

● LIVE

สร้าง AI Chatbots สำหรับองค์กร



LangChain ร่วมกับ Next.JS และ supabase



มีวิดีโอบันทึกการอบรม
ย้อนหลังให้ทุกวัน



สถาบันไอทีเนียส

4 วัน
12 ชั่วโมงเต็ม



Samit Koyom
สถาบันไอทีเนียส





ร่วมกับ
และ

LangChain
Next.JS 


วิทยากร



อ.สา米ตร โภยม (ปาน)

ปริญญาโท คณะเทคโนโลยีและสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ



สถาบันไอทีเนี่ยส

▶ Frontend

Angular, React, Vue, Next, Nuxt, Bootstrap, Tailwind CSS

▶ Backend

PHP, Python, Java, Kotlin, Go, Rust, NodeJS, NestJS, .NET

▶ Database

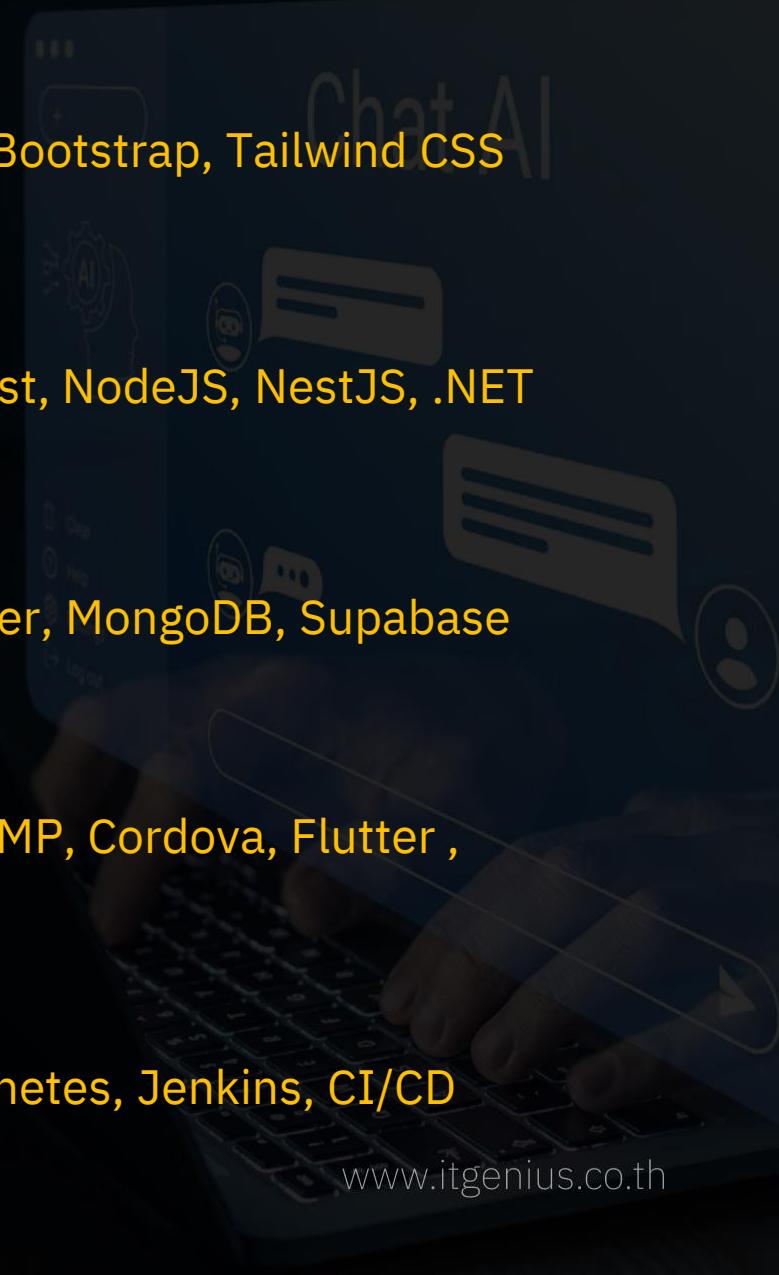
MySQL, PostgreSQL, MS SQL Server, MongoDB, Supabase

▶ Mobile

Java, Kotlin, Objective C, Swift, KMP, Cordova, Flutter ,
React Native, Expo

▶ DevOps

Git, Github, Gitlab, Docker, Kubernetes, Jenkins, CI/CD



www.itgenius.co.th

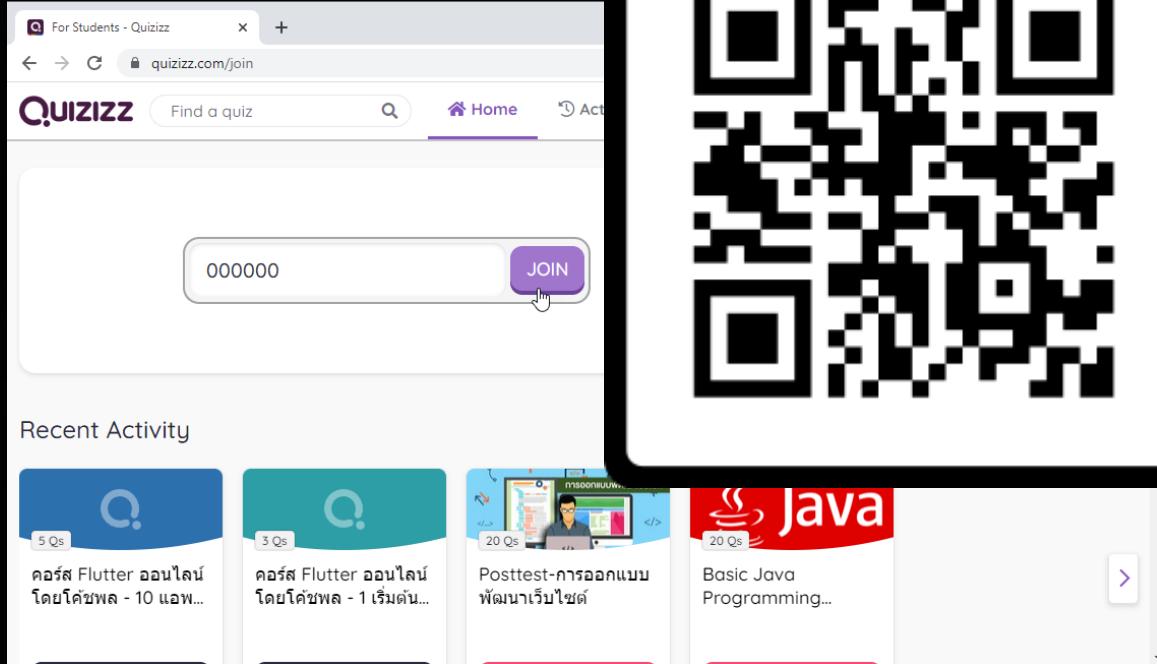
แบบทดสอบก่อนอบรม



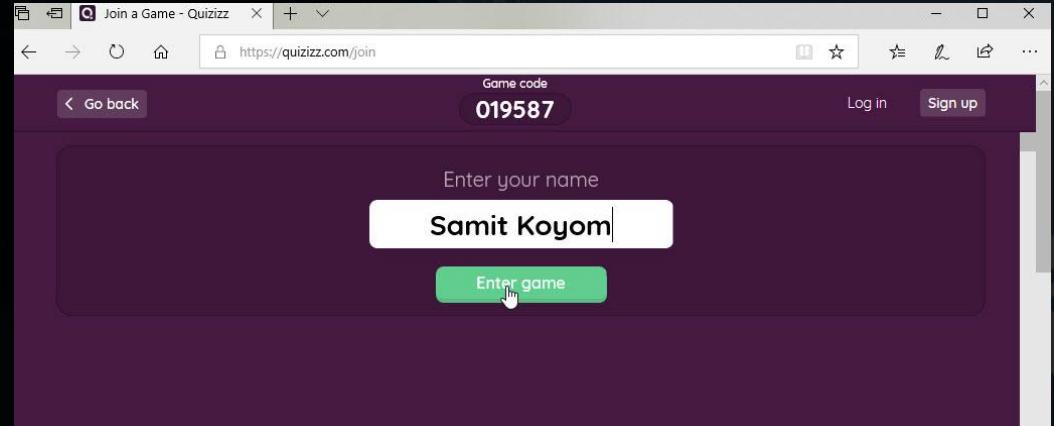
Pretest ทำแบบทดสอบก่อนเรียน

STEP 1: เข้าทำแบบทดสอบกี่เลิ�ก์ ป้อนรหัสเข้าห้องสอบ

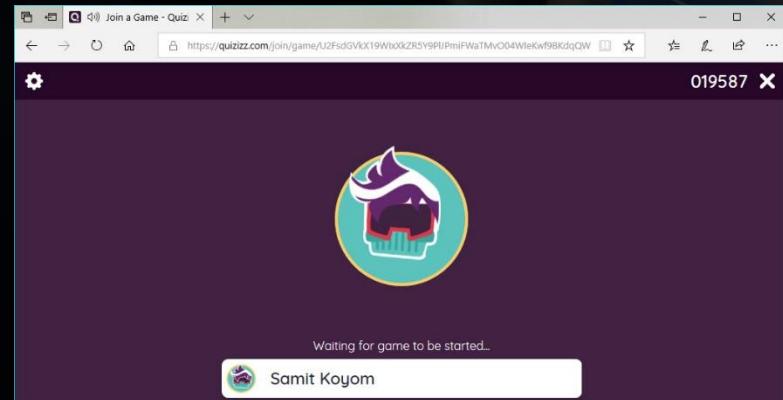
quizizz.com/join



STEP 2: ป้อนชื่อ



STEP 3: รอผู้สอน Start ข้อสอบ



สถาบันไอทีเจเนียส

www.itgenius.co.th



LangChain
ร่วมกับ **Next.JS** 
และ  **supabase**

ดาวน์โหลดเอกสารประกอบการอบรม

bit.ly/aichatbot-langchain



สถาบันไอทีจีเนียส

www.itgenius.co.th

Course Outline



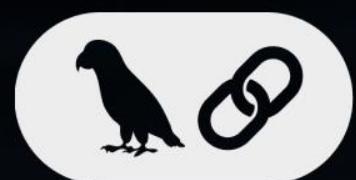
1. ภาพรวม AI Chatbot กับ Langchain.js
2. การพัฒนา Rest API ใน Next.js เพื่อใช้งานกับ Langchain.js
3. พื้นฐาน Langchain.js เชื่อมต่อกับ Gen AI
4. ระบบยืนยันตัวตนด้วย Supabase Auth
5. UI Chatbot ด้วย Prompt-kit-UI Shadcn/UI
6. AI Chatbot มีการเก็บประวัติ (Chat History)
7. เชื่อมต่อ AI กับเครื่องมือภายนอก (Tool Calling)
8. Document Loader, Embedding , Vector Store
9. พัฒนา RAG เพื่อให้ AI ตอบคำถามจากข้อมูลในเอกสารขององค์กร
10. การเผยแพร่ (Deployment) โปรเจ็คต์ไปใช้งานจริง

• LIVE

อบรมออนไลน์



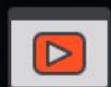
สร้าง AI Chatbots สำหรับองค์กร



LangChain

ร่วมกับ **Next.JS**

และ **supabase**



มีวิดีโอบันทึกการอบรม
ย้อนหลังให้ทุกวัน



สอนสดผ่าน Zoom
รับจำนำเน็ตกัด

Chat AI วันที่ 1

The slide features a large blue speech bubble containing icons related to AI, such as a brain with 'AI', a message bubble, and a person icon. A hand is shown typing on a laptop keyboard. To the right, there is a large red circle with a white stylized 'I' inside.

Samit Koyom
สถาบันไอทีจีเนียส

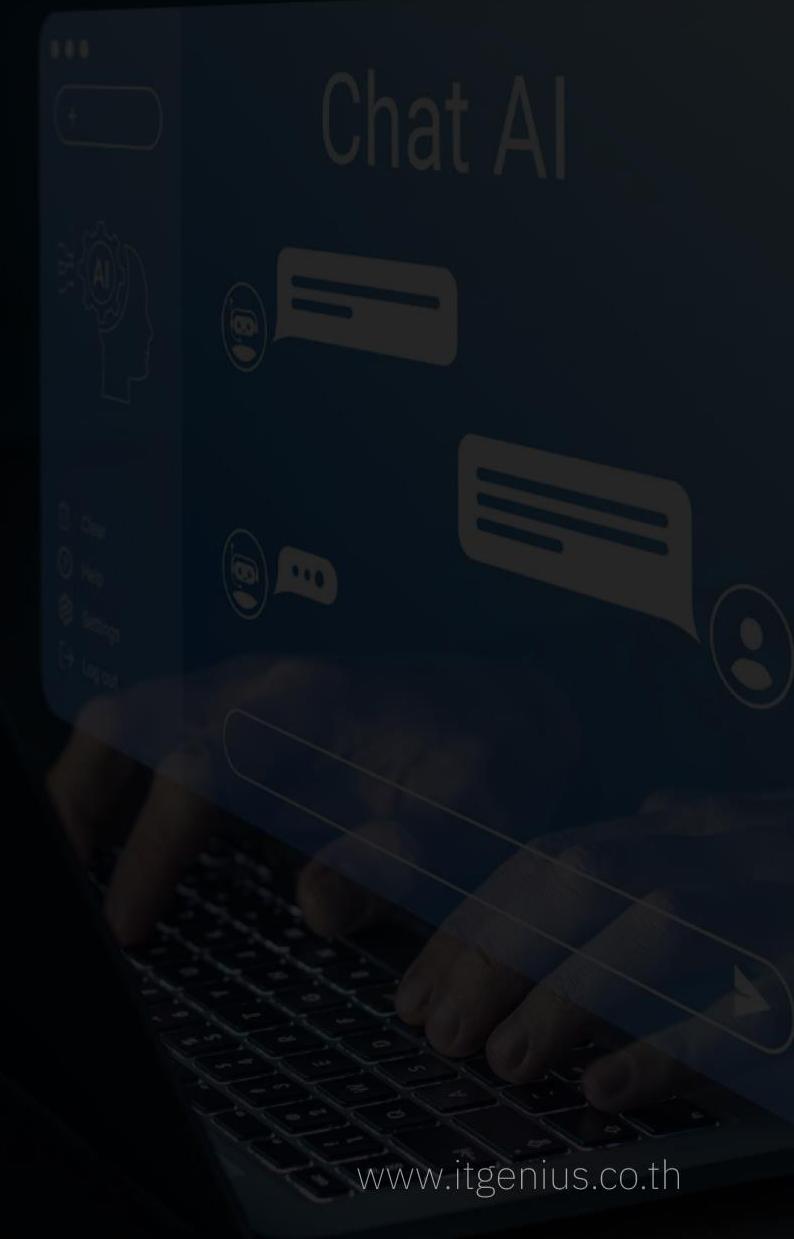


Day 1

1. การรวม AI Chatbot กับ Langchain.js
2. การพัฒนา Rest API ใน Next.js เพื่อใช้งานกับ Langchain.js
3. พื้นฐาน Langchain.js เชื่อมต่อกับ Gen AI

Workshop

AI Chatbot with LangChain and NextJS



สถาบันไอทีจีเนียส

www.itgenius.co.th

AI Genius

● บันทึกเรื่องราว AI, RAG, และ Document Loader

Genius AI Chatbot

สุดยอดแชทบอทจัดการข้อมูลด้วย AI, RAG (Retrieval-Augmented Generation), Document Loader & Vector Embeddings และ Tool Calling ค้นหาข้อมูลจากเอกสาร, ฐานข้อมูล, และซอฟต์แวร์ที่กำหนดเองด้วย LangChain.js, Next.js, Supabase และ OpenAI GPT-4o-mini เพื่อประสิทธิภาพและสนับสนุนทั่วโลก ปลอดภัย และรวดเร็ว

+ เริ่มต้นใช้งาน + เข้าสู่ระบบ

RAG & Document Search
ผลลัพธ์การค้นหาข้อมูลจากเอกสาร, ฐานข้อมูล และ AI agent
Vector Search + Structured Data

Tool Calling & Smart Query
ค้นหาข้อมูลเชิงลึก, ย้อนกลับ, และจัดการข้อมูลเชิงลึกจากฐานข้อมูล Supabase รองรับภาษาไทย-จีนตัวเต็ม, Partial Matching

Security & Modern UI
บล็อกผู้ใช้งาน Supabase Auth, RLS และ UI สวยงามเชิงลึก Responsive, TypeScript, Edge Runtime

Chat History System
บันทึกและจัดการประวัติการสอบถามแบบแยก session
Auto-title, Real-time, ลบ/รีเซ็ตหน้า

Advanced Memory Management
บันทึกและรีใช้ประวัติการสอบถามเพื่อเพิ่ม Context Window, Token Counting, Smart Summarization

Modern UI & Responsive Design
UI สวยงาม รับสมัย รองรับทุกแพลตฟอร์ม shadcn/ui, Tailwind CSS, Mobile Friendly

จุดเด่นที่ผู้ใช้ไว้วางใจ

10K+ ผู้ใช้งานจริง

99.9% Uptime & Reliability

5+ ระบบจัดการ (RAG, Tool Calling, Document Loader, Security, UI)

พร้อมให้บริการ 24/7 | รองรับภาษาไทย-จีนตัวเต็ม | ปลอดภัยและกันไวรัส

Landing Page

© 2025 Genius AI Chatbot. สงวนลิขสิทธิ์ | AI และ AI

AI Genius AI

สวัสดีครับ ร้านสิ่งของยุคใหม่ครับ

AI Chatbot พังก์ชันพื้นฐานครอบคลุม

ร้าน "เทคโนโลยี สมาร์ท สโตร์" ตั้งอยู่ที่:
ที่อยู่: 123 ถนนเทอดไท แขวงบางซื่อ เขตดินแดง กรุงเทพมหานคร 10400
โทรศัพท์: 02-123-4567
อีเมล: info@smarttechstore.com
เว็บไซต์: www.smarttechstore.com

หากต้องการข้อมูลเพิ่มเติบสามารถสอบถามได้โดยครับ!

iPhone 15 Pro Max ราคาเท่าไหร่?

iPhone 15 Pro Max ราคา 12,990 บาท

Macbook Pro ขายอยู่ที่ไหนครับ?

MacBook Pro M3 ขายอยู่ที่ร้าน Apple Store สาขาสุขุมวิท กรุงเทพฯ

MacBook ขายไปแล้วที่เครื่อง?

"MacBook Pro M3" ขายไปแล้วทั้งหมด 2 ชิ้น ครับ โดยมีรายละเอียดการขายดังนี้:

วันที่ขาย	จำนวนที่ขาย (ชิ้น)	ยอดขาย (บาท)
20/8/2568	1	79,900
25/8/2568	1	79,900

ยอดขายรวมทั้งหมดคือ 159,800 บาท หากต้องการข้อมูลเพิ่มเติบสามารถสอบถามได้โดยครับ!

Continue the conversation...

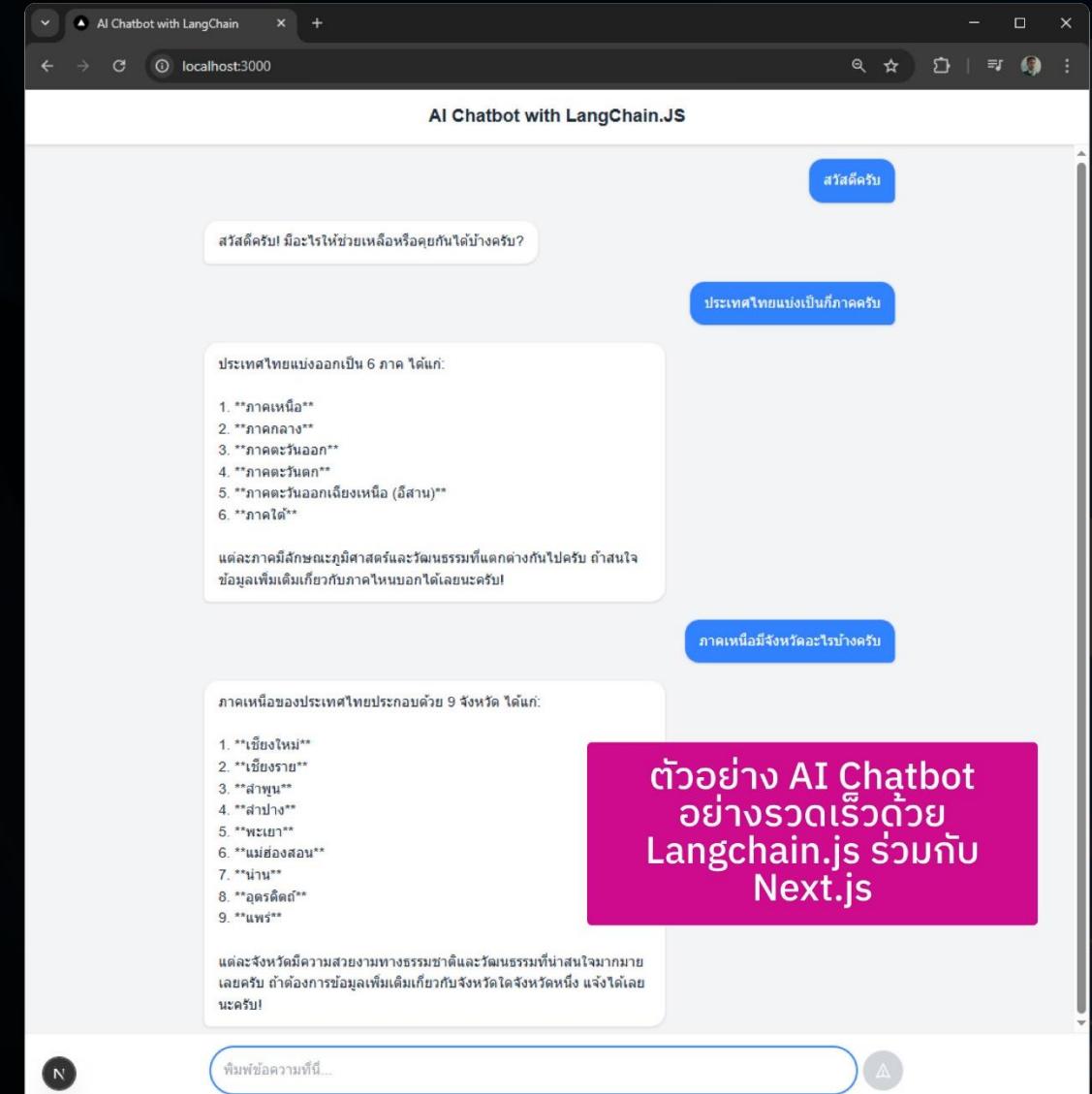
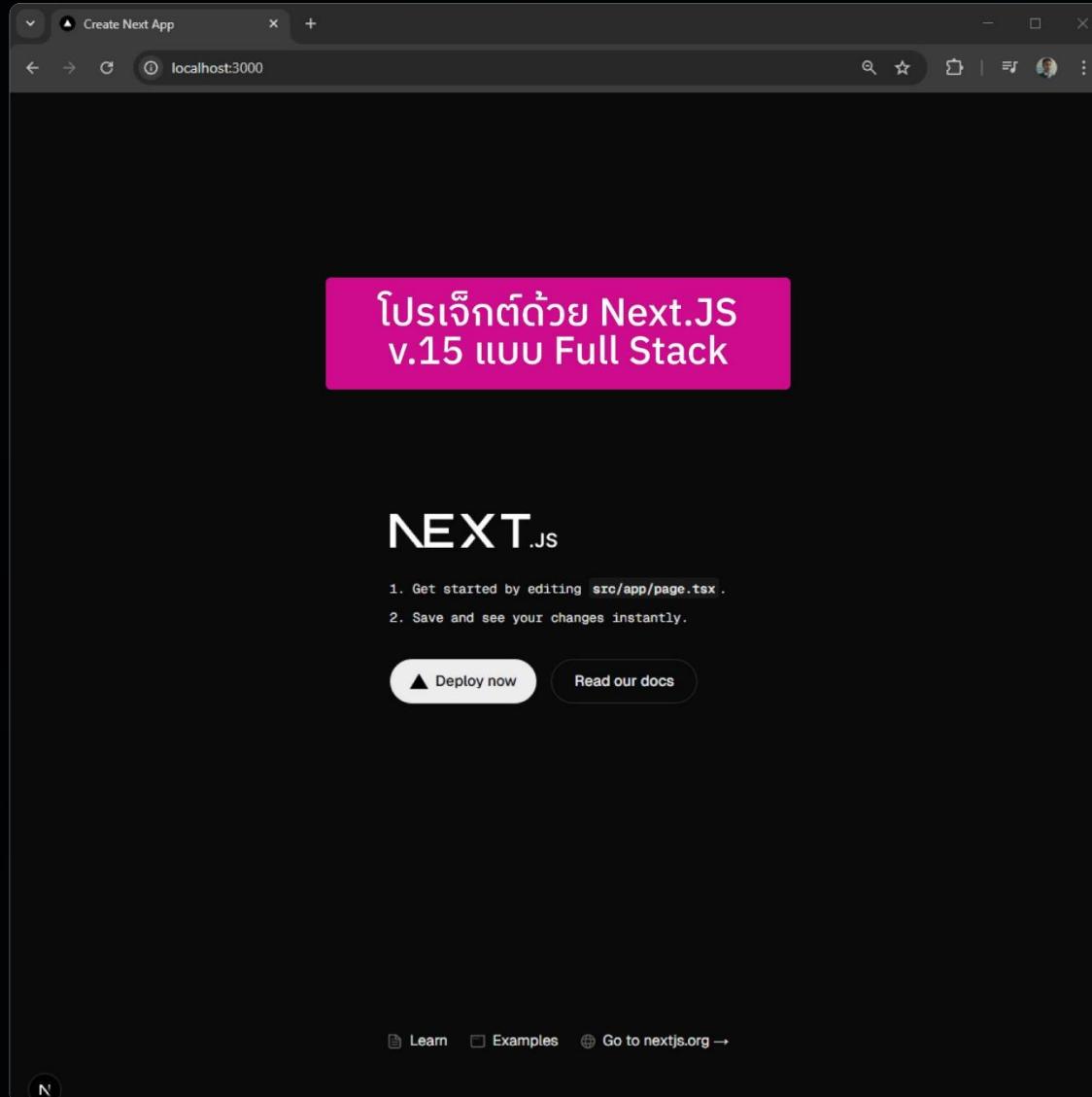
Samit samit@email.com

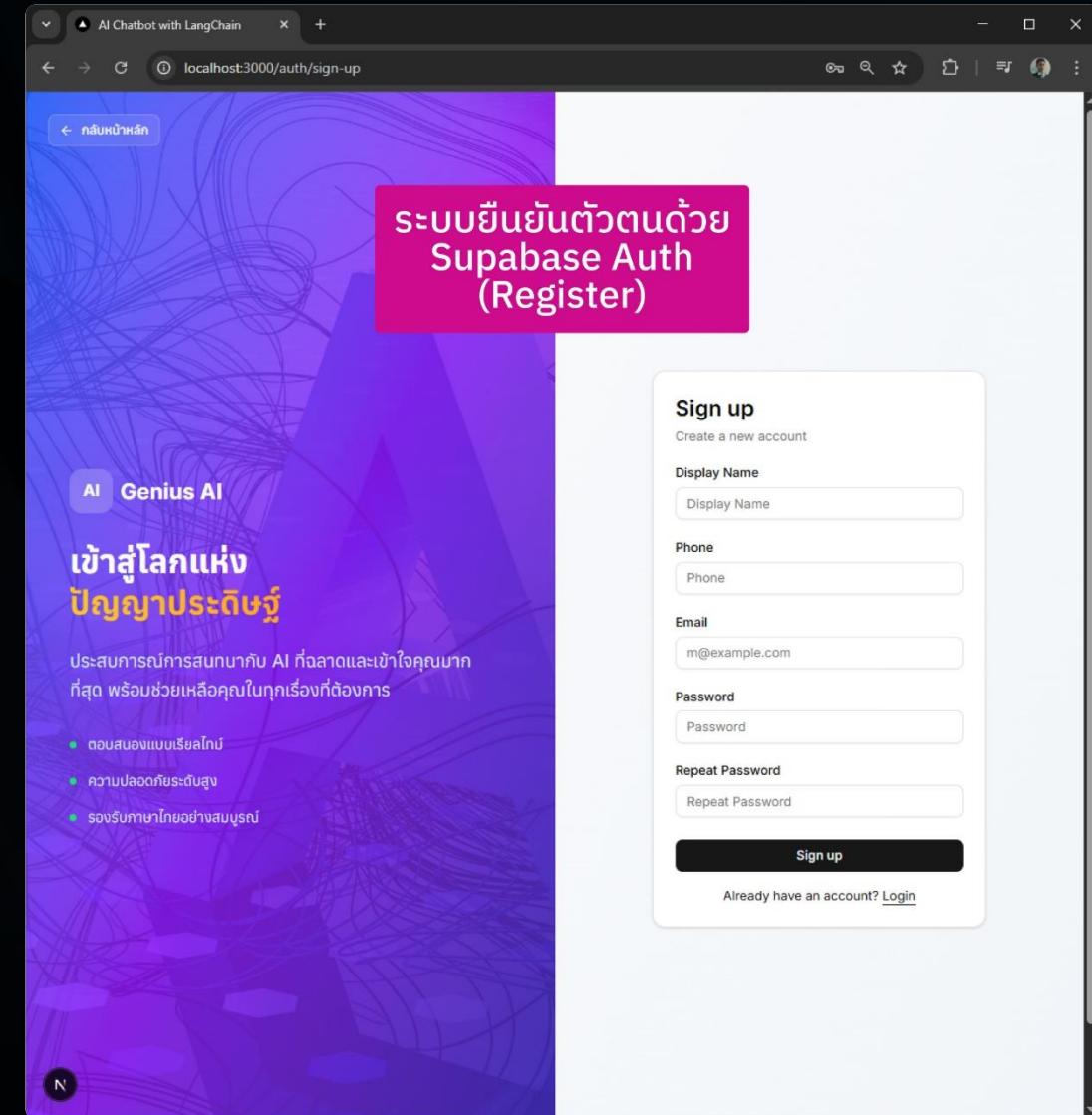
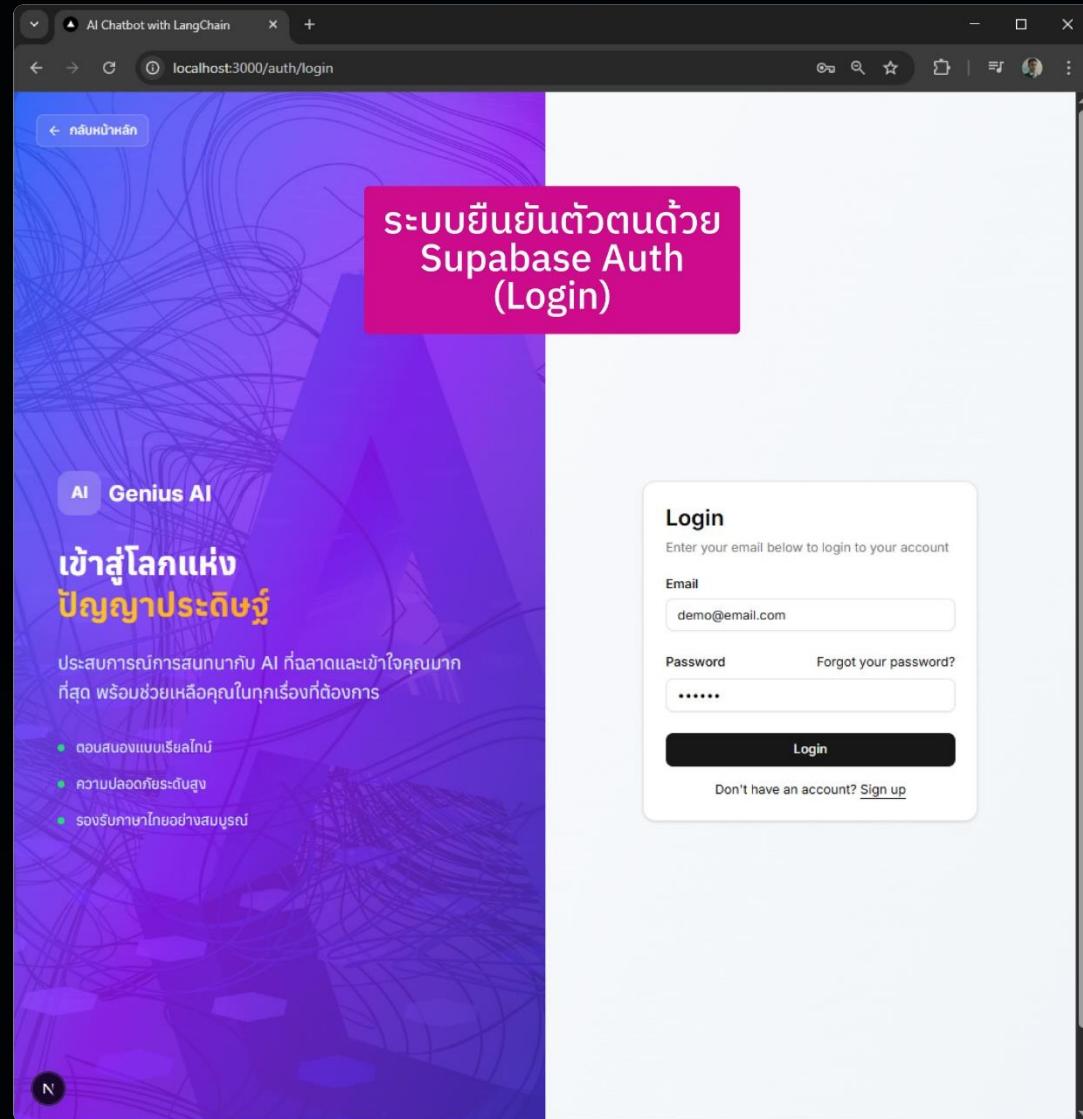
localhost:3000/chat/70e49f15-ed2b-4f16-9b14-1c0d86e2ee9a



สถาบันโปรแกรมเมอร์

www.itgenius.co.th





AI Chatbot with LangChain

localhost:3000/chat

Genius AI Chatbot

สวัสดีค่ะ! บอชเช่นเดียวกัน

สวัสดีคุณ!

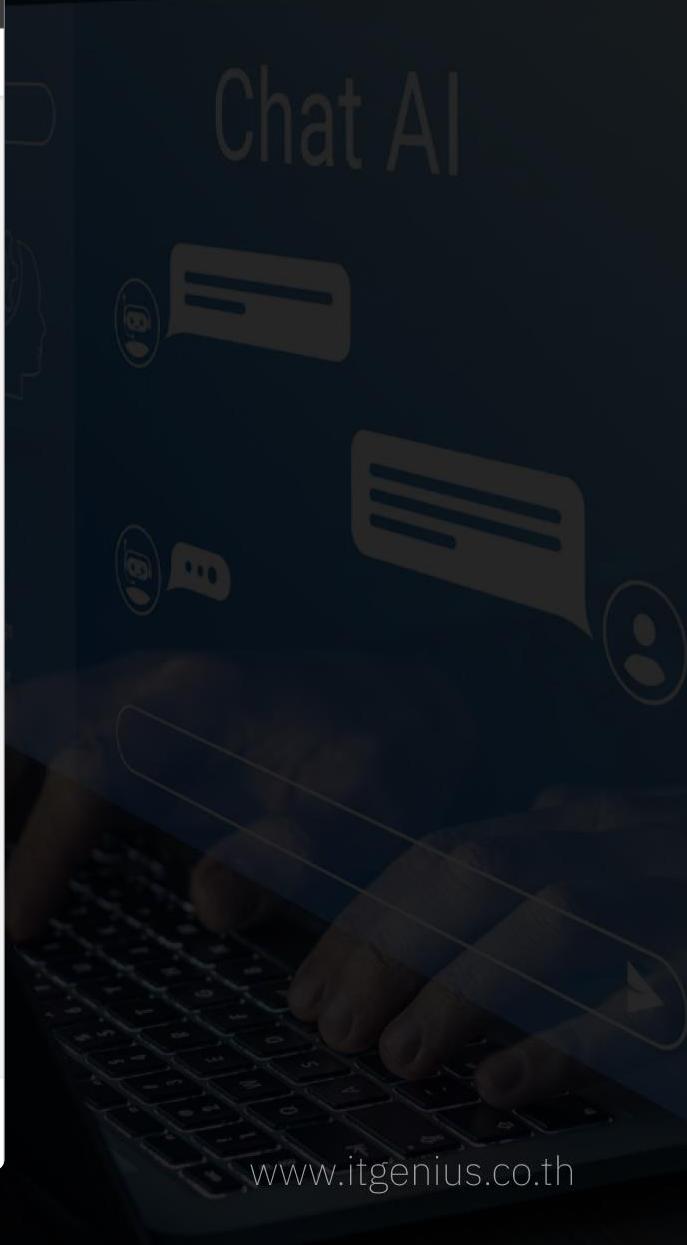
แนะนำอาหารไทยสัก 5 เมนูยอดนิยม

แนะนำอาหารไทยสัก 5 เมนูยอดนิยมที่คุณไม่ควรพลาด:

1. **ผัดไทย** - เส้นกุ้ง กุ้ง หรือไก่ พร้อมต้มยำ กับข้าวโรยด้วยถั่วสังข์
2. **แกงเขียวหวาน** - แกงเขียวหวาน กับไก่ กับข้าวโรยด้วยถั่วสังข์
3. **ต้มยำกุ้ง** - ซุปเปอร์แซดดี้เผ็ดและเปรี้ยว กับกุ้ง กับข้าวโรยด้วยถั่วสังข์
4. **ส้มตำ** - ส้มตำแซ่บๆ กับข้าวโรยด้วยถั่วสังข์ กับกุ้ง กับข้าวโรยด้วยถั่วสังข์
5. **ข้าวมันไก่** - ข้าวมันไก่ที่หุงจากน้ำต้มไก่ เสิร์ฟพร้อมไก่ต้มและน้ำจิ้น กับกุ้ง กับข้าวโรยด้วยถั่วสังข์

ดึงข้อมูลผู้ใช้ มาแสดงใน Chatbot AI

พิมพ์ข้อความที่นี่...



สถาบันไอทีเจเนียส

www.itgenius.co.th

prompt-kit.com/blocks

Full chat app

Preview Code

```
display: grid;
grid-template-columns: repeat(auto-fit, minmax(250px, 1fr));
gap: 1rem;
```

This creates a grid where:

- Columns automatically fit as many as possible
- Each column is at least 250px wide
- Columns expand to fill available space
- There's a 1rem gap between items

Would you like me to explain more about how this works?

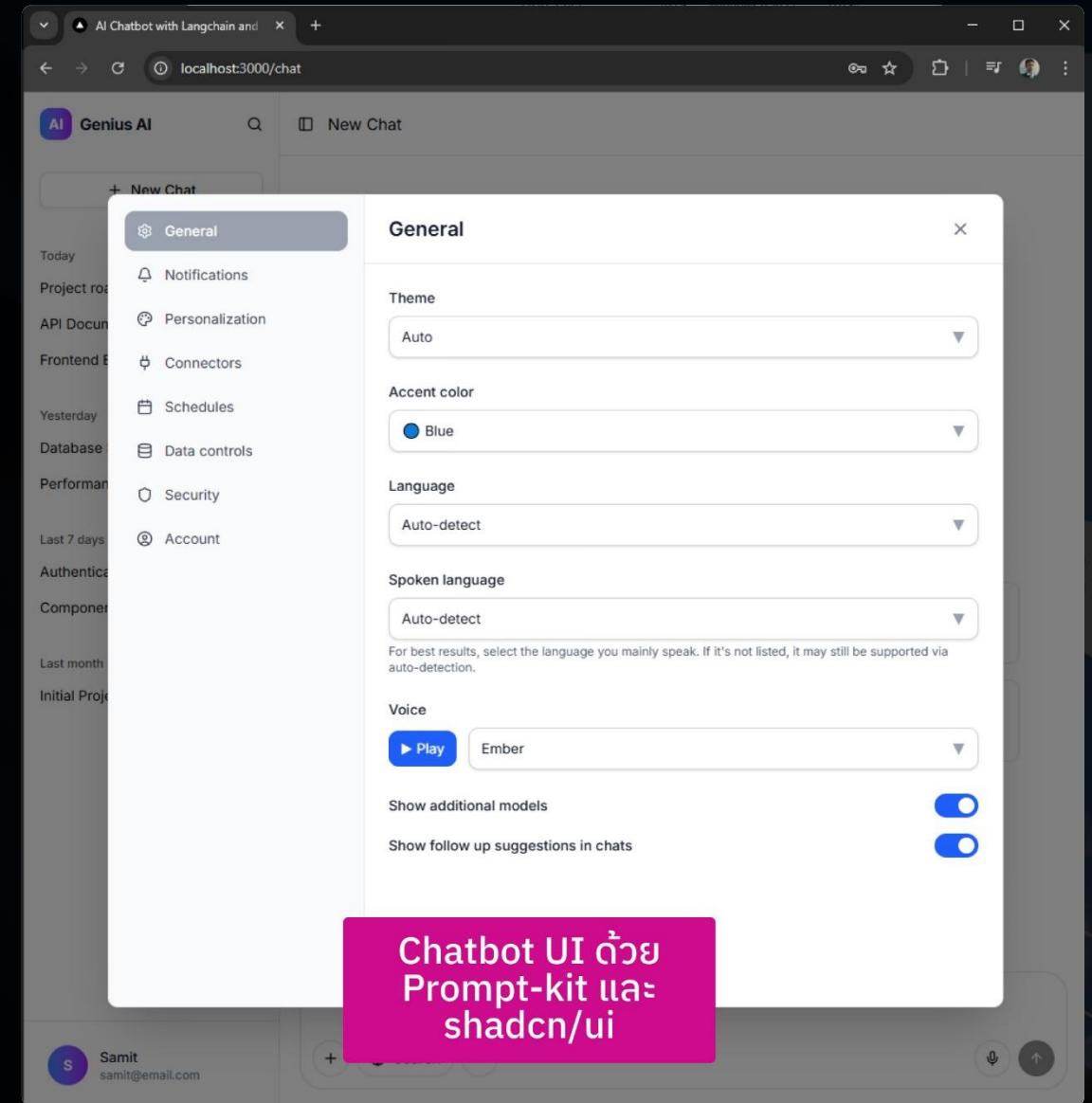
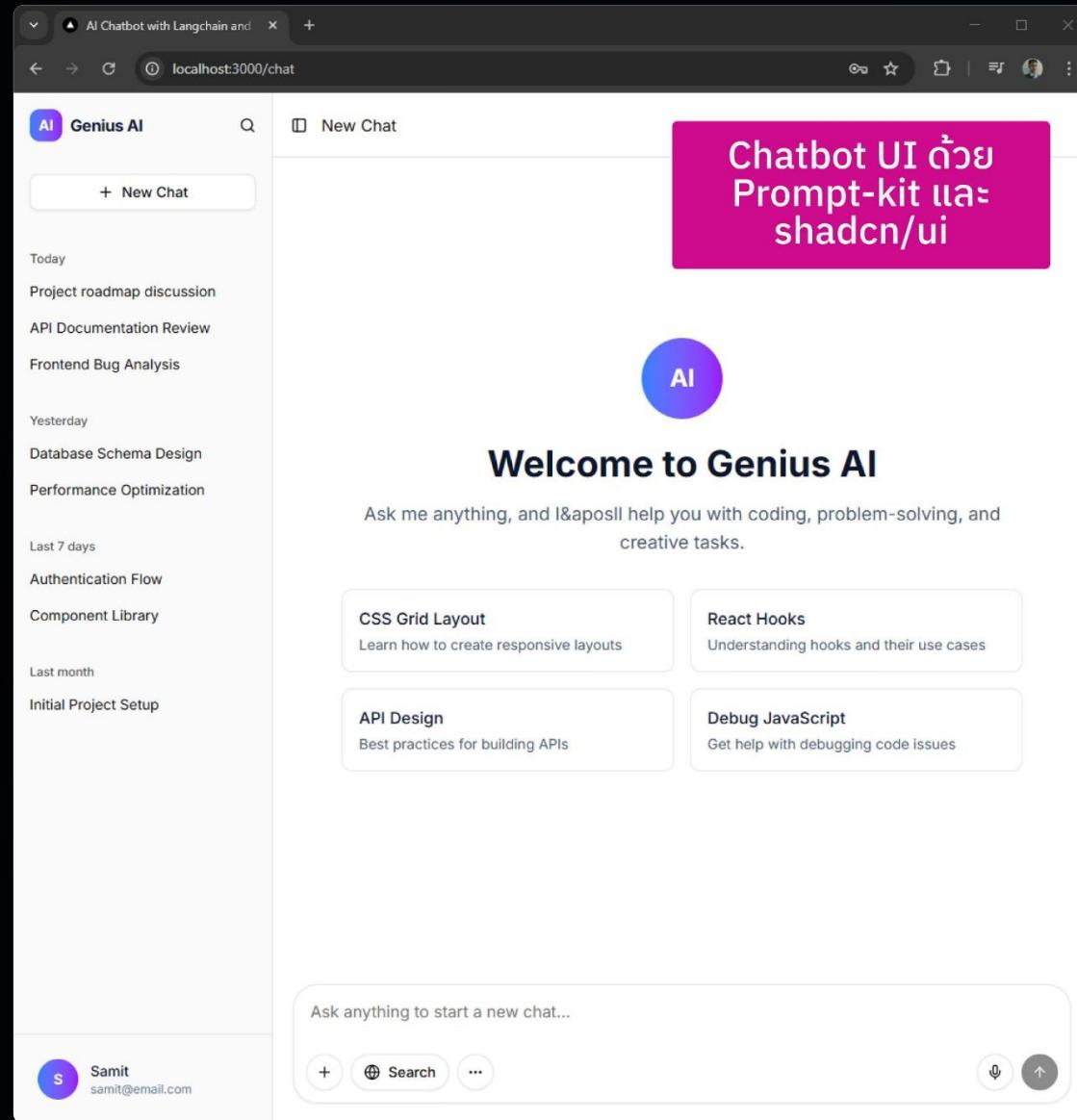
Ask anything

+ Search ...

↑ ↓

ปรับแต่ง Chatbot UI ด้วย prompt-kit





The screenshot shows a browser window with the URL `localhost:3000/chat/project-roadmap-discussion`. The interface has a sidebar on the left with a navigation menu:

- Today:
 - Project roadmap discussion (selected)
 - API Documentation Review
 - Frontend Bug Analysis
- Yesterday:
 - Database Schema Design
 - Performance Optimization
- Last 7 days:
 - Authentication Flow
 - Component Library
- Last month:
 - Initial Project Setup

A central panel displays a conversation:

Hello! Can you help me with a coding question?

Of course! I'd be happy to help with your coding question. What would you like to know?

How do I create a responsive layout with CSS Grid?

Creating a responsive layout with CSS Grid is straightforward. Here's a basic example:

```
// src/app/page.tsx
import { Button } from '@/components/ui/button'

export default function Page() {
  return (
    <main className='p-6'>
      <h1 className='text-2xl font-semibold'>Hello shadcn/ui</h1>
      <Button className='mt-4'>Press me</Button>
    </main>
  )
}
```

This creates a grid where:
Columns automatically fit as many as possible
Each column is at least 250px wide
Columns expand to fill available space
There is a gap between items

Ask anything

List top 5 frontend frameworks show in table

Chatbot UI ด้วย Prompt-kit และ shadcn/ui

Samit samit@email.com

localhost:3000/chat/project-roadmap-discussion



สถาบันไอทีเจเนียส

www.itgenius.co.th

New Chat

Welcome to Genius AI Chatbot

Ask me anything, and I'll help you with coding, problem-solving, and creative tasks.

CSS Grid Layout
Learn how to create responsive layouts

React Hooks
Understanding hooks and their use cases

API Design
Best practices for building APIs

Ask anything to start a new chat...

+ Search ...

AI Genius AI Q

+ New Chat

Today

- Project roadmap discussion
- API Documentation
- Frontend Bug Analysis

Yesterday

- Database Schema Design

Samit samit@email.com

- Upgrade plan
- Customize Genius AI
- Settings
- Log out

General Notifications Personalization

General

รอรับการแสดงผลบน Mobile Size

Blue

Language Auto-detect

Spoken language Auto-detect

Voice



AI Genius AI Q Project roadmap discussion

Light Dark System Today API Documentation Review Frontend Bug Analysis Yesterday Database Schema Design Performance Optimization Last 7 days Authentication Flow Component Library Last month Initial Project Setup

เลือกเปลี่ยน Theme Light / Dark / System ได้

Can you help me with a coding question?

Of course! I'd be happy to help with your coding question. What would you like to know?

How do I create a responsive layout with CSS Grid?

Creating a responsive layout with CSS Grid is straightforward. Here's a basic example:

```
// src/app/page.tsx
import { Button } from '@/components/ui/button'

export default function Page() {
  return (
    <main className='p-6'>
      <h1 className='text-2xl font-semibold'>Hello shadcn/ui</h1>
      <Button className='mt-4'>Press me</Button>
    </main>
  )
}
```

This creates a grid where:
Columns automatically fit as many as possible
Each column is at least 250px wide
Columns expand to fill available space
There's a 1rem gap between items
Would you like me to explain more about how this works?

List top 5 frontend frameworks show in table

Ask anything

+ Search ...

s Samit samit@email.com

AI Genius AI Q Project roadmap discussion

+ New Chat Today API Documentation Review Frontend Bug Analysis Yesterday Database Schema Design Performance Optimization Last 7 days Authentication Flow Component Library Last month Initial Project Setup

การแสดงผลแบบ Dark mode

Can you help me with a coding question?

Of course! I'd be happy to help with your coding question. What would you like to know?

How do I create a responsive layout with CSS Grid?

Creating a responsive layout with CSS Grid is straightforward. Here's a basic example:

```
// src/app/page.tsx
import { Button } from '@/components/ui/button'

export default function Page() {
  return (
    <main className='p-6'>
      <h1 className='text-2xl font-semibold'>Hello shadcn/ui</h1>
      <Button className='mt-4'>Press me</Button>
    </main>
  )
}
```

This creates a grid where:
Columns automatically fit as many as possible
Each column is at least 250px wide
Columns expand to fill available space
There's a 1rem gap between items
Would you like me to explain more about how this works?

List top 5 frontend frameworks show in table

Ask anything

+ Search ...

s Samit samit@email.com



AI Chatbot with Langchain and

localhost:3000/chat/9dab3657-10a3-4b06-93b4-164c9adaf19e

Genius AI

+ New Chat

Today

สวัสดีครับ รับต่อ อ.สาบีต นะครับ โปรดเรียกผมตัวย่อ...

การดูบ่มีค่าธรรมนิใช้เด็กช่วยเหลือในการเดินทาง พร้อมการเรียบเร้นคำศัพท์ใหม่ ๆ ภาษาอังกฤษ และการแท็ปญหา "My Little Pony":

การดูบ่มีสอนเรื่องมิตตรภาพ ความร่วมมือ และการแท็ปญหาผ่านตัวละครที่บ่ารัก "Peppa Pig":

การดูบ่มีบันการเรียนรู้เกี่ยวกับวิธีดูบ่มี ความสันพันธ์ในครอบครัว และการสื่อสาร "Avatar: The Last Airbender":

การดูบ่มีเมื่อเรื่องสักขีงเกี่ยวกับการเดินทาง การเดินดู และการรับผิดชอบ มีการสอนเรื่องคุณธรรมและค่าอ่อนน้อมถ่อมตน "Sesame Street":

รายการที่มีการสอนภารกิจการอ่าน การเขียน และการคำนวณ รวมถึงการสอนเรื่องความน่ารักและการทำงานร่วมกัน "Kazoops!":

การดูบ่มีเป็นการสร้างสรรค์และการคิดสร้าง "Tayo the Little Bus":

การดูบ่มีสอนเรื่องการเดินทาง การเดินทาง และการเดินทางไปยังสถานที่ต่างๆ ที่มีความสนุกสนาน หากต้องการการแนะนำเพิ่มเติมเกี่ยวกับการดูบ่มีคำนับบุ๊ตๆ ขึ้นดีช่วยเสมอครับ!

Continue the conversation...

Samit
samit@email.com

my-langchain-chatbot

route.ts

```

p > api > chat_06_history_optimize > route.ts > ...
/** 
 * =====
 * API Route สำหรับ Chat ที่มีการเก็บประวัติและ Optimize
 * =====
 *
 * ที่ใช้รหัส :
 * - เก็บประวัติการสนทนาใน PostgreSQL
 * - ทำ Summary เพื่อประหยัด Token
 * - Trim Messages เพื่อไม่ให้เกิน Token Limit
 * - Streaming Response สำหรับ Real-time
 * - จัดการ Session ID อัตโนมัติ
 */
import { NextRequest } from 'next/server'
import { ChatOpenAI } from '@langchain/openai'
import { ChatPromptTemplate, MessagesPlaceholder } from '@langchain/core/prompts'
import { toUIMessageStream } from '@ai-sdk/langchain'
import { createUIMessageStreamResponse, UIMessage } from 'ai'
import { PostgresChatMessageHistory } from '@langchain/community/stores/messages'
import { Pool } from 'pg'

import { BaseMessage, AIMessage, HumanMessage, SystemMessage, MessageContent, messages }
import { trimMessages } from '@langchain/core/messages'
import { StringOutputParser } from '@langchain/core/output_parsers'
import { encodingForModel } from '@langchain/core/utils/tiktoken'

export const dynamic = 'force-dynamic'
export const maxDuration = 30

// =====
// การตั้งค่า PostgreSQL Connection Pool

```

Analyzing 'route.ts' and its dependencies

In 8, Col 1 (34 selected) Spaces

มีการรำ Summary Chat เพื่อประหยัด Token ในการเรียก AI



AI Chatbot with Langchain and Streamlit

localhost:3000/chat/be480662-1f14-43a5-801f-8d59faf0b5b6

ให้ AI เชื่อมต่อ กับ Tool ภายนอก เช่นฐานข้อมูลของเรา เพื่อสอบถามข้อมูลของคุณ

เก็บสืบเนื่องมาสู่ห้องครับ

Today

สวัสดีครับ ร้านดังอยู่ที่ไหนครับ

Yesterday

Gaming Mouse ขายไปได้กี่ชิ้นแล้วครับ

มีมาส์สำหรับเกมส์บีบหรือไม่ครับ?

MacBook Pro M3 ราคาเท่าไหร่...

ขอรายชื่อประเภทพิมพ์เมื่อหลัง...

การ Hello World ในภาษา Asse...

Explain React hooks and wh...

ขอ best practice ในการออกแบบ...

Last 7 days

เฉลี่ยราคาสต็อกแบบสัปดาห์ ให้ฟ...

ขอเนื้อเพลง hotel califonia ขอ...

How do I create a responsiv...

ขอรายชื่อบัตรเดบิตในไทยตอนโล...

Help me debug this JavaScr...

How do I create a responsiv...

แนะนำอาหารไทยยอดนิยม 5 รายการ...

สวัสดีครับ แผนที่ สถานที่ บาร์...

ประวัติการขายของสินค้า "Gaming Mouse" มีดังนี้:

วันที่ขาย	จำนวนที่ขาย (ชิ้น)	ยอดขาย (บาท)
21/8/2568	2	5,600
26/8/2568	5	14,000

สรุป:

ขายรวม: 7 ชิ้น

ยอดขายรวม: 19,600 บาท

หากต้องการข้อมูลเพิ่มเติมหรือสอบถามค่าใช้จ่ายอื่นๆ สามารถสอบถามได้ค่ะ!

ยอดขายระหว่าง MacBook กับ mouse ต่างกันเท่าไหร่ครับ

ยอดขายรวมของสินค้าแต่ละรายการมีดังนี้:

MacBook Pro M3: 159,800 บาท

Gaming Mouse: 19,600 บาท

การคำนวณความแตกต่าง:

159,800 - 19,600 = 140,200 บาท

ตั้งนี้ ยอดขายระหว่าง MacBook Pro M3 กับ Gaming Mouse ต่างกัน 140,200 บาทค่ะ หากมีคำาณเพิ่มเติม สามารถสอบถามได้ค่ะ!

Continue the conversation...

+ Search ...

S Samit samit@email.com

localhost:3000/chat/be480662-1f14-43a5-801f-8d59faf0b5b6

AI Genius AI

สวัสดีครับ ร้านดังอยู่ที่ไหนครับ

+ New Chat

Today

สวัสดีครับ ร้านดังอยู่ที่ไหนครับ

Yesterday

Gaming Mouse ขายไปได้กี่ชิ้นแล้วครับ

มีมาส์สำหรับเกมส์บีบหรือไม่ครับ ...

MacBook Pro M3 ราคาเท่าไหร่...

ขอรายชื่อประเภทพิมพ์เมื่อหลัง...

การ Hello World ในภาษา Asse...

Explain React hooks and wh...

ขอ best practice ในการออกแบบ...

Last 7 days

เฉลี่ยราคาสต็อกแบบสัปดาห์ ให้ฟ...

ขอเนื้อเพลง hotel califonia ขอ...

How do I create a responsiv...

ขอรายชื่อบัตรเดบิตในไทยตอนโล...

Help me debug this JavaScr...

How do I create a responsiv...

แนะนำอาหารไทยยอดนิยม 5 รายการ...

สวัสดีครับ แผนที่ สถานที่ บาร์...

ร้าน "เทคโนโลยี สมาร์ท สโตร์" ตั้งอยู่ที่:
ที่อยู่: 123 ถนนเทคโนโลยี แขวงบึงกุ่ม เขตดินถัง กรุงเทพมหานคร 10400
โทรศัพท์: 02-123-4567
อีเมล: info@smarttechstore.com
เว็บไซต์: www.smarttechstore.com

หากต้องการข้อมูลเพิ่มเติมสามารถสอบถามได้เลยครับ!

iPhone 15 Pro Max ราคาเท่าไหร่?

ขออภัยครับ ไม่พบข้อมูลเกี่ยวกับสินค้า "iPhone 15 Pro Max" ในฐานข้อมูล หากต้องการสอบถามเกี่ยวกับสินค้ารุ่นอื่น ๆ ห้ามข้อมูลเพิ่มเติม สามารถแจ้งได้เลยครับ!

ขอ Macbook Pro บี้กครับ ราคาเท่าไหร่

สามารถ AI ให้ไปค้น
ข้อมูลจากเอกสารองค์กรได้

Continue the conversation...

+ Search ...

S Samit samit@email.com

www.itgenius.co.th



ITGenius Default project

Dashboard Docs API S

Settings Your profile Organization General API keys Admin keys People Projects Billing Limits Usage Data controls Project General API keys Webhooks People Limits

API keys

+ Create new secret key

You have permission to view and manage all API keys in this organization.

Do not share your API key with others or expose it in the browser or other client-side code. To protect your account's security, OpenAI may automatically disable any API key that has leaked publicly.

View usage per API key on the [Usage page](#).

NAME	SECRET KEY	PROJECT ACCESS	CREATED BY	PERMISSIONS
ai-chatbot-langchain	sk-...MJgA	Default project	Samit Koyom	All
n8n-api-key	sk-...Ve8A	SampleOpenAP...	Samit Koyom	All
n8n sample	sk-...blcA	Default project	Samit Koyom	All
n8n test	sk-...iIYA	Default project	Samit Koyom	All

ใช้ Open AI หรือ Model Opensource ฟรีก็ได้

<> Cookbook

坛坛 Forum

OpenRouter Search /

Models Chat Rankings Docs

Meta: Llama 3.3 70B Instruct (free)

meta-llama/llama-3.3-70b-instruct:free

Created Dec 6, 2024 | 65,536 context | \$0/M input tokens | \$0/M output tokens

The Meta Llama 3.3 multilingual large language model (LLM) is a pretrained and instruction tuned generative model in 70B (text in/text out). The Llama 3.3 instruction tuned text only model is optimized for multilingual dialogue use cases and outperforms many of the available open source and closed

Free Model weights

Overview Providers Apps Activity Uptime API

ใช้ Open AI หรือ Model Opensource ฟรีก็ได้

Providers for Llama 3.3 70B Instruct (free)

OpenRouter routes requests to the best providers that are able to handle your prompt size and parameters, with fallbacks to maximize uptime.

Sort by

Venice	Latency	Throughput	Uptime
fp8	0.77s	121.1tps	
Total Context Max Output Input Price Output Price Cache Read Cache Write Input Audio Input Audio Cache	65.5K 65.5K \$0 \$0 -- -- -- --		

Together	Latency	Throughput	Uptime
US fp8	0.88s	146.2tps	
Total Context Max Output Input Price Output Price Cache Read Cache Write Input Audio Input Audio Cache	131.1K 2.0K \$0 \$0 -- -- -- --		

Meta	Latency	Throughput	Uptime



A screenshot of a web browser window titled "Download Ollama on Windows". The URL in the address bar is "ollama.com/download/windows". The page content includes a navigation bar with links to "Models", "GitHub", "Discord", and "Turbo", and a search bar labeled "Search models". The main section is titled "Download Ollama" and features icons for "macOS", "Linux", and "Windows". The "Windows" icon is highlighted with a light gray background. Below the icons is a large black button with white text that says "Download for Windows". A small note below the button states "Requires Windows 10 or later".

Download Ollama

macOS

Linux

Windows

Download for Windows

Requires Windows 10 or later



Qwen/Qwen2.5-7B-Instruct · [Hu](#)

huggingface.co/Qwen/Qwen2.5-7B-Instruct

Hugging Face Search models, datasets, users...

Models Datasets Spaces Docs Pricing

Qwen/Qwen2.5-7B-Instruct like 778 Follow Qwen 48.1k

Text Generation Transformers Safetensors English qwen2 chat conversational text-generation-inference

arxiv:2309.00071 arxiv:2407.10671 License: apache-2.0

Model card Files xet Community 22 Train Deploy Use this model

Qwen2.5-7B-Instruct

[Qwen Chat](#)

Introduction

Qwen2.5 is the latest series of Qwen large language models. For Qwen2.5, we release a number of base language models and instruction-tuned language models ranging from 0.5 to 72 billion parameters. Qwen2.5 brings the following improvements upon Qwen2:

- Significantly **more knowledge** and has greatly improved capabilities in **coding** and **mathematics**, thanks to our specialized expert models in these domains.

Downloads last month
11,201,081

Safetensors Model size 7.62B params Tensor type BF16 Chat template Files info

Inference Providers NEW Together AI Text Generation Examples

Run 15,000+ Models Instantly

Inference Providers let you run inference on thousands of models served by our partners using a simple, unified, OpenAI-compatible serverless API ([Learn more](#)).



Samit Koyom's projects Hobby ai-chatbot-langchain

Find... Feedback

Overview Deployments Analytics Speed Insights Logs Observability Firewall Storage Flags Settings

ai-chatbot-langchain

Repository Usage Domains Visit

Production Deployment

Deployment
ai-chatbot-langchain-6y2q9lbc-samit-koyoms-projects.vercel.app

Status Created
Ready 4h ago by Ia

Domains
ai-chatbot-langchain.vercel.app

Source
main
d740a66 initial commit

Deployment Configuration Fluid Compute Deployment Protection Skew Protection Cold Start Prevention

To update your Production Deployment, push to the main branch.

Deployments

Firewall 24h Enable Bot Protection Firewall is active

Observability 6h Edge Requests 161 Function Invocations 40 Error Rate 0%

Analytics Track visitors and page views Enable

Active Branches

ເພີ້ມແປ່ວ່າ (Deployment) ບໍລິຫານ Vercel Cloud ພົມ

ai-chatbot-langchain.vercel.app/chat/96da7afa-73c1-46eb-b129-fb6408232da3

Genius AI

+ New Chat

Today

ມີ macbook pro ລາຄາເກົ່າໄຮ້
ສະເລັດຄົນ ສອບຄານເຮືອງແນວຕົດຄ...

ມີເມາສີເກີນບັນໄຣສາຍ ລາຄາ 2,800 ບາກ ຈຳນວນໃນສັດຖຸ 50 ຊົ່ວໂມງ ໂດຍບໍ່ມາລະເຊີຍດວ່າເປັນເມາສີທີ່ອຸບສອນຈັນໄວ້ຄວົນ

ຍອດຂາຍເມາສີເປັນອ່າງໂປ່ານ

ຂອງກົຍຄົນ ຂະນະນີ້ໄປສາມາຄເຫົ່າເຖິງຂ້ອນບຸລຍອດຂາຍຂອງເມາສີເກີນບັນໄຣສາຍໄດ້ ກຽບາລອງໃໝ່ອົກຄົ້ນໃນກາຍເລັ່ນ

ຍອດຂາຍຂອງ MacBook Pro M3 ມີປະວັດກາຍທີ່ໜັດ 2 ຮາຍຄາ ດັ່ງນີ້:

ວັນທີຂາຍ	ຈຳນວນທີ່ຂາຍ (ຊັ້ນ)	ຍອດຂາຍ (ບາກ)
20/8/2568	1	79,900
25/8/2568	1	79,900

ສຽບ:
ຂາຍຮວບ: 2 ຊົ່ວໂມງ
ຍອດຂາຍຮວບ: 159,800 ບາກ ຄົນ

ເພີ້ມແປ່ວ່າ (Deployment) ບໍລິຫານ Vercel Cloud ພົມ

Continue the conversation...

+ Search ...

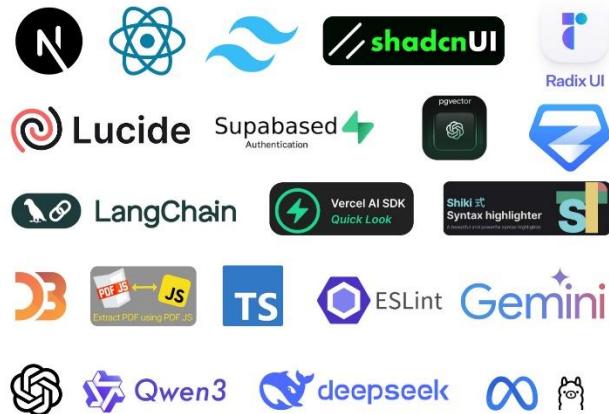


ສຕາບັນໄວ້ກົງເນື້ຍສ

www.itgenius.co.th



Tech Stack



- **Frontend:** Next.js 15.5.0 (App Router)
- **UI Framework:** React 19.1.0
- **Styling:** Tailwind CSS 4
- **UI Components:**
 - shadcn/ui (New York style)
 - Prompt-kit UI Components
 - Radix UI (@radix-ui/react-*)
 - Lucide React Icons
- **Authentication:**
 - Supabase Auth (@supabase/supabase-js, @supabase/ssr)
 - Password-based Authentication
- **Database:**
 - Supabase PostgreSQL with pgvector extension
 - Chat History Storage
 - User Session Management
 - Product & Sales Data (Tool Calling)
 - Vector Embeddings Storage (Document Search)
- **AI Integration:**
 - AI SDK v5 (@ai-sdk/react, @ai-sdk/langchain)
 - LangChain (@langchain/core, @langchain/openai, @langchain/community)
- **Tool Calling:** LangChain Agents with Supabase Tools
- **Document Processing:** Text & PDF loaders, Vector embeddings
- **Vector Search:** pgvector with cosine similarity
- **Language Model:** OpenAI GPT-4o-mini & text-embedding-3-small
- **Utilities:**
 - class-variance-authority (Component variants)
 - clsx & tailwind-merge (Conditional styling)
 - marked & react-markdown (Markdown rendering)
 - shiki (Syntax highlighting)
 - zod (Schema validation สำหรับ Tools)
 - remark-gfm & remark-breaks (Markdown extensions)
 - use-stick-to-bottom (Auto-scroll behavior)
 - d3-dsv (CSV parsing สำหรับ Document Loader)
- **TypeScript:** v5
- **Development:** ESLint 9



แก้ branch ให้เรียบตามได้อย่างง่าย

my-langchain-chatbot

EXPLORER: MY-LA... PROBLEMS PORTS OUTPUT DEBUG CONSOLE GITLENS TERMINAL

COMMIT GRAPH: MY-LANGCHAIN-CHATBOT

All Branches Search commits using natural language (11 for history), e.g. Show my commits from last month No results

BRANCH / TAG	GRAPH	COMMIT MESSAGE	AUTHOR	CHANGES	COMMIT DATE / TIME
08-document-loader-embedding-pgve...		update readme document loader	You	1	2 hours ago
09-rag		update welcome page	You	1	6 hours ago
10-deployment	✓	update front page	You	1	yesterday
08-document-loader-embedding-pgvector-text-csv-pc		10-deployment-finished	You	16	2 days ago
09-rag-finished		09-rag-finished	You	8	2 days ago
08-document-loader-embedding-pgvector-text-csv-pc		08-document-loader-embedding-pgvector-text-csv-pc	You	1	2 days ago
08-document-loader-embedding-pgvector-text-csv		08-document-loader-embedding-pgvector-text-csv	You	4	2 days ago
07-tool-calling-finihed		07-tool-calling-finihed	You	9	2 days ago
07-tool-calling-start		07-tool-calling-start	You	1	3 days ago
06-chat-history-optimize-update-readme		06-chat-history-optimize-update-readme	You	1	3 days ago
06-chat-history-optimize-add-comment		06-chat-history-optimize-add-comment	You	6	3 days ago
05-dark-theme-ui-complete update comment		05-dark-theme-ui-complete update comment	You	3	3 days ago
06-chat-history-update-comment		06-chat-history-update-comment	You	11	3 days ago
06-chat-history-optimize-complete		06-chat-history-optimize-complete	You	11	3 days ago
06-chat-history-complete-update-linter		06-chat-history-complete-update-linter	You	12	3 days ago
06-chat-history-complete-update-readme		06-chat-history-complete-update-readme	You	1	4 days ago
06-chat-history-complete		06-chat-history-complete	You	20	4 days ago
06-integrate-api-to-ui-new-chat-compleate		06-integrate-api-to-ui-new-chat-compleate	You	3	4 days ago
05-dark-theme-ui-complete #1		05-dark-theme-ui-complete #1	You	2	4 days ago
05-dark-theme-ui-complete		05-dark-theme-ui-complete	You	7	4 days ago
04-prompt-kit-ui-complte		04-prompt-kit-ui-complte	You	38	4 days ago



Home Workspaces API Network

Search Postman Ctrl K

Invite Upgrade No environment

Collections: samit

Environments

Flows

History

APIs:

- AIChatbotLangchain
 - 01_Next_API
 - 02_Langchain_Basic
 - POST 01_/api/chat_01_start
 - POST 02_/api/chat_02_request
 - POST 03_/api/chat_03_template
 - POST 04_/api/chat_04_stream
 - 05_Chat_History
 - POST 01_/api/chat_05_history?sessionId=4ebe30a5-0c01-4912-927d-78227102dee9
 - GET 02_/api/chat_05_history?sessionId=4ebe30a5-0c01-4912-927d-78227102dee9 Copy
 - 06_Chat_History_Optimize
 - POST 01_/api/chat_06_history_optimize?sessionId=4ebe30a5-0c01-4912-927d-78227102dee9
 - GET 02_/api/chat_06_history_optimize?sessionId=4ebe30a5-0c01-4912-927d-78227102dee9 Copy
 - 07_Tool_Calling
 - POST 01_/api/chat_07_tool_calling?sessionId=4ebe30a5-0c01-4912-927d-78227102dee9
 - GET 02_/api/chat_07_tool_calling?sessionId=4ebe30a5-0c01-4912-927d-78227102dee9 Copy
 - Document_Loader_EMBEDDING_pgVector
 - GET 01_/api/document_loader_embedding_pgvector/text_csv
 - POST 02_/api/document_loader_embedding_pgvector/text_csv
 - DEL 03_/api/document_loader_embedding_pgvector/text_csv
 - PUT 04_/api/document_loader_embedding_pgvector/text_csv
 - GET 05_/api/document_loader_embedding_pgvector/text_csv_pdf
 - 08_RAG
 - POST 01_/api/chat_08_rag?sessionId=4ebe30a5-0c01-4912-927d-78227102dee9
 - GET 02_/api/chat_08_rag?sessionId=4ebe30a5-0c01-4912-927d-78227102dee9 Copy
- DjangoWebSocket
- gofiber
- QR Menu App API

POST 04_/api/chat_04_stream

Save Share

Send

Params Authorization Headers (9) Body Scripts Settings Cookies Beautify

Body Type: raw

```
1 {  
2   "messages": [  
3     {  
4       "id": "chat-id-001",  
5       "role": "user",  
6       "parts": [  
7         {  
8           "type": "text",  
9           "text": "สวัสดีครับ บริษัทของเรารักษาความปลอดภัยให้แผนกใหญ่มากที่สุดครับ"  
10        }  
11      ]  
12    }  
13  ]  
14 }
```

Response Hist Click Send to get a response

แมก API end point ໄວ້ໃຫ້ກົດສອບງ່າຍ

Postbot Runner Start Proxy Cookies Vault Trash



1. ภาพรวม AI Chatbot กับ Langchain.js



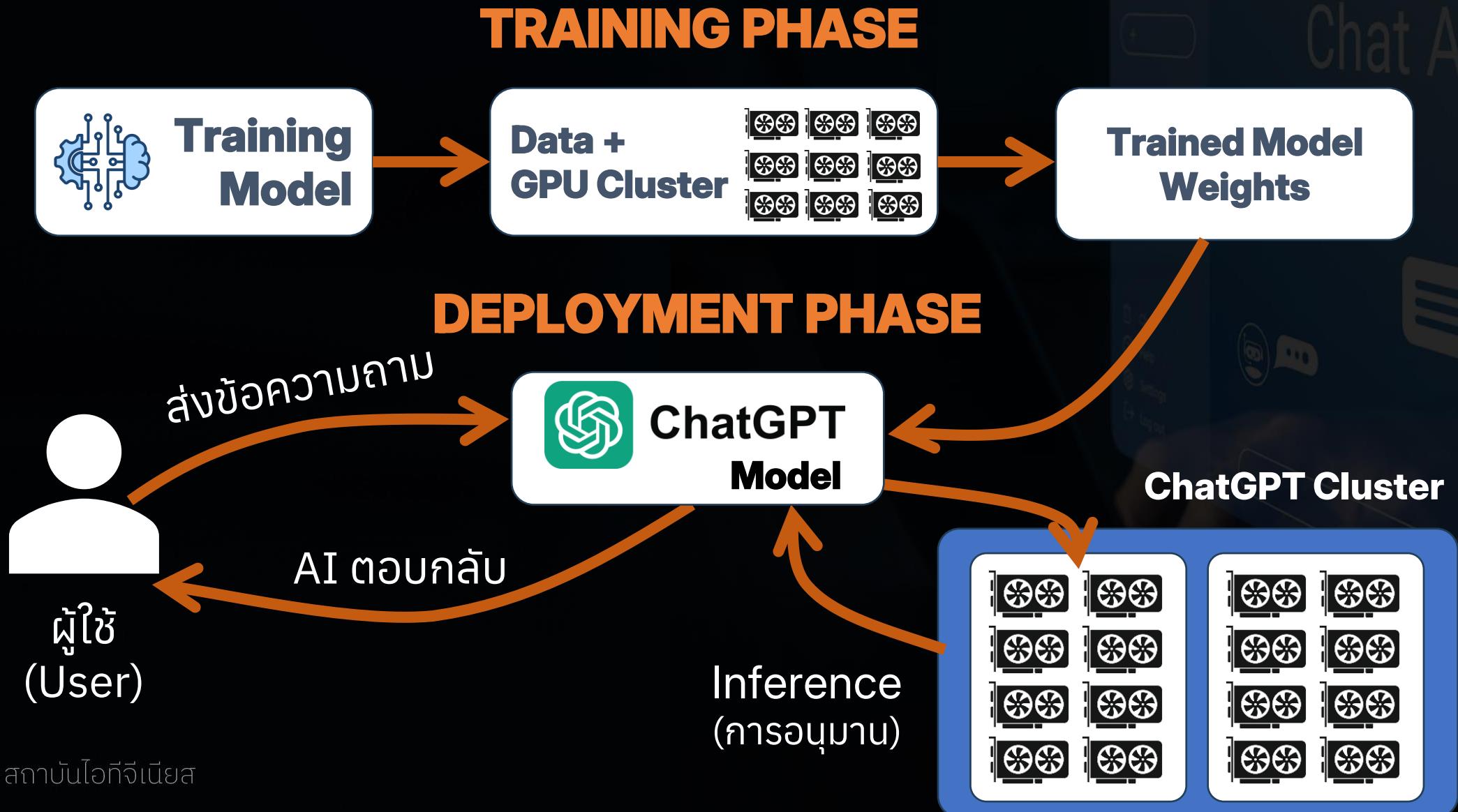


1. การรวม AI Chatbot กับ Langchain.js
2. การพัฒนา Rest API ใน Next.js เพื่อใช้งานกับ Langchain.js
3. พื้นฐาน Langchain.js เชื่อมต่อกับ Gen AI

การสร้างและการทำงานของ Gen AI Model



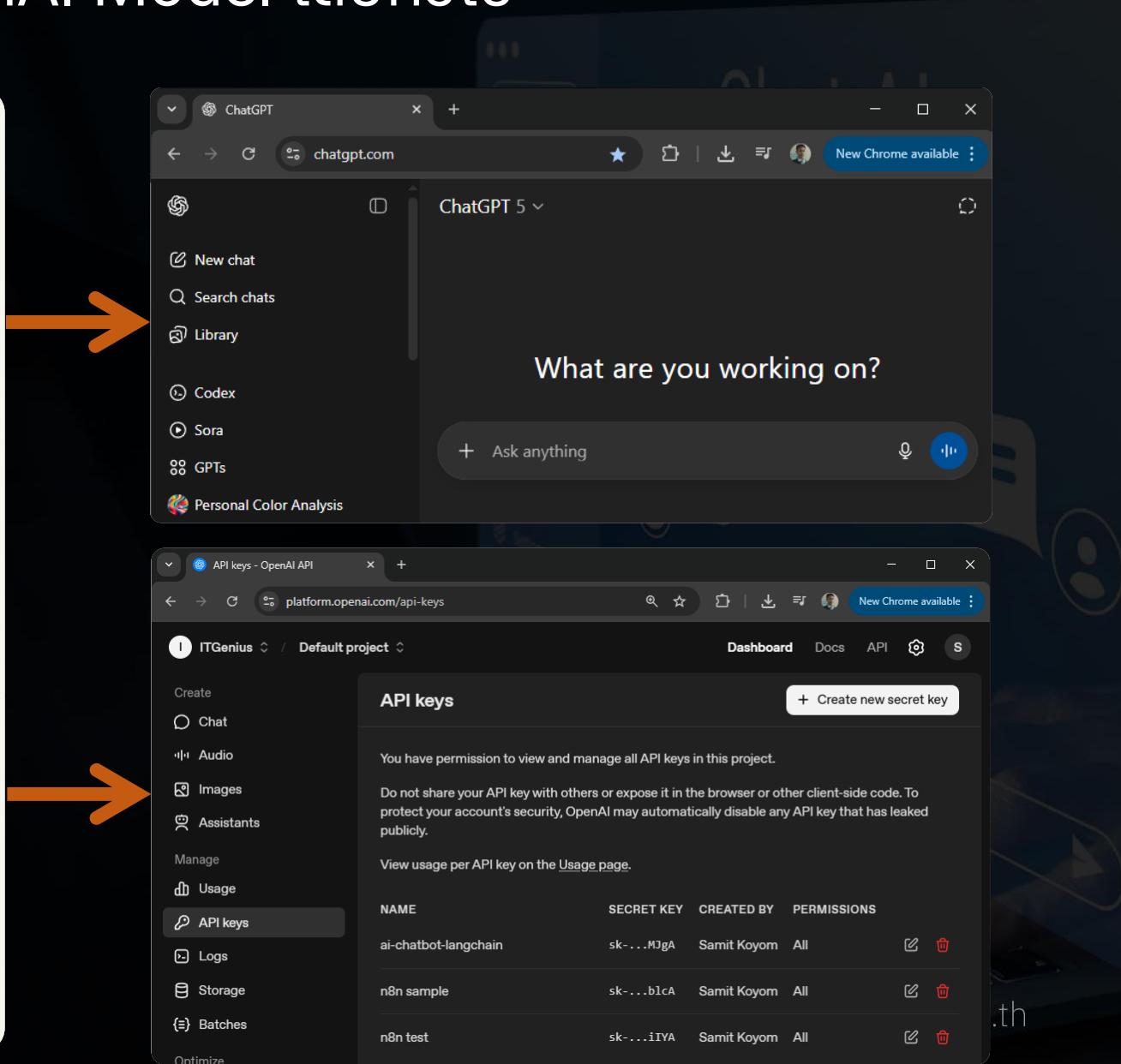
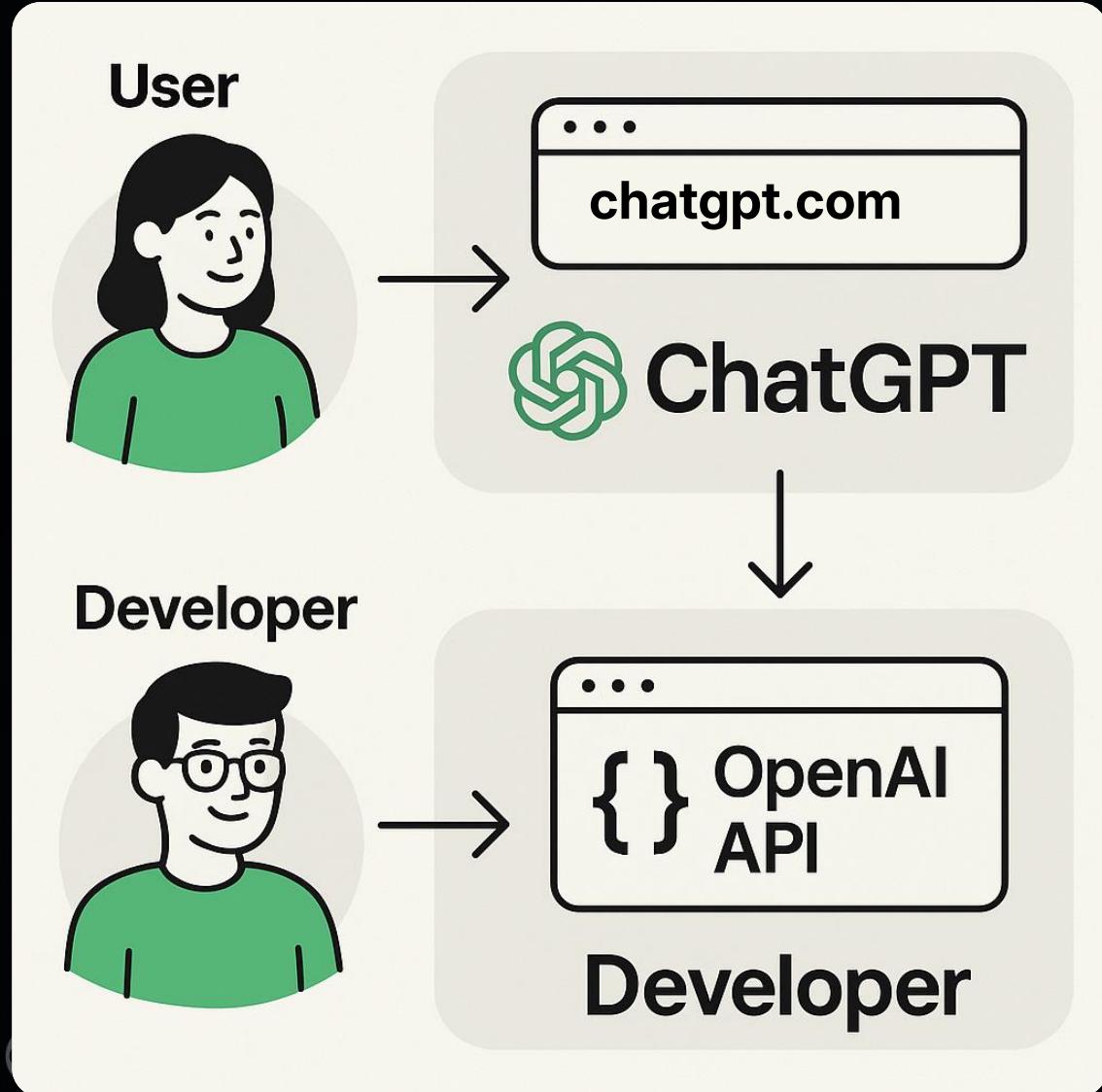
การสร้างและการทำงานของ AI Model



รูปแบบการเรียนรู้ใช้งาน Gen AI Model



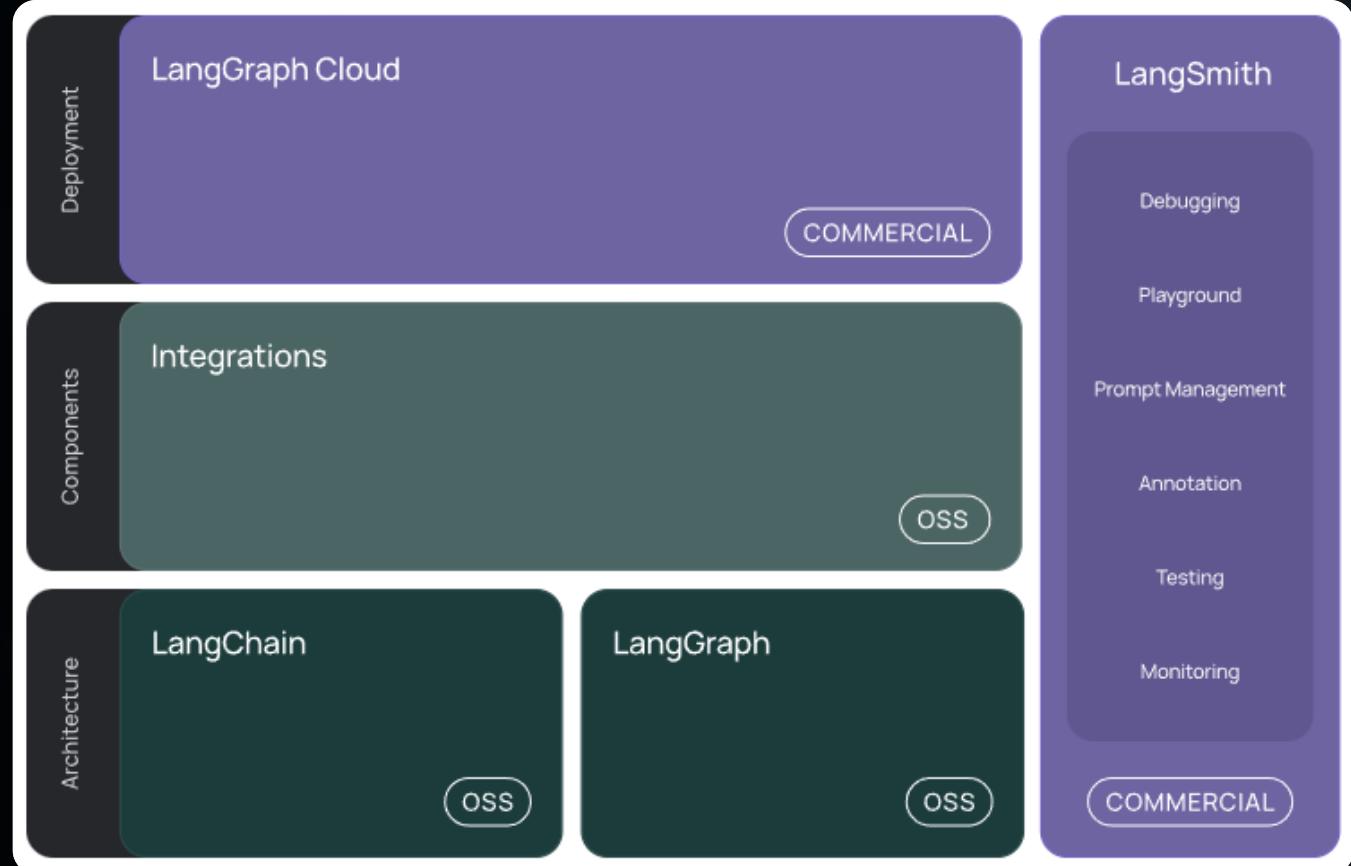
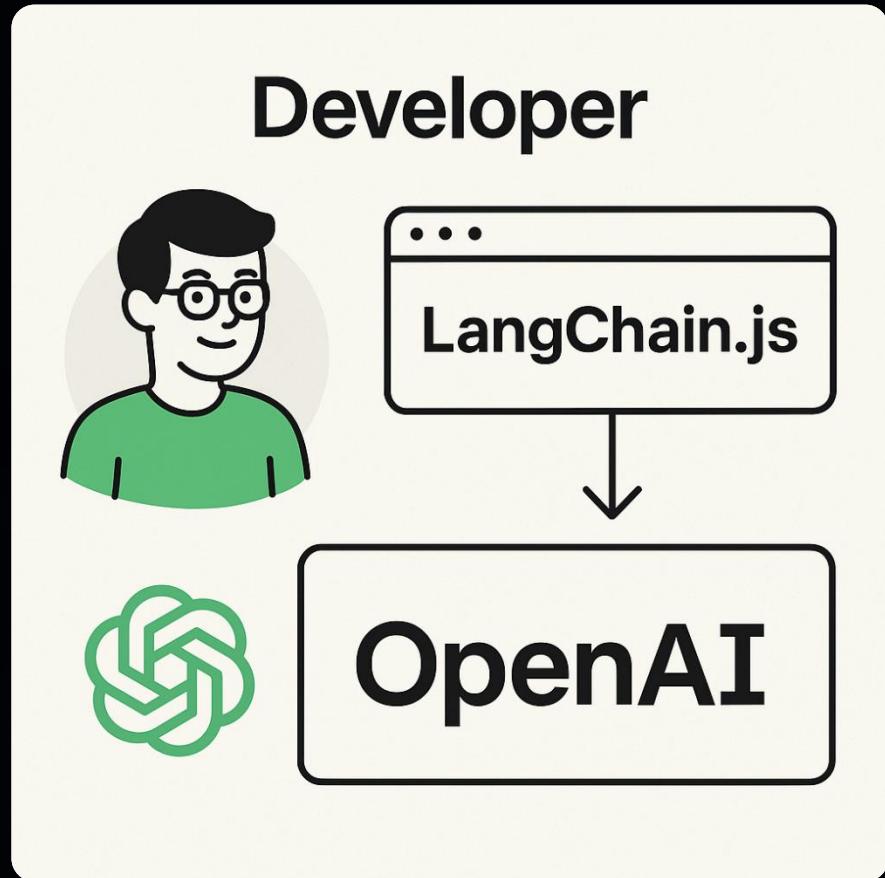
การเข้าใช้งาน GenAI Model โดยก้าวไป

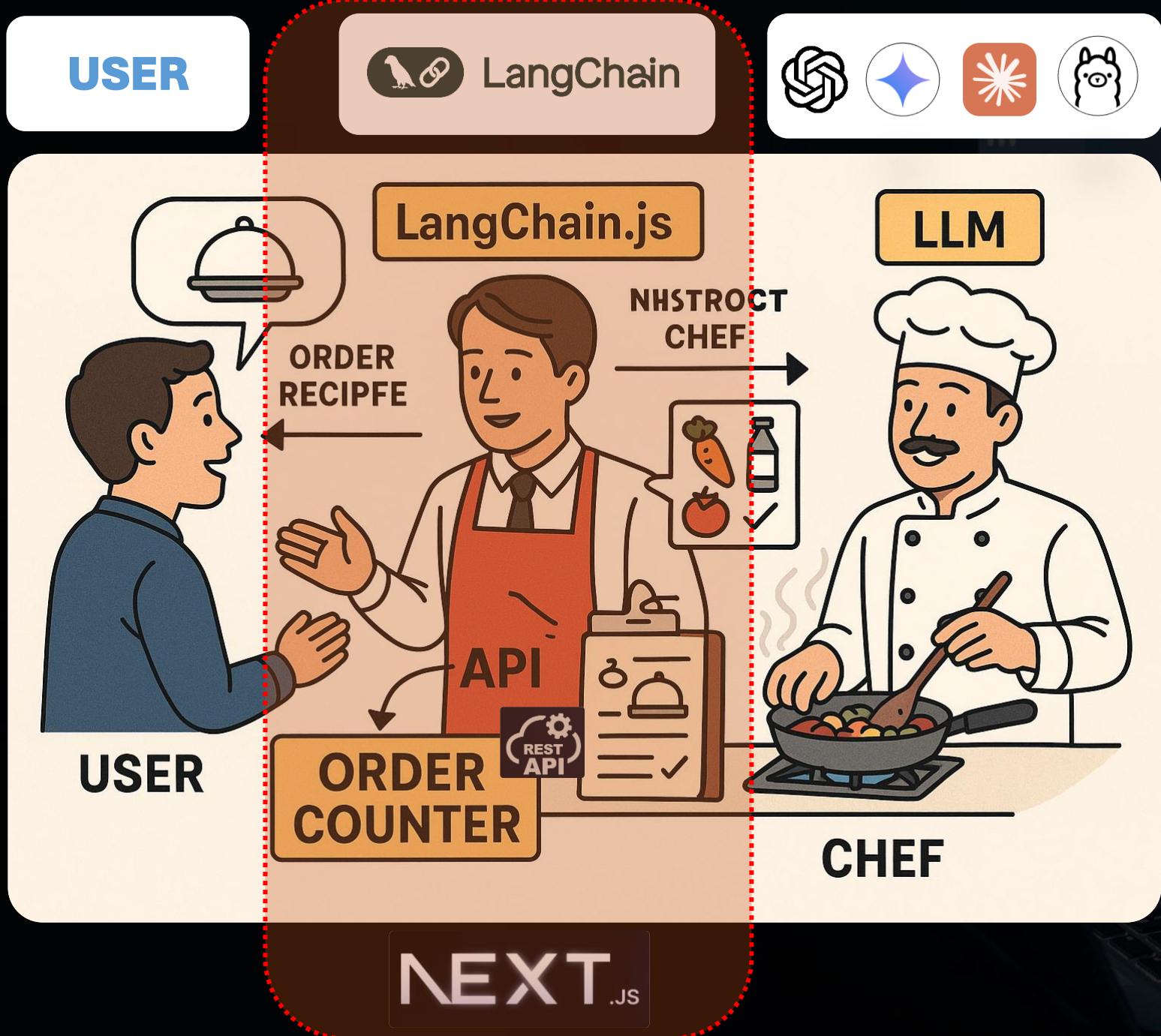


การทำงานของ Langchain กับ AI Model



นักพัฒนาเรียกทำงานกับ GenAI Model ผ่าน Langchain





AI SDK UI



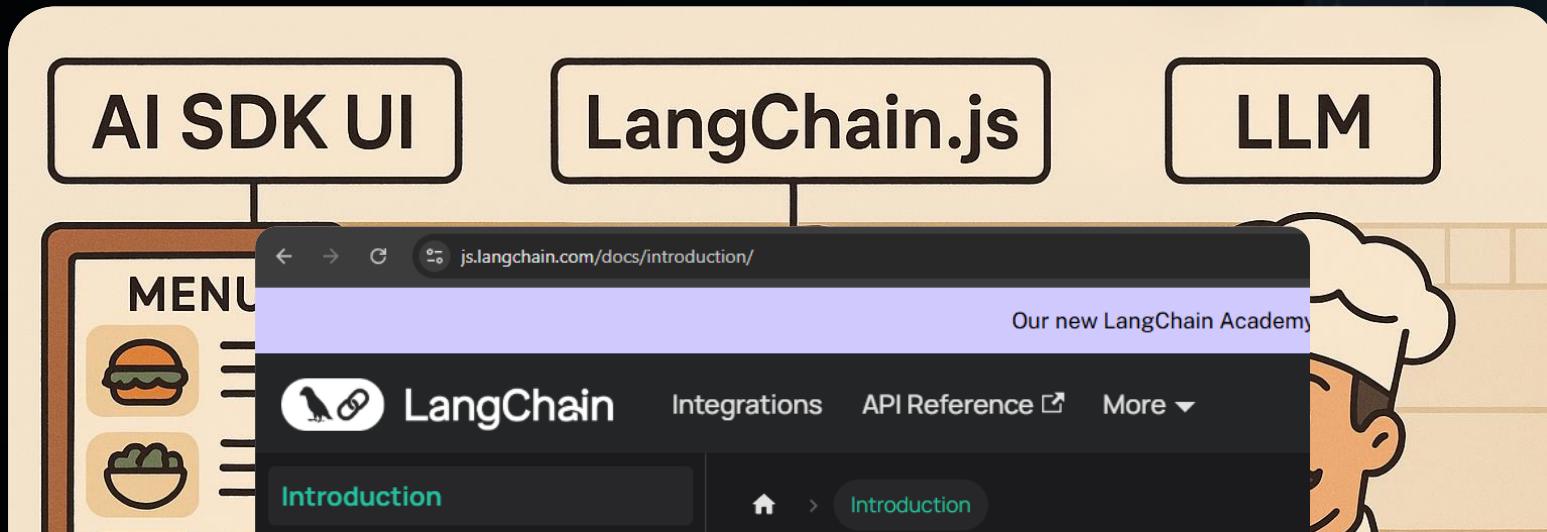
LangChain



AI SDK UI

LangChain.js

LLM



The AI Toolkit for TypeScript

From the creators of Next.js, the AI SDK is a free open-source library that gives you the tools you need to build AI-powered products.

Get Started \$ npm i ai Visit Playground

Trusted by builders at

This screenshot shows the AI SDK UI documentation page. It features a dark header with the title "The AI Toolkit for TypeScript". Below the header, there is a brief description of the AI SDK as a free open-source library. At the bottom, there are three buttons: "Get Started", "\$ npm i ai", and "Visit Playground". The footer contains the text "Trusted by builders at" followed by a list of logos for various companies.

Introduction

LangChain is a framework for building AI applications.

LangChain simplifies every stage of AI development:

- **Development:** Build your applications with simple, composable integrations. Use LangChain's powerful API to connect to any AI service.
- **Productionization:** Use LangChain's built-in infrastructure to host your AI applications at scale.

Our new LangChain Academy is now available!

API keys - OpenAI API

ITGenius / Default project

Create

- Chat
- Audio
- Images
- Assistants

Manage

- Usage
- API keys
- Logs
- Storage
- Batches

+ Create new secret key

API keys

You have permission to view and manage all API keys in this project.

Do not share your API key with others or expose it in the browser or other client-side code. To protect your account's security, OpenAI may automatically disable any API key that has leaked publicly.

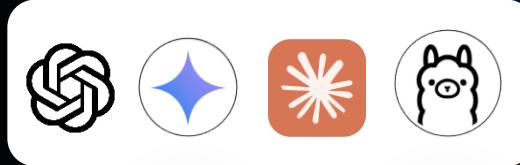
View usage per API key on the Usage page.

NAME	SECRET KEY	CREATED BY	PERMISSIONS
ai-chatbot-langchain	sk-...MjgA	Samit Koyom	All
n8n sample	sk-...blcA	Samit Koyom	All
n8n test	sk-...iiYA	Samit Koyom	All

Optimize

This screenshot shows the LangChain.js Introduction page and the OpenAI API keys dashboard. The introduction page features a large "Introduction" heading and a section on simplifying AI development. The OpenAI API keys dashboard shows a list of API keys with columns for NAME, SECRET KEY, CREATED BY, and PERMISSIONS. The "API keys" section is currently selected. The dashboard also includes sections for Usage, Logs, Storage, and Batches.

Local LLM vs LLM providers



สถาบันไอทีจีเนียส

www.itgenius.co.th

Local LLM



สถาบันไอทีจีเนียส

www.itgenius.co.th



Ollama

Thought for 1.8 seconds

New Chat

Settings

This week

ประเทศไทยมีกี่จังหวัด

สวัสดีครับ

จังหวัดของภาคกลาง (ประเทศไทย) – ตาราง 21 จังหวัด (รวมกรุงเทพมหานครเป็นเขตบริหารพิเศษ)

ลำดับ	ชื่อจังหวัด (อังกฤษ)	ชื่อจังหวัด (ไทย)
1	Nakhon Pathom	จังหวัดนนทบุรี
2	Nonthaburi	จังหวัดนนทบุรี
3	Pathum Thani	จังหวัดปทุมธานี
4	Samut Prakan	จังหวัดสมุทรปราการ
5	Samut Sakhon	จังหวัดสมุทรสาคร
6	Samut Songkhram	จังหวัดสมุทรสงคราม
7	Kamphaeng Phet	จังหวัดกำแพงเพชร
8	Nakhon Sawan	จังหวัดนนทบุรี
9	Phetchabun	จังหวัดเพชรบูรณ์
10	Phichit	จังหวัดพิจิตร
11	Phitsanulok	จังหวัดพิษณุโลก
12	Sukhothai	จังหวัดสุโขทัย

Send a message

⊕ Turbo gpt-oss:20b ↑



Qwen: Qwen3 8B

Local Host

Home Workspaces API Network

Ctrl K

Invite Settings Refresh Upgrade No environment

POST 01/_api/chat_01_start

HTTP AlChatbotLangchain / 02_Langchain_Basic / 01/_api/chat_01_start

POST {{baseURL}} /api/chat_01_start

Params Authorization Headers (16) Body Scripts Settings

Body: none form-data x-www-form-urlencoded raw binary GraphQL

This request does not have a body.

Cookies

Send

History

Collections Environments Flows History

Samit

Search collections

AIChatbotLangchain

- 01_Next_API
- 02_Langchain_Basic
 - POST 01/_api/chat_01_start
 - POST 02/_api/chat_02_request
 - POST 03/_api/chat_03_template
 - POST 04/_api/chat_04_stream
- 05_Chat_History
- 06_Chat_History_Optimize
- 07_Tool_Calling
- Document_Loader_EMBEDDING_pgVector
- 08_RAG

DjangoWebSocket

gofiber

QR Menu App API

Body Cookies (1) Headers (6) Test Results

200 OK 2 m 2.52 s 3.38 KB Save Response

JSON Preview Visualize

```

1 "content": "<think>\nOkay, the user is asking about the budget for this year. I need to provide a clear and concise answer. Let me start by acknowledging their question. Then, I should outline the main aspects of the budget, like revenue targets, cost management, and key areas of investment. It's important to mention the overall direction, maybe something about growth and efficiency. Also, I should invite them to ask more questions if they need specific details. Keep the tone friendly and professional. Let me structure it step by step to make sure all key points are covered without being too technical.\n</think>\n\nสวัสดีครับ 😊\n\nปัจจุบันเราได้รับงบประมาณ 10-15% เมื่อเทียบกับปีที่แล้ว โดยแบ่งการรายจ่ายเป็นส่วนแบ่งตลาดในธุรกิจหลัก\n\n**การจัดการงบประมาณ**\n\n- คาดการณ์ว่าจะมีการเพิ่มรายรับ 10-15% เมื่อเทียบกับปีที่แล้ว โดยแบ่งการรายจ่ายเป็นส่วนแบ่งตลาดในธุรกิจหลัก\n\n**การจัดการต้นทุน**\n\n- ควบคุมต้นทุนอย่างเข้มงวด ลดต้นทุนโดยการซื้ออุปกรณ์และซอฟต์แวร์ที่มีประสิทธิภาพ\n\n**การลงทุน**\n\n- ลงทุนในเทคโนโลยีใหม่ๆ เช่น AI และ Big Data เพื่อเพิ่มประสิทธิภาพการดำเนินงาน\n\n**ความยั่งยืนทางการเงิน**\n\n- รักษาสภาพคล่องทางการเงินที่ดี พร้อมสำรองเงินสดสำหรับสถานการณ์突發\n\n\n"
2
3
4

```

Content-Type: application/json



GPT-OSS

Local Host

The screenshot shows the Postman interface with the following details:

- Header Bar:** Home, Workspaces, API Network, Search Postman, Ctrl + K, Invite, Settings, Notifications (1), Help, Upgrade.
- Left Sidebar:** Collections (Search collections), AIChatbotLangchain, 01_Next_API, 02_Langchain_Basic (selected), POST 01/_api/chat_01_start, POST 02/_api/chat_02_request, POST 03/_api/chat_03_template, POST 04/_api/chat_04_stream, 05_Chat_History, 06_Chat_History_Optimize, 07_Tool_Calling, Document_Loader_EMBEDDING_pgVector, 08_RAG, DjangoWebSocket, gofiber, QR Menu App API.
- Request Details:** Method: POST, URL: {{baseURL}}/_api/chat_01_start, Headers: 16, Body: none (selected), Params, Authorization, Scripts, Settings.
- Response Preview:** Status: 200 OK, Time: 3 m 35.82 s, Size: 3.42 KB, Save Response, Body (JSON):

```
1 {  
2   "content": "สวัสดีครับ\ngานบริษัทเรา มีภาระรวมดังนี้ครับ\ng|\t รายการ | จำนวน (บาท) | เปอร์เซ็นต์ของบรวม |\n|\t-----|  
|\t**งบประมาณรวม** | 1 200 000 000 | 100 % |\n|\t**รายได้** | 1 350 000 000 | 112 % |\n|\t**ค่าใช้จ่าย** |  
1 200 000 000 | 100 % |\n|\t**กำไรสุทธิ** | 150 000 000 | 12 % |\n|\tการจัดสรรงบประมาณ\ng|\tแผนก | งาน (บาท) | เปอร์เซ็นต์ |\n|\t-----|  
|\t การตลาด | 250 000 000 | 20 % |\n|\t การผลิต | 300 000 000 | 25 % |\n|\t การวิจัย & พัฒนา | 150 000 000 | 12.5 % |\n|\t ฝ่ายบริหาร &  
กการเงิน | 100 000 000 | 8.3 % |\n|\t ฝ่ายหัวหน้าบุคคล | 80 000 000 | 6.7 % |\n|\t ค่าธรรมเนียม & ค่าบริการ | 120 000 000 | 10 % |\n|\t ค่าตอบแทนพัสดุฯ  
(ใบน้ำส าล่า) | 100 000 000 | 8.3 % |\n|\t **อื่น ๆ** | 100 000 000 | 8.3 % |\n|\t-----|  
|\t**จดเด่นและความเปลี่ยนแปลง\ng1. **เพิ่มงบการตลาด 15 %**  
เพิ่อสนับสนุนความยั่งยืนด้วยก้าวแรกของการขยายตลาดใหม่ ๆ \n2. **ลดค่าใช้จ่ายด้านการผลิต 5 %** ผ่านการปรับปรุงกระบวนการผลิตให้เกิดประโยชน์\n3.  
**เพิ่มงบ R&D 10 %** เพื่อพัฒนาผลิตภัณฑ์ใหม่และขยายตัวผลิตภัณฑ์ที่มีอยู่现有的\n4. **กำไรสุทธิเพิ่มขึ้น 12 %** จากการควบคุมต้นทุนและเพิ่มยอดขาย\ng|\t-----|  
ค่าแนะนำสำหรับพนักงาน\ng|\t  หากต้องการรายละเอียดเพิ่มเติมเกี่ยวกับงบประมาณของแผนกของคุณ หรือมีข้อเสนอแนะในการปรับใช้หัวหน้าการ  
กรุณาติดต่อฝ่ายการเงินโดยตรง \n|\t- โปรดตรวจสอบว่าการใช้จ่ายของคุณสอดคล้องกับงบประมาณที่กำหนดไว้  
และส่งรายงานการใช้จ่ายตามกำหนดเวลา\ng|\tหากมีคำขอเพิ่มเติมหรือยกเว้นลักษณะใดในส่วนใดเป็นพิเศษ ยินดีช่วยเสมอครับ!",  
3 "usedModel": "gpt-oss:latest"  
4 }
```

Local LLM



สถาบันไอทีจีเนียส

www.itgenius.co.th



Create GPU Droplet

Choose a datacenter region

 New York • Datacenter 2 • NYC2

Choose an image

OS

1-click Models

Custom Images

AI/ML Ready

Recommended: Linux bundled with [required GPU Drivers](#)

Inference Optimized

Deploy any model faster with production-grade-performance

Ubuntu

24.04 (LTS) x64

Fedora

42 x64

Debian

12 x64

CentOS

9 Stream x64

AlmaLinux

AlmaLinux 9

Rocky Linux

9 x64

Summary

GPU

Type: NVIDIA RTX4000 ADA

GPU: 1

VRAM: 20 GB

vCPU: 8

RAM: 32 GB

Boot Disk: 500 GB NVMe SSD

\$0.76/hr

Total cost

\$0.76/hour

Create GPU Droplet

Want to maximize efficiency and cost?

Programmatically manage GPUs in a repeatable and re-usable way with our API. [Create via API](#)



< > 1 of 2 open incidents Droplet Connectivity To learn more, [check our status page](#).

Search by resource name or public IP (Ctrl+B)

Create ? My Team Estimated costs: \$11.11

PROJECTS

MANAGE

App Platform

Agent Platform New

Droplets

GPU Droplets New

Functions

Kubernetes

Volumes Block Storage

Databases

Spaces Object Storage

Container Registry

Backups & Snapshots

Networking

Monitoring

SaaS Add-Ons

By DigitalOcean

Billing

Support

Settings

ubuntu-gpu-4000adax

← Back to GPU Droplets

Web Console Actions

Active • Playground • TOR1 • default-tor1 • Ubuntu 24.04 (LTS) x64

Getting to know your GPU Droplet

Even if your GPU Droplet is powered off, you will still be billed as GPU, CPU, and associated resources are still reserved. To avoid charges, be sure to destroy the instance if it's no longer in use. Before destruction, you can take a snapshot to return back to.

Overview Insights Networking Volumes Backups & Snapshots Activity Settings

TOTAL GPU DROPLET COST

\$0.76 / hour

This is an estimation based on current configuration, you can see the [cost to date in billing](#).

Configuration Details

AUTOMATED BACKUPS

Reduce the stress of failure with automated droplet backups. They can be used for restoring lost or corrupt data and creating new Droplets.

Setup Automated Backups

Connection Details

Public IPv4 138.197.136.10

Private IP 10.118.0.2

Having trouble connecting? [Check out our troubleshooting guide](#).

Enhance Your Droplet's Network:

Improve security and ensure IP Address portability



pwsh in Samit x root@ubuntu-gpu-4000adax: +

(env) root@ubuntu-gpu-4000adax:/home/vllm_project# nvidia-smi

Mon Sep 8 11:31:30 2025

NVIDIA-SMI 535.247.01			Driver Version: 535.247.01		CUDA Version: 12.2		
GPU	Name	Persistence-M	Bus-Id	Disp.A	Volatile	Uncorr.	ECC
Fan	Temp	Pwr:Usage/Cap		Memory-Usage	GPU-Util	Compute M.	MIG M.
0	NVIDIA RTX 4000 Ada Gene...	Off	00000000:01:00.0	0MiB / 20475MiB	0%	Default	Off
30%	42C	P0	31W / 130W				N/A

Processes:

GPU	GI	CI	PID	Type	Process name	GPU Memory Usage
ID	ID					
No running processes found						
(env) root@ubuntu-gpu-4000adax:/home/vllm_project#						



```
(env) root@ubuntu-gpu-4000adax:/home/vllm_project# python -m vllm.entrypoints.openai.api_server --model "Qwen/Qwen2.5-7B-Instruct"
INFO 09-08 16:20:51 [__init__.py:241] Automatically detected platform cuda.
(APIServer pid=9806) INFO 09-08 16:20:54 [api_server.py:1805] vLLM API server version 0.10.1.1
(APIServer pid=9806) INFO 09-08 16:20:54 [utils.py:326] non-default args: {'model': 'Qwen/Qwen2.5-7B-Instruct'}
(APIServer pid=9806) INFO 09-08 16:21:06 [__init__.py:711] Resolved architecture: Qwen2ForCausalLM
(APIServer pid=9806) `torch_dtype` is deprecated! Use `dtype` instead!
(APIServer pid=9806) INFO 09-08 16:21:07 [__init__.py:1750] Using max model len 32768
(APIServer pid=9806) INFO 09-08 16:21:12 [scheduler.py:222] Chunked prefill is enabled with max_num_batched_tokens=2048.
INFO 09-08 16:21:17 [__init__.py:241] Automatically detected platform cuda.
(EngineCore_0 pid=9886) INFO 09-08 16:21:21 [core.py:636] Waiting for init message from front-end.
(EngineCore_0 pid=9886) INFO 09-08 16:21:21 [core.py:74] Initializing a V1 LLM engine (v0.10.1.1) with config: model='Qwen/Qwen2.5-7B-Instruct', speculative_config=None, tokenizer='Qwen/Qwen2.5-7B-Instruct', skip_tokenizer_init=False, tokenizer_mode=auto, revision=None, override_neuron_config={}, tokenizer_revision=None, trust_remote_code=False, dtype=torch.bfloat16, max_seq_len=32768, download_dir=None, load_format=auto, tensor_parallel_size=1, pipeline_parallel_size=1, disable_custom_all_reduce=False, quantization=None, enforce_eager=False, kv_cache_dtype=auto, device_config=cuda, decoding_config=DecodingConfig(backend='auto', disable_fallback=False, disable_any whitespace=False, disable_additional_properties=False, reasoning_backend=''), observability_config=ObservabilityConfig(show_hidden_metrics_for_version=None, otlp_traces_endpoint=None, collect_detailed_traces=None), seed=0, served_model_name=Qwen/Qwen2.5-7B-Instruct, enable_prefix_caching=True, chunked_prefill_enabled=True, use_async_output_proc=True, pooler_config=None, compilation_config={"level":3,"debug_dump_path":","cache_dir":","backend":","custom_ops":[],"splitting_ops":["vllm.unified_attention","vllm.unified_attention_with_output","vllm.mamba_mixer2"],"use_inductor":true,"compile_sizes":[],"inductor_compile_config":{"enable_auto_functionalized_v2":false}, "inductor_passes":{},"cudagraph_mode":1,"use_cudagraph":true,"cudagraph_num_of_warmups":1,"cudagraph_capture_sizes":[512,504,496,488,480,472,464,456,448,440,432,424,416,408,400,392,384,376,368,360,352,344,336,328,320,312,304,296,288,280,272,264,256,248,240,232,224,216,208,200,192,184,176,168,160,152,144,136,128,120,112,104,96,88,80,72,64,56,48,40,32,24,16,8,4,2,1],"cudagraph_copy_inputs":false,"full_cuda_graph":false,"pass_config":{},"max_capture_size":512,"local_cache_dir":null}}
(EngineCore_0 pid=9886) INFO 09-08 16:21:23 [parallel_state.py:1134] rank 0 in world size 1 is assigned as DP rank 0, PP rank 0, TP rank 0, EP rank 0
(EngineCore_0 pid=9886) WARNING 09-08 16:21:23 [topk_topp_sampler.py:61] FlashInfer is not available. Falling back to the PyTorch-native implementation of top-p & top-k sampling. For the best performance, please install FlashInfer.
(EngineCore_0 pid=9886) INFO 09-08 16:21:23 [gpu_model_runner.py:1953] Starting to load model Qwen/Qwen2.5-7B-Instruct...
(EngineCore_0 pid=9886) INFO 09-08 16:21:23 [gpu_model_runner.py:1985] Loading model from scratch...
(EngineCore_0 pid=9886) INFO 09-08 16:21:23 [cuda.py:328] Using Flash Attention backend on V1 engine.
(EngineCore_0 pid=9886) INFO 09-08 16:21:24 [weight_utils.py:296] Using model weights format ['*.safetensors']
Loading safetensors checkpoint shards: 0% Completed | 0/4 [00:00<?, ?it/s]
Loading safetensors checkpoint shards: 25% Completed | 1/4 [00:00<00:01, 2.10it/s]
Loading safetensors checkpoint shards: 50% Completed | 2/4 [00:01<00:01, 1.98it/s]
Loading safetensors checkpoint shards: 75% Completed | 3/4 [00:01<00:00, 2.02it/s]
Loading safetensors checkpoint shards: 100% Completed | 4/4 [00:02<00:00, 1.98it/s]
Loading safetensors checkpoint shards: 100% Completed | 4/4 [00:02<00:00, 1.99it/s]
```



```
pwsh in Samit      x  root@ubuntu-gpu-4000adax: ~ + ^
```

(APIServer pid=9806) INFO 09-08 16:21:41 [launcher.py:44] Route: /v1/rerank, Methods: POST
(APIServer pid=9806) INFO 09-08 16:21:41 [launcher.py:44] Route: /v2/rerank, Methods: POST
(APIServer pid=9806) INFO 09-08 16:21:41 [launcher.py:44] Route: /scale_elastic_ep, Methods: POST
(APIServer pid=9806) INFO 09-08 16:21:41 [launcher.py:44] Route: /is_scaling_elastic_ep, Methods: POST
(APIServer pid=9806) INFO 09-08 16:21:41 [launcher.py:44] Route: /invocations, Methods: POST
(APIServer pid=9806) INFO 09-08 16:21:41 [launcher.py:44] Route: /metrics, Methods: GET
(APIServer pid=9806) INFO: Started server process [9806]
(APIServer pid=9806) INFO: Waiting for application startup.
(APIServer pid=9806) INFO: Application startup complete.
(APIServer pid=9806) WARNING: Invalid HTTP request received.
(APIServer pid=9806) INFO: 27.145.115.105:2124 - "GET / HTTP/1.1" 404 Not Found
(APIServer pid=9806) INFO: 27.145.115.105:2124 - "GET /favicon.ico HTTP/1.1" 404 Not Found
(APIServer pid=9806) INFO: 27.145.115.105:2103 - "GET / HTTP/1.1" 404 Not Found
(APIServer pid=9806) INFO: 27.145.115.105:2103 - "GET / HTTP/1.1" 404 Not Found
(APIServer pid=9806) INFO: 27.145.115.105:2123 - "GET /v1/chat/completions HTTP/1.1" 405 Method Not Allowed
(APIServer pid=9806) INFO: 27.145.115.105:2463 - "GET / HTTP/1.1" 404 Not Found
(APIServer pid=9806) INFO 09-08 16:26:25 [chat_utils.py:470] Detected the chat template content format to be 'string'. You can set `--chat-template-content-format` to override this.
(APIServer pid=9806) INFO: 27.145.115.105:2464 - "POST /v1/chat/completions HTTP/1.1" 200 OK
(APIServer pid=9806) INFO 09-08 16:26:31 [loggers.py:123] Engine 000: Avg prompt throughput: 4.2 tokens/s, Avg generation throughput: 5.8 tokens/s, Running: 0 reqs, Waiting: 0 reqs, GPU KV cache usage: 0.0%, Prefix cache hit rate: 0.0%
(APIServer pid=9806) INFO 09-08 16:26:41 [loggers.py:123] Engine 000: Avg prompt throughput: 0.0 tokens/s, Avg generation throughput: 0.0 tokens/s, Running: 0 reqs, Waiting: 0 reqs, GPU KV cache usage: 0.0%, Prefix cache hit rate: 0.0%

Postman API Testing Environment

Samit Collection

HTTP POST http://138.197.136.10:80

vLLMSample / http://138.197.136.10:8000/v1/chat/completions

POST http://138.197.136.10:8000/v1/chat/completions

Params Authorization Headers (9) Body Scripts Settings Cookies Beautify

Body (raw JSON)

```
1 {  
2   "model": "Qwen/Qwen2.5-7B-Instruct",  
3   "messages": [  
4     {"role": "user", "content": "สวัสดี! คุณคือใคร?"}  
5   ]  
6 }
```

Body Cookies Headers (4) Test Results

200 OK 3.87 s 990 B Save Response

Body (JSON) Preview Visualize

```
1 {  
2   "id": "chatmpl-47cb50135d85450087ecd80ee6ecbe1a",  
3   "object": "chat.completion",  
4   "created": 1757348785,  
5   "model": "Qwen/Qwen2.5-7B-Instruct",  
6   "choices": [  
7     {  
8       "index": 0,  
9       "message": {  
10         "role": "assistant",  
11         "content": "สวัสดี! ฉันเป็น Qwen เป็นโมเดล AI ที่สร้างโดย Alibaba Cloud ฉันสามารถช่วยตอบคำถาม สนทนา และให้ข้อมูลค่าๆ ได้ครับ/ค่ะ",  
12         "refusal": null,  
13         "annotations": null,  
14         "audio": null,  
15         "function_call": null,  
16         "tool_calls": [],  
17         "reasoning_content": null  
18       },  
19       "logprobs": null,  
20       "finish_reason": "stop",  
21       "stop_reason": null  
22     },  
23   ],  
24   "service_tier": null,  
25   "system_fingerprint": null,
```

Online Find and replace Console Postbot Runner Start Proxy Cookies Vault Trash



LLM providers



OpenRouter



สถาบันไอทีจีเนียส

www.itgenius.co.th



OpenRouter



Ollama



สถาบันไอทีจีเนียส

www.itgenius.co.th



Qwen: Qwen3 8B

 OpenRouter

Home Workspaces API Network

Search Postman Ctrl K

Invite Settings Notifications Upgrade No environment

Collections Environments Flows History

Samit

POST 01/_api/chat_01_start

HTTP AlChatbotLangchain / 02_Langchain_Basic / 01/_api/chat_01_start

POST {{baseURL}}/_api/chat_01_start

Params Authorization Headers (16) Body Scripts Settings Cookies

none form-data x-www-form-urlencoded raw binary GraphQL

This request does not have a body

Send

POST 01/_api/chat_01_start

POST 02/_api/chat_02_request

POST 03/_api/chat_03_template

POST 04/_api/chat_04_stream

05_Chat_History

06_Chat_History_Optimize

07_Tool_Calling

Document_Loader_EMBEDDING_pgVector

08_RAG

DjangoWebSocket

gofiber

QR Menu App API

Body Cookies (1) Headers (6) Test Results

200 OK 20.49 s 3.64 KB Save Response

{ } JSON Preview Visualize

```

1 {
2   "content": "สวัสดีครับ! ขอบคุณที่ส่องความเกี่ยวกับงบประมาณปีนี้ครับ สำหรับปีนี้บริษัทได้จัดทำงบประมาณรวมทั้งหมด **1,200 ล้านบาท**  

  ซึ่งเพิ่มขึ้นจากปีที่แล้วประมาณ **10%** เพื่อรองรับเป้าหมายการเติบโตและโอกาสใหม่ๆ ที่เกิดขึ้นในตลาด。/\ก/\ก/## รายละเอียดงบประมาณหลัก:  

  **ค่าใช้จ่ายในการดำเนินงาน (Operational Costs):** \k - เพิ่มขึ้น **8%** เมื่อจากต้นทุนวัสดุคงเหลือและค่าแรงเพิ่มขึ้นจากภาวะเงินฟื้น  

  รวมถึงการลงทุนในระบบดิจิทัลเพิ่มเติมของบริษัทฯ ที่ต้องใช้เงินจำนวนมาก。/\n2. **การวิจัยและพัฒนา (R&D):** \k - ได้รับงบประมาณเพิ่มขึ้น **15%**  

  เพื่อสนับสนุนโครงการวิจัยและพัฒนาอย่างต่อเนื่อง และการพัฒนาผลิตภัณฑ์ที่ตอบโจทย์ลูกค้าในอนาคต。/\n3. **การตลาด (Marketing):** \k - เพิ่มขึ้น **12%**  

  เพื่อยกระดับการตลาดในช่องทางดิจิทัล และเพิ่มการรับรู้แบรนด์ในตลาดใหม่。/\n4. **การขยายตัว (Expansion):** \k - แผนขยายสาขาในกรุงเทพฯ  

  และการสร้างศูนย์เทคโนโลยีในจังหวัดเชียงใหม่ ได้รับงบประมาณ **200 ล้านบาท** สำหรับการลงทุนในโครงสร้างพื้นฐาน。/\n/\k/## จุดเด่นสำคัญ: \k- **ความยั่งยืน (Sustainability):** \k บริษัทมุ่งมั่นการลดต้นทุนพลังงานและเพิ่มประสิทธิภาพการใช้ทรัพยากร โดยจัดสรรงบประมาณ **10%** สำหรับโครงการค่าสิ่งแวดล้อม。/\n-  

  **พัฒนาทีมงาน:** งบประมาณส่วนการพัฒนาทีมงานและห้องปฏิบัติการเพิ่มขึ้น **15%** เพื่อเสริมสร้างศักยภาพทีม  

  /\n/\k/## ขอขอบคุณสำหรับข้อมูลเพิ่มเติมเกี่ยวกับแผนการใช้งบประมาณในแผนกของคุณ หรือต้องการรายละเอียดเฉพาะเรื่อง สามารถสอบถามได้โดยครับ  

  ยินดีช่วยเหลือทุกเรื่อง! 😊",
3 "usedModel": "qwen/qwen3-8b:free"
4 }

```

LLM providers



API keys - OpenAI API

platform.openai.com/api-keys

ITGenius / Default project

Dashboard Docs API S

Create Chat Audio Images Assistants

Usage API keys Logs Storage Batches Optimize

API keys

+ Create new secret key

You have permission to view and manage all API keys in this project.

Do not share your API key with others or expose it in the browser or other client-side code. To protect your account's security, OpenAI may automatically disable any API key that has leaked publicly.

View usage per API key on the [Usage page](#).

NAME	SECRET KEY	CREATED BY	PERMISSIONS
ai-chatbot-langchain	sk-...MJgA	Samit Koyom	All
n8n sample	sk-...blcA	Samit Koyom	All
n8n test	sk-...iIYA	Samit Koyom	All



GPT-4o mini



OpenAI

paces ▾ API Network

Search Postman

Ctrl K

Invite



New Import

POST 01/_api/chat_01_start



HTTP AIChatbotLangchain / 02_Langchain_Basic / 01/_api/chat_01_start

POST

`{baseUrl} /api/chat_01_start`

Params

Authorization

Headers (16)

Body

Scripts

Settings

none

form-data

x-www-form-urlencoded

raw

binary

GraphQL

This request does not have a body

Body

Cookies (1)

Headers (6)

Test Results

200 OK

5.95 s

1.11 KB

{ } JSON

Preview

Visualize

```
1  {
2      "content": "สวัสดีครับ งบประมาณเป็นเรามีการจัดสรรงบประมาณตามแผนกลยุทธ์ของบริษัท โดยเน้นการลงทุนในด้านการพัฒนาผลิตภัณฑ์ใหม่ การตลาด และการฝึกอบรมพนักงาน เพื่อเพิ่มประสิทธิภาพในการทำงานและขยายตลาด/ก/หากคุณต้องการรายละเอียดเพิ่มเติมเกี่ยวกับงบประมาณในแต่ละแผนกหรือโครงการใด ๆ สามารถสอบถามได้เลยครับ",
3      "usedModel": "gpt-4o-mini-2024-07-18"
4 }
```

LLM providers



สถาบันไอทีจีเนียส

www.itgenius.co.th

Google AI Studio

aistudio.google.com/apikey

API Keys

+ Create API key

Quickly test the Gemini API

API quickstart guide

Code

```
curl "https://generativelanguage.googleapis.com/v1beta/models/gemini-2.0-flash:generateContent" \
-H 'Content-Type: application/json' \
-H 'X-goog-api-key: GEMINI_API_KEY' \
-X POST \
-d '{
  "contents": [
    {
      "parts": [
        {
          "text": "Explain how AI works in a few words"
        }
      ]
    }
]'
```

Your API keys are listed below. You can also view and manage your project and API keys in Google Cloud.

Look up API Key for project

Project number	Project name	API key	Created	Plan

Studio

Dashboard

API keys

Usage & Billing

Changelog

Documentation

Get API key

View status

Settings

Samitkoyom@gmail.com

AI

genius.co.th

LLM providers



Azure AI Studio



สถาบันไอทีจีเนียส

www.itgenius.co.th

portal.azure.com/#view/Microsoft_Azure_CostManagement/Menu/~/subscriptions

Microsoft Azure

Search resources, services, and docs (G+)

Copilot

New Chrome available

Home > Cost Management: Samit Koyom

Cost Management: Samit Koyom | Azure subscriptions

Billing account

Search Add Refresh Export to CSV Troubleshoot Feedback

To make it easier to view all your subscriptions together we're removing this page and listing Azure subscriptions in the new All billing subscriptions page instead. To view your Azure subscriptions, select All billing subscriptions in the left pane, and then select Usage based / Azure subscriptions tab.

Overview Change scope Access control Diagnose and solve problems Reporting + analytics Monitoring Optimization Settings Billing Invoices Payment methods Azure subscriptions Help

View Azure subscriptions billed to your account. The charges shown below are estimated amounts based on your Azure usage and do not include tax. The amount excludes Azure reservations and marketplace transactions.

Text search Invoice section : All invoice sections Billing profile : All billing profiles Status : Any status

Showing 1 to 1 of 1 subscriptions.

Name ↑↓	ID ↑↓	Plan ↑↓	Invoice section ↑↓	Billing profile ↑↓	Status ↑↓	Service tenant ID	Month-to-date charges	Last month's charges
Azure subscription 1	9afdd173-1a1e-42d0-b...	Microsoft Azure Plan	Samit Koyom	Samit Koyom	Active	a0047fc6-f8fa-4f4c-96c...	0.00	0.00 ***

< Previous Page 1 of 1 Next >

Add or remove favorites by pressing Ctrl + Shift + F



สถาบันไอทีเจเนียส

www.itgenius.co.th

Microsoft Azure

Search resources, services, and docs (G+)

Copilot

samit90daytalk@hotmail.com DEFAULT DIRECTORY

Home > SamitApp

SamitApp | Keys and Endpoint

Azure OpenAI

Search Regenerate Key1 Regenerate Key2

Overview Activity log Access control (IAM) Tags Diagnose and solve problems Resource visualizer

Resource Management

Keys and Endpoint (selected)

Encryption Pricing tier Networking Identity Cost analysis Properties

Security Monitoring Automation Help

KEY 1

These keys are used to access your Azure AI Foundry API. Do not share your keys. Store them securely—for example, using Azure Key Vault. We also recommend regenerating these keys regularly. Only one key is necessary to make an API call. When regenerating the first key, you can use the second key for continued access to the service.

Show Keys

KEY 1

KEY 2

Location/Region

eastus

Endpoint

https://samitapp.openai.azure.com/

<https://portal.azure.com/#@samit90daytalk@hotmail.onmicrosoft.com/resource/subscriptions/9afdd173-1a1e-42d0-b157-3c57455b4b30/resourceGroups/SampleApp/providers/Microsoft.CognitiveServices/accounts/SamitApp/cskeys>



ai.azure.com/resource/models?wsid=/subscriptions/9afdd173-1a1e-42d0-b157-3c57455b4b30/resourceGroups/sampleapp/providers/Microsoft.CognitiveServices/accounts/samit-me4hlzb5-eastus2&tid=a0047fc6-f8fa-4f4c-96c5-a53e... ☆ 🔍 🌐 🌐 🌐 New Chrome available :

Azure AI Foundry / samit-me4hlzb5-eastus2 / Model catalog

Find the right model to build your custom AI solution

Announcements

Introducing GPT-audio & GPT-realtime

GPT-Realtime boosts speech-to-speech instruction following; GPT-Audio powers...

[Check out models](#) [Read blog ↗](#)

Introducing Mistral Document AI (25.05)

Document processing with state-of-the-art OCR and structured data extraction

[Check out model](#) [Read blog ↗](#)

Introducing GPT-5 models

GPT-5 unifies frontier reasoning and advanced coding with high-performance...

[Check out models](#) [Read blog ↗](#)

Announcing new gpt-oss models

Push the open model frontier with gpt-oss-120b and gpt-oss-20b, released und...

[Check out models](#) [Read blog ↗](#)

Model Router with GPT-5

Model router fast-tracks efficient adoption of the GPT-5 models for your business.

[Check out model](#) [Read blog ↗](#)

Industry Capabilities Inference tasks Fine-tuning tasks Licenses

Search Models 98

gpt-5-mini Chat completion	gpt-5-nano Chat completion	gpt-5-chat Chat completion	FLUX-1.1-pro Text to image	FLUX.1-Kontext-pro Text to image, Image to image	codex-mini Responses
DeepSeek-R1-0528 Chat completion	sora Video generation	grok-3 Chat completion	grok-3-mini Chat completion	model-router Chat completion	o4-mini Chat completion
MAI-DS-R1 Chat completion	gpt-4.1 Chat completion	gpt-4.1-mini Chat completion	gpt-4.1-nano Chat completion	mistral-medium-2505 Chat completion, Image classific...	Phi-4-reasoning Chat completion
Phi-4-mini-reasoning Chat completion	Llama-4-Scout-17B-16E-Inst... Chat completion	Llama-4-Maverick-17B-128E... Chat completion	cohere-command-a Chat completion	embed-v-4-0 Embeddings, Summarization	o3-mini Chat completion
DeepSeek-V3-0324 Chat completion	gpt-4o-mini-tts Text to speech	gpt-4o-transcribe Speech to text	gpt-4o-mini-transcribe Speech to text	DeepSeek-V3 Chat completion	DeepSeek-R1 Chat completion
Phi-4-mini-instruct Chat completion	Phi-4-multimodal-instruct Chat completion	Phi-4 Chat completion	mistral-small-2503 Chat completion, Image classific...	gpt-4o-mini-audio-preview Audio generation	gpt-4o-mini-realtime-preview Audio generation



ai.azure.com/resource/models/gpt-5-mini/version/2025-08-07/registry/azure-openai?wsid=/subscriptions/9afdd173-1a1e-42d0-b157-3c57455b4b30/resourceGroups/sampleapp/providers/Microsoft.CognitiveServices/accounts/samit...

Azure AI Foundry / samit-me4hlzb5-eastus2 / Models / gpt-5-mini

gpt-5-mini

Use this model Fine-tune

Overview Model catalog Playgrounds Build and customize Agents Templates Fine-tuning Observe and optimize Monitoring Protect and govern Azure OpenAI Evaluation Guardrails + controls Risks + alerts Governance Azure OpenAI Stored completions Batch jobs Assistant vector stores Data files My assets Models + endpoints Web apps More Management center

Details Existing deployments License

gpt-5-mini powers low cost and fast experiences such as real-time agents, orchestrating tool calls in response to customer support requests.

Key Capabilities

- A lightweight version for cost-sensitive applications.
- Now supporting minimal reasoning, a new verbosity setting, and the "customs" tool for raw text output.
- Supports new "allowed tools" tool choice that enables you to specify multiple tools in the tool choice instead of just one
- supports new "preamble" support, allowing the model to "think" before calling a tool. This is always enabled and controlled through prompting.
- gpt-5-mini supports multimodal inputs, real-time streaming and full tool support for smarter, more dynamic user experiences

See more

Model Versions

Learn more about region availability

East US 2

Model ID	Deployment type	Lifecycle	Max request	Retirement Date
2025-08-07	Global Standard, Data Zone Standard	Generally available	Input: N/A Output: N/A	Fri, Aug 7, 2026

Quick facts

gpt-5-mini

Chat completion Direct from Azure

Training data last updated May 2024

Pricing See direct from Azure pricing

Model ID

Reference this model ID when deploying the model in code

azureml://registries/azure-openai/models/gpt-5-mini/versions/2025-08-07



Azure AI Foundry / samit-me4hlzb5-eastus2 / Models / gpt-5-mini

Overview Model catalog Playgrounds Build and customize Agents Templates Fine-tuning Observe and optimize Monitoring Protect and govern Azure OpenAI Evaluation Guardrails + controls Risks + alerts PREVIEW Governance PREVIEW Azure OpenAI Stored completions Batch jobs Assistant vector stores Data files My assets Models + endpoints Web apps ... More Management center

gpt-5-mini

Use this model Fine-tune

Details Existing deployments License

Key Capabilities

- A lightweight version for cost-sensitive applications.
- Now supporting minimal reasoning, a new verbosity setting.
- Supports new "allowed tools" tool choice that enables you to...
- supports new "preamble" support, allowing the model to "th...
- gpt-5-mini supports multimodal inputs, real-time streaming

See more

Model Versions

Learn more about region availability

East US 2

Model ID	Deployment type
2025-08-07	Global Standard, Data Zone Standard

Data, media and languages

Property	Description
Supported data types	Inputs text, image
Outputs	text
Supported languages	en

Deploy gpt-5-mini

Deployment name * gpt-5-mini-2

Deployment type Global Standard

Global Standard: Pay per API call with the highest rate limits. Learn more about Global deployment types. Data might be processed globally, outside of the resource's Azure geography, but data storage remains in the AI resource's Azure geography. Learn more about data residency.

Deployment details

Model version 2025-08-07

AI resource samit-me4hlzb5-eastus2

Capacity 100K tokens per minute (TPM)

Resource location East US 2

Content safety DefaultV2

Version upgrade policy Once a new default version is available

Customize Deploy Cancel

Quick facts

gpt-5-mini Chat completion Direct from Azure

Training data last updated May 2024

Pricing See direct from Azure pricing

Model ID Reference this model ID when deploying the model in code

azureml://registries/azure-openai/models/gpt-5-mini/versions/2025-08-07



Azure AI Foundry / samit-me4hlzb5-eastus2 / Deployments / gpt-5-mini-2

Overview Model catalog Playgrounds Build and customize Agents Templates Fine-tuning Observe and optimize Monitoring Protect and govern Azure OpenAI Evaluation Guardrails + controls Risks + alerts Governance Azure OpenAI Stored completions Batch jobs Assistant vector stores Data files My assets Models + endpoints Web apps ... More Management center

← gpt-5-mini-2

Details Metrics

[Open in playground](#) Request quota Edit Delete

Endpoint

Target URI
https://samit-me4hlzb5-eastus2.cognitiveservices.azure.com/openai/responses...

Deployment info

Name	gpt-5-mini-2	Provisioning state	Succeeded
Deployment type	Global Standard	Created on	2025-09-09T09:35:38.393495Z
Created by	samit90daytalk@hotmail.com	Modified on	Sep 9, 2025 4:35 PM
Modified by	samit90daytalk@hotmail.com	Version upgrade policy	Once a new default version is available
Rate limit (Tokens per minute)	100,000	Rate limit (Requests per minute)	100
Model name	gpt-5-mini	Model version	2025-08-07
Life cycle status	GenerallyAvailable	Date created	Aug 7, 2025 7:00 AM
Date updated	Aug 7, 2025 7:00 AM	Model retirement date	Aug 8, 2026 7:00 AM

Monitoring & safety

Content filter DefaultV2

Language: Javascript SDK: OpenAI SDK Authentication type: Key Authentication

Get Started

Below are example code snippets for a few use cases. For additional information about Azure OpenAI SDK, see full [documentation](#) and [samples](#).

1. Authentication using API Key

For OpenAI API Endpoints, deploy the Model to generate the endpoint URL and an API key to authenticate against the service. In this sample endpoint and key are strings holding the endpoint URL and the API Key.

The API endpoint URL and API key can be found on the Deployments + Endpoint page once the model is deployed.

To create a client with the OpenAI SDK using an API key, initialize the client by passing your API key to the SDK's configuration. This allows you to authenticate and interact with OpenAI's services seamlessly.

```
const api_key = "<your-api-key>";
const endpoint = "https://samit-me4hlzb5-eastus2.cognitiveservices.azure.com/openai/v1";
const modelName = "gpt-5-mini";
const deployment_name = "gpt-5-mini-2";

const client = new OpenAI({
  baseURL: endpoint,
  apiKey: api_key
});
```

2. Install dependencies

- Install Node.js
- Copy the following lines of text and save them as a file package.json inside your folder.

```
{
  "type": "module"
```

Azure AI Foundry / samit-me4hlzb5-eastus2 / Deployments / gpt-5-mini-2

Docs All resources ⚙️ 😊 samit-me4hlzb5-eastus2 (eastus2, S0) SK

← gpt-5-mini-2

Overview Model catalog Playgrounds Build and customize Agents Templates Fine-tuning Observe and optimize Monitoring Protect and govern Azure OpenAI Evaluation Guardrails + controls Risks + alerts Governance Azure OpenAI Stored completions Batch jobs Assistant vector stores Data files My assets Models + endpoints Web apps More Management center

Details Metrics

Open in playground Request quota Edit Delete

Endpoint

Target URI
https://samit-me4hlzb5-eastus2.cognitiveservices.azure.com/openai/responses...

Key
.....

Deployment info

Name gpt-5-mini-2	Provisioning state Succeeded
Deployment type Global Standard	Created on 2025-09-09T09:35:38.393495Z
Created by samit90daytalk@hotmail.com	Modified on Sep 9, 2025 4:35 PM
Modified by samit90daytalk@hotmail.com	Version upgrade policy Once a new default version is available
Rate limit (Tokens per minute) 100,000	Rate limit (Requests per minute) 100
Model name gpt-5-mini	Model version 2025-08-07
Life cycle status GenerallyAvailable	Date created Aug 7, 2025 7:00 AM
Date updated Aug 7, 2025 7:00 AM	Model retirement date Aug 8, 2026 7:00 AM

Language

REST curl Key Authentication

Get Started

1. Authentication using API Key

For Serverless API Endpoints, deploy the Model to generate the endpoint URL and an API key to authenticate against the service. In this sample endpoint and key are strings holding the endpoint URL and the API Key. The API endpoint URL and API key can be found on the Deployments + Endpoint page once the model is deployed.

If you're using bash:

```
export AZURE_API_KEY=<your-api-key>
```

If you're in powershell:

```
$Env:AZURE_API_KEY = "<your-api-key>"
```

If you're using Windows command prompt:

```
set AZURE_API_KEY = <your-api-key>
```

2. Run a basic code sample

Paste the following into a shell

```
curl -X POST "https://samit-me4hlzb5-eastus2.cognitiveservices.azure.com/openai/responses?api-version=2025-04-01-preview" \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $AZURE_API_KEY" \
-d '{
  "messages": [
    {
      "role": "user",
      "content": "Hello, how are you?"
    }
  ]
}'
```

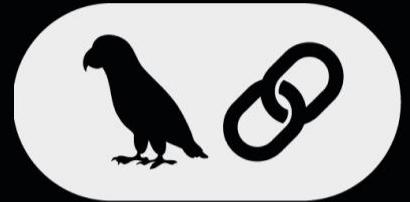


```
pwsh in Samit x + v
Samit $env:AZURE_API_KEY = "5KcA$GXATX"
Samit $uri = "https://samit-me4hlzb5-eastus2.cognitiveservices.azure.com/openai/responses?api-version=2025-04-01-preview"
Samit $headers = @{
    "Content-Type" = "application/json"
    "Authorization" = "Bearer $env:AZURE_API_KEY"
}
Samit $body = @{
    "input": "I am going to Thailand, what should I see?",
    "model": "gpt-5-mini-2"
}
Samit Invoke-RestMethod -Uri $uri -Method 'POST' -Headers $headers -Body $body

id : resp_68bfffba565a08190813c449c9d3d98f105209028245e369f
object : response
created_at : 1757412261
status : completed
background : False
content_filters :
error :
incomplete_details :
instructions :
max_output_tokens :
max_tool_calls :
model : gpt-5-mini-2
output : {@{id=rs_68bfffba5b9388190909f928e8620efa605209028245e369f; type=reasoning; summary=System.Object[]},@{id=msg_68bfffba1220819093fcdc6d65d801f405209028245e369f; type=message; status=completed; content=System.Object[]; role=assistant}}
parallel_tool_calls : True
previous_response_id :
prompt_cache_key :
reasoning : @{effort=medium; summary=}
safety_identifier :
service_tier : default
store : True
temperature : 1.0
text : @{format=}
tool_choice : auto
tools : {}
top_p : 1.0
```



สร้าง AI Chatbots สำหรับองค์กร



LangChain
ร่วมกับ Next.JS
และ ➡ supabase



เมวิด์ไอบันทึกการอบรม
ย้อนหลังให้ทุกวัน



สถาบันไอทีจีเนียส

วัน
4 12
ชั่วโมงเต็ม



Samit Koyom
สถาบันไอทีจีเนียส

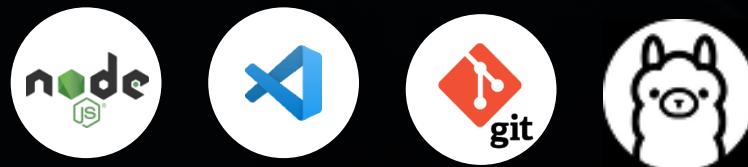


โปรแกรม (Tool and Editor) ที่ใช้บرم

แนะนำ หลักสูตรนี้ใช้ Node.js เวอร์ชัน 20 ขึ้นไป

1. Node.js 22.x
2. Visual Studio Code
3. Git
4. Ollama (Optional) - ไม่บังคับ

แนะนำ Ollama เป็นเครื่องมือรัน AI model แบบ Local เมน้ำสำหรับเครื่องที่ VGA แยก และคอมพิวเตอร์ควรมี Spec สูงพอควร ไม่จำเป็น และไม่บังคับให้ต้องติดตั้งหากเครื่องไม่พร้อม





1. ติดตั้ง Node JS



Download Node.JS V.22.x

<https://nodejs.org/en/>

The screenshot shows the official Node.js download page at <https://nodejs.org/en/download>. At the top, there's a banner stating "New security releases to be made available Wednesday, May 14, 2025". The main heading is "Download Node.js®". Below it, a dropdown menu shows "Get Node.js® v22.15.0 (LTS)" selected, along with "for Windows" and "using fpm" and "with npm". A code block displays a PowerShell script for installing Node.js via fpm:

```
1 # Download and install fpm:  
2 winget install Schniz.fpm  
3  
4 # Download and install Node.js:  
5 fpm install 22  
6  
7 # Verify the Node.js version:  
8 node -v # Should print "v22.15.0".  
9  
10 # Verify npm version:  
11 npm -v # Should print "10.9.2".
```

A "Copy to clipboard" button is available for the script. Below the script, a note says "'fpm' is a cross-platform Node.js version manager. If you encounter any issues please visit [fpm's website](#)". Further down, it says "Or get a prebuilt Node.js® for Windows running a x64 architecture." with two download buttons: "Windows Installer (.msi)" and "Standalone Binary (.zip)". At the bottom, there's a link to the "changelog" and "blog post" for this version.

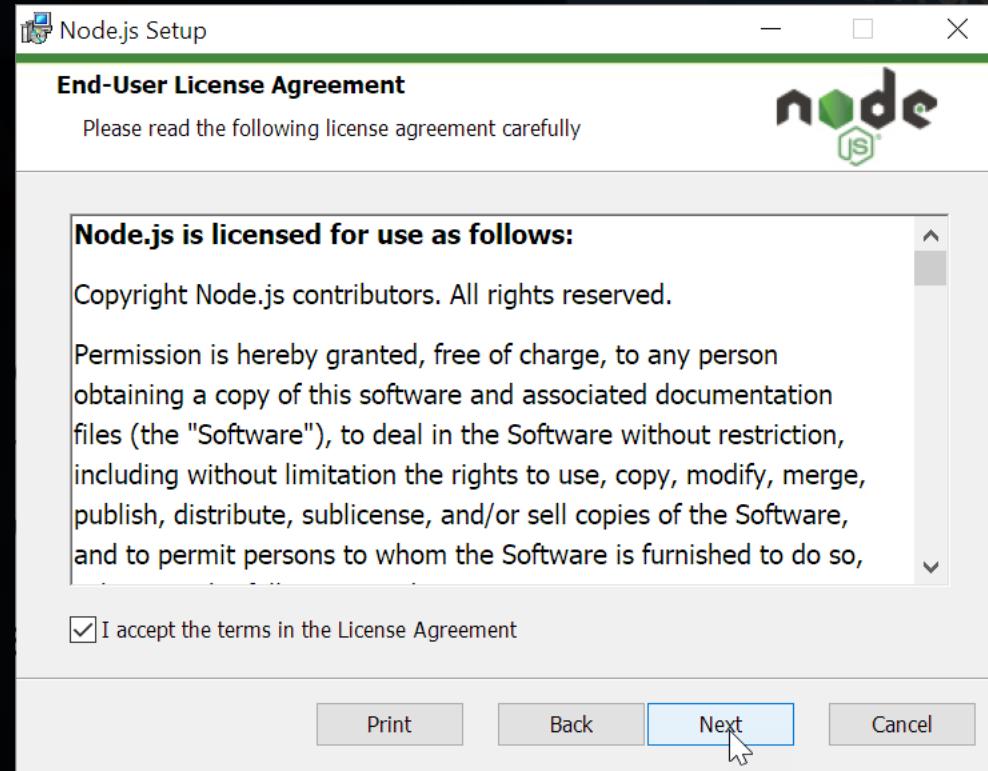
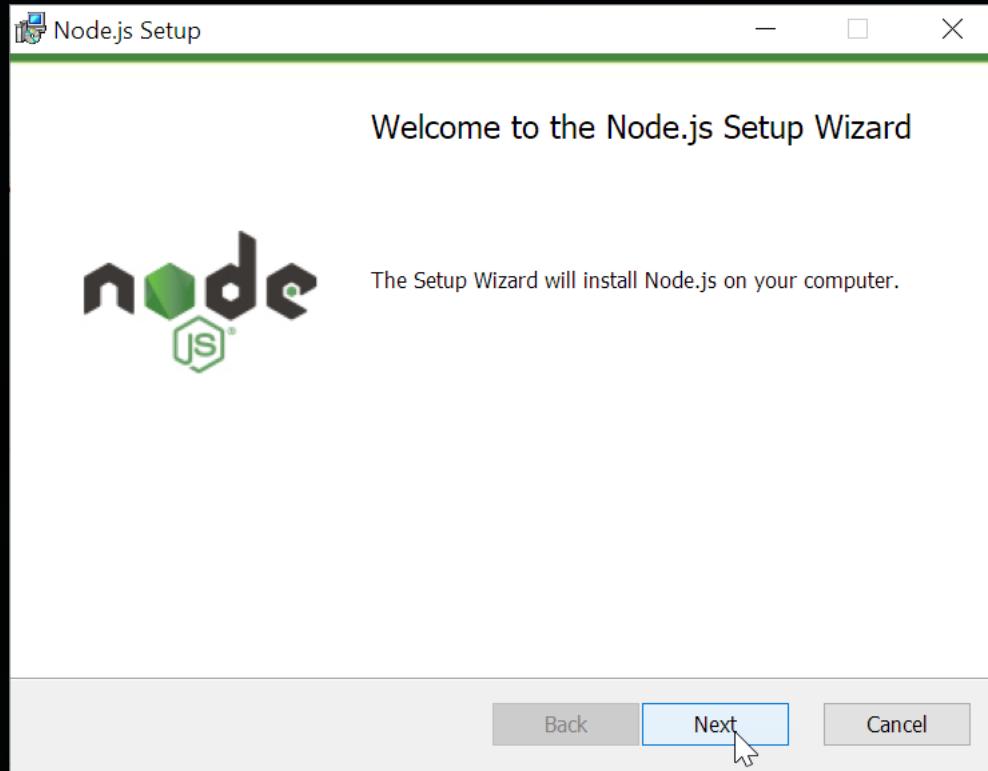


หมายเหตุ การอัปเดตของ Node.js 20 ขึ้นไป สามารถใช้ Node.js 21, 22, 23 หรือ 24 ก็ได้

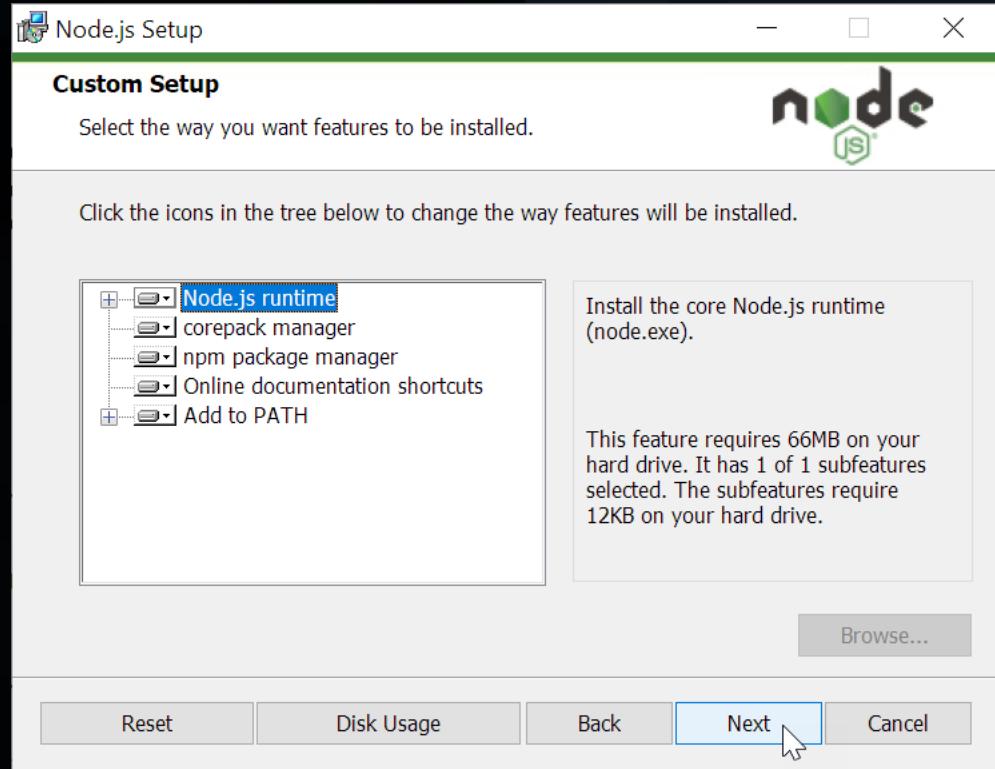
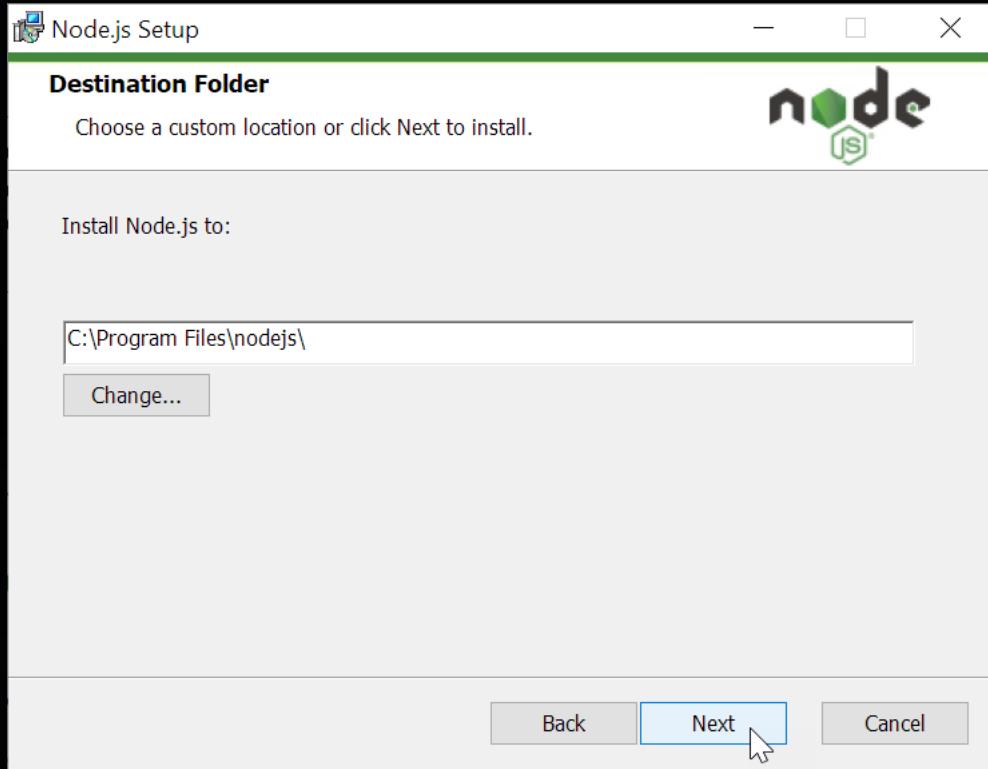
สถาบันไอทีจีนียส

www.itgenius.co.th

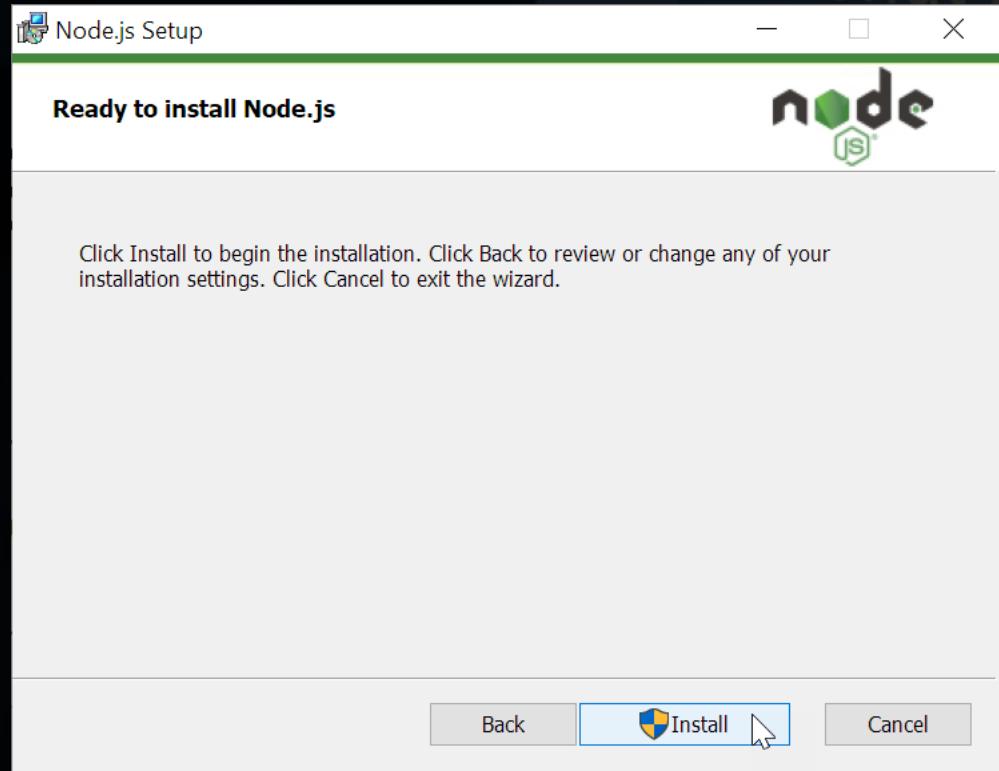
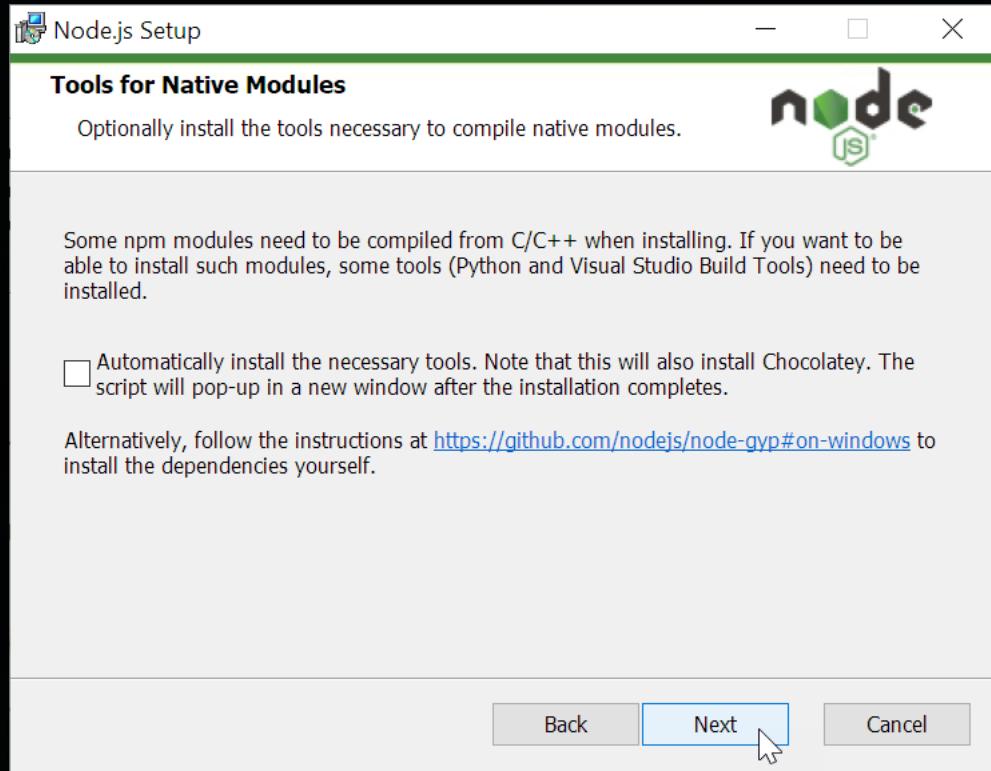
ติดตั้ง Node.JS V.22.x



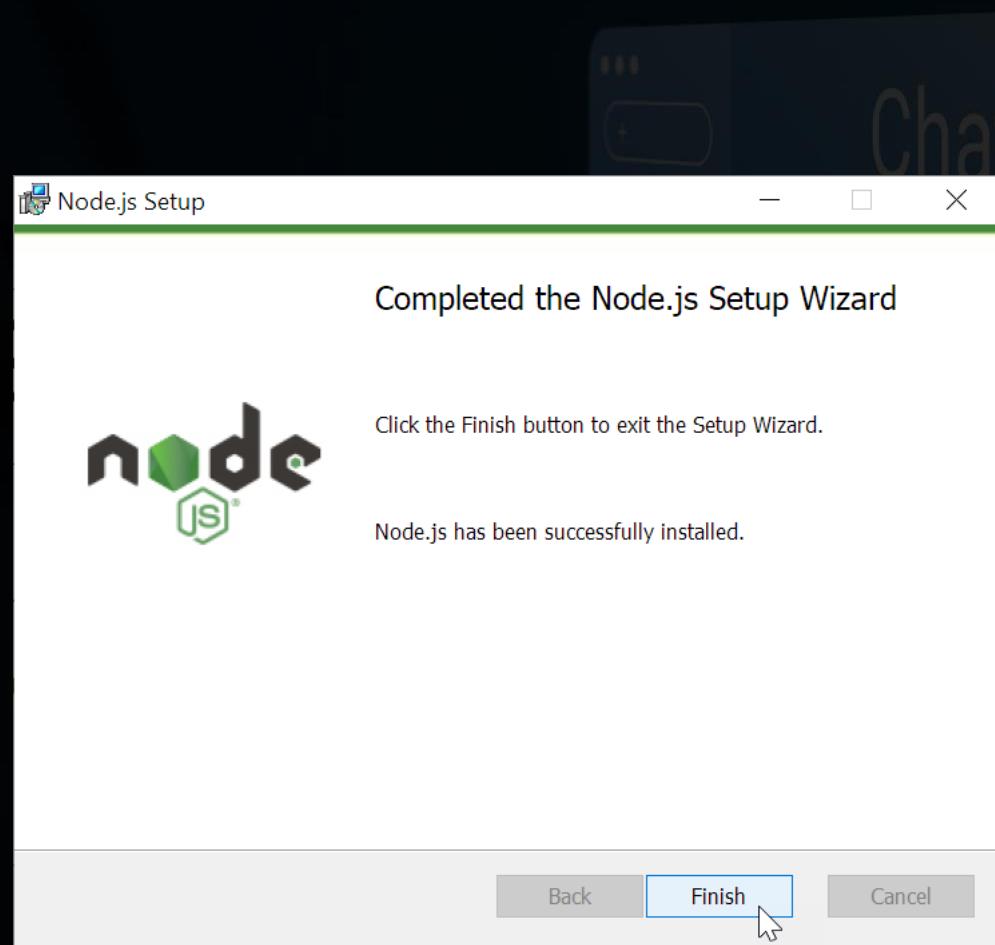
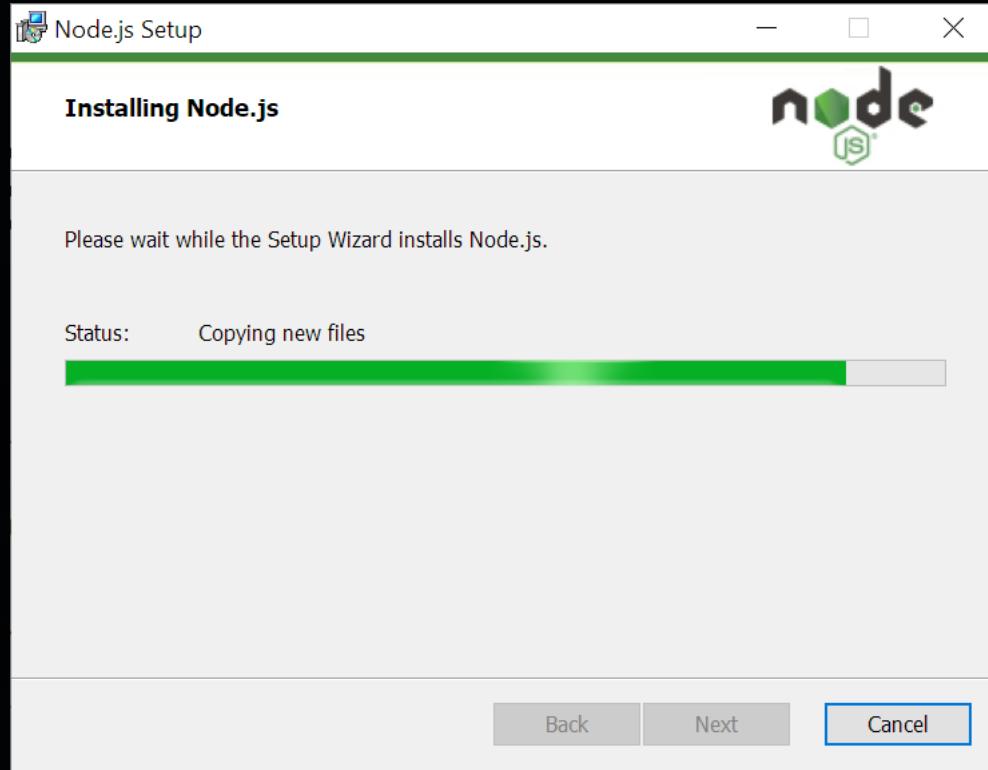
ติดตั้ง Node.js V.22.x



ติดตั้ง Node.js V.22.x



ติดตั้ง Node.JS V.22.x



ทดสอบหลังติดตั้งเสร็จ

```
node -v
```

```
C:\Users\Samit>node -v  
v22.14.0
```

```
C:\Users\Samit>
```

```
npx -v
```

```
C:\Users\Samit>npx -v  
10.9.2
```

```
C:\Users\Samit>
```

```
npm -v
```

```
C:\Users\Samit>npm -v  
10.9.2
```

```
C:\Users\Samit>
```

หมายเหตุ การอัปเดต Node.js 20 ขึ้นไป สามารถใช้ Node.js 21, 22, 23 หรือ 24 ก็ได้

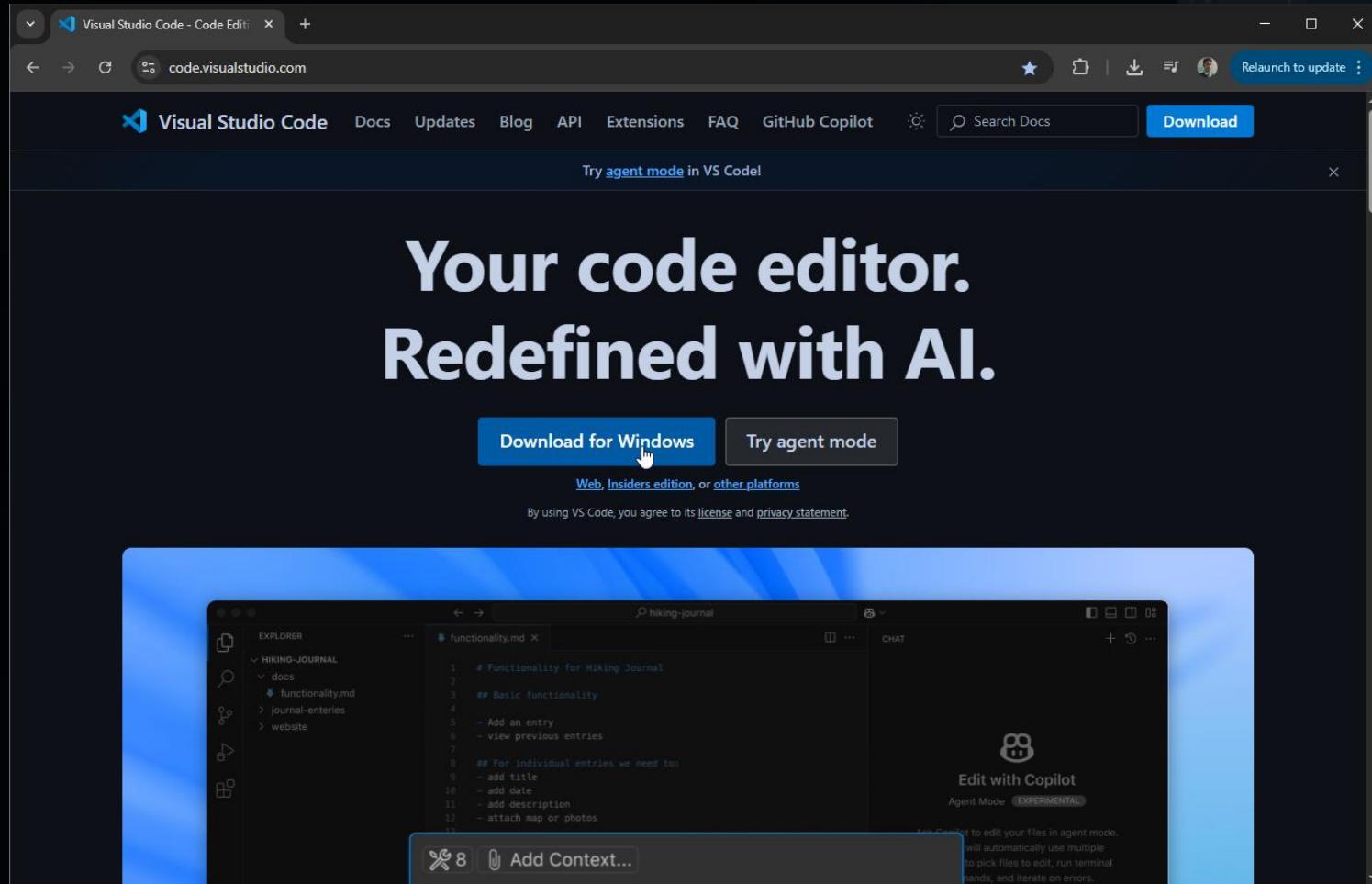




2. ติดตั้ง Visual Studio Code



ຕັດຕັ້ງ Visual Studio Code ວຽກຄ່າສ່ວນເສຣິມທີ່ຈຳເປັນ



ເຂົ້າໄປລາວນີ້ໂລດ Visual Studio Code ໄດ້ທີ່ <https://code.visualstudio.com>

www.vtgenius.co.th

การติดตั้งส่วนเสริม (Extension) ของ Visual Studio Code



รายชื่อ Extensions ที่แนะนำสำหรับ VS Code

- 1. ES7+React/Redux/React-Native snippets** by dsznajder
- 2. Auto Import – ES6, TS, JSX, TSX** by Sergey Korenuk
- 3. Color Picker** by anseki
- 4. Material Icon Theme** by Philipp Kief
- 5. Tailwind CSS IntelliSense** by Tailwind Labs
- 6. Prettier – Code formatter** by Prettier
- 7. One Dark Pro** by binaryify

NEXT.JS





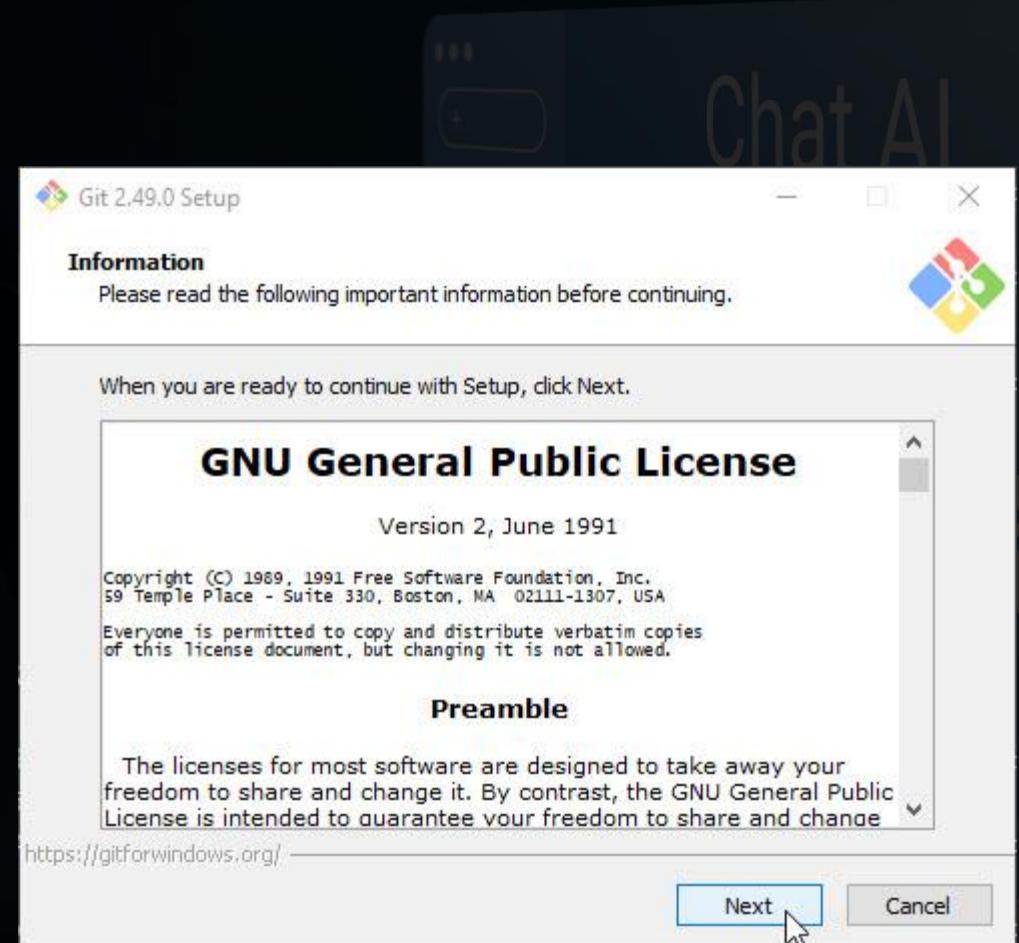
3. ติดตั้ง Git

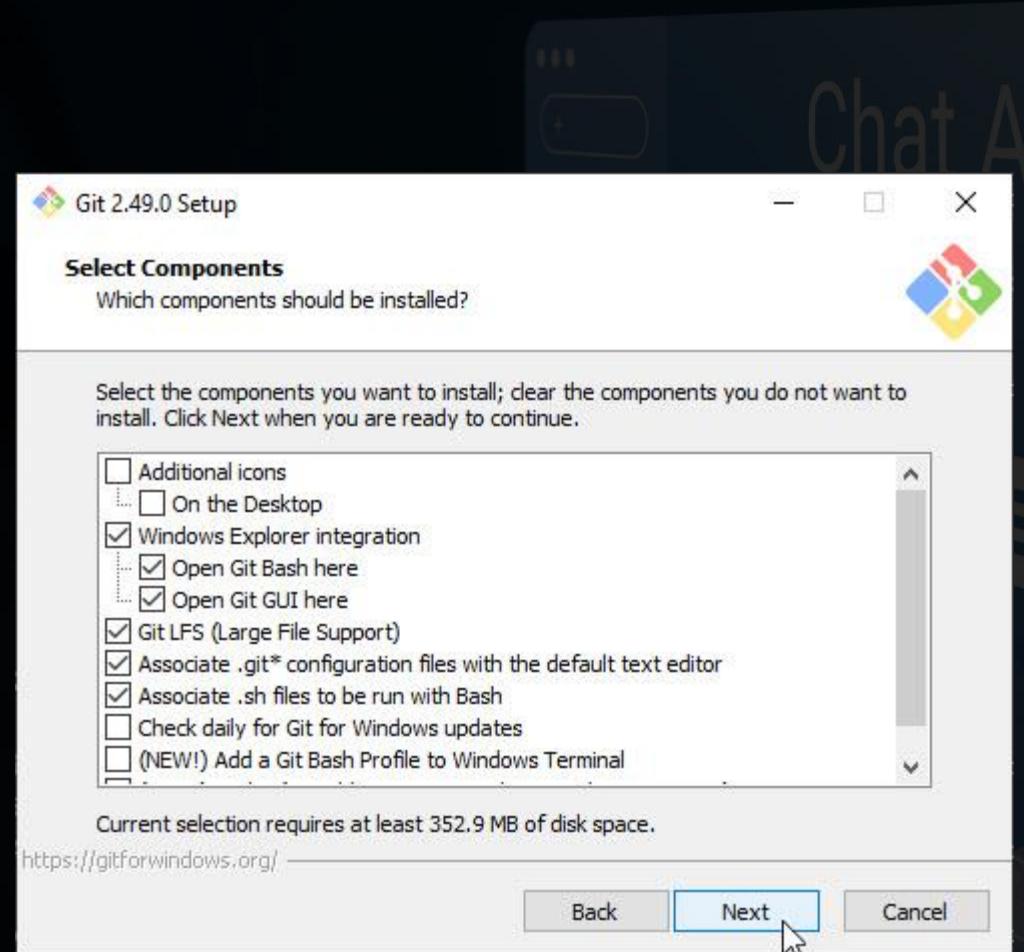
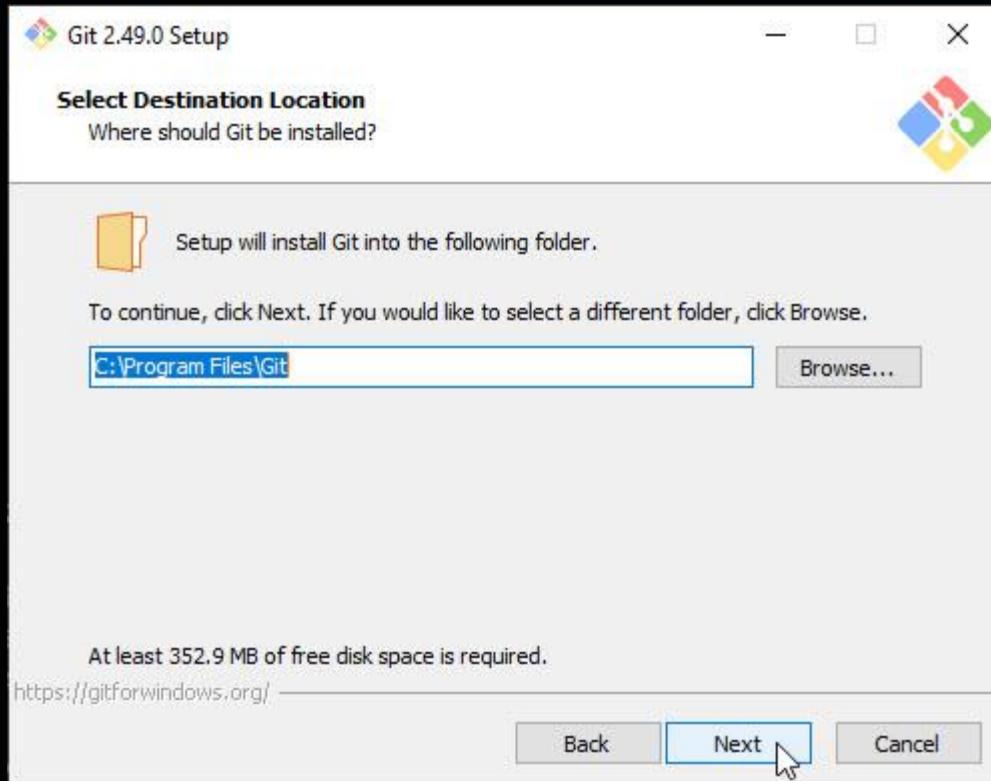


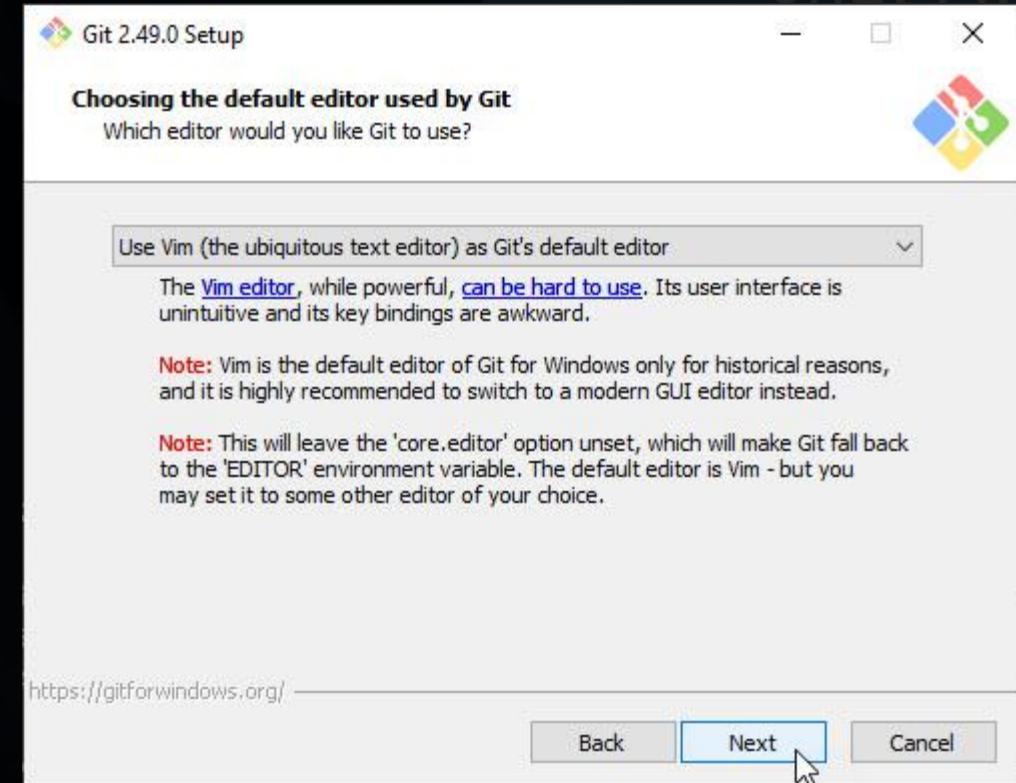
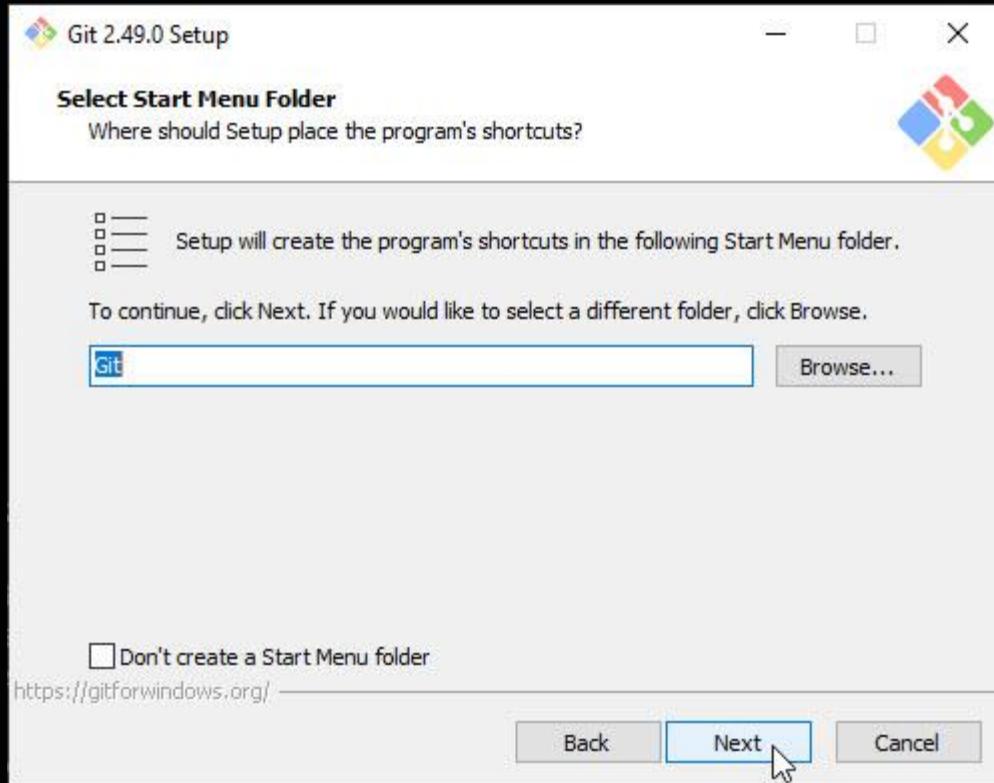
ดาวน์โหลดไฟล์ติดตั้ง Git ได้ที่ <https://git-scm.com/>

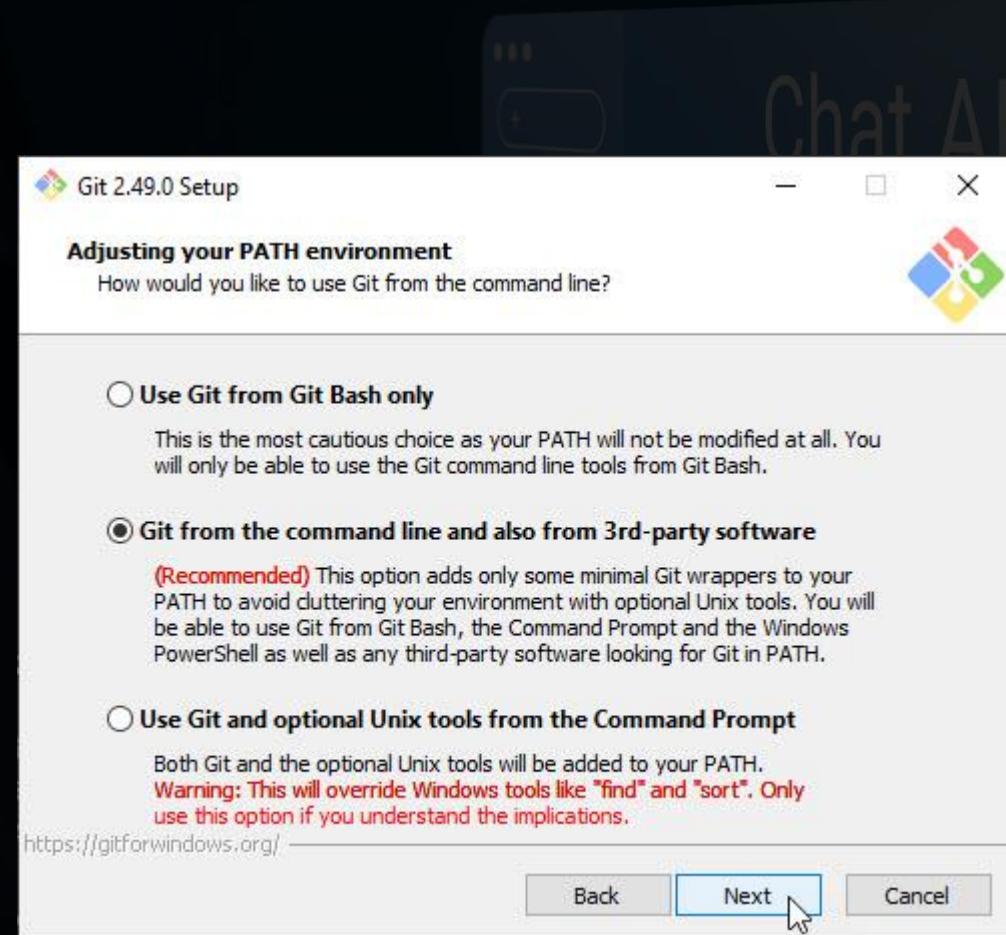
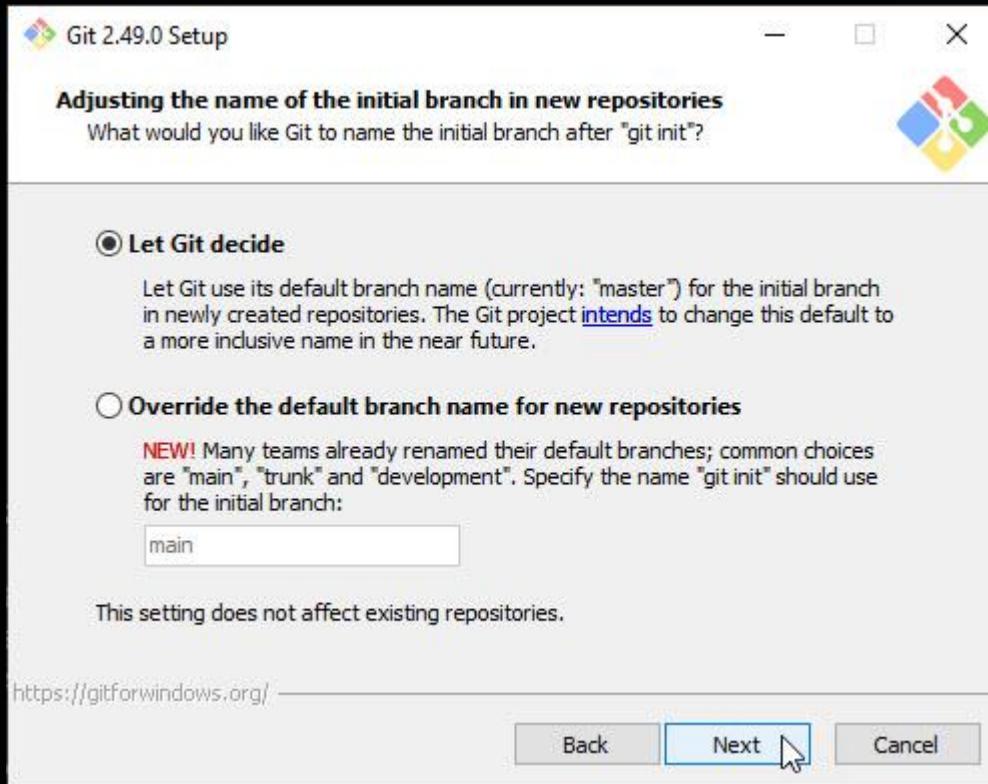
The screenshot shows the official website for Git (<https://git-scm.com/>). The page features a dark background with orange text highlights. At the top, there's a brief introduction to Git as a free and open-source distributed version control system. Below this, a diagram illustrates the concept of distributed branching between multiple repositories. The main navigation menu includes links for "About", "Documentation", "Downloads", and "Community". A prominent section on the right displays the "Latest source Release" (version 2.49.0) with a "Download for Windows" button. Other download options like "Windows GUIs", "Mac Build", "Tarballs", and "Source Code" are also listed.

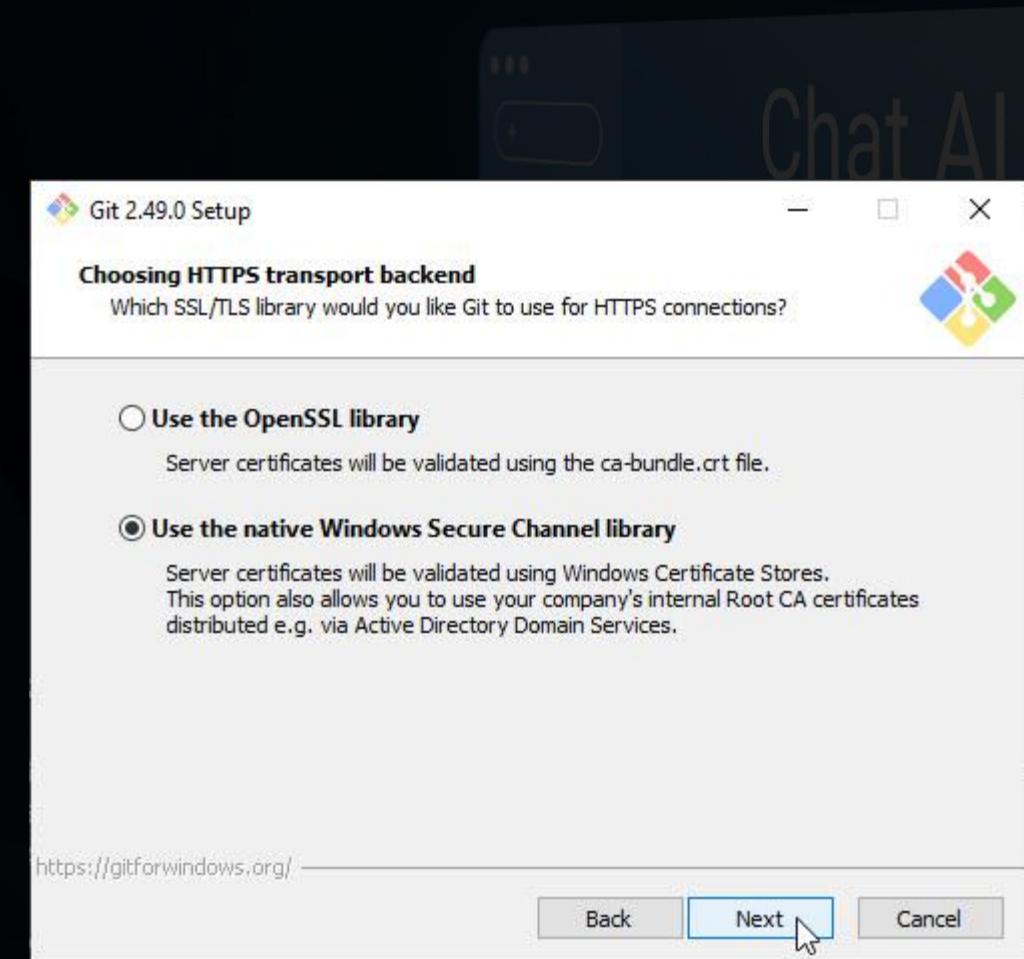
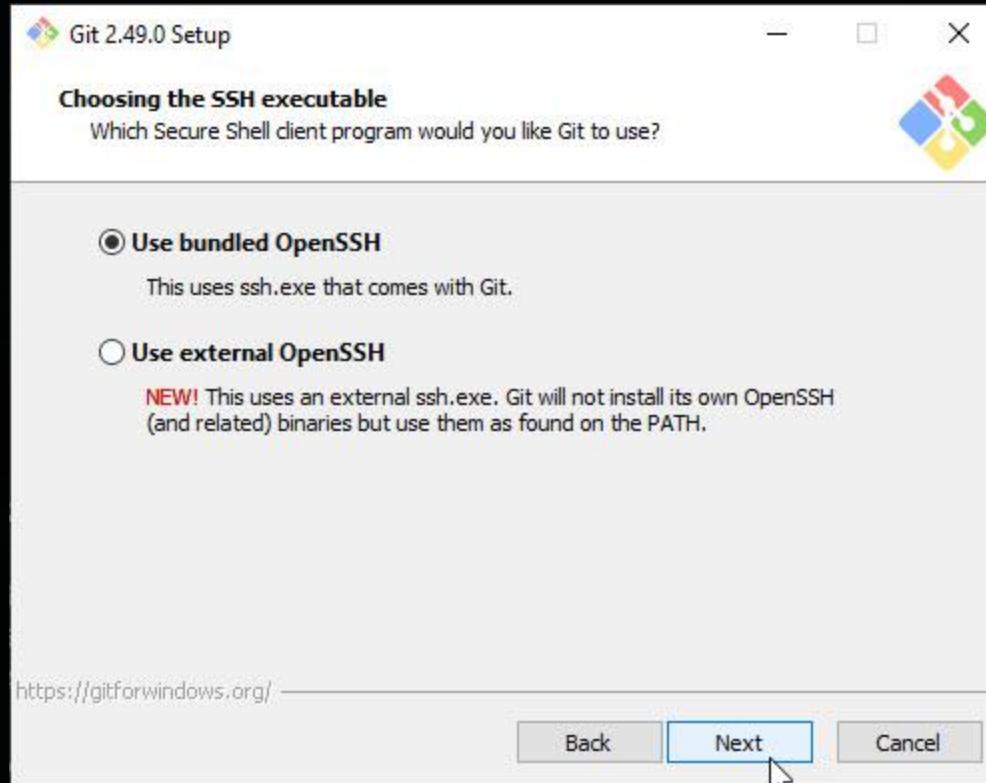


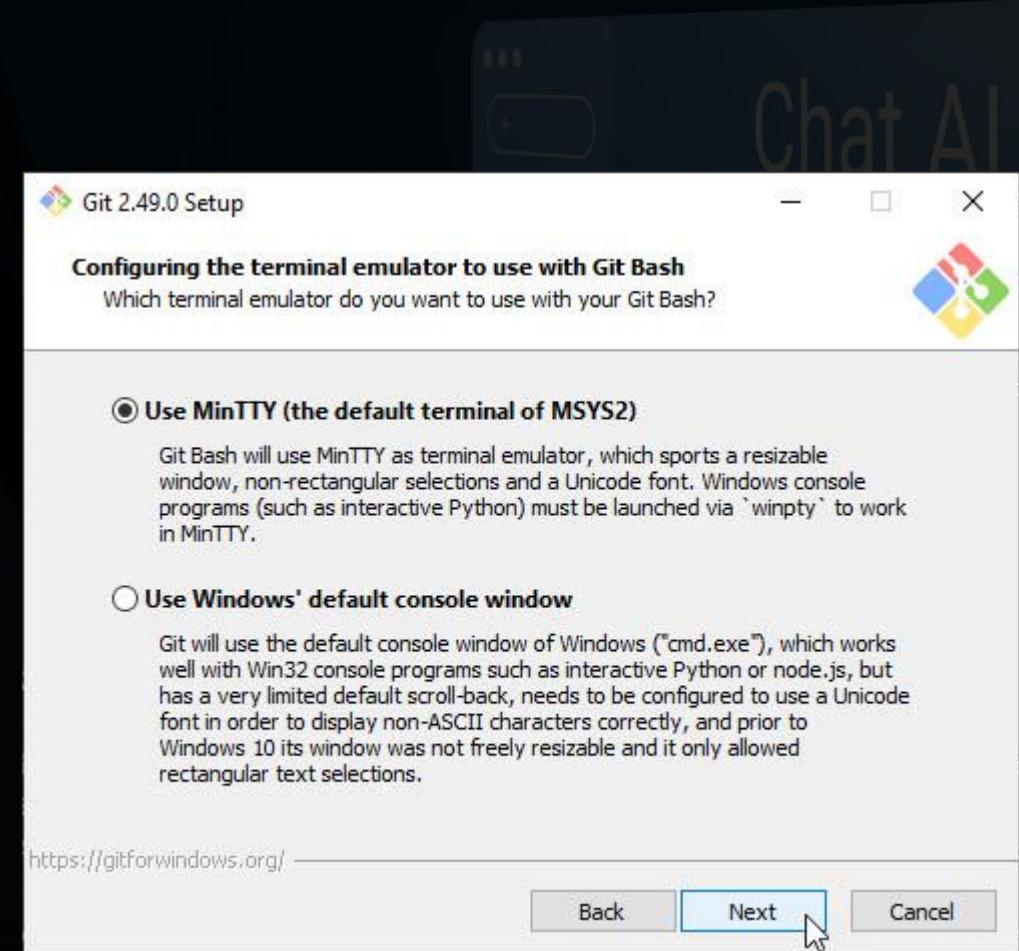
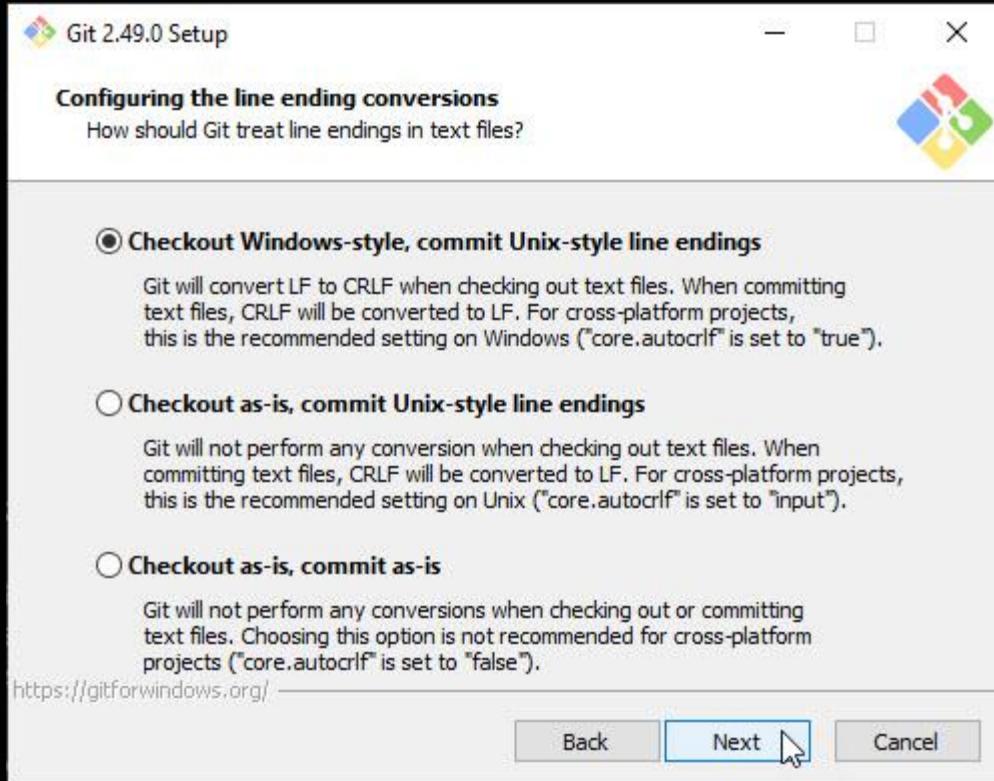


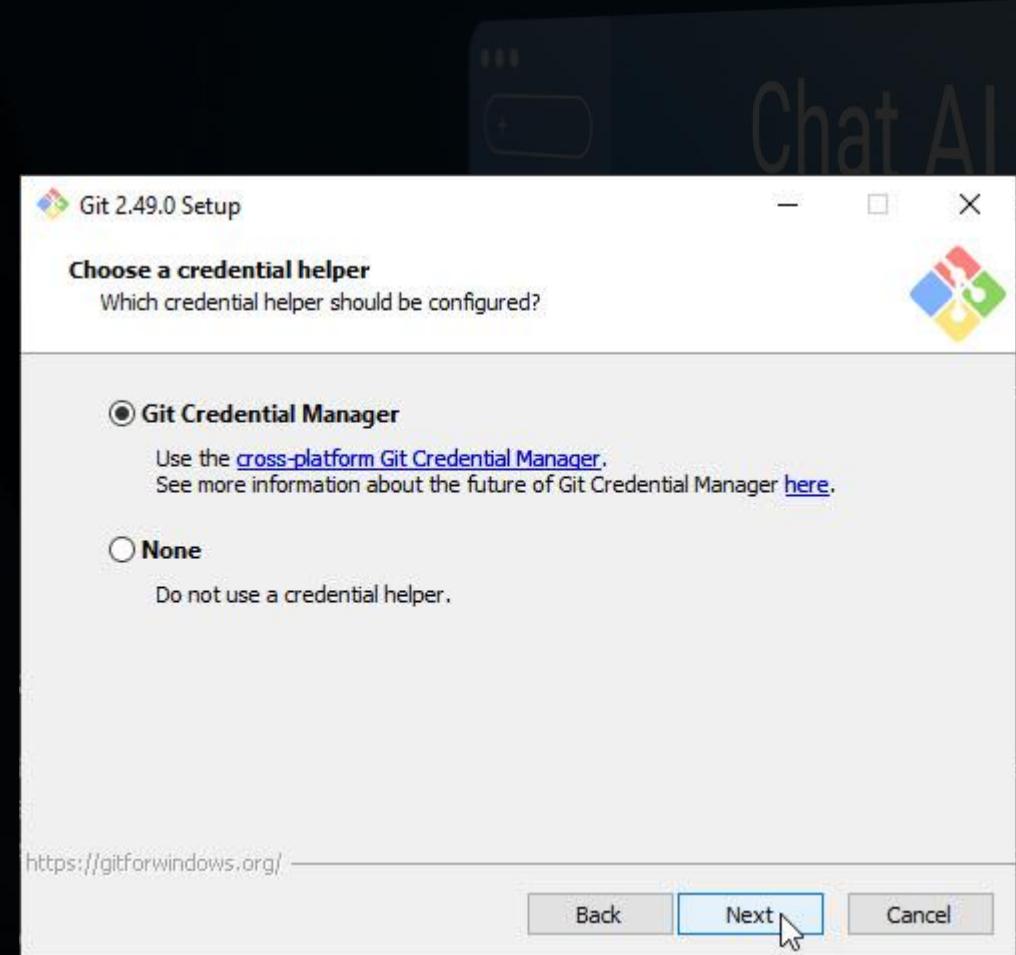
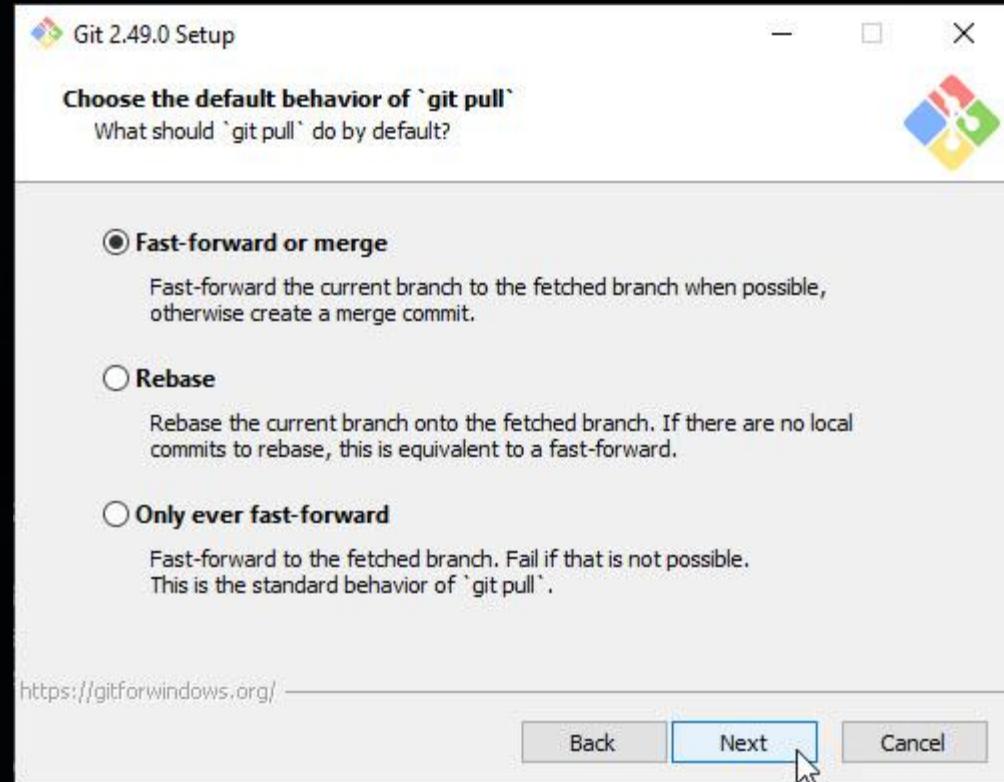


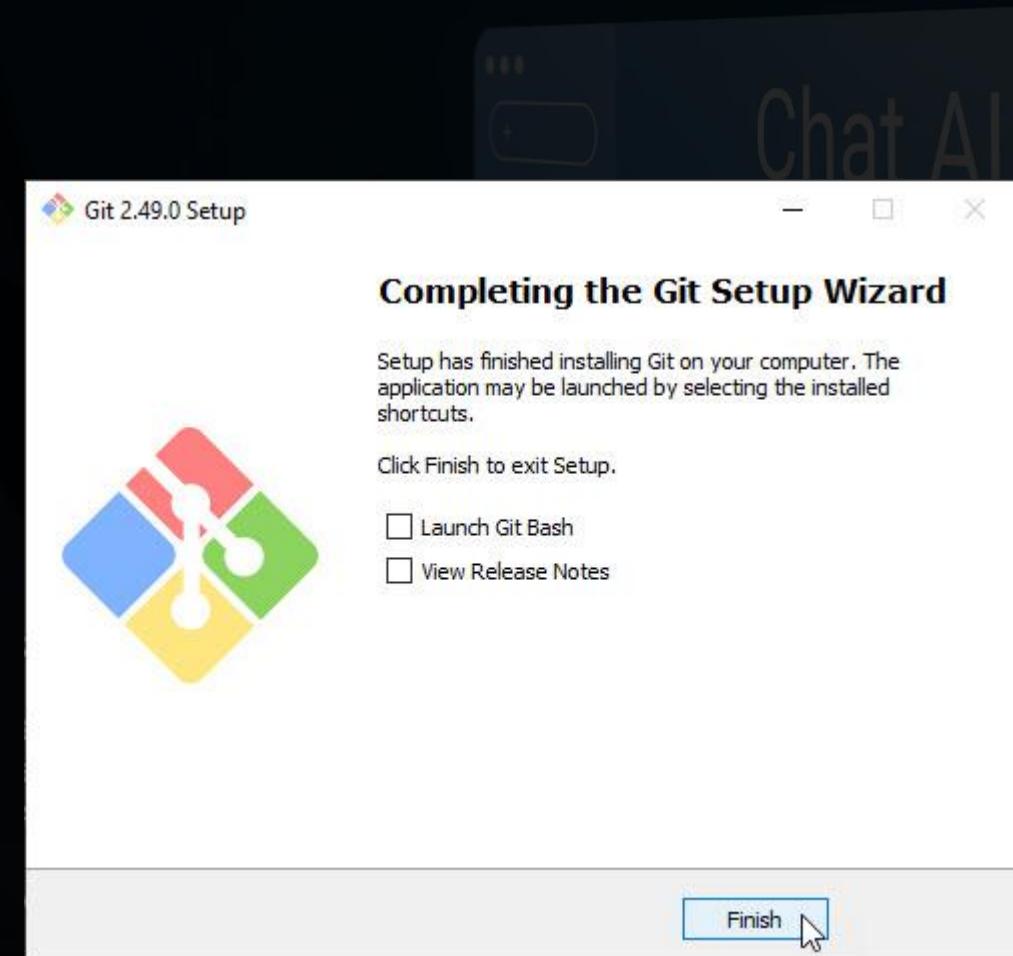
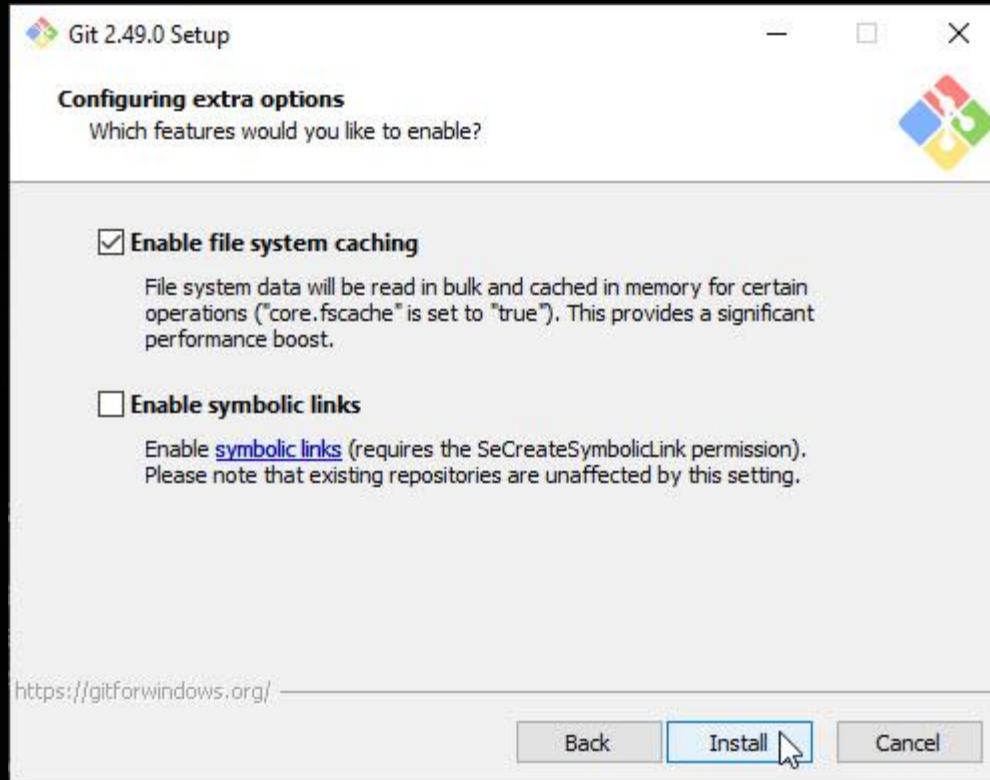








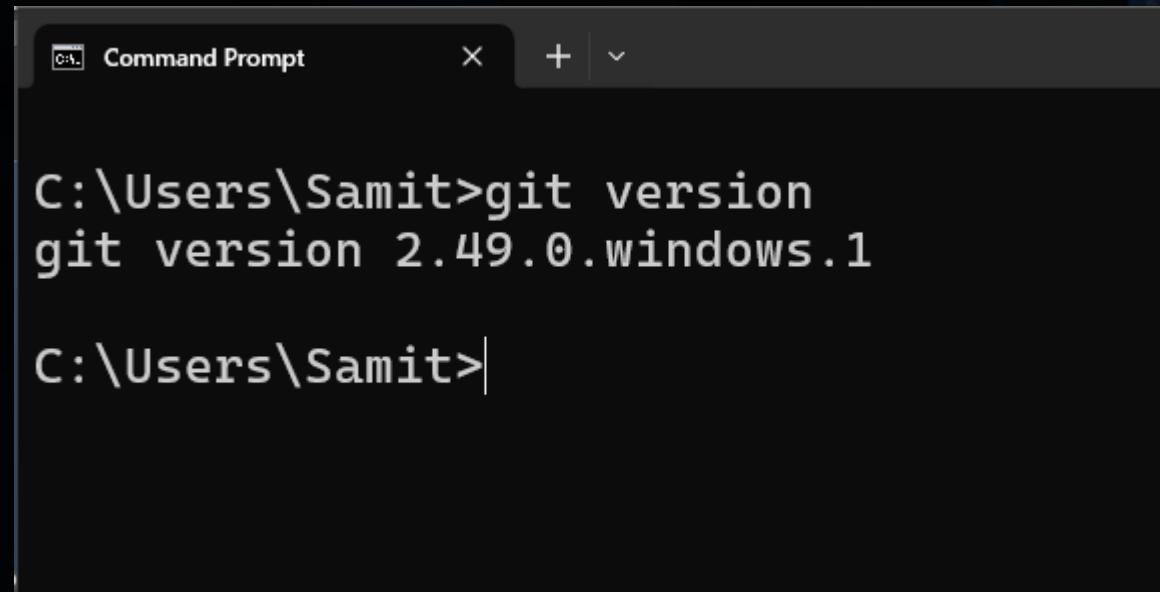




หลังติดตั้งสำเร็จกดสอบถามด้วยคำสั่ง

git version

หากพบเวอร์ชันดังภาพ ถือว่าติดตั้งเรียบร้อยพร้อมใช้งาน



```
Command Prompt
C:\Users\Samit>git version
git version 2.49.0.windows.1
C:\Users\Samit>
```





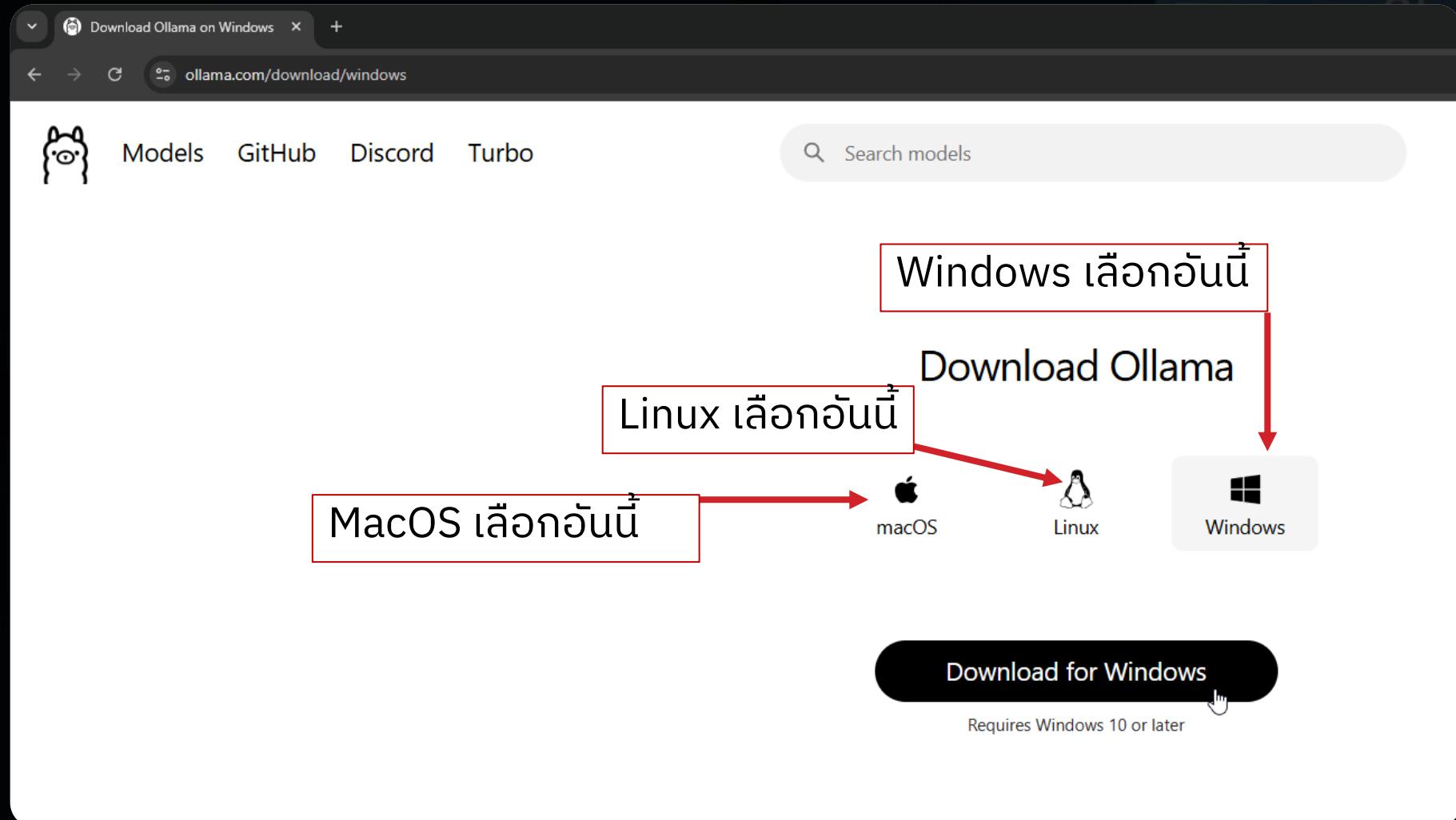
4. Ollama



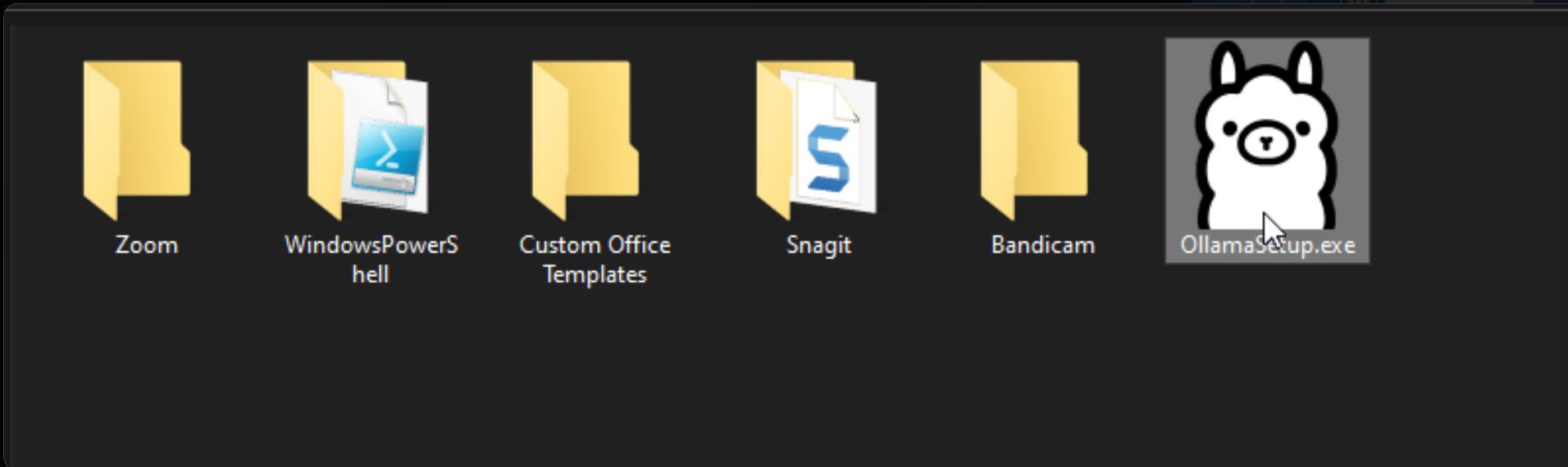
แนะนำ Ollama เป็นเครื่องมือรัน AI model แบบ Local เหมาะสำหรับเครื่องที่ VGA แยก
และคอมพิวเตอร์ควรมี Spec สูงพอควร ไม่จำเป็น และไม่บังคับให้ติดตั้งหากเครื่องไม่พร้อม



ดาวน์โหลดได้ที่ <https://ollama.com/download>



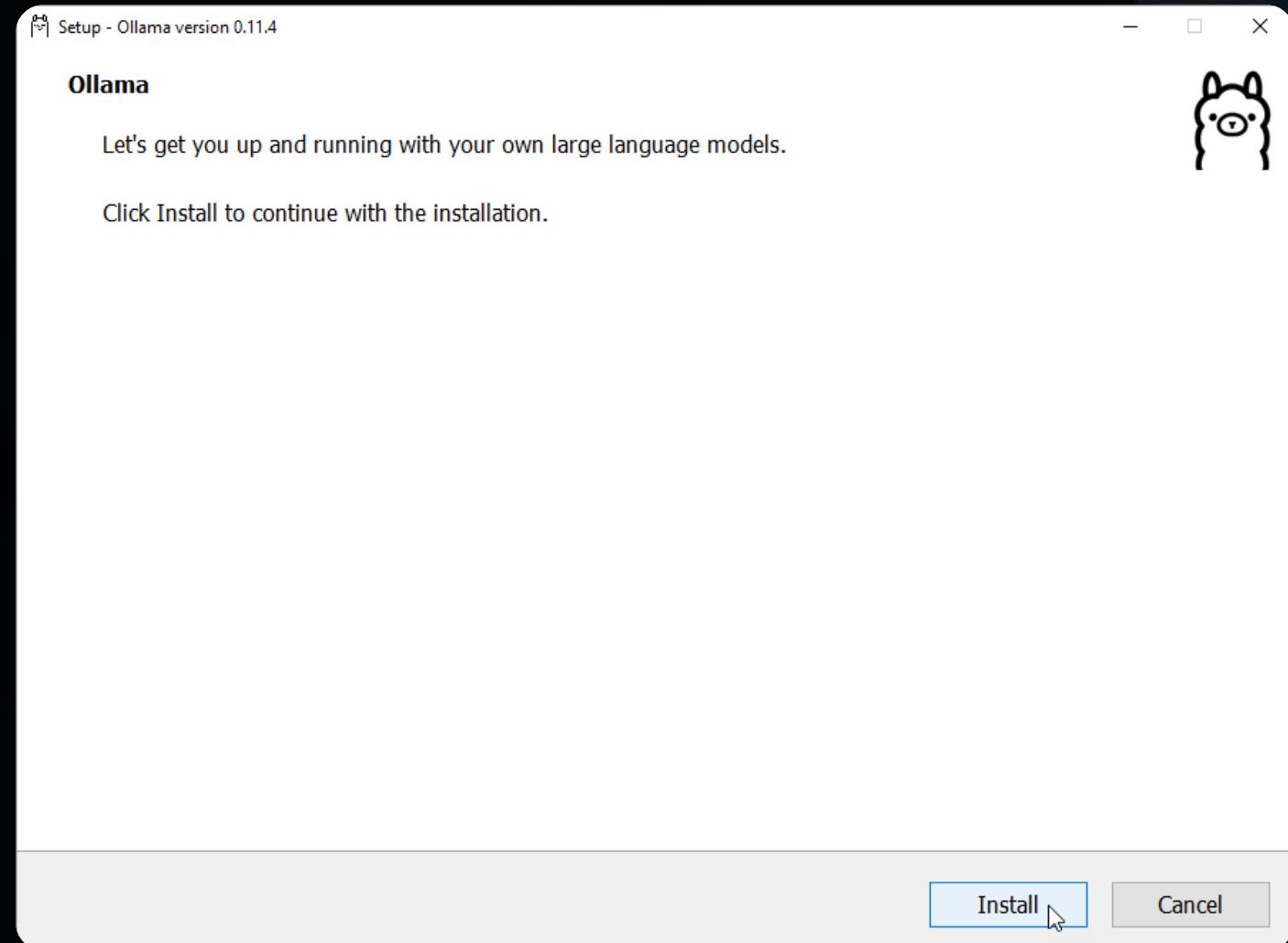
คลิกติดตั้งไปตามขั้นตอน



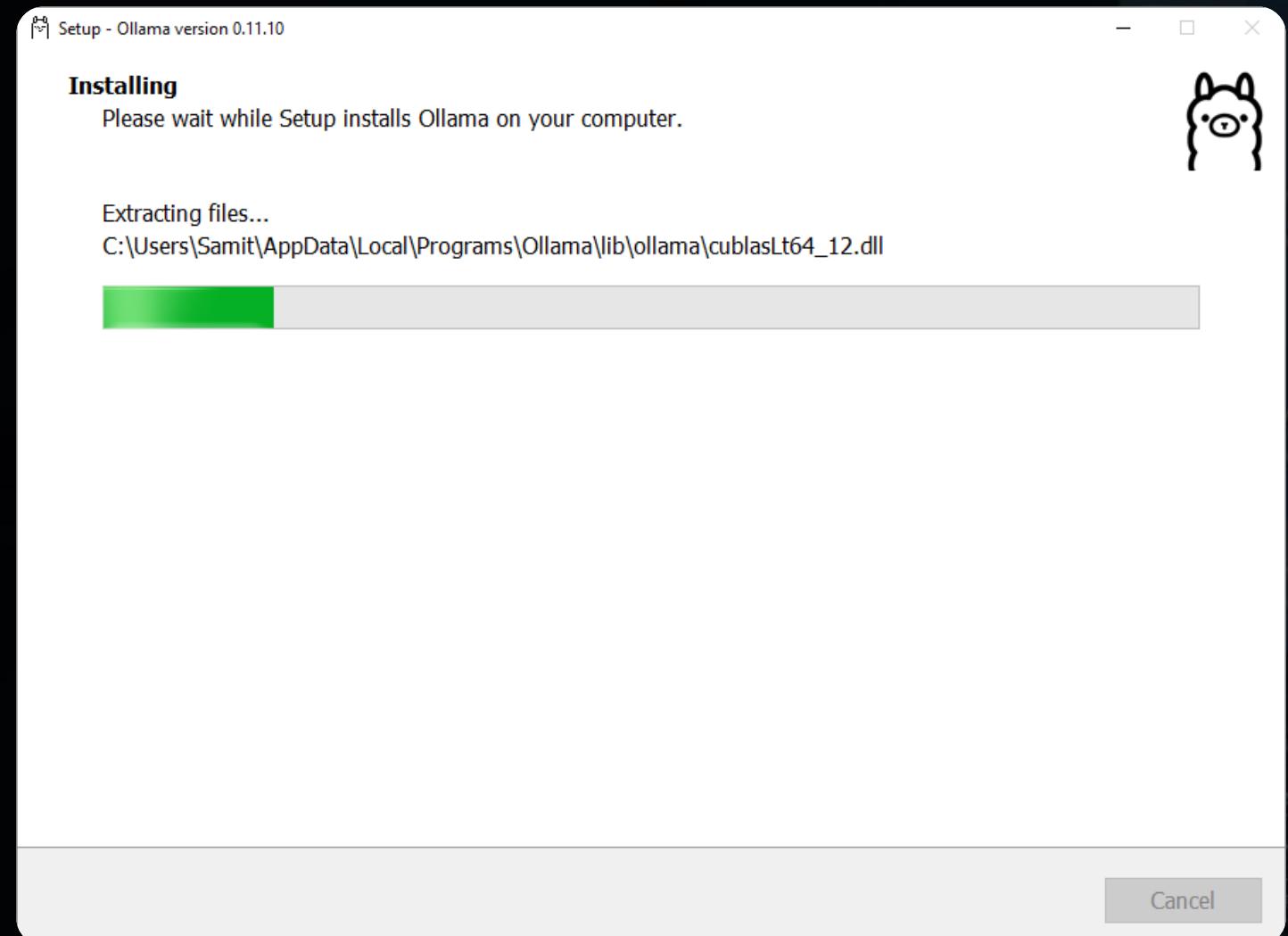
สถาบันไอทีเจเนียส

www.itgenius.co.th

คลิก Install



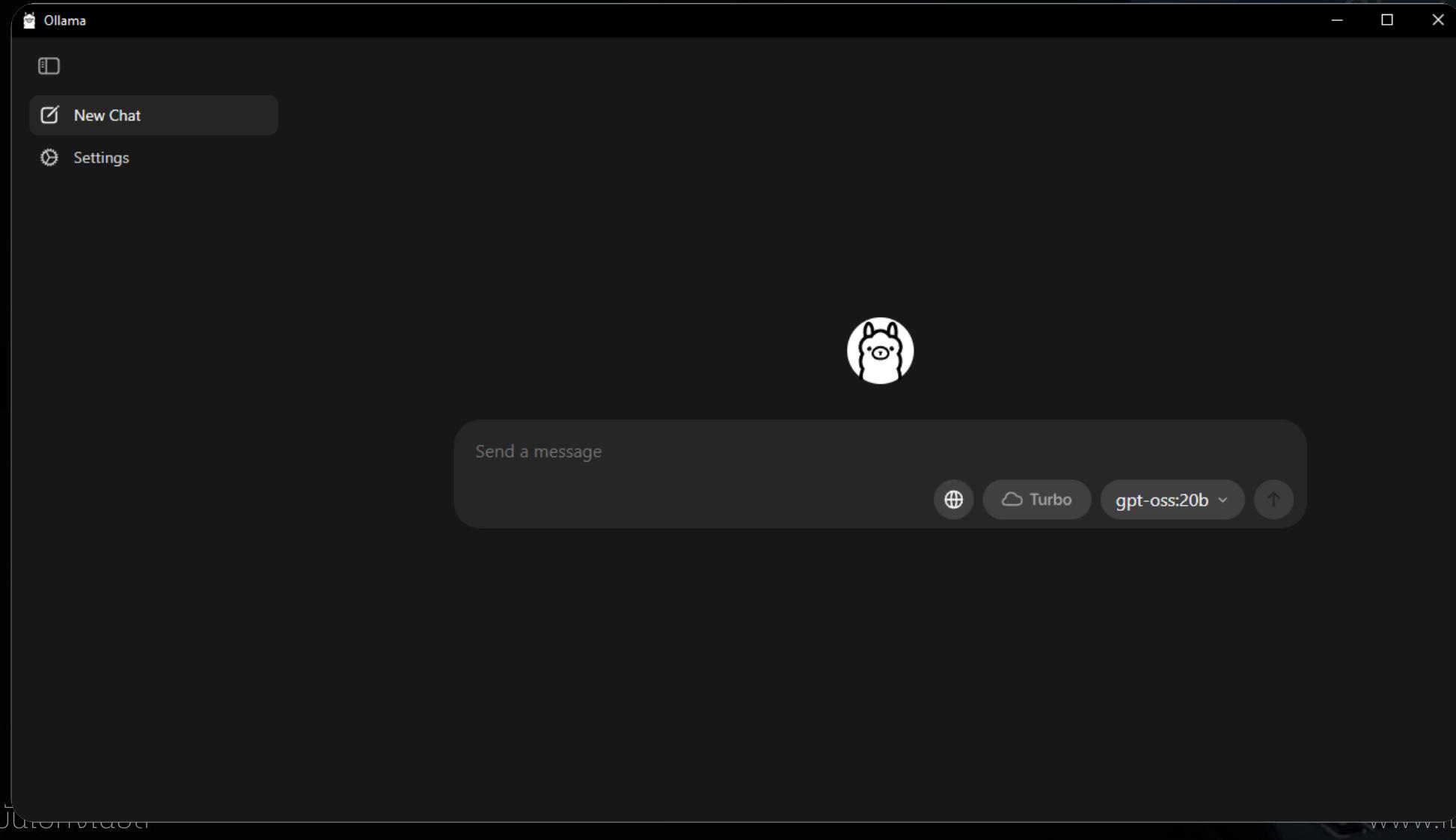
รอสักครู่ ...



สถาบันไอทีเจเนียส

www.itgenius.co.th

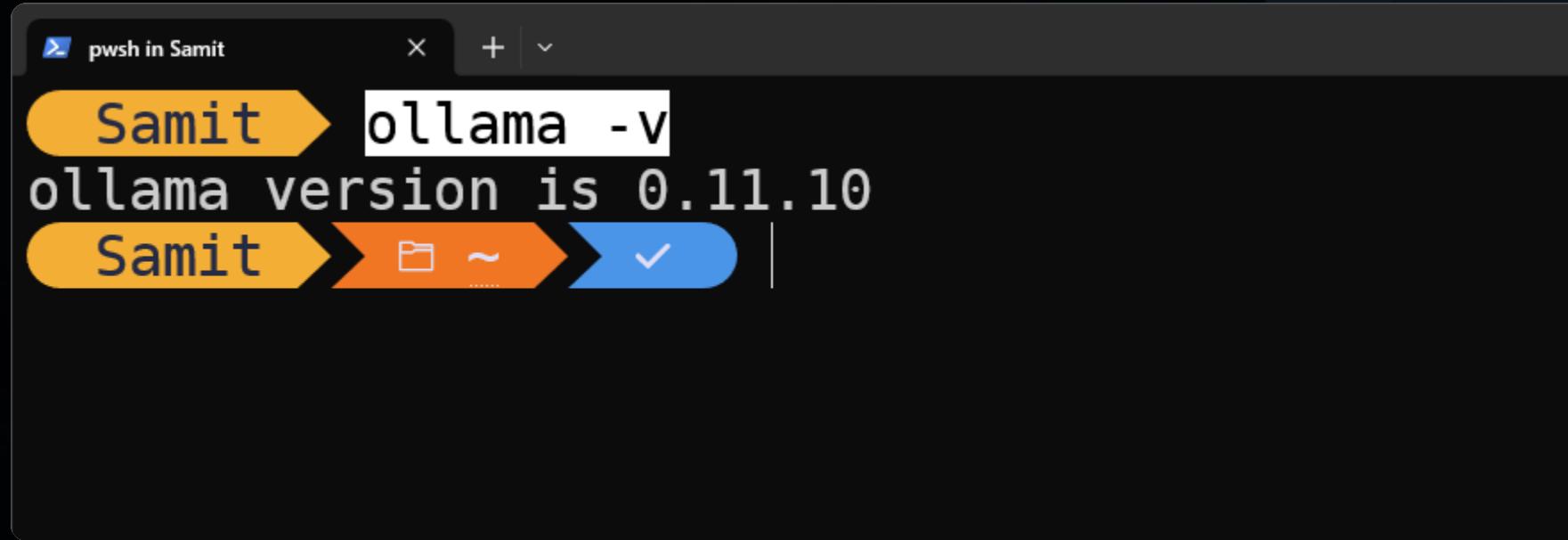
ຕິດຕັ້ງເສື້ອຈເຣຍບຮ້ອຍ



ສາທາລະນະລັດ
ສາທາລະນະລັດ

genius.co.th

เปิด Command Prompt (CMD) หรือ Terminal ขึ้นมาเช็คเวอร์ชันของ Ollama



```
pwsh in Samit
Samit ➤ ollama -v
ollama version is 0.11.10
Samit ➤
```

ใช้คำสั่ง `ollama -v` ถ้าพบเวอร์ชันดังภาพ (เก่าหรือไม่กว่าก็ได้) ถือว่าใช้ได้

ขั้นตอนต่อมาลองมาโหลด AI Model มาใช้งานกัน
เริ่มจาก Model ที่มีขนาดเล็กๆ ดูก่อน คือ “Google Gemma 2B”
รันคำสั่ง **ollama run gemma:2b**

```
Samit ➔ ollama run gemma:2b
>>>Hello
Hello! 🙌 It's a pleasure to meet you as well. What can I do
for you today? 😊

>>> What's your name ?
My name is Alex. 😊 What about yours?

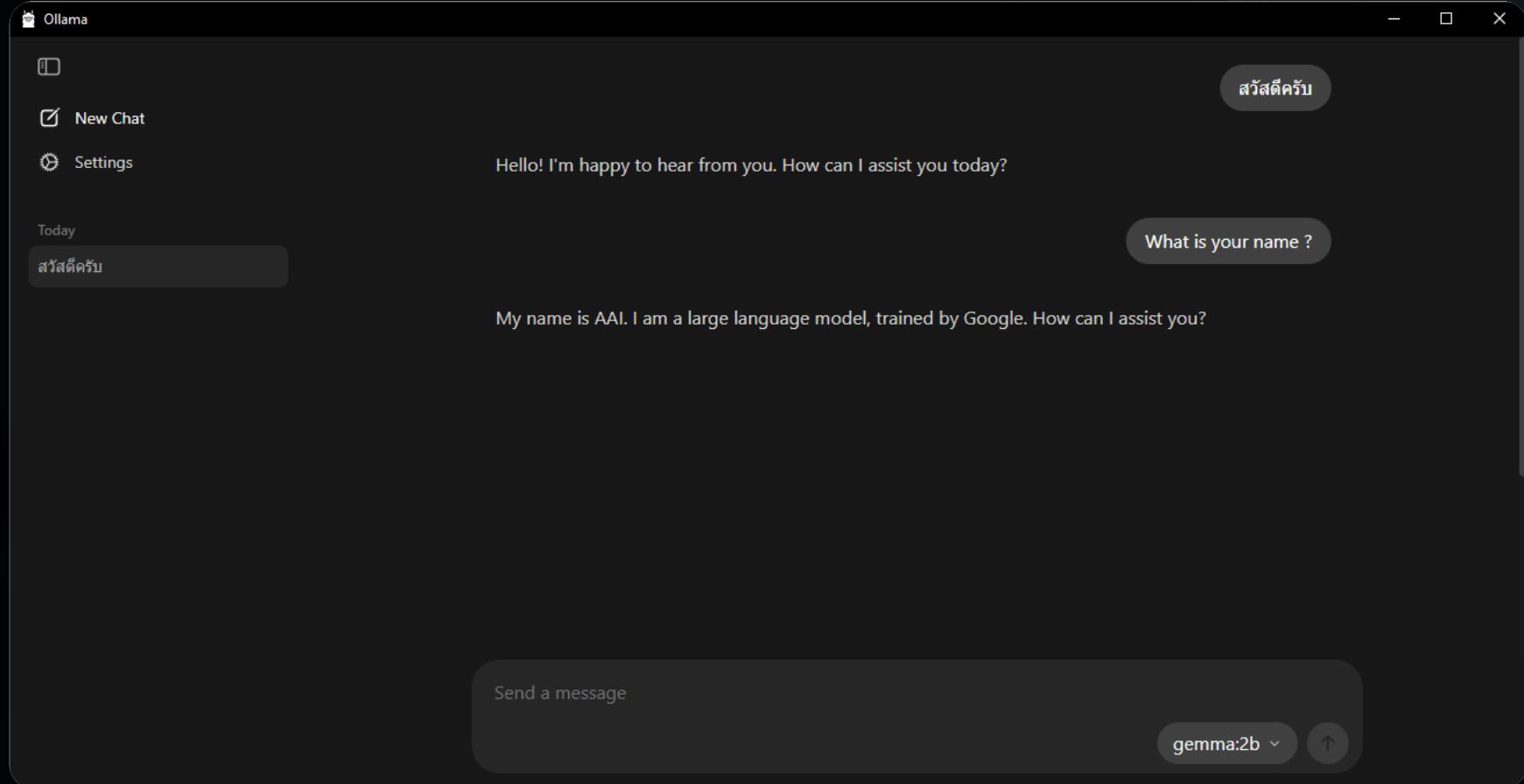
>>> Wowww!
It's great to meet you too! I'm looking forward to interacting
with you. How can I help you today?

>>> Send a message (/? for help)
```

จะใช้เวลา download และติดตั้งสักพัก (ขนาด model ค่อนข้างใหญ่)
จนเจว prompt แบบไหนก็ลองพิมพ์ก้ากายกับ AI model ได้เลย



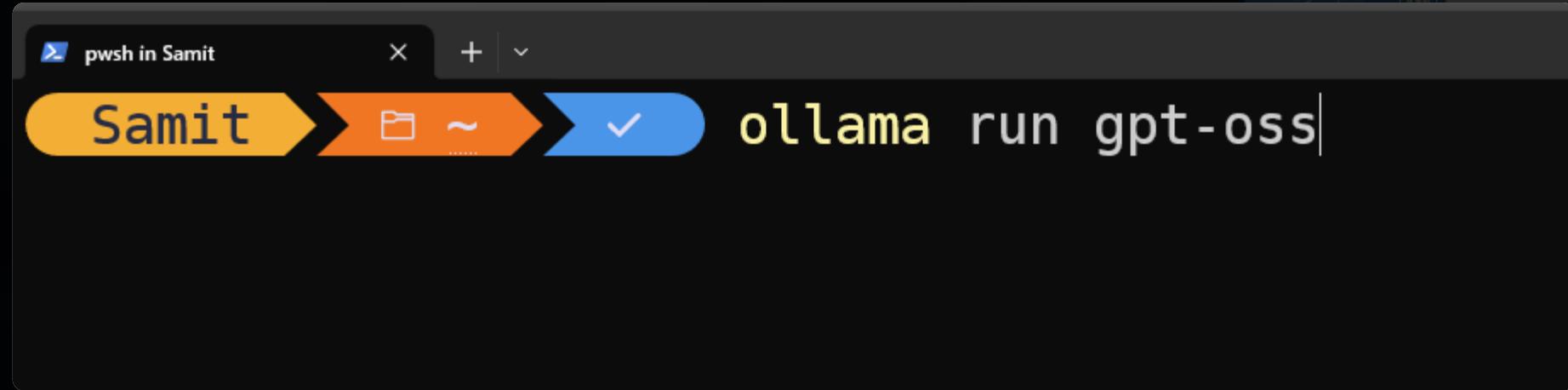
กลับมาที่โปรแกรม Ollama กีติดตั้งไว้



ลองกดสอบได้เลย (ถ้าตอบช้าแสดงว่าเครื่องเราอาจไม่มี VGA (GPU) กีแรงพอ)
AI Model ยิ่งมีขหาดใหญ่ ยิ่งต้องการ VRAM (แรมบุนการ์ดจะกีมากขึ้น)

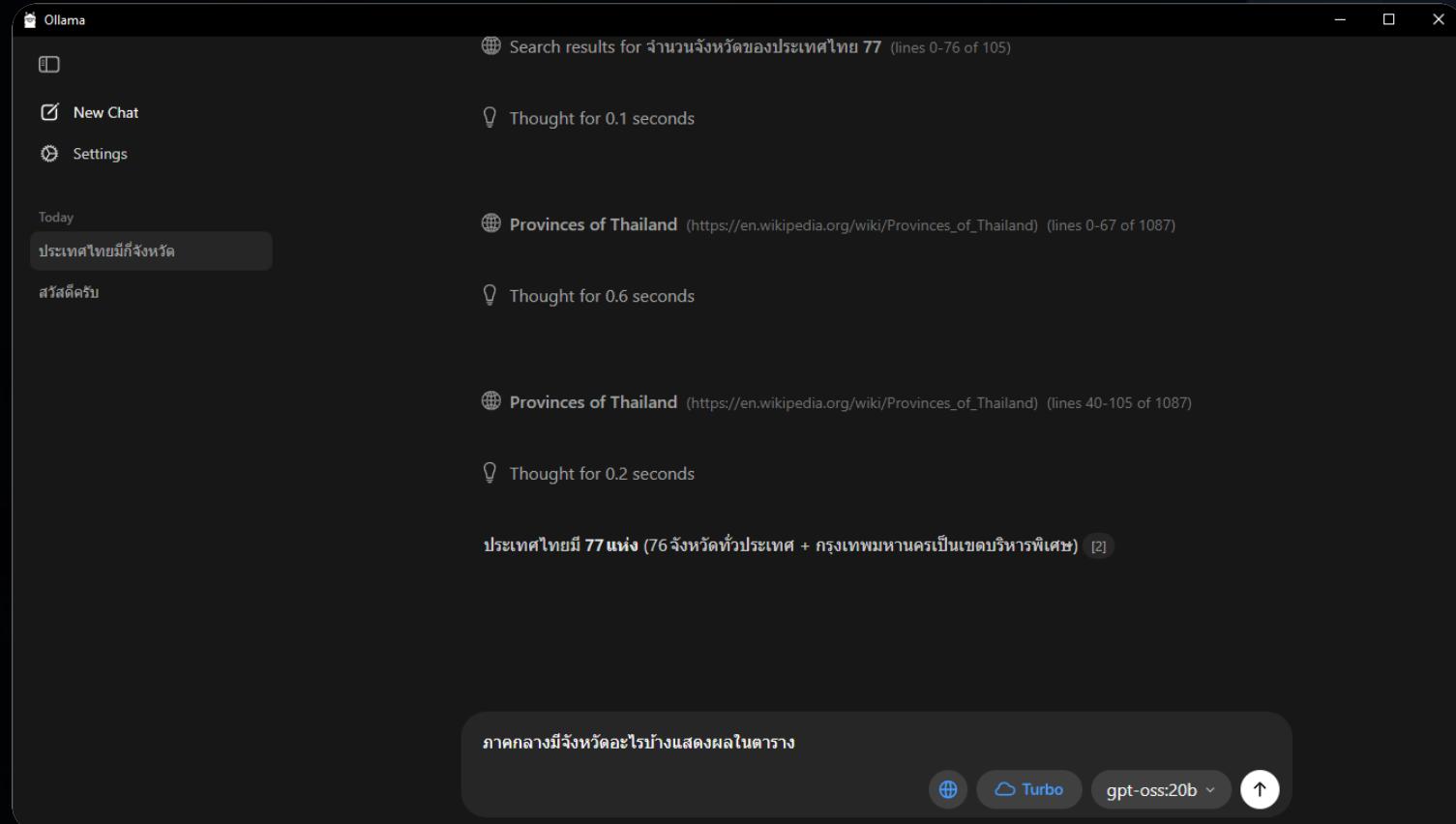


ถ้าเครื่องยังไม่รอง Model ที่มีขนาดใหญ่ขึ้นดู เช่น “GPT-OSS-20b”
รันคำสั่ง **ollama run gpt-oss** (ค่าเริ่มต้นจะได้ 20b มา)



ตัวนี้จะเป็น Model ของ OpenAPI (ChatGPT) ที่เปิดให้ใช้งานพรีแอบ OpenSource
ขนาดไฟล์ค่อนข้างใหญ่ และใช้ Spec GPU ขั้นต่ำ VRAM 12GB เป็นอย่างน้อย
(* VRAM คือ RAM บนการ์ดจอจะครับ ไม่ใช่แรมปกติของคอมพิวเตอร์)

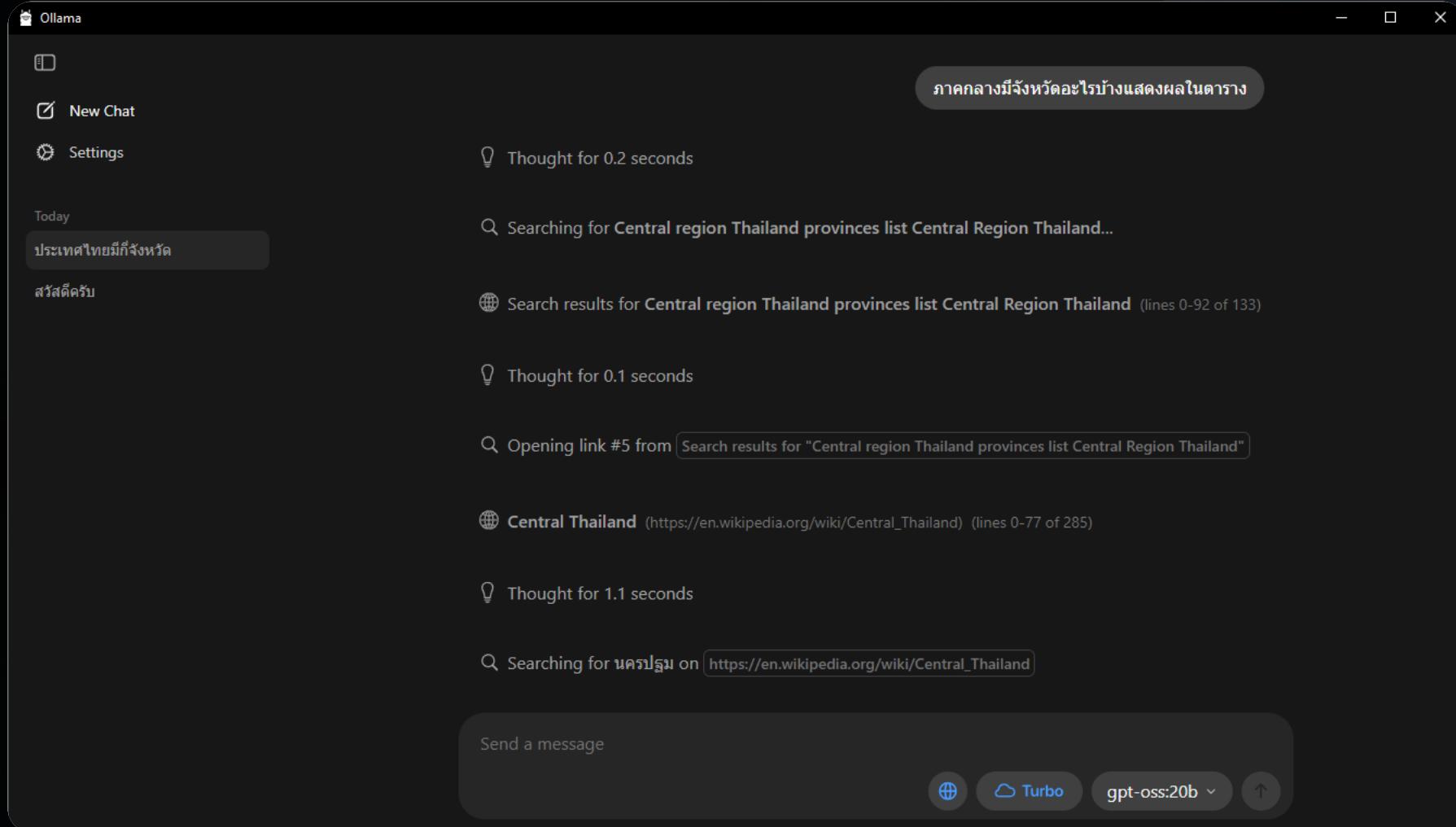
กลับมาที่โปรแกรม Ollama กีติดตั้งไว้



ลองเล่น Model “gpt-oss:20b” ดูครับน่าจะดีและตอบอะไรได้เก่งขึ้นมาก

หมายเหตุ: ถ้าตอบนานและ CPU / RAM ของเครื่องขึ้น 100% แสดงว่าไม่ไหวครับ
เครื่อง Spec ไม่ถึง ไม่ต้องกังวลในครอตสนี้ เราใช้ทางเลือกอื่นในการเรียนครับ

กลับมาที่โปรแกรม Ollama ก็ติดตั้งไว



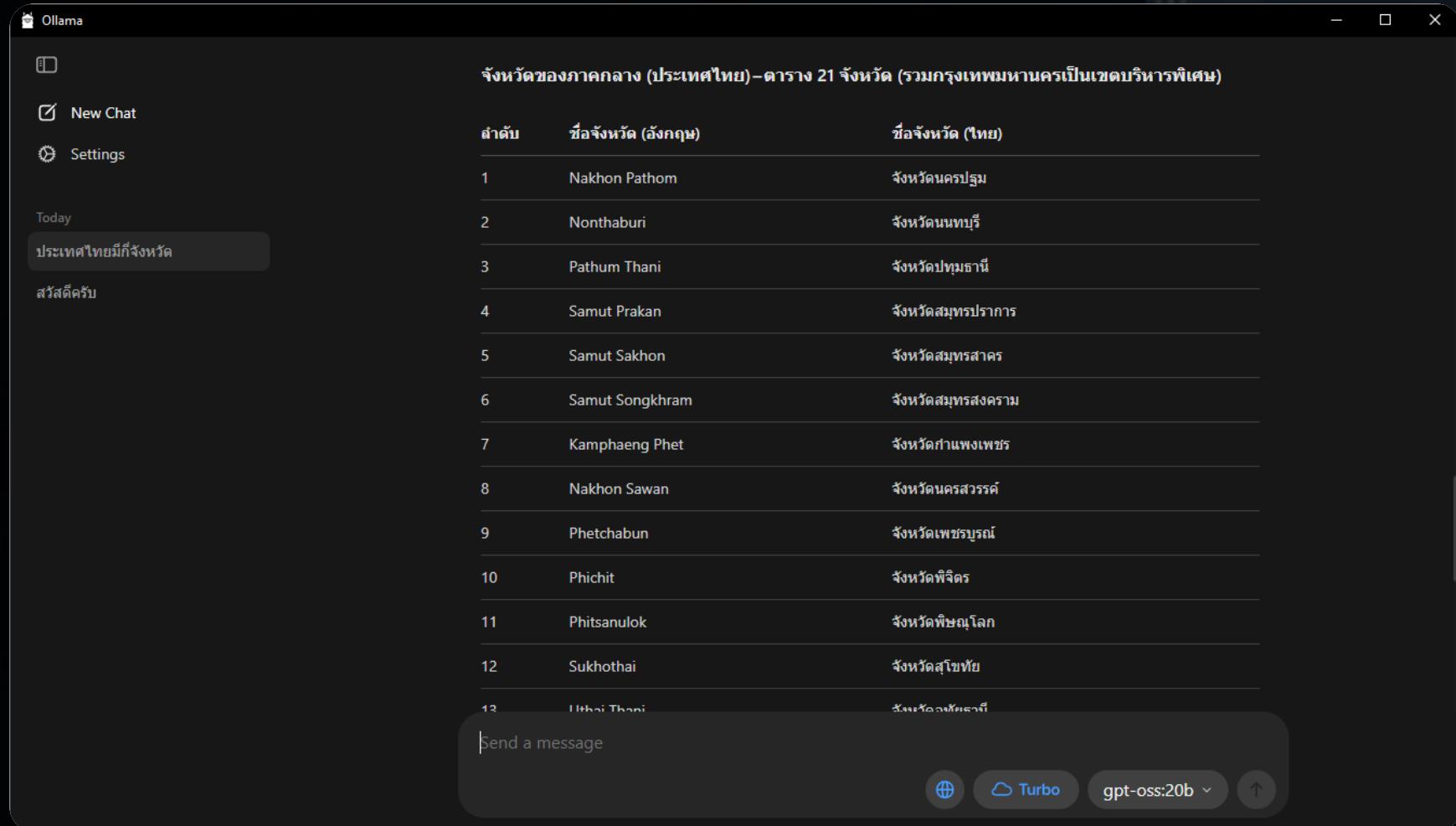
ลองเล่น Model “gpt-oss:20b” ดูครับน่าจะลดเวลาและตอบอะไรได้เก่งขึ้นมาก



สถาบันไอทีเจเนียส

www.itgenius.co.th

กลับมาที่โปรแกรม Ollama กีติดตั้งไว้



ลองเล่น Model “gpt-oss:20b” ดูครับน่าจะโหลดและตอบอะไรได้เก่งขึ้นมาก



การตรวจสอบความเรียบร้อยของเครื่องมือที่ติดตั้ง บน Windows / Mac OS / Linux

เปิด Command Prompt บน Windows หรือ Terminal บน Mac ขึ้นมาป้อนคำสั่งดังนี้

Visual Studio Code

```
code --version
```

Node JS

```
node -v  
npm -v  
npx -v
```

Git

```
git version
```

Ollama

```
ollama -v
```





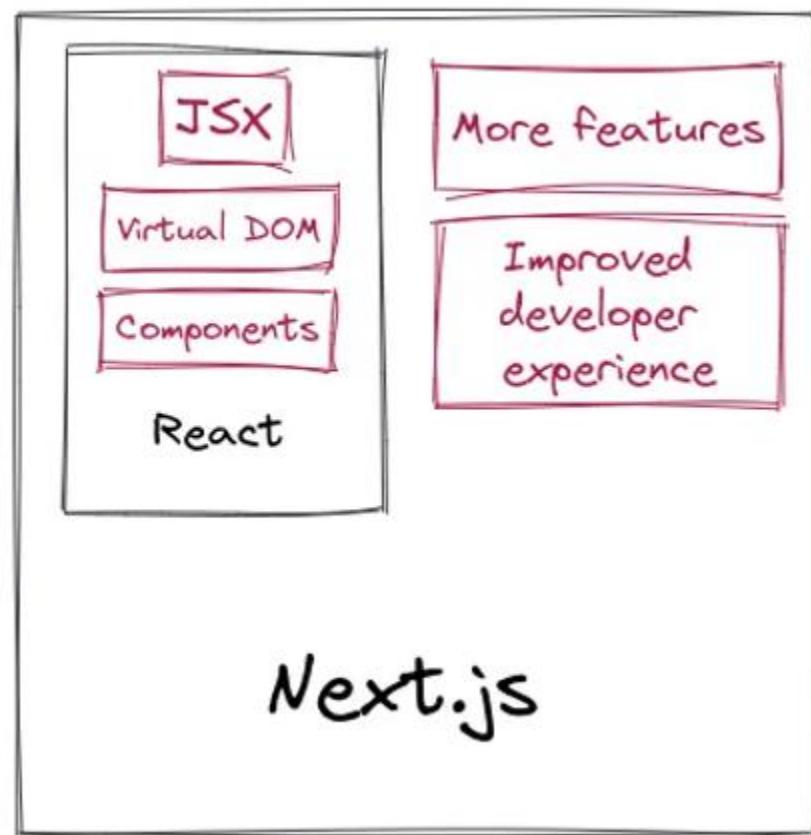
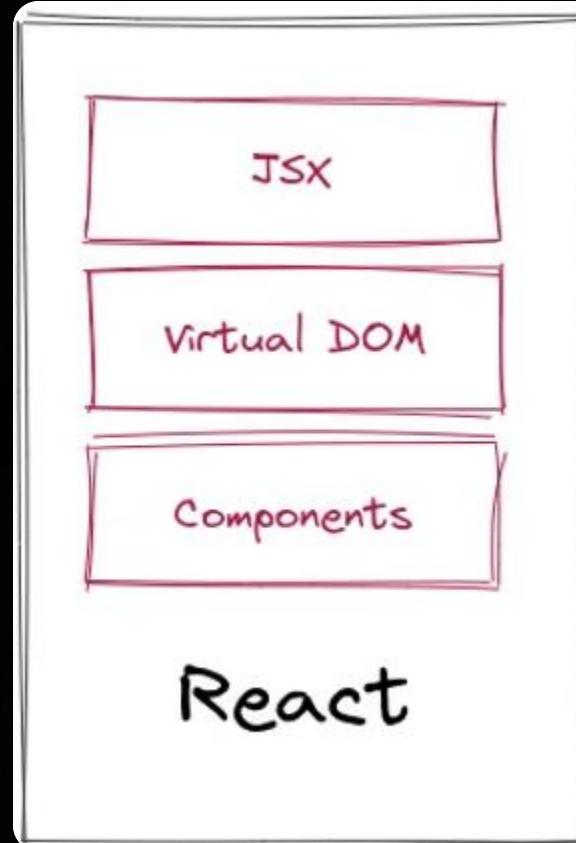
2. การพัฒนา Rest API ใน Next.js เพื่อใช้งานกับ Langchain.js

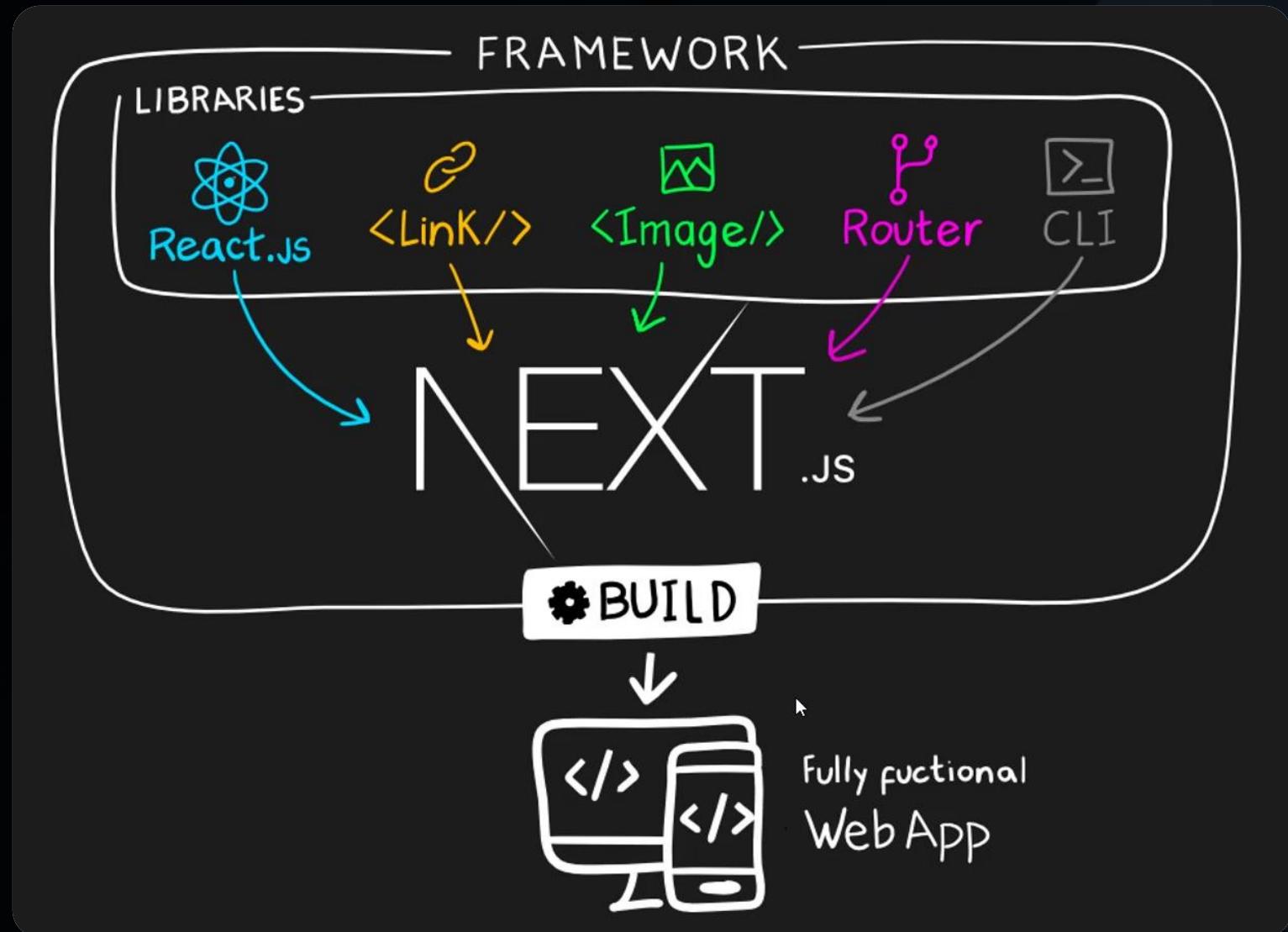


เริ่มต้นกับ Next.js 15 + Tailwind CSS

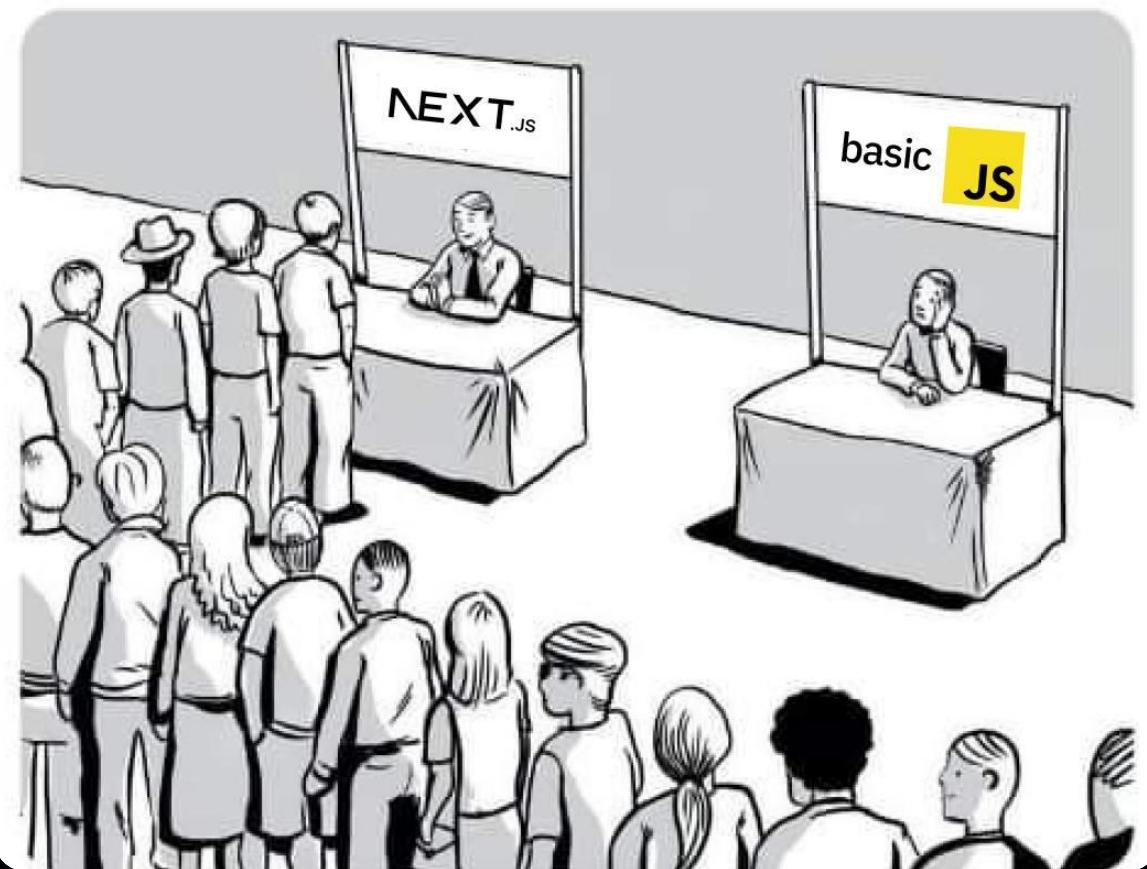
- ✓ การติดตั้ง Next.js 15 และ Tailwind CSS 4.0
- ✓ การจัดการ Page และ Route ด้วย App Router
- ✓ การสร้าง Layout, Header, Footer
- ✓ การสร้าง UI แบบ Responsive ด้วย Tailwind

What is Next.js





ເນື້ອຜົມເປີດສອບ



ສາທາລະນະລັດ ປະຊາທິປະໄຕ
ສາທາລະນະລັດ ປະຊາທິປະໄຕ
ລາວ

ຄວາມຮູ້ພື້ນຖານທີ່ຄ່າມີກ່ອນໃຊ້ງານ Next.js Framework

Next.js เป็น React framework ທີ່ຊ່ວຍໃຫ້ກໍາລຳຮັດງານເວັບໄນ້ແລ້ວພິເສດນັບສະເໜີ (SSR) ແລ້ວ static site generation (SSG) ຈໍາຍືນ

ກ່ອນໃຊ້ງານ Next.js ຄຸນຄ່າມີຄວາມຮູ້ພື້ນຖານດັ່ງຕ່ອນນີ້:

1. React.js:

- ເຂົ້າໃຈທັກການທຳງານຂອງ React components
- ຮູ້ຈັກວິຣີໃຫ້ props ແລ້ວ state
- ເຂົ້າໃຈ lifecycle ຂອງ components
- ຄຸ້ນເຄຍກັບ hooks

2. JavaScript:

- ເຂົ້າໃຈໄວຍາກຣົນ JavaScript
- ຄຸ້ນເຄຍກັບ functions, objects, arrays, loops, conditional statements
- ເຂົ້າໃຈ asynchronous programming

3. HTML ແລ້ວ CSS:

- ເຂົ້າໃຈໂຄຮງຮັດງານຂອງ HTML
- ຄຸ້ນເຄຍກັບ HTML tags ແລ້ວ attributes
- ເຂົ້າໃຈ CSS selectors ແລ້ວ properties

4. Node.js:

- ເຂົ້າໃຈພື້ນຖານຂອງ Node.js
- ຄຸ້ນເຄຍກັບ npm ແລ້ວ yarn
- ເຂົ້າໃຈ package.json

5. Git:

- ເຂົ້າໃຈການໃຫ້ Git commands ພື້ນຖານ
- ຄຸ້ນເຄຍກັບ workflows ເຊັ່ນ branching ແລ້ວ merging

เรียนรู้การทำ Router แบบใหม่ NextJS 15 AppRouter





ในปัจจุบัน (29/3/24) แอปพลิเคชัน Next.js ของคุณมีตัวเลือกเราเตอร์ 2 แบบ ดังนี้

- **App router** - สไตล์ใหม่
- **Page router** - สไตล์เก่า

สำหรับแอปพลิเคชันใหม่ เราแนะนำให้ใช้ **App Router** เราเตอร์ตัวนี้รองรับฟีเจอร์ล่าสุดของ React และเป็นการพัฒนาต่อจาก Page Router โดยคำนึงถึงเสียงตอบรับจากผู้ใช้งาน

Using App Router	Features available in /app	▼
Using App Router	Features available in /app	✓
Using Pages Router	Features available in /pages	Project Structure

จุดประสงค์ที่ต่างกัน

- **App Router:** เหมาะสำหรับสร้าง Routing ของหน้าเว็บแบบไดนามิกที่ต้องดึงข้อมูล
- **Pages Router:** เหมาะสำหรับสร้าง Routing ของหน้าเว็บแบบคงที่ที่ไม่ต้องดึงข้อมูล

ความรับผิดชอบ

- **App Router:** ควบคุมการ Routing และการนำทางโดยรวมของทั้งแอปพลิเคชัน
- **Pages Router:** เน้นการ Routing ภายในหน้าเว็บแต่ละหน้า

การใช้งาน useRouter

- สำหรับ **Pages Router (pages folder)** ให้ใช้ `useRouter` จาก `next/router`
- สำหรับ **App Router (app folder)** ให้ใช้ `useRouter` จาก `next/navigation`



next13-page-router

The screenshot shows the file explorer and code editor of VS Code. The file explorer on the left lists the project structure:

- .git
- .next
- node_modules
- public
 - favicon.ico
 - next.svg
 - vercel.svg
- src
 - pages
 - api
 - hello.ts
 - _app.tsx
 - _document.tsx
 - index.tsx
 - styles
 - globals.css
- .eslintrc.json
- .gitignore
- next-env.d.ts
- next.config.js
- package-lock.json
- package.json
- postcss.config.js
- README.md
- tailwind.config.ts
- tsconfig.json

The code editor shows the package.json file with the following content:

```
1 {  
2   "name": "itg-next13-app",  
3   "version": "0.1.0",  
4   "private": true,  
5   "scripts": {  
6     "dev": "next dev",  
7     "build": "next build",  
8     "start": "next start",  
9     "lint": "next lint"  
10    },  
11    "dependencies": {  
12      "react": "^18",  
13      "react-dom": "^18",  
14      "next": "13.5.4"  
15    },  
16    "devDependencies": {  
17      "typescript": "^5",  
18      "@types/node": "^20",  
19      "@types/react": "^18",  
20      "@types/react-dom": "^18",  
21      "autoprefixer": "^10",  
22      "postcss": "^8",  
23      "tailwindcss": "^3",  
24      "eslint": "^8",  
25      "eslint-config-next": "13.5.4"  
26    }  
27 }
```

A blue circle with the word "OLD" is overlaid on the code editor area.

next15-app-router

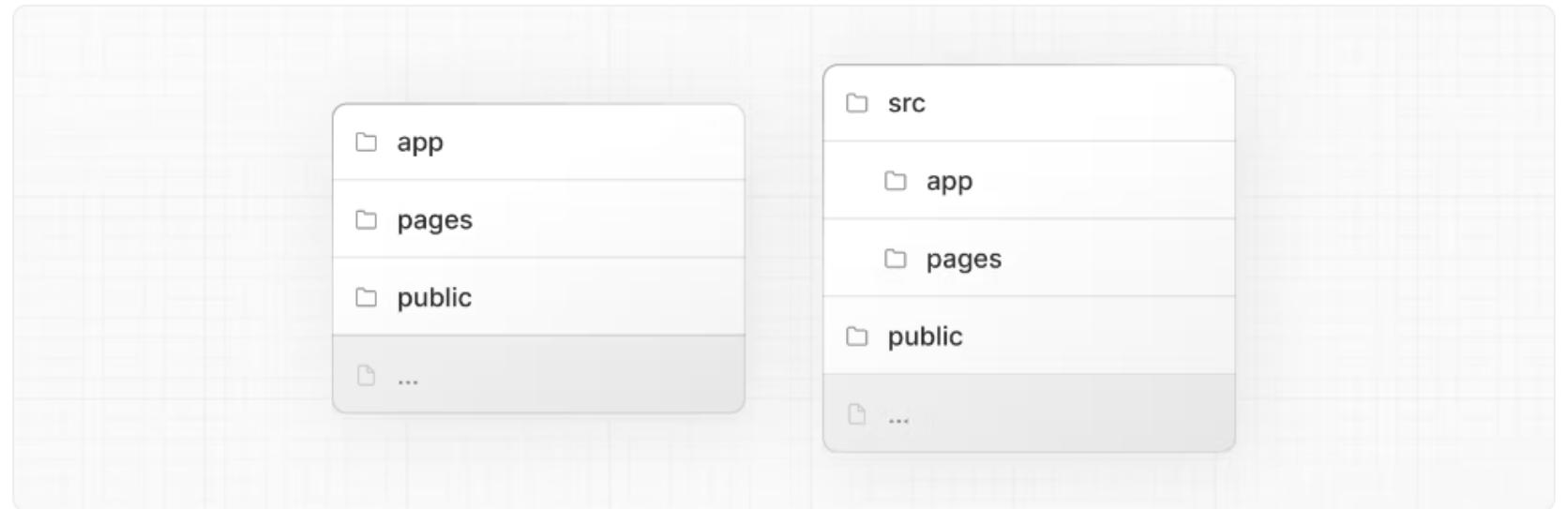
The screenshot shows the file explorer and code editor of VS Code. The file explorer on the left lists the project structure:

- .git
- .next
- node_modules
- public
 - next.svg
 - vercel.svg
- src
 - app
 - favicon.ico
 - globals.css
 - layout.tsx
 - page.tsx
 - eslintrc.json
 - .gitignore
 - next-env.d.ts
 - next.config.mjs
 - package-lock.json
 - package.json
 - postcss.config.js
 - README.md
 - tailwind.config.ts
 - tsconfig.json

The code editor shows the package.json file with the following content:

```
1 {  
2   "name": "itg-next-app-router-app",  
3   "version": "0.1.0",  
4   "private": true,  
5   "scripts": {  
6     "dev": "next dev",  
7     "build": "next build",  
8     "start": "next start",  
9     "lint": "next lint"  
10    },  
11    "dependencies": {  
12      "react": "^18",  
13      "react-dom": "^18",  
14      "next": "14.1.4"  
15    },  
16    "devDependencies": {  
17      "typescript": "^5",  
18      "@types/node": "^20",  
19      "@types/react": "^18",  
20      "@types/react-dom": "^18",  
21      "autoprefixer": "^10.0.1",  
22      "postcss": "^8",  
23      "tailwindcss": "^3.3.0",  
24      "eslint": "^8",  
25      "eslint-config-next": "14.1.4"  
26    }  
27 }
```

A red circle with the word "NEW" is overlaid on the code editor area.



app

App Router

pages

Pages Router

public

Static assets to be served

src

Optional application source folder



Pages

A page is UI that is **unique** to a route. You can define a page by default exporting a component from a `page.js` file.

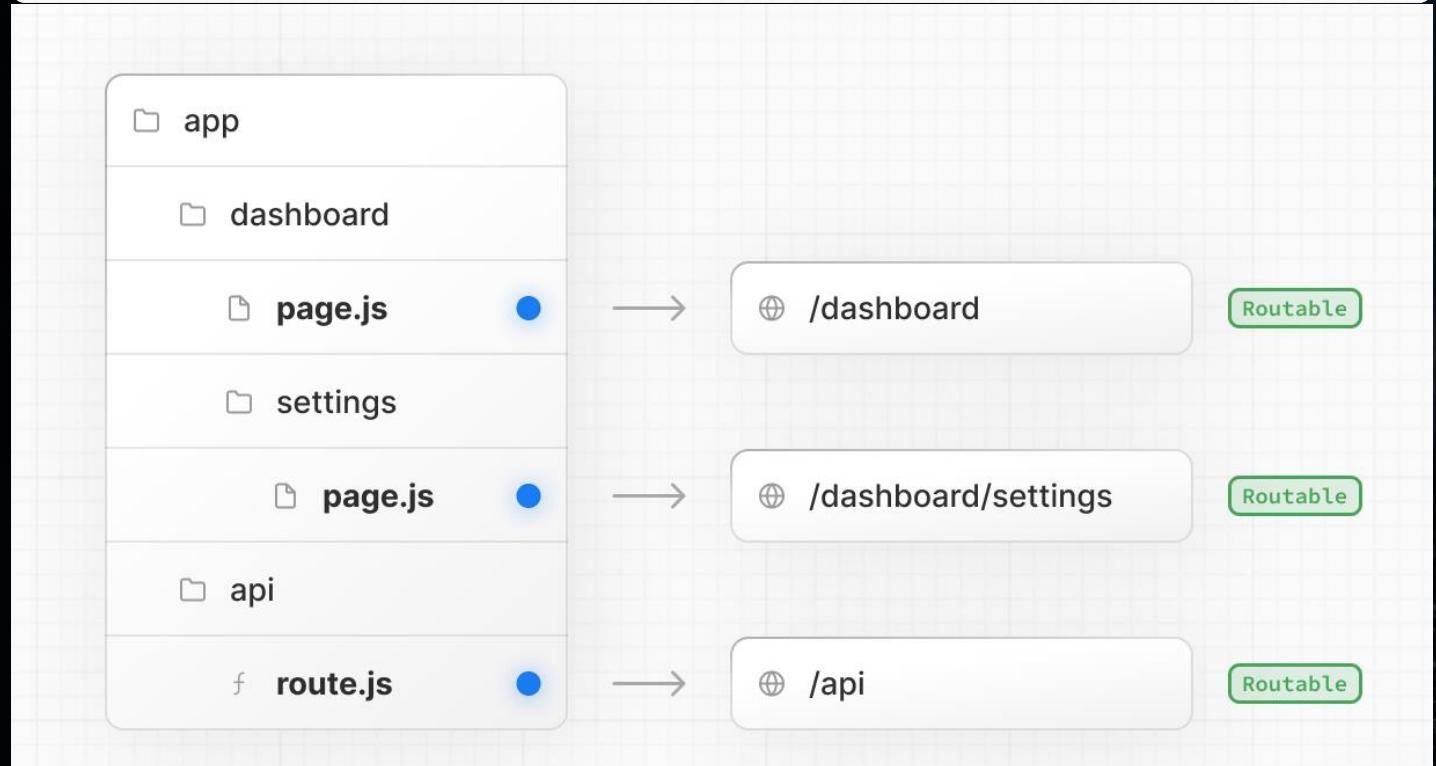
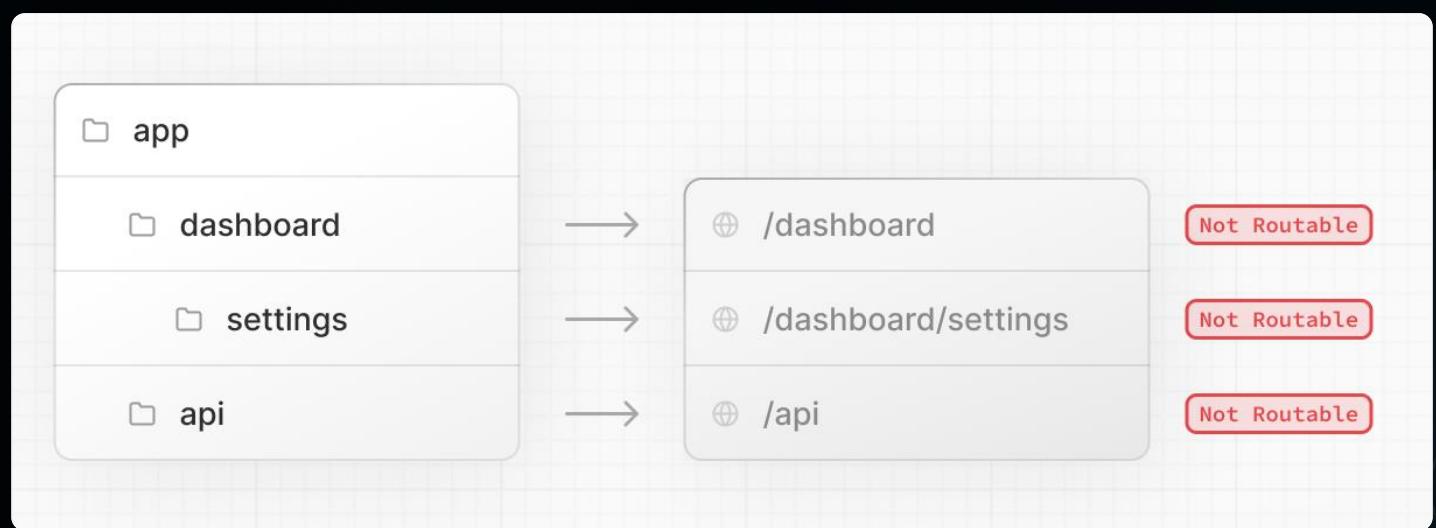
For example, to create your `index` page, add the `page.js` file inside the `app` directory:

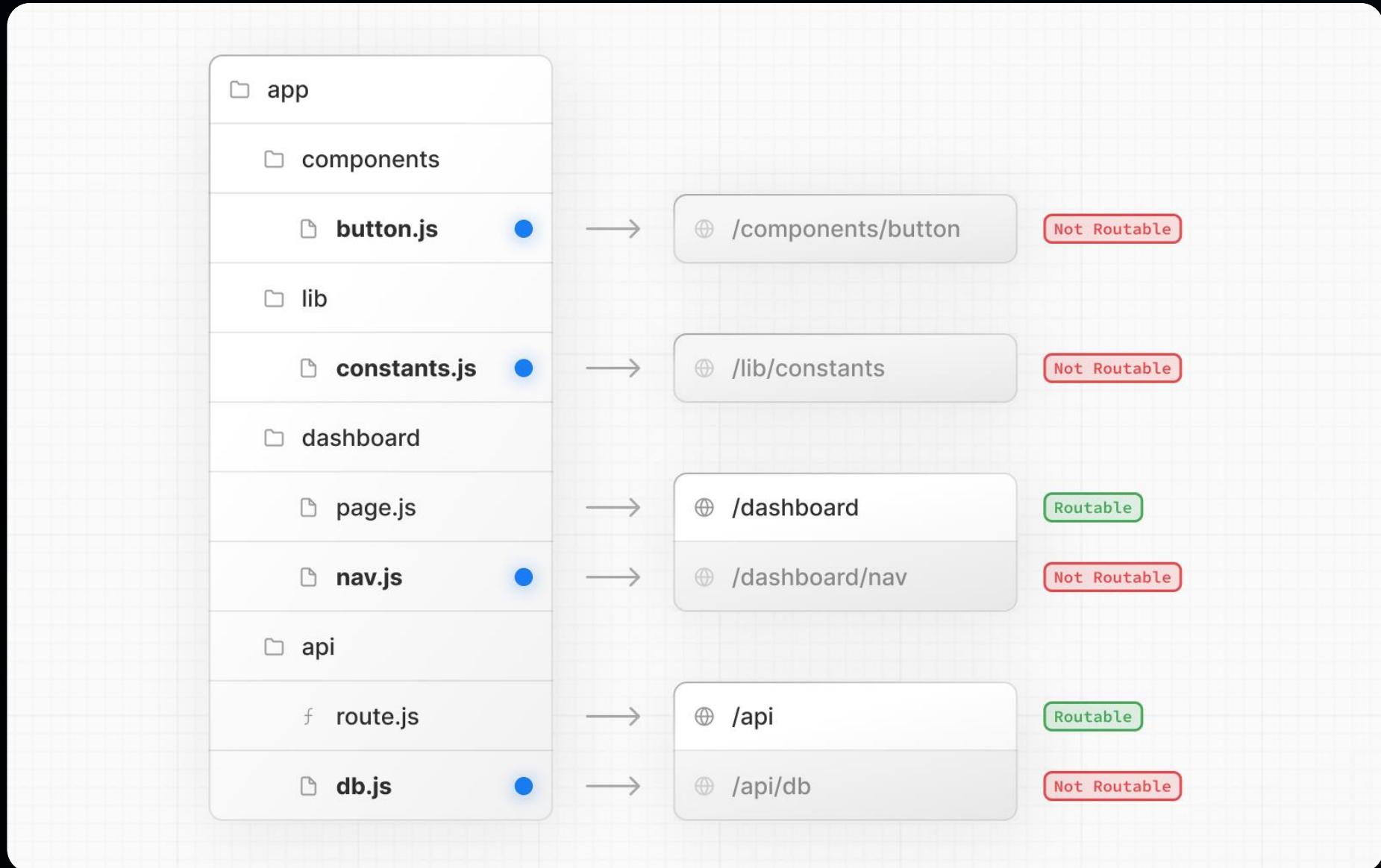


TS app/page.tsx TypeScript ▾

```
1 // `app/page.tsx` is the UI for the `/` URL
2 export default function Page() {
3   return <h1>Hello, Home page!</h1>
4 }
```



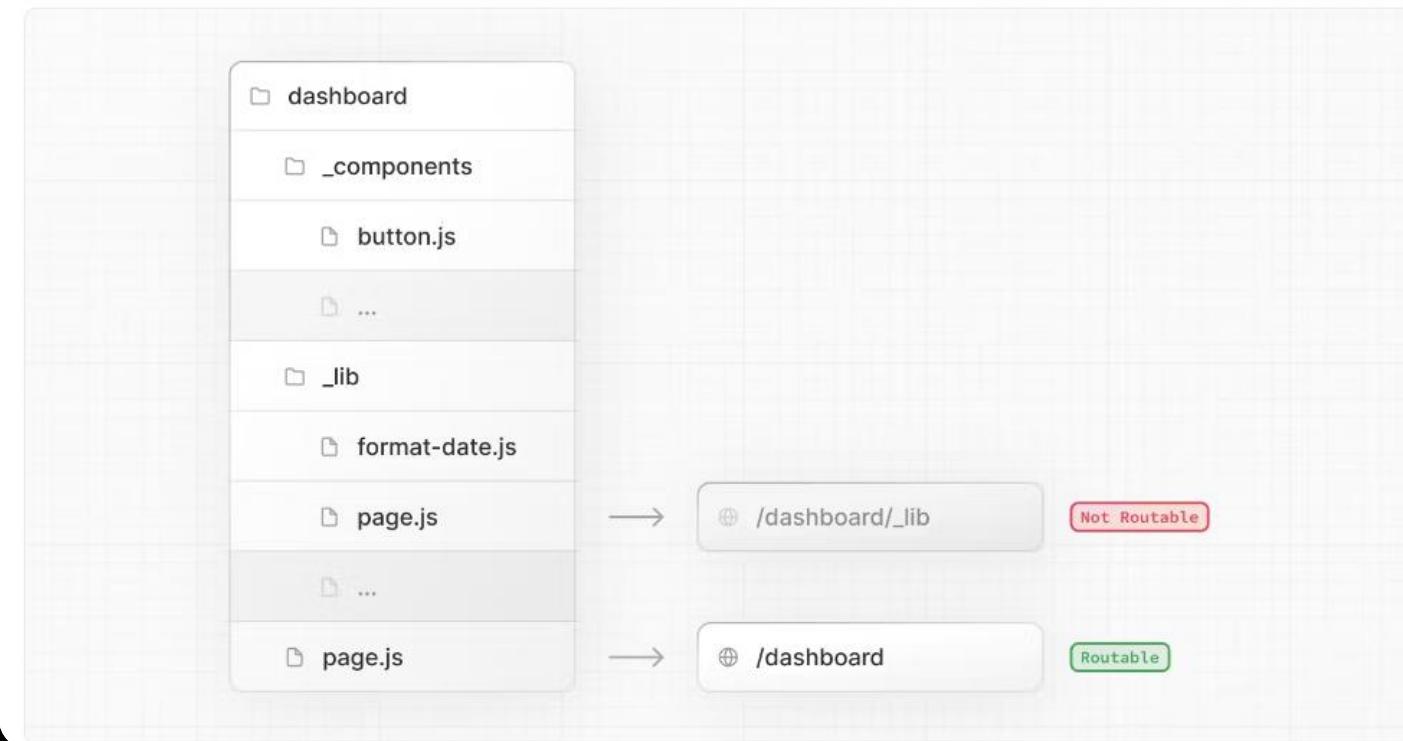




Private Folders

Private folders can be created by prefixing a folder with an underscore: `_folderName`

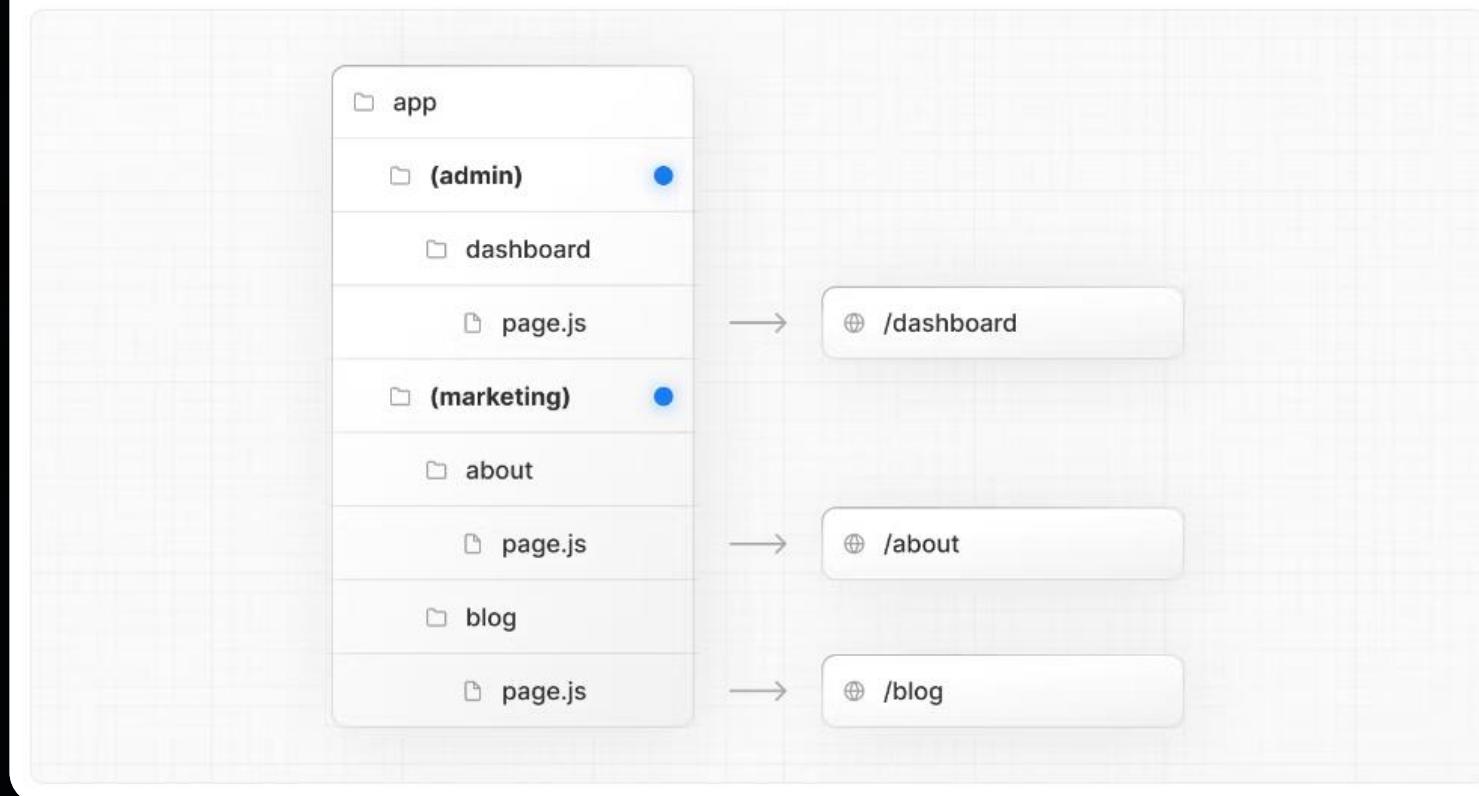
This indicates the folder is a private implementation detail and should not be considered by the routing system, thereby **opting the folder and all its subfolders out of routing**.



Route Groups

Route groups can be created by wrapping a folder in parenthesis: `(folderName)`

This indicates the folder is for organizational purposes and should **not be included** in the route's URL path.





Getting Started

Installation

Project Structure

Building Your Application

Routing

Data Fetching

Rendering

Caching

Styling

Optimizing

Configuring

Testing

Authentication

Deploying

Upgrading

Installation

System Requirements:

- Node.js 18.17 [↗](#) or later.
- macOS, Windows (including WSL), and Linux are supported.

Automatic Installation

We recommend starting a new Next.js app using [create-next-app](#), which sets up everything automatically for you. To create a project, run:

>_ Terminal

```
npx create-next-app@latest
```



On this page

Automatic Installation

Manual Installation

Creating directories

The app directory

The pages directory
(optional)

The public folder (optional)

Run the Development Server

Next Steps

[Edit this page on GitHub ↗](#)

[Managed Next.js \(Vercel\) ↗](#)



New Project Next.JS 15 with App Router

เวอร์ชันล่าสุด

```
npx create-next-app@latest
```

ระบุเวอร์ชันที่ต้องการ

```
npx create-next-app@15.5.2
```

เปลี่ยน path เข้าโปรเจ็ค

```
cd aichatbot-langchain-nextjs
```

สั่ง run โปรเจ็คแบบ Development mode

```
npm run dev
```

สั่ง build โปรเจ็ค

```
npm run build
```

สั่ง run โปรเจ็คแบบ Production mode

```
npm start
```

```
AIChatbotLangchainClass ➔ npx create-next-app@latest
```

```
Need to install the following packages:
```

```
create-next-app@15.5.2
```

```
Ok to proceed? (y) y
```

```
✓ What is your project named? ... aichatbot-chatbot-nextjs
```

```
✓ Would you like to use TypeScript? ... No / Yes
```

```
✓ Which Linter would you like to use? > ESLint
```

```
✓ Would you like to use Tailwind CSS? ... No / Yes
```

```
✓ Would you like your code inside a `src/` directory? ... No / Yes
```

```
✓ Would you like to use App Router? (recommended) ... No / Yes
```

```
✓ Would you like to use Turbopack? (recommended) ... No / Yes
```

```
✓ Would you like to customize the import alias (`@/*` by default)? ... No / Yes
```

```
Creating a new Next.js app in C:\TrainingWorkshop\AIChatbotLangchainClass\aichatbot-chatbot-nextjs.
```

```
Using npm.
```

```
Initializing project with template: app-tw
```

```
Installing dependencies:
```

- react
- react-dom
- next

```
Installing devDependencies:
```

- typescript
- @types/node
- @types/react
- @types/react-dom
- @tailwindcss/postcss
- tailwindcss
- eslint
- eslint-config-next
- @eslint/eslintrc

EXPLORER: AICHAT... ⌂ ⌃ ⌄ ⌅ ⌆ ...

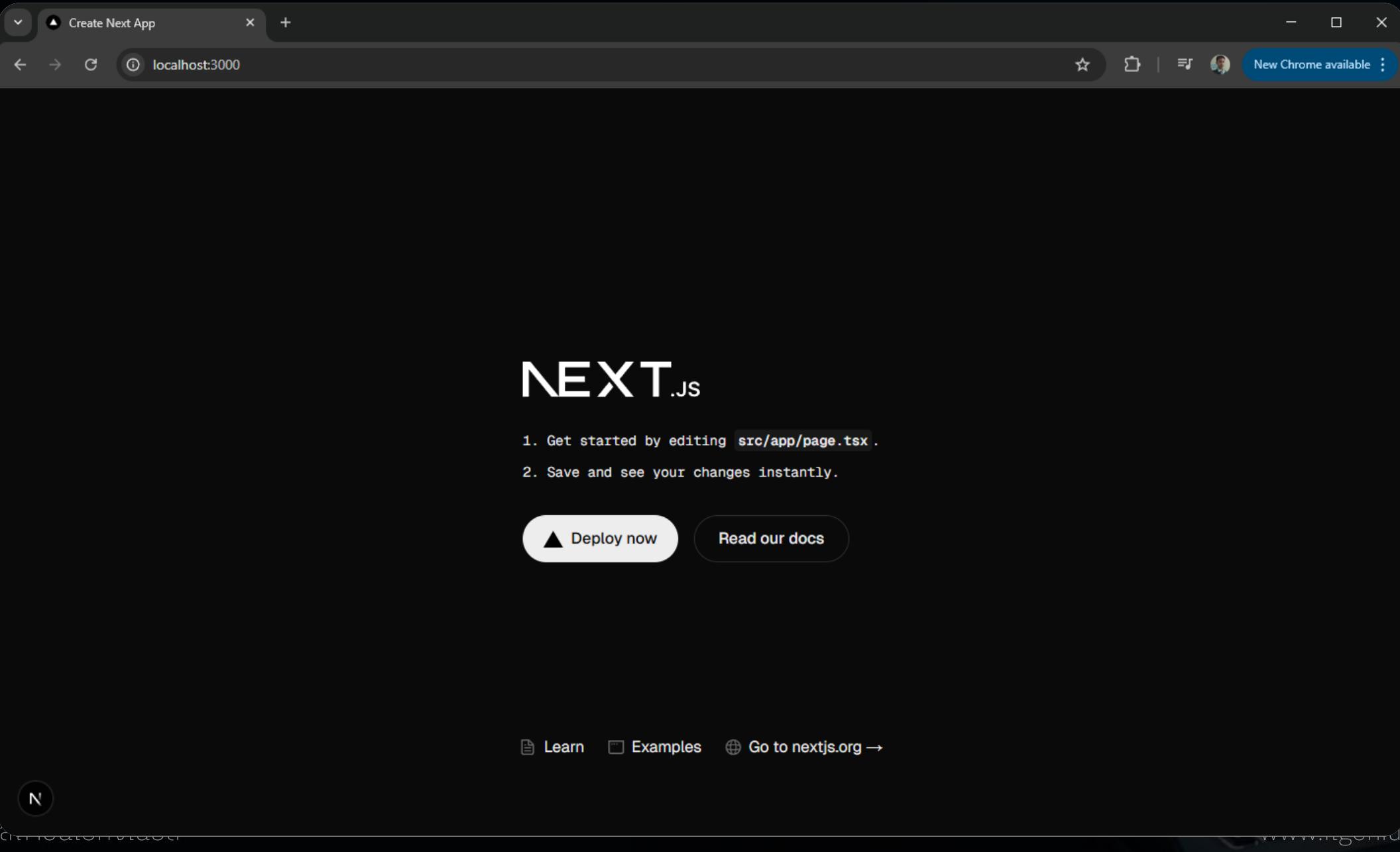
- > 📁 .git
- > 📁 node_modules
- > 🗂️ public
- > 🗂️ src
 - ❖ .gitignore
 - ⌚ eslint.config.mjs
 - ⌚ next-env.d.ts
 - ⌚ next.config.ts
 - ⌚ package-lock.json
 - ⌚ package.json
 - ⌚ postcss.config.mjs
 - ⌚ README.md
 - ⌚ tsconfig.json

package.json ✘

package.json > ...
You, 1 minute ago | 1 author (You)

```
1  {  
2      "name": "aichatbot-chatbot-nextjs",  
3      "version": "0.1.0",  
4      "private": true,  
5      "scripts": {  
6          "dev": "next dev",  
7          "build": "next build",  
8          "start": "next start",  
9          "lint": "eslint"  
10     },  
11     "dependencies": {  
12         "react": "19.1.0",  
13         "react-dom": "19.1.0",  
14         "next": "15.5.2"  
15     },  
16     "devDependencies": {  
17         "typescript": "^5",  
18         "@types/node": "^20",  
19         "@types/react": "^19",  
20         "@types/react-dom": "^19",  
21         "@tailwindcss/postcss": "^4",  
22         "tailwindcss": "^4",  
23         "eslint": "^9",  
24         "eslint-config-next": "15.5.2",  
25         "@eslint/eslintrc": "^3"  
26     }  
27 }  
28 }
```





Create Next App

localhost:3000

Elements Console Sources Network Lighthouse >

5:39:42 PM - localhost:3000

http://localhost:3000/

NEXT.js

1. Get started by editing `src/app/page.tsx`.

2. Save and see your changes instantly.

[Deploy now](#) [Read our docs](#)

Learn Examples Go to nextjs.org →

Performance

Values are estimated and may vary. The [performance score is calculated](#) directly from these metrics. [See calculator.](#)

▲ 0–49 ■ 50–89 ● 90–100



Home Workspaces API Network

Search Postman Ctrl K

Invite Upgrade No environment

Samit

Collections + Search collections

AIChatbotLangchain

- 01_Next_API
- 02_Langchain_Basic
 - POST 01./api/chat_01_start
 - POST 02./api/chat_02_request
 - POST 03./api/chat_03_template
 - POST 04./api/chat_04_stream
- 05_Chat_History
 - POST 01./api/chat_05.history?sessionId=4ebe30a5-0c01-4912-927d-78227102dee9
 - GET 02./api/chat_05.history?sessionId=4ebe30a5-0c01-4912-927d-78227102dee9 Copy
- 06_Chat_History_Optimize
 - POST 01./api/chat_06.history_optimize?sessionId=4ebe30a5-0c01-4912-927d-78227102dee9
 - GET 02./api/chat_06.history_optimize?sessionId=4ebe30a5-0c01-4912-927d-78227102dee9 Copy
- 07_Tool_Calling
 - POST 01./api/chat_07.tool_calling?sessionId=4ebe30a5-0c01-4912-927d-78227102dee9
 - GET 02./api/chat_07.tool_calling?sessionId=4ebe30a5-0c01-4912-927d-78227102dee9 Copy
- Document_Loader_Embbeding_pgVector
 - GET 01./api/document_loader_embbeding_pgvector/text_csv
 - POST 02./api/document_loader_embbeding_pgvector/text_csv
 - DEL 03./api/document_loader_embbeding_pgvector/text_csv
 - PUT 04./api/document_loader_embbeding_pgvector/text_csv
 - GET 05./api/document_loader_embbeding_pgvector/text_csv_pdf
- 08_RAG
 - POST 01./api/chat_08.rag?sessionId=4ebe30a5-0c01-4912-927d-78227102dee9
 - GET 02./api/chat_08.rag?sessionId=4ebe30a5-0c01-4912-927d-78227102dee9 Copy

DjangoWebSocket

gofiber

QR Menu App API

HTTP 02_Langchain_Basic / 04./api/chat_04_stream

POST {{baseUrl}} /api/chat_04_stream

Params Authorization Headers (9) Body Scripts Settings Cookies Beautify

Body raw binary GraphQL JSON

```
1  {
2     "messages": [
3         {
4             "id": "chat-id-001",
5             "role": "user",
6             "parts": [
7                 {
8                     "type": "text",
9                     "text": "สวัสดีครับ บริษัทของเรารักษาความปลอดภัยให้มากที่สุดครับ"
10                }
11            ]
12        }
13    ]
14 }
```

Response Click Send to get a response

Send

Save Share

API end point ไว้ให้ทดสอบง่าย





3. พื้นฐาน Langchain.js เชื่อมต่อกับ Gen AI

AI Chatbot with LangChain

localhost:3000

AI Chatbot with LangChain.JS

สวัสดีครับ! มีอะไรให้ช่วยเหลือหรือคุยกันได้บ้างครับ?

ประเภทไทยแม่งออกเป็น 6 ภาค ได้แก่:

- **ภาคเหนือ**
- **ภาคกลาง**
- **ภาคตะวันออก**
- **ภาคตะวันตก**
- **ภาคตะวันออกเฉียงเหนือ (อีสาน)**
- **ภาคใต้**

แต่ละภาคมีลักษณะภูมิศาสตร์และวัฒนธรรมที่แตกต่างกันไปครับ ถ้าสนใจข้อมูลเพิ่มเติมเกี่ยวกับภาคไหนบอกได้เลยนะครับ!

ภาคเหนือมีซึ่งหน้าดูไวน้ำงครับ

ภาคเหนือของประเทศไทยประกอบด้วย 9 จังหวัด ได้แก่:

- **เชียงใหม่**
- **เชียงราย**
- **ลำปาง**
- **ลำพูน**
- **พะเยา**
- **แม่ฮ่องสอน**
- **น่าน**
- **อุตรดิตถ์**
- **แพร่**

แต่ละจังหวัดมีความสวยงามทางธรรมชาติและวัฒนธรรมที่น่าสนใจมากมาย เลยครับ ถ้าต้องการข้อมูลเพิ่มเติมเกี่ยวกับจังหวัดใดจังหวัดหนึ่ง แจ้งได้เลย นะครับ!

พิมพ์ที่ข้อความที่... ▲

**ตัวอย่าง AI Chatbot
อย่างรวดเร็วด้วย
Langchain.js ร่วมกับ
Next.js**



สถาบันไอทีเจเนียส

Chat AI

www.itgenius.co.th



LangChain
ร่วมกับ
Next.JS 
และ  supabase

Thank you



ว.สาเมตร โภยม
สถาบันไอทีจีเนียส



สถาบันไอทีจีเนียส

www.itgenius.co.th