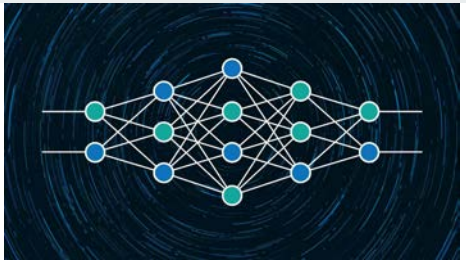
Використання великих мовних моделей на мобільній платформі для генерації медіа

Керівник наукової роботи
ст. викл. Франків О. О.
Виконала студентка Іваненко В. А.

Вступ



- Стрімка еволюція штучного інтелекту (ШІ) призвела до трансформаційних змін у різних галузях, а **великі мовні моделі (Large language models, LMMs),** такі як GPT-3 (OpenAI) та її нащадки, LaMDA (Google AI) [2], та інші показують **безпрецедентні можливості в обробці природної мови** (NLP).
- Типове розгортання цих моделей через хмарні API створює значні проблеми, включаючи **затримку**, **проблеми з конфіденційністю** та **залежність від підключення до Інтернету.**
- Проте, використання великої мовної моделі безпосередньо на пристрої теж **не позбавлене викликів**, головний з яких - **обмежені ресурси, особливо, якщо йдеться про мобільну платформу**.

Завдання роботи

1. Огляд теоретичних основ і поточного стану LLM та їх застосування.
2. Виявлення та вирішення проблем, пов'язаних з роботою з LLM на пристрої.
3. Розробка та реалізація розширення клавіатури на iOS, яке використовує LLM на пристрої для генерації тексту.

Розділ 1. Теоретичні відомості про великі мовні моделі

1.1 Великі мовні моделі та технології, що лежать в їх основі

1.2 Дослідження можливостей великих мовних моделей

1.3 Переваги та недоліки великих мовних моделей

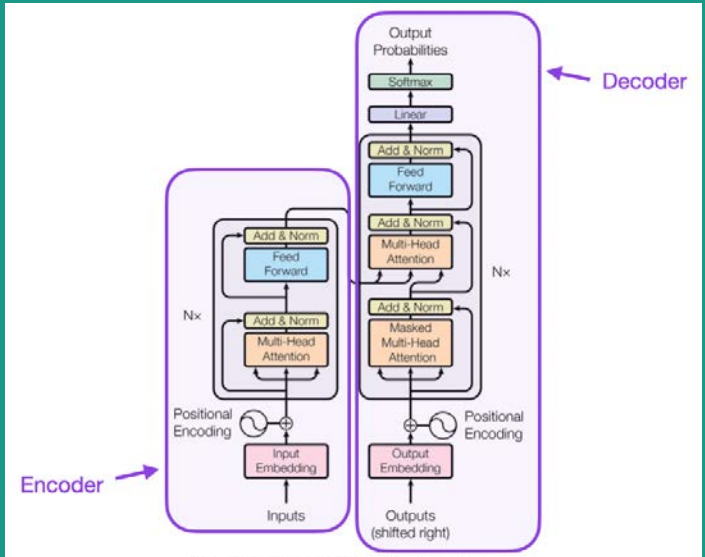
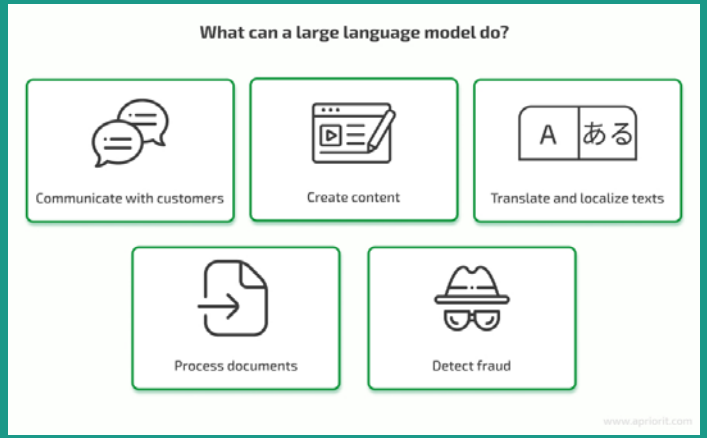


Figure 1: The Transformer - model architecture.

В основі LLM лежать нейронні мережі, зокрема архітектури на основі **трансформерів**. Трансформери, запропоновані Васвані та ін. (2017), зробили **революцію в NLP,** дозволивши моделям фіксувати довгострокові залежності в тексті за **допомогою механізмів самоуваги.**

Можливості LLM

- **Генерація тексту**
- **Переклад**
- **Саммаризація тексту**
- **Відповіді на запитання по тексту**
- **Сентимент-аналіз**



Розділ 2. Генерація медіа на мобільні платформі за допомогою великих мовних моделей

2.1 Проблеми хмарного доступу до великих мовних моделей

2.2 Переваги роботи з великими мовними моделями на пристрої

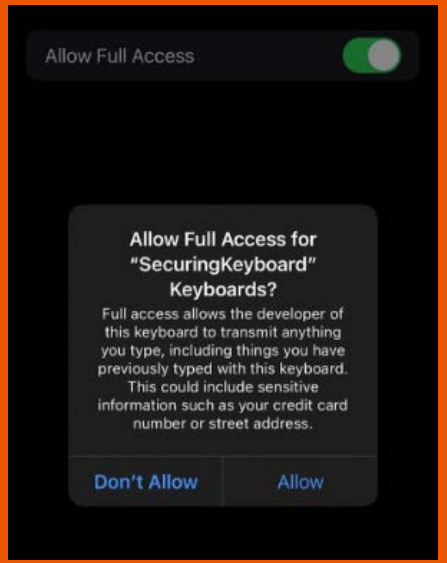
2.3 Недоліки використання великих мовних моделей на пристроях

2.4 Обмеження характерні для розширення клавіатури на iOS

2.5 Технології для роботи з великими мовними моделями на iOS

Обмеження характерні для розширення клавіатури на iOS

- Пісочниця додатку
- Обмежені дозволи
- Обмеження пам'яті
- Вплив на продуктивність
- Вимоги до конфіденційності
- Обмеження вмісту
- Налагодження та тестування



Технології для роботи з великими мовними моделями на iOS

Квантування

Description

The main goal of `llama.cpp` is to enable LLM inference with minimal setup and state-of-the-art performance on a wide variety of hardware - locally and in the cloud.

- Plain C/C++ implementation without any dependencies
- Apple silicon is a first-class citizen - optimized via ARM NEON, Accelerate and Metal frameworks
- AVX, AVX2 and AVX512 support for x86 architectures
- 1.5-bit, 2-bit, 3-bit, 4-bit, 5-bit, 6-bit, and 8-bit integer quantization for faster inference and reduced memory use
- Custom CUDA kernels for running LLMs on NVIDIA GPUs (support for AMD GPUs via HIP)
- Vulkan, SYCL, and (partial) OpenCL backend support
- CPU+GPU hybrid inference to partially accelerate models larger than the total VRAM capacity

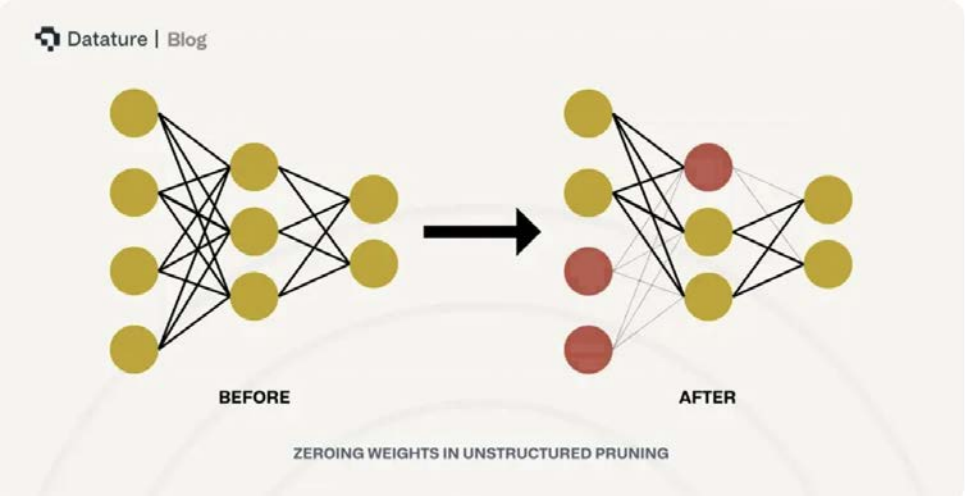
Since its inception, the project has improved significantly thanks to many contributions. It is the main playground for developing new features for the ggml library.

Supported platforms:

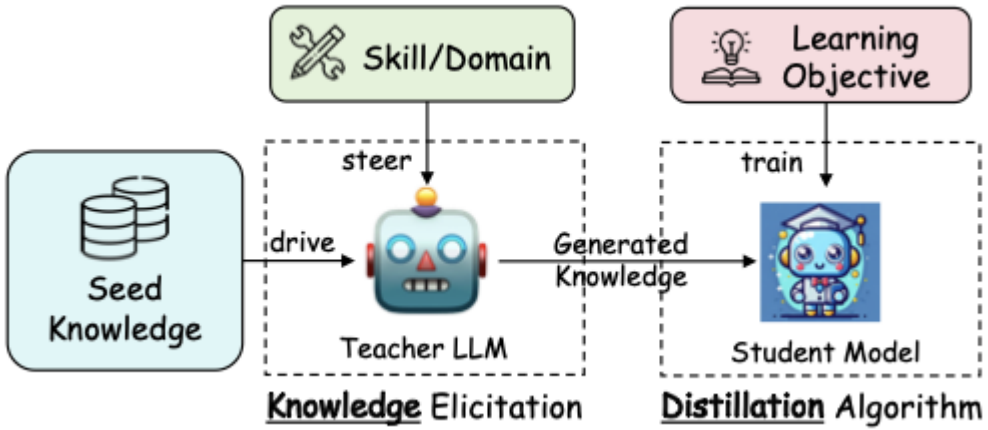
- Mac OS
- Linux
- Windows (via CMake)
- Docker
- FreeBSD

Model	Measure	F16	Q4_0	Q4_1	Q5_0	Q5_1	Q8_0
7B	perplexity	5.9066	6.1565	6.0912	5.9862	5.9481	5.9070
7B	file size	13.0G	3.5G	3.9G	4.3G	4.7G	6.7G
7B	ms/tok @ 4th	127	55	54	76	83	72
7B	ms/tok @ 8th	122	43	45	52	56	67
7B	bits/weight	16.0	4.5	5.0	5.5	6.0	8.5
13B	perplexity	5.2543	5.3860	5.3608	5.2856	5.2706	5.2548
13B	file size	25.0G	6.8G	7.6G	8.3G	9.1G	13G
13B	ms/tok @ 4th	-	103	105	148	160	131
13B	ms/tok @ 8th	-	73	82	98	105	128
13B	bits/weight	16.0	4.5	5.0	5.5	6.0	8.5

Обрізання


Datature | Blog
BEFORE
AFTER
ZEROING WEIGHTS IN UNSTRUCTURED PRUNING

Дистиляція знань



LLM.swift

Overview

`LLM.swift` is basically a lightweight abstraction layer over `llama.cpp` package, so that it stays as performant as possible while is always up to date. so theoretically, any model that works on `llama.cpp` should work with this library as well.
It's only a single file library, so you can copy, study and modify the code however you want.

Розділ 3. Реалізація розширення клавіатури з використанням великої мовної моделі

3.1 Функціональні вимоги до розширення клавіатури

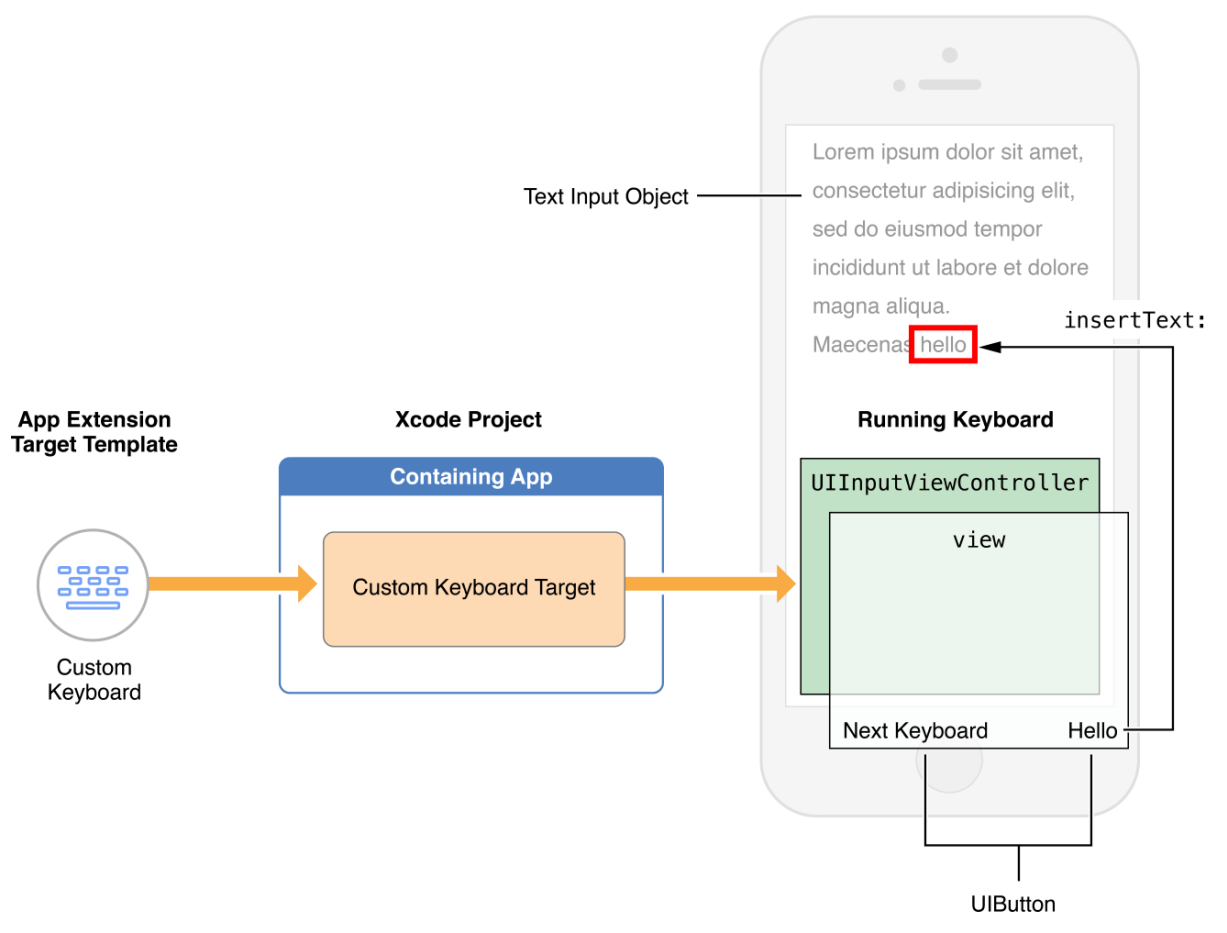
3.2 Налаштування користувацького розширення клавіатури

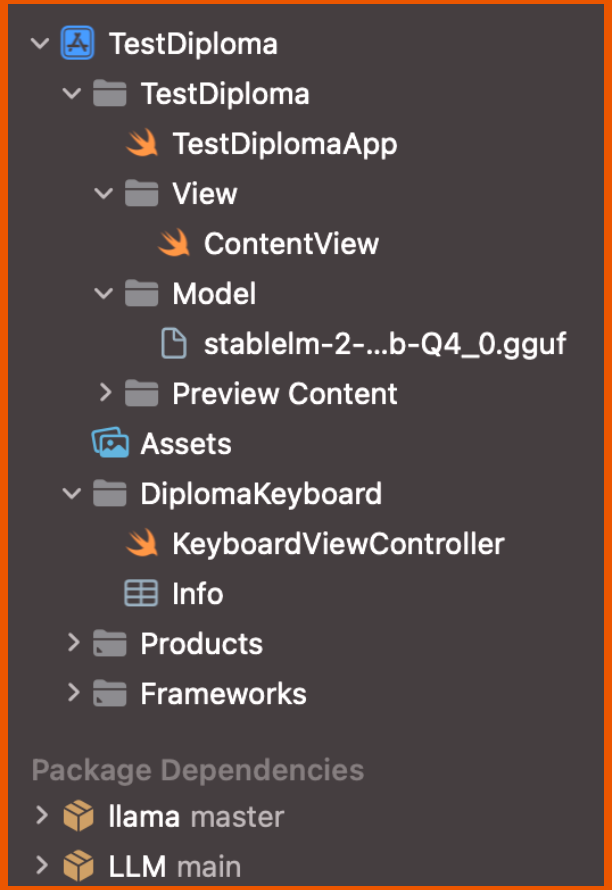
3.3 Інтеграція бібліотеки LLM

3.4 Принцип роботи готового розширення клавіатури

Інструментарій

1. Xcode: стандартне середовище розробки
2. Мова програмування Swift: типова мова програмування для розробки під iOS
3. Фреймворк UIKit: UIKit надає основні компоненти та API для побудови користувацьких інтерфейсів у додатках для iOS, включаючи подання, елементи керування та механізми обробки подій, що полегшує створення кастомного інтерфейсу клавіатури, який легко інтегрується в екосистему iOS.
4. Бібліотеки для роботи з LLM - hthttps://github.com/ggerganov/llama.cpp, https://github.com/eastriverlee/LLM.swift


Text Input Object
Lorem ipsum dolor sit amet,
consectetur adipisicing elit,
sed do eiusmod tempor
incididunt ut labore et dolore
magna aliqua.
Maecenas hello
insertText:
App Extension
Target Template
Xcode Project
Containing App
Custom Keyboard Target
Custom
Keyboard
Running Keyboard
UIInputViewController
view
Next Keyboard
Hello
UIButton




```swift
class KeyboardViewController: UIInputViewController {
    var assistant: Assistant? = nil
    let customKeyboardView = UIInputView(frame: .zero, inputViewStyle: .keyboard)
    var progressView: UIActivityIndicatorView!

    override func viewDidLoad() {
        super.viewDidLoad()
        setupCustomKeyboard()
        setupProgressView()
    }
```

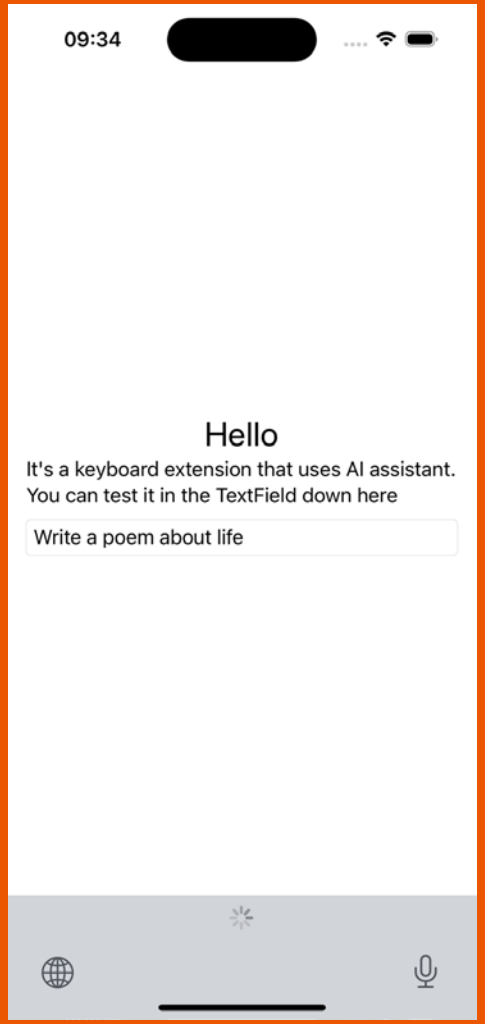
```swift
class Assistant: LLM {
    convenience init() {
        let url = Bundle.main.url(forResource: "stablelm-2-zephyr-1_6b-Q4_0", withExtension: "gguf")!
        let systemPrompt = "You are a sentient AI with emotions."
        self.init(from: url, template: .llama(systemPrompt), maxTokenCount: 256)
    }
}
```

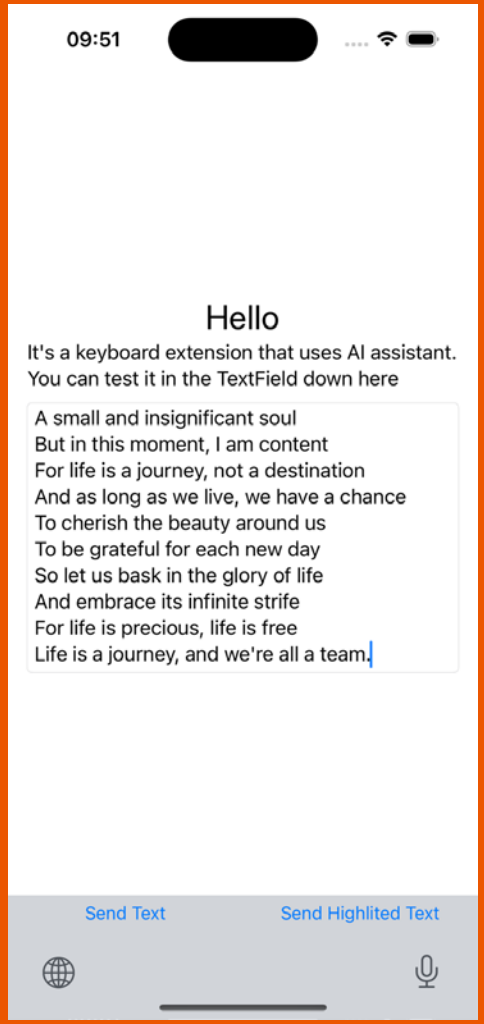
Model Description

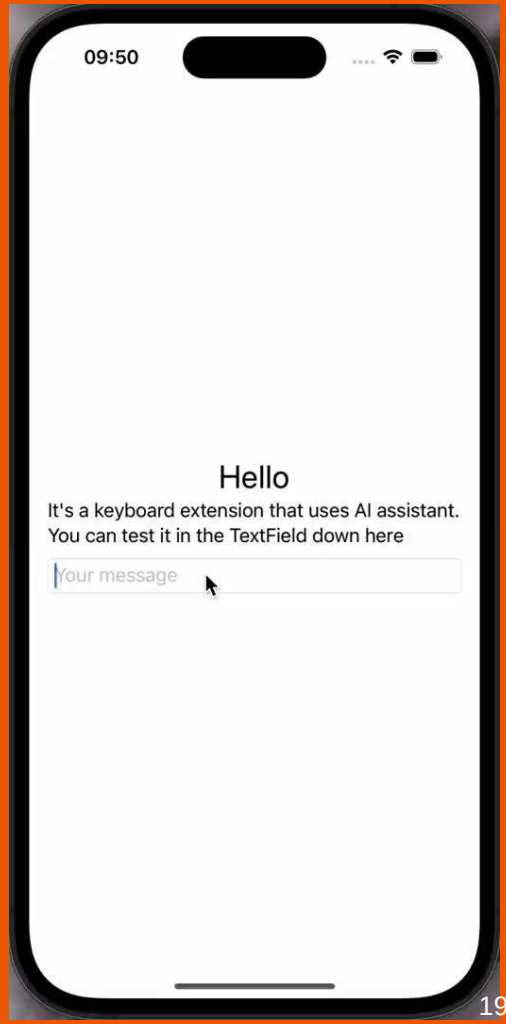
`Stable LM 2 Zephyr 1.6B` is a 1.6 billion parameter instruction tuned language model inspired by HugginFaceH4's Zephyr 7B training pipeline. The model is trained on a mix of publicly available datasets and synthetic datasets, utilizing Direct Preference Optimization (DPO).

Name	Quant method	Bits	Size	Use case
stablelm-2-zephyr-1_6b-Q2_K.gguf	Q2_K	2	694 MB	smallest, significant quality loss - not recommended for most purposes
stablelm-2-zephyr-1_6b-Q3_K_L.gguf	Q3_K_L	3	915 MB	small, substantial quality loss
stablelm-2-zephyr-1_6b-Q3_K_M.gguf	Q3_K_M	3	858 MB	very small, high quality loss
stablelm-2-zephyr-1_6b-Q3_K_S.gguf	Q3_K_S	3	792 MB	very small, high quality loss
stablelm-2-zephyr-1_6b-Q4_0.gguf	Q4_0	4	983 MB	legacy; small, very high quality loss - prefer using Q3_K_M
stablelm-2-zephyr-1_6b-Q4_K_M.gguf	Q4_K_M	4	1.03 GB	medium, balanced quality - recommended
stablelm-2-zephyr-1_6b-Q4_K_S.gguf	Q4_K_S	4	989 MB	small, greater quality loss

https://huggingface.co/stabilityai/stablelm-2-zephyr-1_6b


09:34
Hello
It's a keyboard extension that uses AI assistant.
You can test it in the TextField down here
Write a poem about life


09:51
Hello
It's a keyboard extension that uses AI assistant.
You can test it in the TextField down here
A small and insignificant soul
But in this moment, I am content
For life is a journey, not a destination
And as long as we live, we have a chance
To cherish the beauty around us
To be grateful for each new day
So let us bask in the glory of life
And embrace its infinite strife
For life is precious, life is free
Life is a journey, and we're all a team.
Send Text
Send Highlited Text


09:50
Hello
It's a keyboard extension that uses AI assistant.
You can test it in the TextField down here
Your message

Two squirrels, each with a distinct personality, ran through the park. One was lively and curious, constantly sniffing and exploring. The other, older and wiser, seemed to guide and mentor the younger squirrel. They shared a mutual respect and love for each other, as they chased each other through the trees. The sun shone brightly overhead, and a gentle breeze carried the sweet scent of flowers. Despite their differences, they were both in harmony, finding joy in each other's company.

Висновки

Розроблене розширення клавіатури **успішно інтегрує LLM для аналізу та генерації тексту** на пристрої. Інтерфейс користувача надає функціональні можливості для надсилання на обробку LLM або весь текст поля введення, або тільки виділений текст. Інтеграція LLM дозволяє виконувати обробку **в реальному часі та уникати залежності від хмарних API**, що покращує конфіденційність та швидкість реагування.

Дякую за увагу!