

ВПЛИВ ТЕМПЕРАТУРИ НА ЕФЕКТИВНІСТЬ ВИДОБУТУ СТРУКТУРОВАНИХ ДАНИХ ВЕЛИКИМИ МОВНИМИ МОДЕЛЯМИ СЕРІЙ LLAMA 3 ТА QWEN 2.5

Незважаючи на низку об'єктивних обмежень, включаючи латентність обробки та схильність до галюцинацій, великі мовні моделі (LLM) набувають статусу домінуючого інструментарію в галузі обробки природної мови. Результати такої обробки можуть бути представлені як у форматі неструктурованого тексту, так і структурованих документів (JSON, YAML, XML). Останній варіант є доцільним, якщо подальша обробка спирається на класичні детерміновані засоби. JSON утвердився як поширений стандарт обміну даними в мережевих комунікаціях, що закономірно відобразилося як на структурі тренувальних даних LLM, так і на спектрі супутніх програмних інструментів.

LLM підтримують ряд параметрів, що впливають на ефективність і якість їх роботи в певних сценаріях. До основних відносяться температура, штрафні коефіцієнти (за наявність, частоту, появу певних токенів) та вагові параметри семпсування (top-k, top-p, top-a, min-p). Попри те, що зазначені параметри є фундаментальними для більшості архітектур, їх імплементація значно різниться між провайдерами, навіть у рамках однієї базової моделі. Серед них температура залишається найбільш універсальним показником. Цей параметр регулює міру стохастичності в процесі генерування послідовностей токенів. Підвищені значення температури збільшують дисперсію та розширюють діапазон потенційних відповідей, тоді як знижені значення посилюють детермінованість і передбачуваність результату.

Метою дослідження є встановлення зв'язку між параметром температури та якісними характеристиками екстракції структурованої інформації з великих україномовних документів моделями Llama 3.x-instruct (1B, 3B, 8B, 70B) та Qwen 2.5-instruct (7B, 32B, 72B). Як об'єкт аналізу використано силабус навчальної дисципліни «Організація баз даних та знань» Чернівецького національного університету імені Юрія Федьковича (розробник Ковальчук М.Л., 8 сторінок, 2270 слів). З документа видобувались регламентні дані, загальний опис, тематична структура та список літератури. Загалом досліджувалось 160 інформаційних полів. Для коректної оцінки якості екстракції було розроблено еталонне представлення документа. Для кожної моделі та значення температури проведено серію з 20 експериментальних вимірювань з метою об'єктивної оцінки як якісних показників, так і стабільності результатів.

Для порівняння отриманих даних застосовано метрику Path-based Normalized Similarity Score, що ґрунтується на трансформації ієрархічної JSON-структури в множину пар ключ-значення шляхом заміни ключів на повний шлях доступу (path enumeration), а також конвертації значень-масивів у множину пар ({ключ, індекс}, значення). При зіставленні текстових елементів застосовувалась нормалізована міра Левенштайна, для порядкових величин – пряме порівняння значень. До згенерованих документів попередньо застосовувалась процедура нормалізації з видаленням полів, що містять порожні значення ("", NULL, undefined). Експериментальні дослідження проводились із використанням мета-провайдера OpenRouter.

У таблиці 1 представлені оптимальні параметри для кожної з моделей, де колонка «значення» вказує максимальне з середніх значень та дисперсію серед експериментів, для яких показник коефіцієнта варіації свідчить про стабільність результату (менше за 20%), колонка «Т» вказує на значення показника температури, при яких було досягнуте значення, «робоча область температур» визначалась експериментами, де результат був стабільним. Показник середньої ціни обробки 1000 документів визначався на основі даних OpenRouter та може значно змінюватись залежно від провайдера. Для моделей llama-3.2-3b, llama-3.3-70b жоден з експериментів не показав стабільного

результату. Для моделі qwen-2.5-7b декілька значень коефіцієнта варіації були вище за 20%, але в межах похибки. Рисунки 1 та 2 ілюструють отримані дані. На рисунку 2 не зображені значення більші за 100%.

Таблиця 1. Максимальні показники якості моделей в робочій області

	Максимальна усереднена якість		Робоча область температур	Середня ціна обробки 1000 документів, \$
	Значення	T		
llama-3.2-3b	$*0,14 \pm 0,03$	0	—	0,18
llama-3.1-8b	$0,59 \pm 0,13$	0.05	0.0–0.05	1,76
llama-3.3-70b	$*0,51 \pm 0,27$	0.5	—	1,88
qwen-2.5-7b	$0,57 \pm 0,11$	0	$**0.0–0.4$	1,07
qwen-2.5-coder-32b	$0,87 \pm 0,02$	0.05	0.0–1.0	1,73
qwen-2.5-72b	$0,86 \pm 0,02$	0	0.0–1.0	2,49

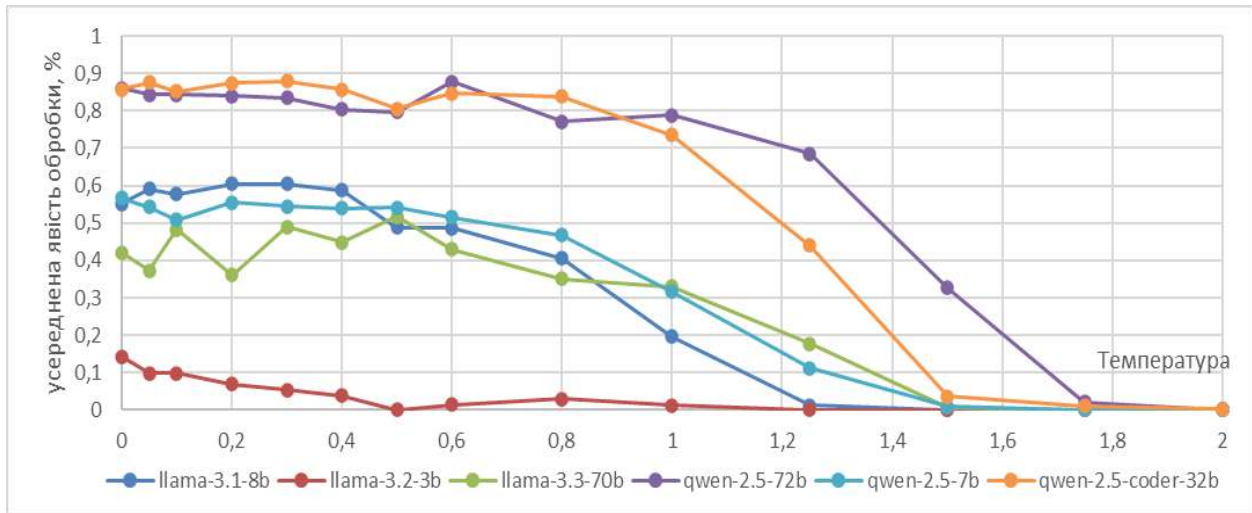


Рис 1. Середні показники якості в залежності від температури

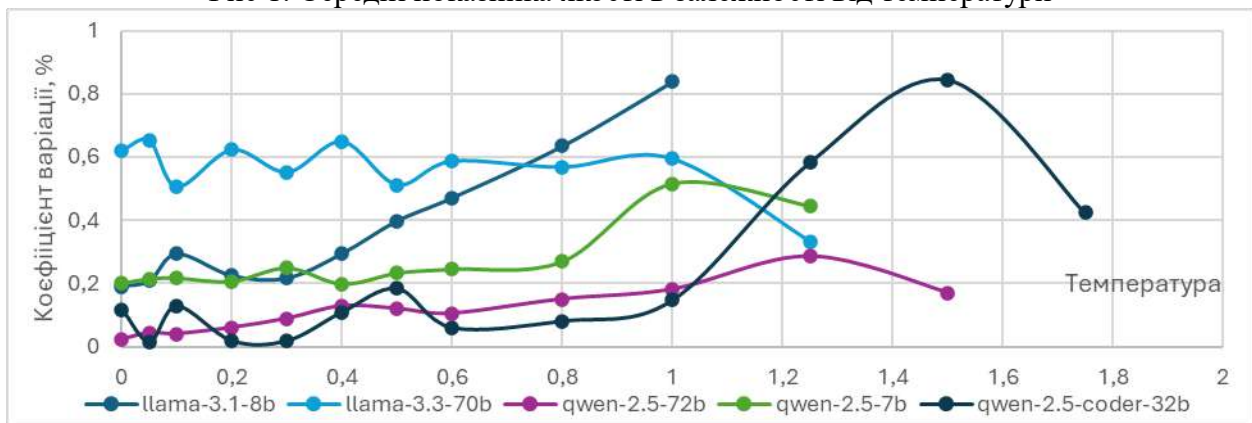


Рис 2. Коефіцієнт варіації значень якості в залежності від температури

Підбиваючи підсумок, оптимальною температурою для задачі екстракції структурованої інформації в моделях серій Llama 3.x та Qwen 2.5 є 0.0-0.05. Моделі Llama менш підходять для заданої задачі, а модель llama-3.2-1b не впоралась із завданням. Згідно до серії експериментів оптимальною як співвідношенням ціна/якість є qwen-2.5-coder-32b, яка відзначилась високою якістю

– $87\% \pm 2$, стабільністю ($1,5\%$ при $t=0.05$) та ціною – $1.73\$$ за 1000 документів, широким діапазоном робочих параметрів.

Науковий керівник: Кузіков Б.О., к.т.н., доцент

*Шулінок І.Е., н.с.,
Панасенко С.В., м.н.с.,
Кікоть Г.Ф., пров. екон.*

*Інститут кібернетики імені В.М. Глушкова НАНУ
irina.shulinok@gmail.com*

ДЕЯКІ АСПЕКТИ МОДЕЛЮВАННЯ РЕСУРСНОГО ЗАБЕЗПЕЧЕННЯ МІЖГАЛУЗЕВОЇ ДИНАМІКИ

Аналітично-дослідні задачі збалансування чи перерозподілу ресурсного потенціалу країни викликані як економічними глобальними чинниками (уповільнення світової економіки внаслідок руйнації налагоджених товарно-збутових ланцюгів, зростання цін на сировину та енергоресурси, зокрема підвищення загальних міжнародних ризиків, нарощення рівня світової конфронтації та оборонних витрат), так і міжгалузевим зменшенням ресурсного забезпечення окремих видів економічної діяльності (ВЕД) через руйнування виробничих потужностей, зменшення національного земельного фонду, руйнування чи порушення інфраструктурних зв'язків, скорочення робочої сили, зменшення обсягу споживчих витрат; переорієнтація державних видатків та ряду інфляційних процесів. Всі ці чинники, що визначені в Постанові Президії НАН України від 30 березня 2022 року [1] суттєво впливають на зміну внутрішньогалузевої структури економіки країни та потребують визначення можливих сценаріїв оптимального перерозподілу ключових виробничих ресурсів.

Метою розроблених в Інституті кібернетики методів та програмних засобів [2 - 5] є визначення таких підходів, що в умовах порушення міжгалузевих взаємозв'язків та загального макроекономічного спаду, дозволяють досліджувати варіанти ресурсного забезпечення для визначених видів економічної діяльності (ВЕД), які при оптимальному використанні чи перерозподілі ресурсів сприятимуть відновленню відносної структурної збалансованості.

Розрахунки за економетричною моделлю ТВВ проводились на основі динамічних рядів побудованих із статистичних даних ТВВ, що наводяться в щорічних Бюлетенях "Національні рахунки України". Для розрахунків створено відповідний модельний інструментарій з дослідження виробничих та соціально-економічних процесів за звітними таблицями «витрати-випуск» та показниками Національних рахунків України, що дозволяє аналізувати та прогнозувати як структурні зрушення в межах проміжного споживання між ВЕД, так і відповідний вплив цих зрушень на макроекономічні показники, зокрема на обсяг випуску продукції в постійних цінах (рисунок 1).