізація в Керівництвах із систем управління безпекою на морських судах виявились надмірно складними для судових екіпажів [2 – 4]. Тому основна увага зосереджується не на якості аналізу, а на формальному заповненні звітних форм. Непоодинокими є випадки, коли на судах звітні форми з аналізу ризиків заповнюються після проведення судових операцій [5].

У публікаціях [6 – 8] запропоновано уніфікувати заходи для оцінювання ризиків та покращення адміністрування роботи з ризиками. Однак ці заходи не забезпечують кардинального покращення оцінювання операційних ризиків на судах [5].

Найбільш результативним заходом щодо покращення оцінювання операційних ризиків на судах є використання мовних моделей штучного інтелекту (ШІ) [2, 5]. Експериментальне дослідження показало, що використання такого заходу значно підвищує якість оцінювання ризиків та суттєво скорочує трудовитрати.

Однак виявилася **практична проблема**: незважаючи на зростання інтересу до застосування ШІ у судновій практиці [9], судноплавні компанії висловлюють сумніви щодо надійності цієї технології під час оцінювання операційних ризиків. Основні аргументи полягають у тому, що:

- моделі ШІ можуть генерувати галюцинації та помилкові твердження;

- відсутня впевненість, що запропоновані ШІ рекомендації повністю відповідають вимогам регуляторних актів ІМО та національним правилам;

- немає достатнього підтвердження надійності експертних оцінок, сформованих засобами ШІ.

У зв'язку з цим **метою дослідження** є порівняльне оцінювання відомих мовних моделей ШІ для підтримки оцінювання операційних ризиків на судах з акцентом на тих властивостях, які визначають практичну надійність їх застосування.

Для порівняльного аналізу необхідно визначити набір критеріїв оцінювання їхніх відповідей. Кількість таких критеріїв має бути достатньою для забезпечення змістовної повноти аналізу, але водночас не надмірною, щоб не ускладнювати процедуру аналізу без істотного підвищення його результативності. Отже, доцільно орієнтуватися на оптимальний баланс між повнотою та простотою системи критеріїв. Рациональність добору критеріїв повинна визначатися метою порівняння і практичною значущістю критеріїв для прийняття рішення щодо вибору моделі, а не лише формальним інтересом до широти її можливостей. Якщо певний критерій не впливає на підсумковий висновок про доцільність використання тієї чи іншої моделі, його варто розглядати як надмірний. У цьому контексті може бути застосоване таке практичне правило: якщо видалення критерію не змінює вибір моделі, то цей критерій, імовірно, не має вирішального значення для аналізу.

У межах цього дослідження порівняння мовних моделей під час оцінювання операційних ризиків на судах запропоновано здійснювати за кількістю суттєвих ризиків, виявлених моделями, а також на основі комплексу критеріїв із визначеною інтерпретацією балів, що охоплює:

- кількість фраз із ознаками галюцинації* (фактично помилкових тверджень, вигаданих посилань) –

1 – багато хибних, вигаданих тверджень, які роблять висновки ненадійними,

2 – помилкових чи сумнівних тверджень помітно більше допустимого, частина їх впливає на розуміння ризиків,

3 – є окремі сумнівні або слабо підтверджені фрази, але основна аргументація залишається правдоподібною,

4 – є поодинокі неточності, які не впливають на результат оцінювання ризиків,

5 – ключові твердження правдоподібні і не містять ознак галюцинації;

операційну небезпеку – наявність порад або висновків, що суперечать вимогам регуляторних актів ІМО чи національних правил –

1 – відповідь містить небезпечні рекомендації або висновки, які прямо суперечать обов'язковим вимогам і здатні призвести до неправильних дій,

2 – є принаймні одна рекомендація або інтерпретація норми, що створює практичний ризик,

3 – в цілому відповідь безпечна, але є спірні або недостатньо точні формулювання, які можуть викликати неправильне тлумачення,

4 – відповідь нормативно та операційно коректна, можливі лише малозначущі відхилення від практичної безпеки,

5 – відповідь повністю безпечна, не містить ризикованих порад та узгоджена із застосовними вимогами;

рівень експертної довіри до прогнозів категорії витрат та імовірності виникнення ризиків –

1 – прогнози виглядають випадковими, внутрішньо непереконливими або явно помилковими,

2 – окремі оцінки мають слабе обґрунтування і виглядають неправдоподібними,

3 – прогнози загалом прийнятні, але спираються на неповне обґрунтування,

4 – прогнози добре аргументовані і переважно переконливі, сумніви стосуються лише окремих оцінок чи деталей, у тому числі через часткове порушення заданої процедури оцінювання ризиків,

5 – прогнози послідовно обґрунтовані, логічні та викликають високий рівень експертної довіри;

емпірична підтверджуваність - частка тверджень, які відповідають даним звітам про аварії, інциденти та спостереженням портового контролю (тобто підтверджених практикою) –

1 – більшість тверджень не підтверджується практикою чи спостереженнями, або суперечить їм,

2 – зв'язок з практикою простежується слабо, підтверджуються лише окремі положення,

3 – значна частина тверджень узгоджується з практикою, але підтвердження охоплює не всі ключові елементи,

4 – більшість ключових тверджень підтверджується матеріалами реальних випадків, інцидентів чи спостережень,

5 – всі ключові твердження підтверджуються практикою або добре узгоджуються з емпіричними даними;

релевантність і глибину відповіді – відповідь дана по суті відповідно до очікувань від запиту та заданої процедури оцінювання ризиків без відхилення у загальні міркування –

1 – відповідь в основному не відповідає запиту та заданій процедурі оцінювання ризиків або майже повністю йде "убік",

2 – пропущені важливі етапи заданої процедури оцінювання ризиків або відповідь є поверховою та/або шаблонною,

3 – відповідь в основному релевантна запиту та заданій процедурі оцінювання ризиків, але глибина розкриття недостатня,

4 – відповідь добре сфокусована на поставлених завданнях, але розкриває не всі аспекти ризиків або частково порушує задану процедуру оцінювання ризиків,

5 – відповідь повністю відповідає запиту та заданій процедурі оцінювання ризиків та не містить загальних міркувань;

консистентність – логічну узгодженість та відсутність суперечностей між частинами відповіді –

1 – відповідь внутрішньо суперечлива і суперечності роблять її непридатною для використання,

2 – є помітні логічні неузгодженості, що впливають на інтерпретацію та висновки,

3 – загальна логіка збережена, але зустрічаються окремі неузгоджені положення,

4 – відповідь майже повністю логічно узгоджена, можливі лише малоістотні розбіжності або пропуски фрагментів, що підтверджують логіку оцінювання ризиків, або через часткове порушення заданої процедури оцінювання ризиків,

5 – відповідь повністю послідовна, логічно зв'язна і не містить внутрішніх протиріч;

структурованість – ступінь логічної організації та впорядкованості інформації –

- 1 – відповідь хаотична, логіка викладу не простежується,
- 2 – є окремі елементи структури, але порядок викладу непослідовний,
- 3 – структура відповіді в цілому зрозуміла, хоча організація матеріалу могла б бути кращою,
- 4 – відповідь в цілому добре організована, логіка викладу ясна, є лише незначні логічні недоліки,
- 5 – відповідь зразково структурована, логічно організована і зручна для аналізу;

грунтовність пояснення ризиків – змістовна обґрунтованість природи ризику, механізмів його виникнення та можливих наслідків –

- 1 – пояснення природи ризиків фактично відсутнє або зводиться до шаблонних фраз,

- 2 – названі окремі причини або наслідки ризиків, але механізм виникнення ризиків розкритий слабо,

- 3 – ризики пояснені прийнятно (зазначено основні причини, механізм та наслідки), але без достатньої деталізації,

- 4 – дано докладне і переважно переконливе пояснення ризиків, хоча окремі ланки аргументації розкрито не в повному обсязі,

- 5 – дано глибоке та змістовне пояснення природи ризиків, механізмів їх виникнення та ймовірних наслідків;

точність застосування професійної, у тому числі морської термінології –

- 1 – термінологія використовується невірно, непрофесійно або з грубим змішуванням понять,

- 2 – терміни частково коректні, але помітна плутанина, що знижує якість аналізу,

- 3 – термінологія в цілому прийнятна і неточності не руйнують сенс відповіді,

- 4 – професійні терміни застосовуються в основному точно і доречно, зустрічаються лише поодинокі неточності або надмірні спрощення,

- 5 – термінологія коректна, професійна і використовується строго за контекстом.

Запропонований комплекс критеріїв не слід розглядати як універсальний. У разі наявності специфічних вимог до оцінювання операційних ризиків наведений перелік може бути розширений шляхом включення додаткових критеріїв.

Для реалізації завдання цього дослідження пропонується провести оцінювання ризиків на прикладі конкретної суднової операції, для якої накопичено фактичний досвід і може бути здійснений експертний аналіз отриманих результатів. Такою судновою операцією обрано роботу, що аналізувалась у публікації [5] для Oil Tanker "**Diamondway**" (delivered 21.06.2016, IMO Nr. 9742895, automation class A1, Length Overall 277 m, Breadth 48 m, Depth 23,1 m, Laden Speed 14,60 Kts, Displacement 79016,2 Mts).

Зміст роботи: *"Ремонт тріщин у корпусі Scrubber методом зварювання. Тріщини знаходяться на висоті до 3 метрів від підлоги."*

Для порівняльного аналізу було відібрано низку сучасних великих мовних моделей, що репрезентують різні підходи до генеративного ШІ та належать до різних екосистем розробників.

До переліку включено як базові версії моделей, так і конфігурації з увімкненими режимами посиленого міркування, глибокого мислення або дослідження. Це дало змогу оцінити не лише загальну якість відповідей, а й вплив спеціалізованих режимів на глибину аналізу, логічну послідовність, точність аргументації та здатність моделей розв'язувати складні прикладні завдання.

Перелік вибраних моделей (у дужках скорочені позначення):

Claude Sonnet 4.6 (CS 4.6) – велика мовна модель, розроблена компанією Anthropic; використовується як базова модель у чат-боті Claude та призначена для діалогу, аналізу текстів, програмування й роботи зі складними запитамі;

Claude Sonnet 4.6 з увімкненим режимом "міркування" (CS46-R) – та сама модель, але в конфігурації з посиленим режимом міркування для глибокого покрокового аналізу завдань;

DeepSeek-R1 (DS-R1) – модель, розроблена китайською компанією DeepSeek з орієнтацією на складні логічні, математичні та програмні задачі;

DeepSeek-R1 з увімкненим режимом "глибоке мислення" (DS-R1-DT) – для більш ретельного аналізу складних запитів;

ERNIE 5.0 (ERNIE 5.0) – велика мультимодальна модель, розроблена китайською компанією Baidu; навчалася працювати не лише з текстом, а й із зображеннями, аудіо та відео в межах єдиної архітектури;

Gemini 3.1 з увімкненим режимом "професійне мислення" (Gem 3.1-PT) – модель, розроблена Google DeepMind, яка використовується в екосистемі Gemini для професійних і аналітичних сценаріїв;

GPT-5.4 (GPT-5.4) – велика мовна модель компанії OpenAI, що лежить в основі чат-бота ChatGPT та використовується для широкого спектра завдань;

GPT-5.4 з увімкненим режимом "міркування" (GPT-5.4-R) – для багатокрокового розв’язання задач;

Grok 4.1 (Grok 4.1) – модель, розроблена компанією xAI, яка використовується в чат-боті Grok та орієнтована на інтерактивне спілкування, пошук інформації й аналітичну підтримку користувача;

Kimi K2.5 з увімкненим режимом "міркування" (Kimi K2.5-R) – модель китайської компанії Moonshot AI, що застосовується в чат-боті Kimi та орієнтована на складні агентні сценарії, міркування, пошук і роботу з великими обсягами інформації;

Microsoft Copilot з увімкненим режимом "глибоке мислення" (Copilot-DT) – не окрема модель, а чат-бот/AI-асистент компанії Microsoft, який використовує великі мовні моделі у власному інтерфейсі; режим «глибоке мислення» призначений для складніших аналітичних задач;

Mistral з увімкненим режимом "міркування" (Mistral-R) – модель, розроблена французькою компанією Mistral AI, використовується в чат-інтерфейсі Le Chat з орієнтацією на генерацію, аналіз і технічні сценарії (без режиму «міркування») та з режимом «глибоке дослідження» модель припустилася грубих помилок, у зв’язку з чим ці режими не були використані для порівняння);

Qwen3.5-Plus (Qwen 3.5+) – велика мовна модель, створена компанією Alibaba Cloud (дочірня структура в Сінгапурі китайської Alibaba Group) для аналітичних і дослідницьких завдань;

Qwen3.5-Plus з увімкненим режимом "глибокого дослідження" (Qwen 3.5+DR) – з акцентом на багатокроковий пошук, аналіз і синтез інформації;

Perplexity Deep Research (PPLX-MM) – не окрема модель, а спеціальний дослідницький режим платформи **Perplexity**, який автоматично використовує **комбінацію моделей** для проведення глибокого аналізу (у цьому режимі платформа не розкриває вибір моделей, які використовувалися для відповіді).

Усі досліджувані моделі виконали оцінювання ризиків для заданої судової операції та в більшості випадків продемонстрували результати, що перевищували якість традиційного "ручного" аналізу. Водночас між моделями спостерігалися суттєві відмінності як у повноті ідентифікації ризиків, так і в якості обґрунтування, нормативній коректності рекомендацій, рівні структурованості та надійності повторної оцінки залишкових ризиків.

Сумарно моделями для заданої суднової операції ідентифіковані ризики:

1. **Пожежа** або локальний вибух у зоні Scrubber внаслідок виконання вогневих робіт;
2. **Отруєння** токсичними **газами** в робочій зоні;
3. **Падіння з висоти** до 3 метрів;
4. **Ураження електричним струмом** від зварювального обладнання;
5. **Опіки** очей і шкіри, термічні травми;
6. **Падіння інструменту** та оснащення на людей, що знаходяться нижче місця виконання робіт;
7. **Пошкодження корпусу** Scrubber через неправильну технологію зварювання;
8. **Перегрів персоналу** під час роботи всередині димової труби;
9. **Забруднення** навколишнього середовища через можливі витoki промивної води.

У табл. 1 наведено перелік ризиків, ідентифікованих досліджуваними моделями під час оцінювання заданої суднової операції.

Аналіз цих даних дає підстави стверджувати, що сучасні мовні моделі достатньо впевнено розпізнають очевидні та безпосередньо пов'язані з описом роботи ризики, тоді як непрямі або менш очевидні небезпеки виявляються ними гірше. Використання режимів міркування, глибокого мислення та глибокого дослідження загалом сприяло збільшенню кількості ідентифікованих ризиків, хоча не завжди супроводжувалося підвищенням якості обґрунтування.

За результатами оцінювання за запропонованими критеріями у відповідях досліджуваних моделей виявлено низку характерних недоліків, які відрізнялися як за змістом, так і за потенційним впливом на практичну надійність застосування моделей у судновій діяльності.

Модель **CS 4.6** після вжитих заходів захисту за деякими ризиками суттєво знижує одразу й прогнозовані втрати і ймовірності виникнення ризиків, що виглядає надто оптимістично без відповідного пояснення; названі небезпеки є практично правдоподібними та конкретні заходи відповідають практиці, але модель не наводить посилання на конкретні документи; поєднанню категорії втрат "дуже низькі" та категорії ймовірності "низька" помилково присвоєна категорія ризику "низький" замість "несуттєвий".

Таблиця 1

Ризики, ідентифіковані моделями

Ризик	CS 4.6	CS 4.6-R	DS-R1	DS-R1-DT	ERNIE 5.0	Gem 3.1-PT	GPT-5.4	GPT-5.4-R	Grok 4.1	Kimi K2.5-R	Copilot-DT	Mistral-R	Qwen 3.5+	Qwen 3.5+DR	PPLX-MM
Пожожа	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Отруєння газами	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Падіння з висоти	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Ураження електричним струмом	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Опіки	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Падіння інструменту	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Попшкодження корпусу	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Перегрів персоналу	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Забруднення	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Модель **CS 4.6-R** аналогічно **CS 4.6** не наводить посилання на конкретні документи щодо названих небезпек та конкретних заходів.

Модель **DS-R1** помилково врахувала, що роботи виконуються всередині Scrubber, тоді як у описі судна зазначено лише "The Scrubber is installed ... inside the ship's smokestack"; ввела спірний сценарій "порушення інертного середовища / вибух у вантажних танках" через роботи біля Scrubber у димовій трубі; залишкові ризики штучно занижені (5 із 6 визнані несуттєвими); зв'язок пропонованих заходів з практикою простежується слабо і модель не наводить посилання на конкретні документи; окремі заходи щодо впливу на ризики викладені недостатньо докладно; модель одночасно вважає категорію ризику "отруєння газами" – "критичним" до заходів, а після заходів знижує його до "несуттєвого" без переконливого пояснення; окремі ланки аргументації ризиків розкрито не в повному обсязі; використаний термін "дедвейт" замість "водотоннажність".

Модель **DS-R1-DT** помилково припускає близьке проходження трубопроводів Inert Gas System біля місця ремонту; невдало рекомендує укривати сусідні трубопроводи "азбестовими полотнами", тоді як сучасні зварювальні blankets зазвичай позиціонуються як asbestos-free; так само як і у моделі **DS-R1** залишкові ризики штучно занижені та відсутні посилання на конкретні документи; ймовірність пожежі та отруєння обґрунтовується існуванням процедур контролю, які потім знову подаються як заходи реагування; окремі ланки аргументації ризиків розкрито не в повному обсязі; використана помилкова термінологія "чергування скрубера".

Модель **ERNIE 5.0** запропонувала надмірно загальний опис заходів впливу на ризики, що потенційно може порушити практичну безпеку; у первинній оцінці ризику є явна помилка: поєднанню категорії втрат "великі" та категорії ймовірності "висока" помилково присвоєна категорія ризику "критичний" замість "високий"; модель неодноразово знижує категорію втрат після заходів без обґрунтування; уся аргументація надана без нормативної чи емпіричної підстави; глибина розкриття ризиків недостатня; консистентність знижена через помилку в розрахунку категорії ризику; відповідь добре структурована, але структура відповіді наповнена недостатньо; причини ризиків і заходи надані загальними фразами без технічної конкретики; є спрощення професійної термінології ("використовувати правильні матеріали та інструменти").

Модель **Gem 3.1-PT** помилково врахувала, що роботи виконуються всередині Scrubber, тоді як у описі судна зазначено лише "The Scrubber is installed ... inside the ship's smokestack"; загальна довіра до прогнозів витрат та імовірності не висока через слабе обґрунтування категорій і надто оптимістичне зменшення втрат, зокрема до рівня "дуже низька" для падіння з висоти; глибина розкриття ризиків недостатня; відповідь добре структурована, але структура відповіді наповнена недостатньо; пояснення ризиків надмірно короткі, механізми розвитку подій розкриті поверхово, а для залишкового ризику майже немає окремого аргументування, чому саме обрана нова категорія втрат і імовірності;

Модель **GPT-5.4** оцінена як якісна, але вона не дотягує до зразкової через те, що частина повторних оцінок після заходів виглядає занадто оптимістичною, особливо там, де знижено не лише імовірність, а й саму тяжкість наслідків.

Модель **GPT-5.4-R** також оцінена як якісна, але вона не дотягує до зразкової через те, що категорія імовірності виникнення отруєння газами після заходів щодо впливу на ризик різко зменшується від "середньої" до "дуже низької". Це виглядає методично припустимо, але досить оптимістично без обґрунтування, чому саме "дуже низька", а не "низька".

Модель **Grok 4.1** дала рекомендацію "провести газовий аналіз простору Scrubber з використанням системи інертного газа", яка є некоректною, оскільки ця система не призначена для газоаналізу Scrubber. Також у відповіді наведено характеристики вантажних насосів, які не мають відношення до оцінювання ризиків при виконанні заданої роботи. Ці неточності віднесені до галюцинацій моделі, а рекомендація "використання системи інертного газа" також до операційної небезпеки. Повторна оцінка категорії ризику "падіння з висоти" – "несуттєвий" – після запропонованих заходів виглядає занадто оптимістичною. Модель не наводить посилання на конкретні документи щодо названих небезпек та конкретних заходів. Окремі заходи щодо впливу на ризики викладені недостатньо докладно. Окремі ланки аргументації ризиків розкрито не в повному обсязі. Використано термін "вибухозахисне зварювальне обладнання", хоча такий клас не існує. Використано термін "вибухозахищене зварювальне обладнання", хоча такий клас не існує. Також використано некоректну аббревіатуру для неруйнівного контролю зварювання.

Модель **Kimi K2.5** окремі заходи щодо впливу на ризики виклала недостатньо докладно; використано термін "гарячий наряд" замість "пожежний наглядовий пост".

Модель **Copilot-DT** допустила "можливий контакт з інертним газом", хоча основа такого контакту відсутня; рекомендація "при необхідності – використання інертного газу для витіснення парів" у контексті зварювальних робіт біля Scrubber є ризикованою, оскільки може створити кисневе збіднення робочого простору; поєднанню категорії втрат "середні" та категорії ймовірності "дуже низька" помилково присвоєна категорія ризику "несуттєвий" замість "низький"; поєднанню категорії втрат "великі" та категорії ймовірності "низька" помилково присвоєна категорія ризику "низький" замість "середній"; модель неодноразово знижує категорію втрат після заходів без обґрунтування; уся аргументація надана без нормативної чи емпіричної підстави; розкрито не всі аспекти ризиків; консистентність знижена через помилки в перерахунку категорій залишкових ризиків; недоречно використаний термін "газозневоднення".

Модель **Mistral-R** запропонувала надмірно загальний опис заходів впливу на ризики, що потенційно може порушити практичну безпеку; після заходів модель без обґрунтування знижує категорію втрат майже для всіх ризиків, зокрема для пожежі/вибуху – з "дуже великих" до "середніх"; аргументація тримається на загальних міркуваннях без емпіричної чи нормативної опори; глибина розкриття ризиків недостатня; окремі заходи щодо впливу на ризики викладені недостатньо докладно; порушено структуру відповіді, яка вимагається процедурою оцінювання ризиків; пояснення причин ризиків і заходів дуже короткі; професійна термінологія надмірно спрощена.

Модель **Qwen 3.5+** ідентифікувала недостатню кількість ризиків, дала надмірно коротку відповідь, аргументація якої тримається на загальних міркуваннях без емпіричної чи нормативної опори; глибина розкриття ризиків недостатня; окремі заходи щодо впливу на ризики викладені недостатньо докладно; є перепустки фрагментів, які підтверджують логіку оцінювання ризиків; недостатньо розкрито етапи оцінювання ризиків; пояснення причин ризиків і заходів дуже короткі; є поодинокі неточності в професійній термінології (втома металу конструкції, наявність залишкових газів вихлопу).

Модель **Qwen 3.5+DR** оформила відповідь у вигляді звіту у форматі pdf на 17 сторінках, 8 з яких – список використаних джерел інформації з інтернет-посиланнями; у докладній відповіді твердження правдоподібні але повідомлено, що роботи виконуються всередині

Scrubber, тоді як у описі судна зазначено лише "The Scrubber is installed ... inside the ship's smokestack"; рекомендація витіснити пари інертним газом у зоні робіт може створювати кисневе збіднення; рівень експертної довіри до прогнозів, релевантність та консистентність відповіді знижено через порушення заданої процедури оцінювання ризиків; використаний невдалий термін "стрічкові платформи" (у відсотковому відношенні цю помилку можна вважати незначною).

Комбінація моделей **PPLX-MM** допустила найменші невідповідності при оцінюванні ризиків, але помилково віднесла до ризиків "порушення вимог системи управління безпекою/охорони праці та регуляторних вимог", хоча відповідно до процедури оцінювання ризиків ризики повинні були формулюватися у термінах можливих негативних подій/наслідків, які можуть статися в майбутньому внаслідок виконання роботи.

У табл. 2 мовні моделі порівняно за допомогою запропонованих критеріїв. З табл. 2 і загалом з дослідження зроблено **висновки**:

1. Проведене дослідження показало, що сучасні мовні моделі ШІ загалом здатні ефективно підтримувати оцінювання операційних ризиків на судах, однак їх практична придатність істотно залежить від конкретної моделі та режиму її використання. За однакової постановки завдання моделі продемонстрували помітні відмінності за рівнем експертної довіри, емпіричною підтверджуваністю, релевантністю, глибиною пояснення ризиків, а також за наявністю ознак галюцинацій і потенційно небезпечних рекомендацій.

2. Найвищі підсумкові результати у дослідженні продемонстрували моделі GPT-5.4-R, PPLX-MM, CS 4.6-R та Kimi K2.5-R, які забезпечили найбільш збалансоване поєднання операційної безпеки, структурованості, консистентності, ґрунтовності пояснення та точності професійної термінології. Особливо слід відзначити PPLX-MM, яка допустила найменше змістових невідповідностей, а також GPT-5.4-R і CS 4.6-R, що забезпечили найбільш збалансоване поєднання якості аналізу і логічної послідовності відповіді.

Основні обмеження застосування моделей GPT-5.4 та GPT-5.4-R пов'язані не з явними помилками чи небезпечними порадами, а переважно з дещо оптимістичним зниженням окремих залишкових ризиків після запровадження заходів контролю, що вимагає додаткової експертної перевірки при практичному застосуванні.

У моделі CS 4.6-R недостатня опора на конкретні нормативні джерела, через що навіть якісно сформована відповідь у практиці потребує додаткової верифікації.

Таблиця 2

Критеріальні оцінки моделей

Критерій	CS 4.6	CS 4.6-R	DS-R1	DS-R1-DT	ERNIE 5.0	Gem 3.1-PT	GPT-5.4	GPT-5.4-R	Grok 4.1	Kimi K2.5-R	Copilot-DT	Mistral-R	Qwen 3.5+	Qwen 3.5+DR	PPLX-MM
Кількість фраз із ознаками галюцинації	-	-	2	1	-	1	-	-	2	-	1	-	-	1	-
Операційна безпека	5	5	5	5	4	5	5	5	4	5	2	4	5	5	5
Рівень експертної довіри	3	5	2	2	2	2	3	4	4	5	2	3	4	4	5
Емпірична підтвердженість	3	3	2	2	2	5	5	5	3	5	2	2	2	5	5
Релевантність і глибина відповіді	5	5	4	5	3	3	5	5	4	4	4	3	3	4	4
Консистентність	4	5	4	4	4	5	5	5	5	5	4	4	4	4	5
Структурованість	5	5	5	5	4	4	5	5	5	5	5	3	3	5	5
Грунтовність пояснення	5	5	4	5	3	2	5	5	4	5	5	3	3	5	5
Точність професійної термінології	5	5	4	4	4	5	5	5	4	4	4	4	4	5	5

3. Дослідження також показало, що режими "міркування", "глибокого мислення" та "глибокого дослідження" загалом підвищують результативність моделей. Однак такий ефект не є автоматичною гарантією якості, оскільки в окремих випадках посилені режими одночасно породжували нові галюцинації, надлишкові припущення або методичні відхилення.

4. Найбільш проблемними у дослідженні виявилися моделі Copilot-DT, ERNIE 5.0, DS-R1, DS-R1-DT, Mistral-R і частково Qwen 3.5+, оскільки саме для них характерні нижчі оцінки за рівнем експертної довіри, емпіричною підтверджуваністю або операційною безпекою. Допущені окремими моделями поодинокі небезпечні поради в сфері судових робіт мають принципове значення, тому використання їх без експертного контролю не може вважатися допустимим.

5. Запропонований набір критеріїв є практично цінним, оскільки дає змогу оцінювати не лише формальну повноту відповіді, а передусім ті властивості моделей, які безпосередньо визначають надійність їх застосування в судовій практиці.

6. Практичне значення отриманих результатів полягає в тому, що мовні моделі доцільно розглядати не як заміну судового командного складу, а як інструмент інтелектуальної підтримки під час підготовки, перевірки та уточнення оцінок ризику. Найбільш доцільним є їх використання для первинної ідентифікації ризиків, формування переліку заходів та прискорення документування результатів.

7. Загальний висновок дослідження полягає в тому, що використання мовних моделей ШІ для оцінювання операційних ризиків на судах є перспективним і може забезпечувати суттєве підвищення якості аналізу. Проте ефективність такого підходу визначається правильним вибором моделі і обов'язковою експертною верифікацією її висновків. Для експертної верифікації можуть бути залучені й інші моделі ШІ.

Зроблені висновки не є статичними, оскільки природа систем ШІ є динамічною: вони постійно навчаються, удосконалюються та адаптуються до вирішуваних завдань [10].

ЛІТЕРАТУРА REFERENCES

1. A Guide to Risk Assessment in Ship Operations – No 127: IACS Rec. 2012/Rev.1 2021. – 6 p.

2. Горб С.І., Сагайдак О.І., Горб О.С. Використання штучного інтелекту для оцінки ризиків на морських судах // Автоматизація суднових технічних засобів: наук.-техн. зб. – 2024. – Вип. 29. – Одеса: НУОМА. – С. 33 – 42.

3. Waldemar Daszuta, Samrat Ghosh. Seafarers' perceptions of competency in risk assessment and management: an empirical study. //WMU Journal of Maritime Affairs. – 2018. – Volume 17. – P. 543 – 564.

4. Samrat Ghosh, Waldemar Daszuta. Failure of risk assessment on ships: factors affecting seafarer practice. //Australian Journal of Maritime & Ocean Affairs. – 2019. – Volume 11. – Issue 3. – P. 185 – 198.

5. Горб С. І. Аналіз операційних ризиків на судах з використанням штучного інтелекту // Автоматизація суднових технічних засобів: наук.-техн. зб. – 2025. – Вип. 30. – Одеса: НУОМА. – С. 21 – 40.

6. Gorb Sergey, Gorb Alex. Measuring risk; a practical approach. //Seaways. – 2013. – March. – P. 11 – 14.

7. Горб С.И., Горб А.С. Оценка рисков в техническом менеджменте судов. //Автоматизация судовых технических средств: науч. -техн. сб. – 2012. – Вып. 18. – Одесса: ОНМА. – С. 13 – 22.

8. Горб С., Горб А. Бег на месте? Оценки рисков в судоходных компаниях. //Порты Украины. – 2012. – № 7. – С. 30 – 32.

9. Irmia Durlik, Tymoteusz Miller, Ewelina Kostecka, Tomasz Tunski Artificial Intelligence in Maritime Transportation: A Comprehensive Review of Safety and Risk Management Applications. // Applied Sciences. – 2024. – Vol. 14. – Iss. 18, 8420. – 32 p.

10. ISO/IEC 23894:2023. Information technology – Artificial intelligence – Guidance on risk management.

Abstract. The study examines the potential for the practical application of modern artificial intelligence language models in the assessment of operational risks on board marine vessels. The relevance of the research lies in the fact that shipping companies continue to treat such technologies with caution because of the risk of hallucinations, possible methodological errors, and the insufficient substantiation of generated conclusions.

The aim of the study is to provide a comparative evaluation of well-known artificial intelligence language models with regard to their suitability for supporting shipboard risk assessment procedures. To this end, a system of criteria is proposed that focuses on properties directly affecting the reliability of practical application, including the presence of hallucination indicators, the operational safety of recommendations, the level of expert

confidence, empirical verifiability, the relevance and depth of the response, consistency, structural coherence, the validity of risk justification, and the accuracy of professional terminology.

The comparison was carried out using a specific shipboard operation involving the welding repair of cracks in the Scrubber casing, which made it possible to compare model outputs with the actual requirements of shipboard practice and accumulated expert experience. The results show that most of the models examined successfully identify the principal obvious risks associated with hot work, exposure to toxic gases, work at height, electric shock, and thermal injury. At the same time, the quality of model outputs varies considerably: alongside strong-performing solutions, cases were identified of unjustifiably optimistic reassessment of residual risk, the use of questionable assumptions, and isolated potentially unsafe recommendations.

The study demonstrates that language models are most appropriately used as an intelligent support tool for the initial identification of hazards, the development of preventive measures, and the acceleration of documentation processes. It is concluded that the use of artificial intelligence in shipboard risk assessment is promising, provided that the model is selected appropriately and that all generated conclusions are subjected to mandatory expert verification.