

Створення середовищ навчання з підкріпленням для постнавчання LLM

Виконав: студент 2-го року навчання,
Спеціальності

121 Інженерія програмного
забезпечення

Холявко Микита Володимирович

Керівник: Глибовець Микола
Миколайович, д-р фіз.-мат. наук,
професор кафедри інформатики

Актуальність дослідження

- Більшість існуючих рішень орієнтовані на однокрокову генерацію коду;
- Фреймворки часто працюють лише з ізольованими функціями;
- Ігноруються проміжні результати редагування;
- Бракують динамічного зворотного зв'язку від користувача.

Мета та задачі роботи

Розробка середовища навчання з підкріпленням, яке забезпечує ітеративну взаємодію агента з реальним програмним забезпеченням.

Ключові

ВИМОГИ:

- Багатокрокова модифікація коду
- Інтерактивний фідбек
- Формування проміжної винагороди
- Безпечна ізоляція

Наукова новизна

Наукова новизна дослідження полягає у розробці архітектури інтерактивного середовища для постнавчання великих мовних моделей, що реалізує механізм надання фідбеку моделі під час вирішення поставленої проблеми та використовує проміжні результати для розрахунку винагороди.

Функції винагороди

Перевірка кінцевого
результату:

$$r = \begin{cases} -1, fail \\ \frac{s}{t}, pass \end{cases}$$

Ітеративна перевірка
результату:

$$r_t = \begin{cases} r_{t-1}, fail \\ r_{t-1} - 0.2, s_n = 0 \\ r_{t-1} + s_n, s_n \neq 0 \end{cases}$$

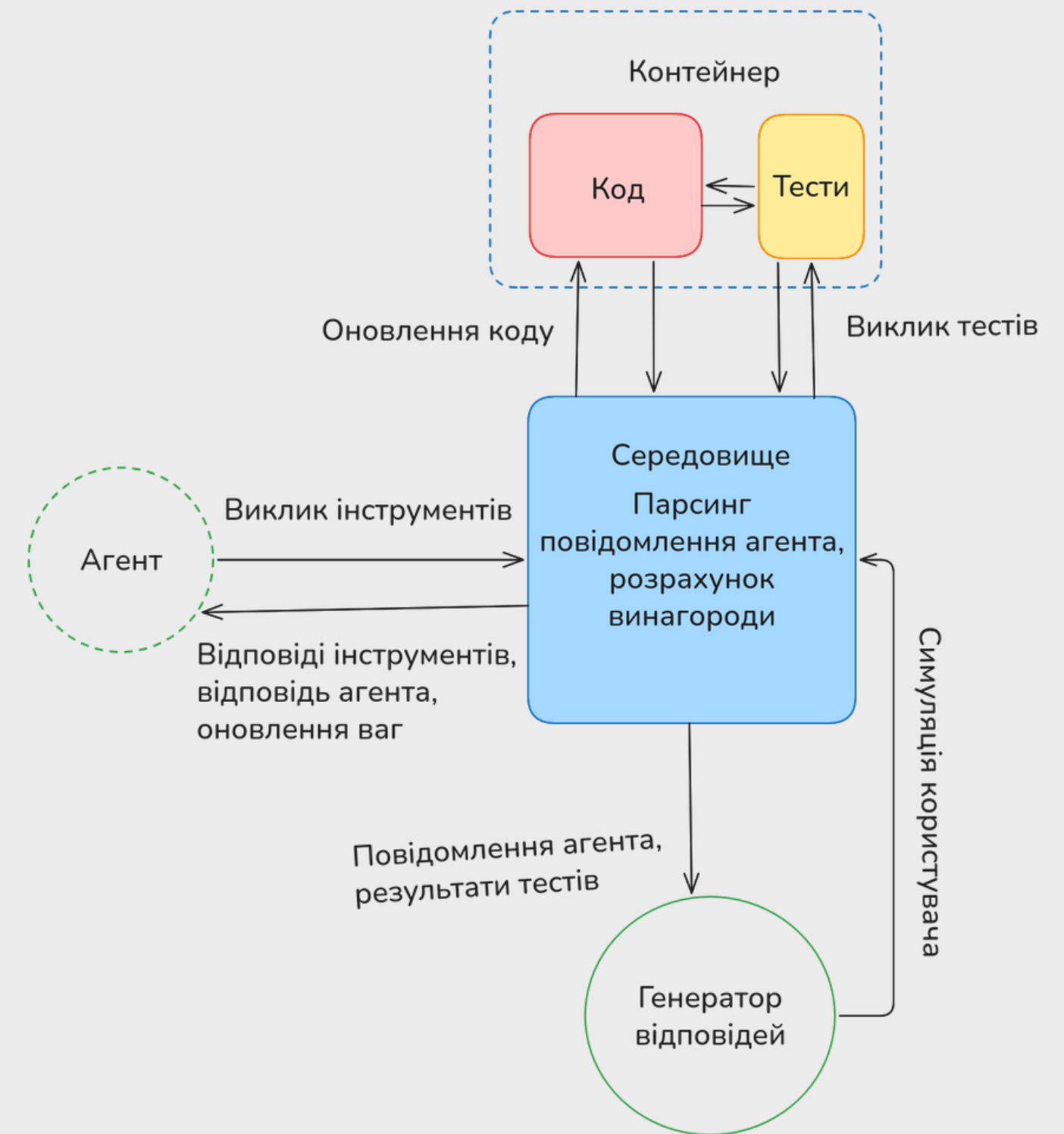
Найкращий
результат:

$$r_t = \begin{cases} \max(r_{t-1}, -1), fail \\ \max(r_{t-1}, \frac{s}{t}), pass \end{cases}$$

Архітектура середовища

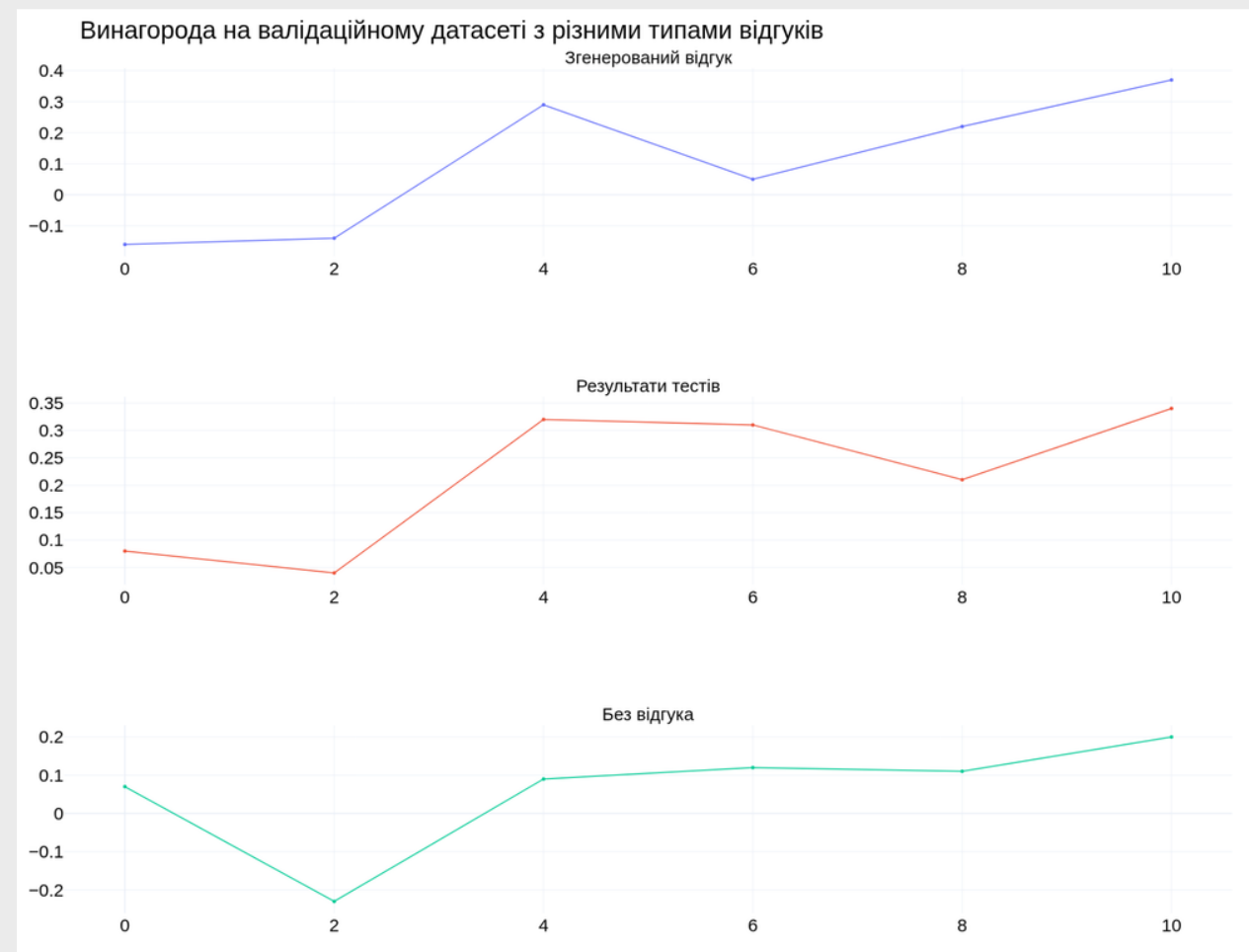
Етапи одного кроку

- Агент формує відповідь із викликом функції
- Усі маніпуляції з файловою системою відбуваються всередині Docker контейнера
- Запуск юніт тестів та аналіз результатів
- Допоміжна модель генерує відгук, що імітує коментарі реального користувача на основі результатів виконання



Аналіз впливу зворотного зв'язку

- Різні типи відгуків показали стабільне підвищення точності моделі
- При відсутності фідбеку якість моделі не покращується



`_lookup_items`, `_keys_len`, and `_products` from the provided `items` mapping.

2. The `get` method needs to retrieve values using multiple keys, handling wildcards (ALL) and using the `_lookup` property and `_iterate_lookup_keys` method.

Here's my implementation:

```
write { "file_path": "mypy_boto3_builder/utis/..." } Tool Call
```

```
> tool #0b043747 None Tool Output
```

user

It looks like the `__init__` method isn't initializing `_keys_len` with the correct value based on the input keys, causing the `_iterate_lookup_keys` method to raise an error. Also, I notice that you didn't update the constructor to determine the key length from the input items, which might be why `_keys_len` is zero. Can you check whether the constructor properly sets `_keys_len` based on the structure of your items?

assistant

You're absolutely right. The issue is that `_keys_len` is not being properly initialized from the structure of the input items. Let me fix this by updating the `__init__` method to properly determine the key length based on the input structure.

The problem is that we need to determine the length of the keys from the actual structure of the items, not just set it to 0. Let me implement the fix:

```
write { "file_path": "mypy_boto3_builder/utis/..." } Tool Call
```

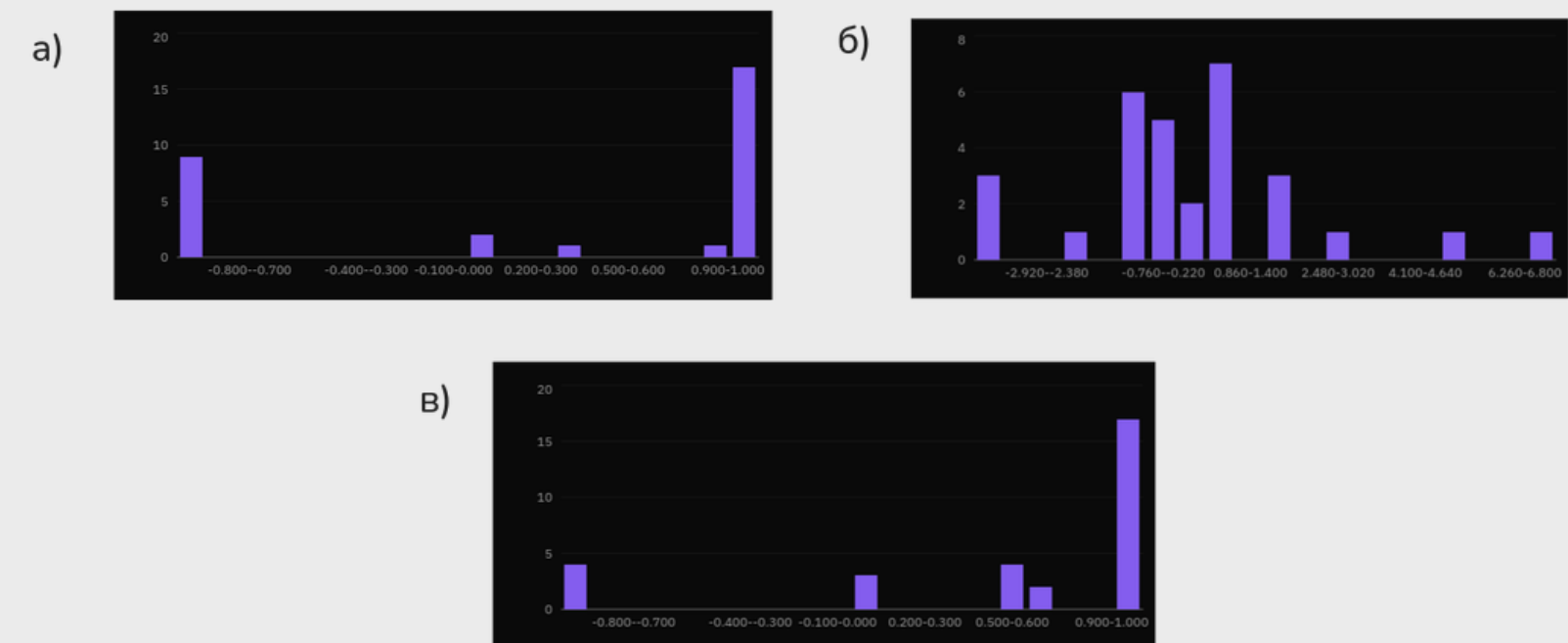
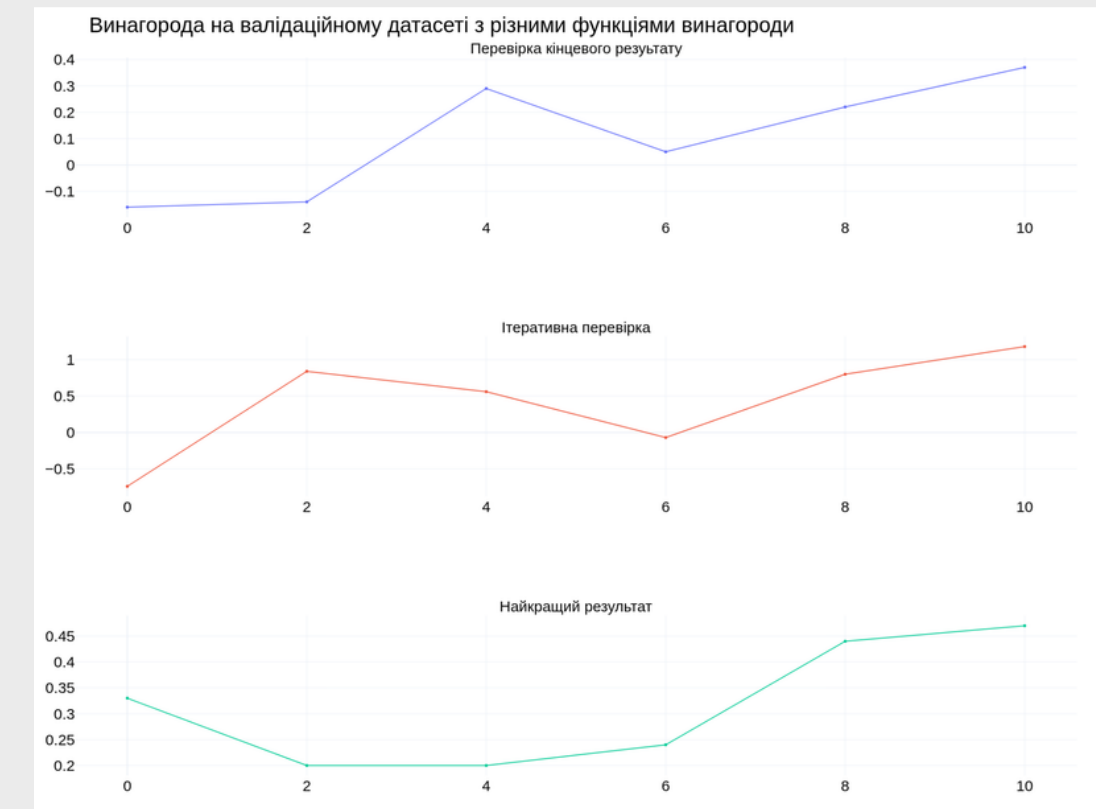
```
> tool #5f135041 None Tool Output
```

user

It looks like `_keys_len` is still not being correctly set based on the actual input items. The error shows that `_keys_len` remains at 0 when `_iterate_lookup_keys` is called, which causes the unexpected `ValueError`. You need to ensure that `_keys_len` is initialized properly after processing the items, based on the structure of the nested data. Make sure to set it to the length of the key tuples derived from the input items, rather than leaving it at 0.

Порівняння функцій винагороди

- Усі досліджені функції винагороди демонструють позитивний тренд навчання
- Функції фінального/найкращого результату мають виражену дискретність: більшість прикладів мають винагороду 1
- Ітеративна функція демонструє більш рівномірний розподіл



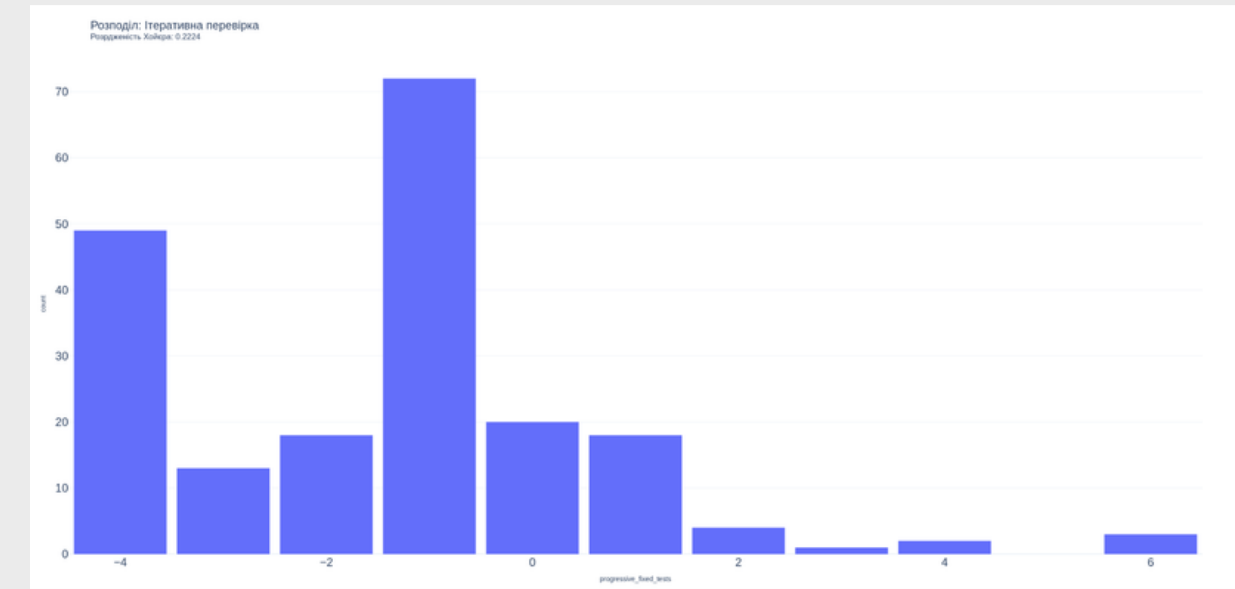
Порівняльний аналіз із існуючими рішеннями

	Багатокрокова модифікація	Інтеграція з реальними репозиторіями	Онлайн навчання з підкріпленням	Ітеративна функція винагороди	Проміжні відгуки користувача
PerfRL	-	-	+	-	-
StepCoder	-	-	+	-	-
RLEF	+	-	+	-	+
SWE-gym	+	+	-	-	-
SWE-dev	+	+	-	-	-
Розроблене середовище	+	+	+	+	+

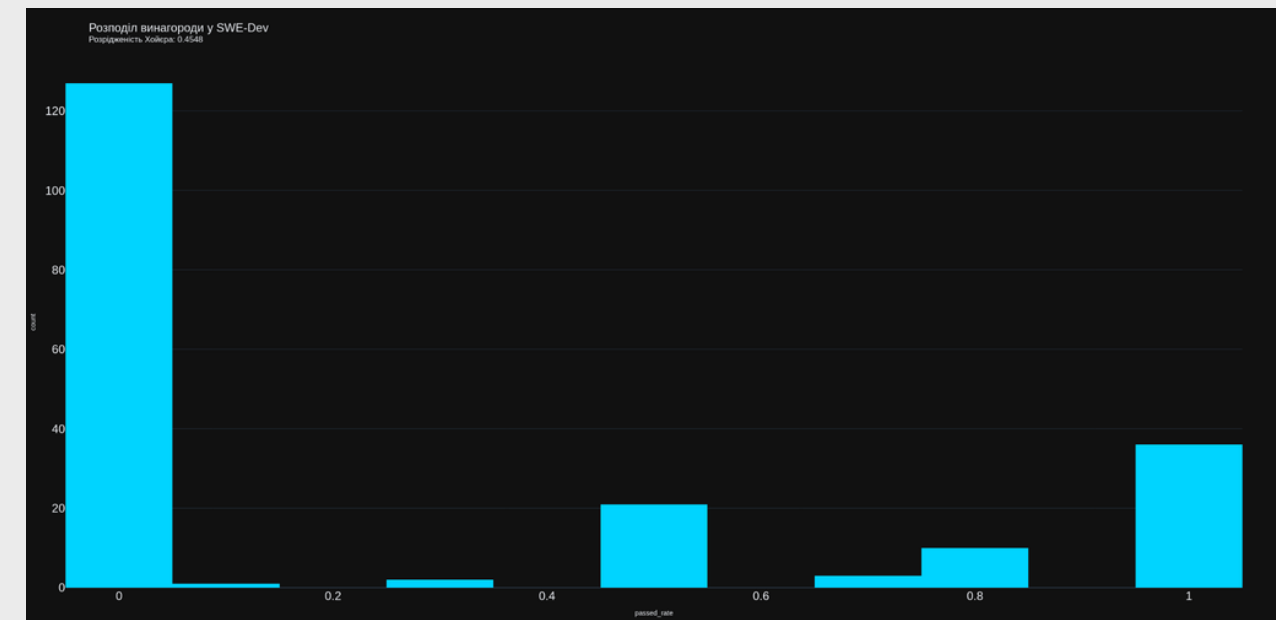
Порівняння винагороди з SWE-Dev

	Метрика	
	Хойєра:	
Розроблене середовище		SWE-Dev
0.2224		0.4548

Розроблене середовище забезпечує агента більш різноманітним сигналом зворотного зв'язку



Розподіл значень винагороди з функцією ітеративної перевірки результату у розробленому середовищі



Розподіл значень винагороди у середовищі SWE-Dev

Висновки

- Розроблено середовище для постнавчання мовних моделей на задачах програмування
- Експериментально підтверджено коректність роботи розробленого середовища
- Підтверджено, що врахування проміжних станів редагування усуває проблему розсіяної винагороди
- Перевірено, що інтеграція динамічного відгуку допомагає генерувати траєкторії для тренування

**Дякую за
увагу!**