การเพิ่มประสิทธิภาพการทำงานองค์กรด้วย Generative AI

ดร.มนัสชัย คุณาเศรษฐ

สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ
กระทรวงการอุดมศึกษา วิทยาศาสตร์ วิจัยและนวัตกรรม

22 ก.ย. 2568

วิทยากร

ดร.มนัสชัย คุณาเศรษฐ

- ผู้ช่วยผู้อำนวยการ สวทช. (2565 - ปัจจุบัน)
- หัวหน้าทีมวิจัย ศูนย์ทรัพยากรคอมพิวเตอร์เพื่อการคำนวณขั้นสูง (2562 - 2565)
- นักวิจัย ศูนย์นาโนเทคโนโลยีแห่งชาติ (2556 - 2561)

ผลงานและรางวัล

- ออกแบบและพัฒนาซูเปอร์คอมพิวเตอร์ “LANTA”
- นักเทคโนโลยีรุ่นใหม่ ประจำปี 2565 มูลนิธิส่งเสริมวิทยาศาสตร์และเทคโนโลยี ในพระบรมราชูปถัมภ์
- รางวัลวิทยานิพนธ์ สภาการวิจัยแห่งชาติ

Topics

- ทำความรู้จัก Generative AI
- การใช้ Generative AI เพื่อการเพิ่มประสิทธิภาพในองค์กร
- กรณีศึกษาในการพัฒนาและประยุกต์ใช้งาน AI

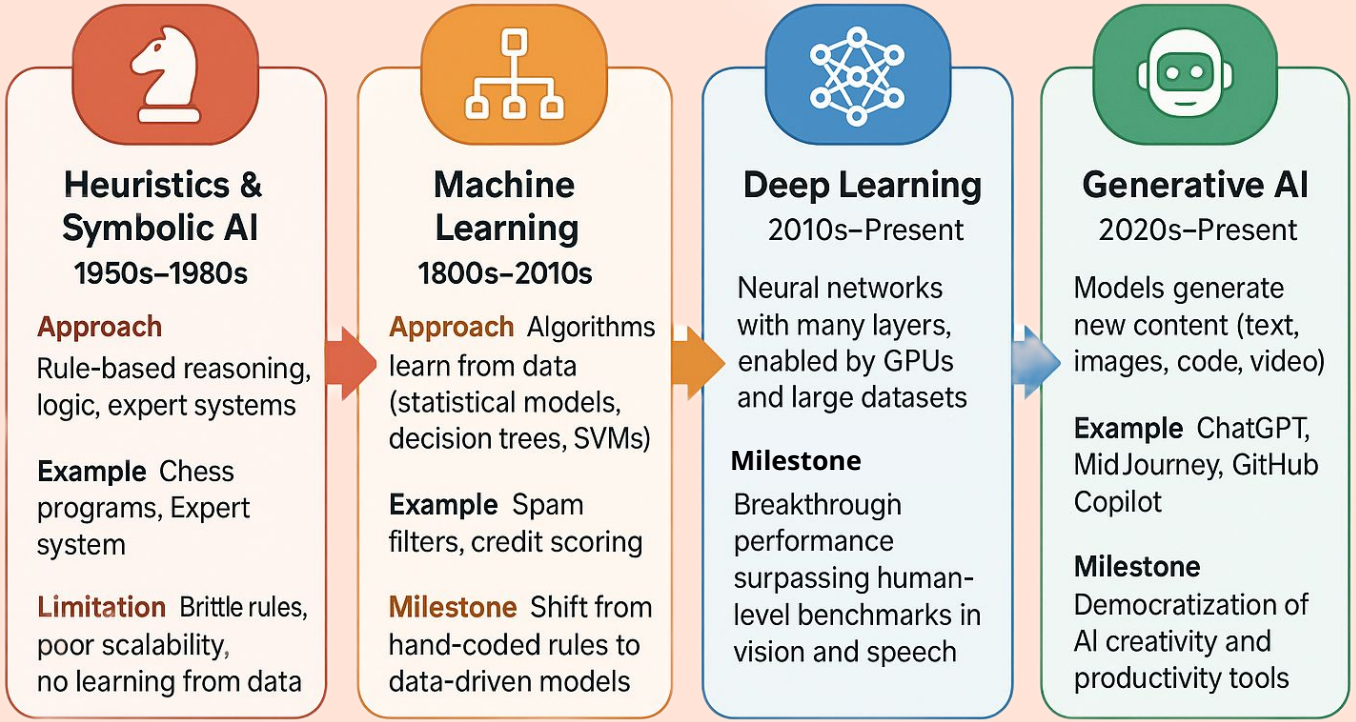
Topics

- ทำความรู้จัก Generative AI
- การใช้ Generative AI เพื่อการเพิ่มประสิทธิภาพในองค์กร
- กรณีศึกษาในการพัฒนาและประยุกต์ใช้งาน AI

NSTDA ChatAI
chatai.nstda.or.th

รู้จัก Generative AI

Age of AI


Heuristics & Symbolic AI
1950s–1980s
Approach
Rule-based reasoning, logic, expert systems
Example Chess programs, Expert system
Limitation Brittle rules, poor scalability, no learning from data
Machine Learning
1800s–2010s
Approach Algorithms learn from data (statistical models, decision trees, SVMs)
Example Spam filters, credit scoring
Milestone Shift from hand-coded rules to data-driven models
Deep Learning
2010s–Present
Neural networks with many layers, enabled by GPUs and large datasets
Milestone
Breakthrough performance surpassing human-level benchmarks in vision and speech
Generative AI
2020s–Present
Models generate new content (text, images, code, video)
Example ChatGPT, Mid Journey, GitHub Copilot
Milestone
Democratization of AI creativity and productivity tools

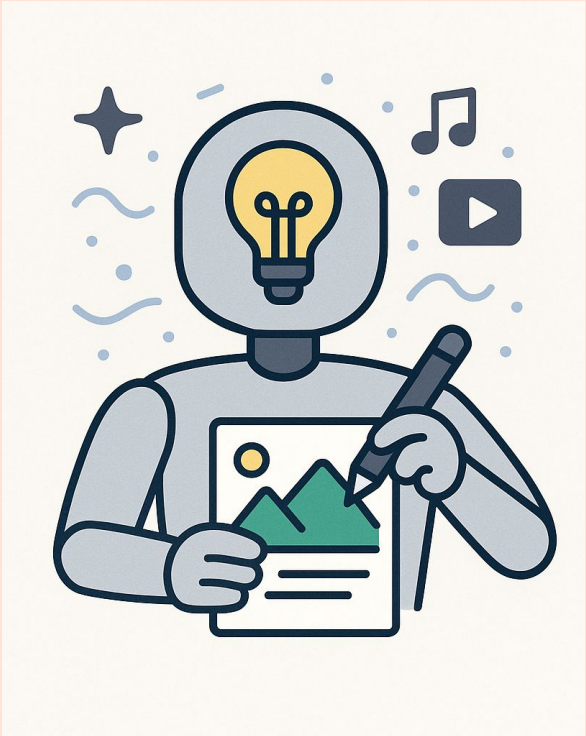
Generative AI คืออะไร

Generative AI เป็น AI ที่มุ่งเน้นความสามารถในการสร้างเนื้อหาใหม่ ๆ เช่น ข้อความ ภาพ เสียง วิดีโอ หรือ Code

โดยใช้การเรียนรู้รูปแบบ โครงสร้าง และความสัมพันธ์จากข้อมูลที่มีอยู่ แล้วนำมาสร้างผลลัพธ์ใหม่ที่มีลักษณะคล้ายคลึงกับสิ่งที่มนุษย์สร้างขึ้น

รูปแบบหนึ่งของ Gen AI ที่นิยมใช้งานคือ

Large Language Model (LLM)



LLM ทำงานอย่างไร

LLM ทำงานโดย “ทำนายคำ (token) ถัดไป” ในประโยค โดยใช้ข้อมูลสถิติจากรูปแบบและความสัมพันธ์ของคำที่เรียนรู้มาจากข้อความ/ตัวอักษรจำนวนมาก ที่เรียนรู้ระหว่างการ train model

Prompt: เช้านี้ฝนตกถนน _____

(✔ลื่น | ✔เปียก | ✔ปิด | ❌แห้ง)

Prompt: ไก่งามเพราะ _____

(✔ขน | ✔พันธุ์ | ❌กุ้ง)

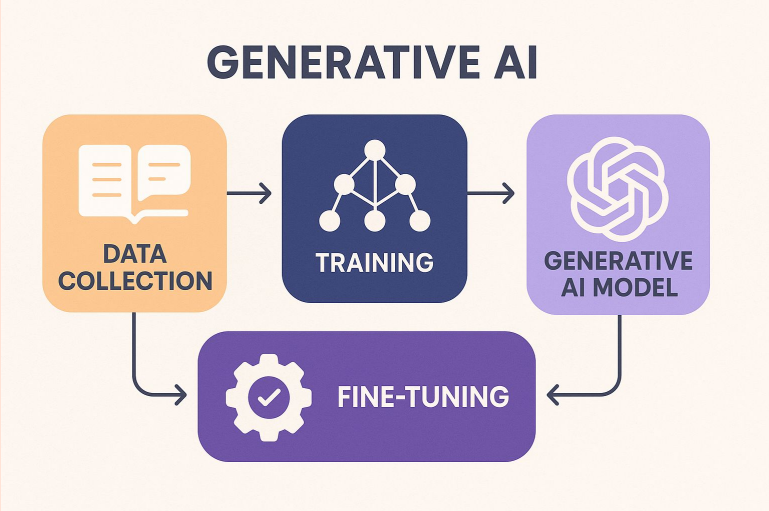
Prompt: เช้านี้ฝนตกปอย _____

(✔ปรอย | ❌เปต)

Prompt: อย่าไปปอย _____

(❌ปรอย | ✔เปต)

ขั้นตอนในการพัฒนา LLM



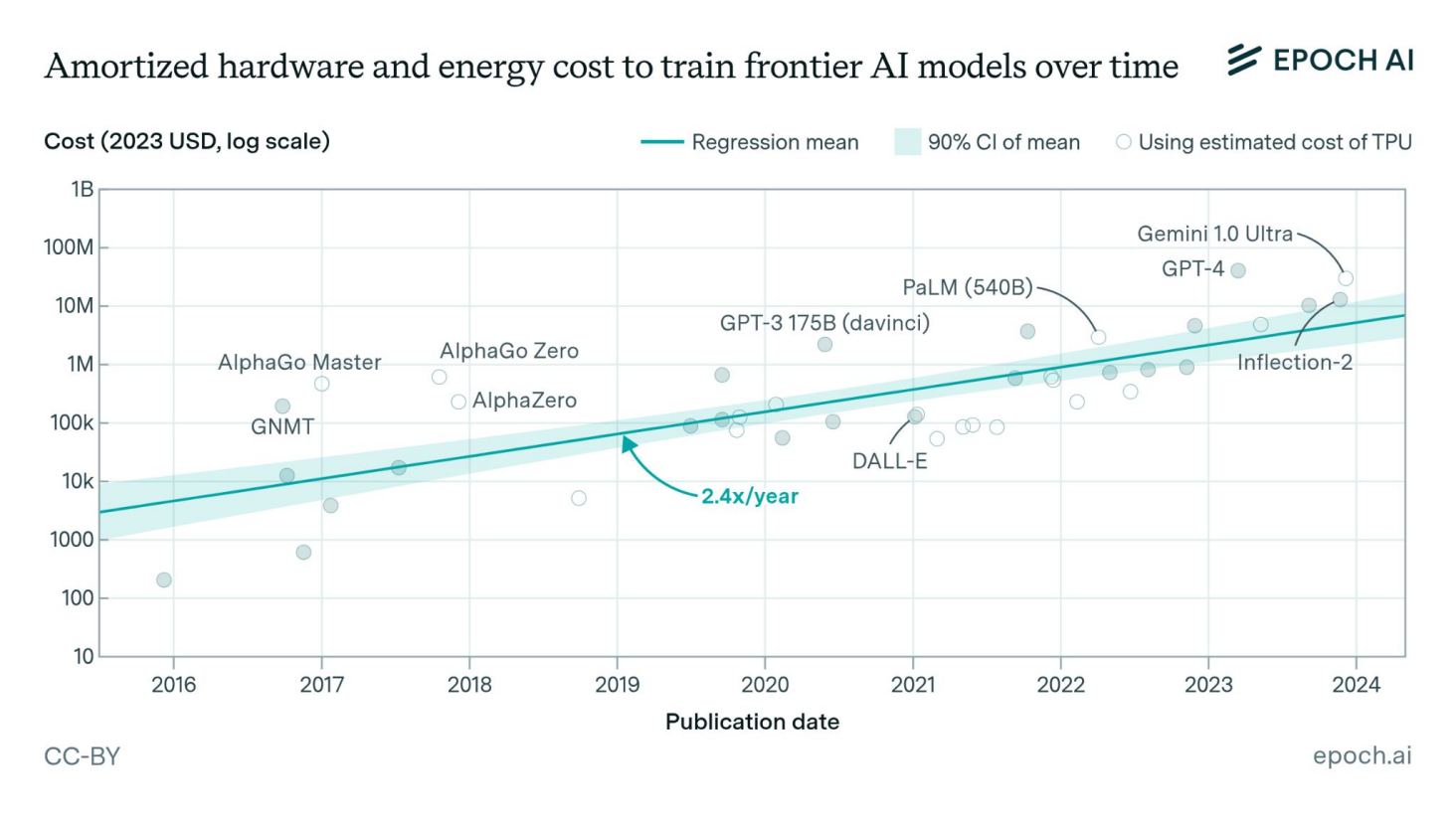
1. **Collect and Prepare Data** – รวบรวมข้อมูล และจัดรูปแบบให้เหมาะสม
2. **Design Architecture*** – ออกแบบโครงสร้าง Neural Model
3. **Training Model*** – ใช้พลังประมวลผลสูงในการเรียนรู้จากข้อมูล
4. **Finetuning** – ปรับให้เหมาะกับงานจริง เพิ่มความถูกต้องและปลอดภัย
5. **Deploy and Evaluate** – ประเมินผลด้วยเกณฑ์มาตรฐาน แล้วเปิดใช้ผ่านระบบหรือ API

***ในระดับองค์กร ขั้นตอนที่ 2 - 3 สามารถตัดได้ โดยการเลือกใช้ Pre-trained model**

ข้อจำกัดของ LLM

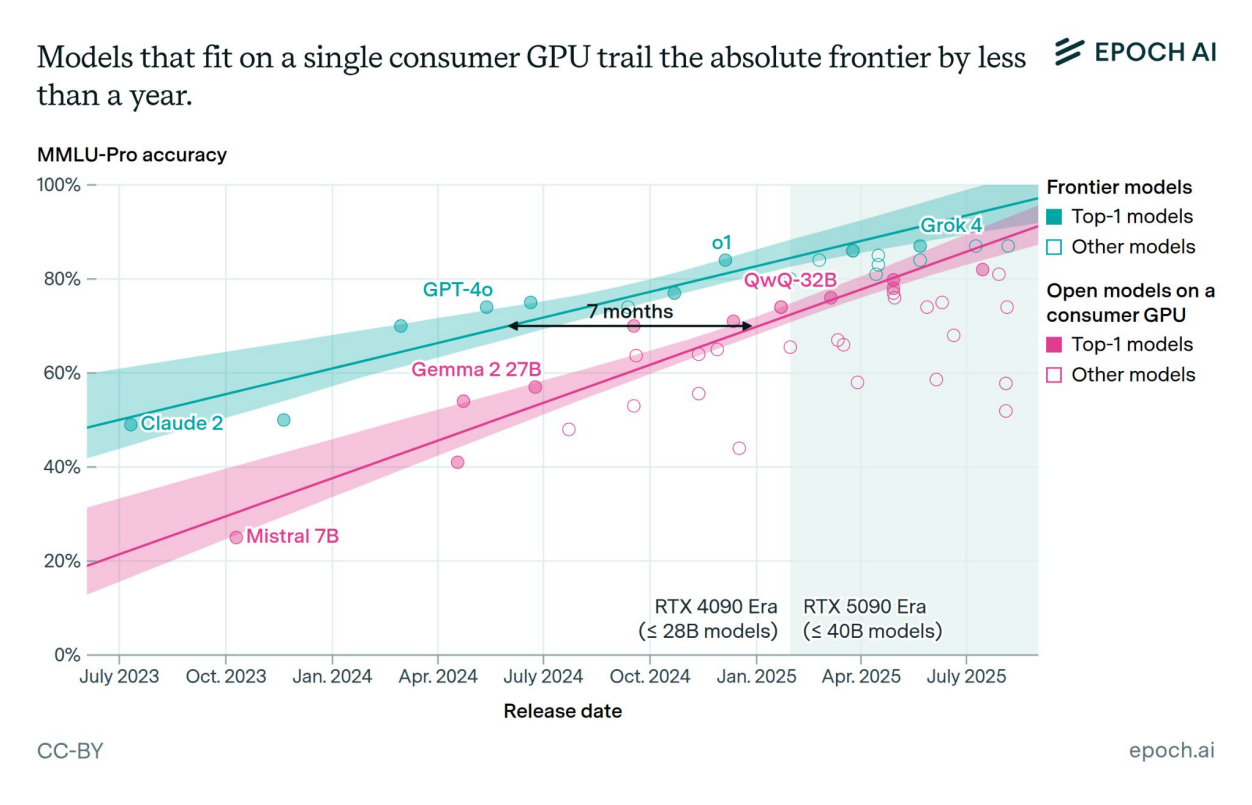
- Accuracy – ความถูกต้องขึ้นกับข้อมูลที่ใช้ train
- Randomness – ผลลัพธ์ไม่เหมือนกันทุกครั้ง ควบคุมไม่ได้ 100%
- Hallucination – อาจสร้างข้อมูลที่ไม่จริง แต่ดูสมเหตุสมผล
- Reasoning limits – มีข้อจำกัดด้านตรรกะและการคิดเชิงเหตุผล
- Knowledge cutoff – ไม่รู้ข้อมูลใหม่ล่าสุด
- Data risk – ต้องระวังเรื่องข้อมูลสำคัญและความเป็นส่วนตัว

การพัฒนา LLM มีต้นทุนสูงมาก



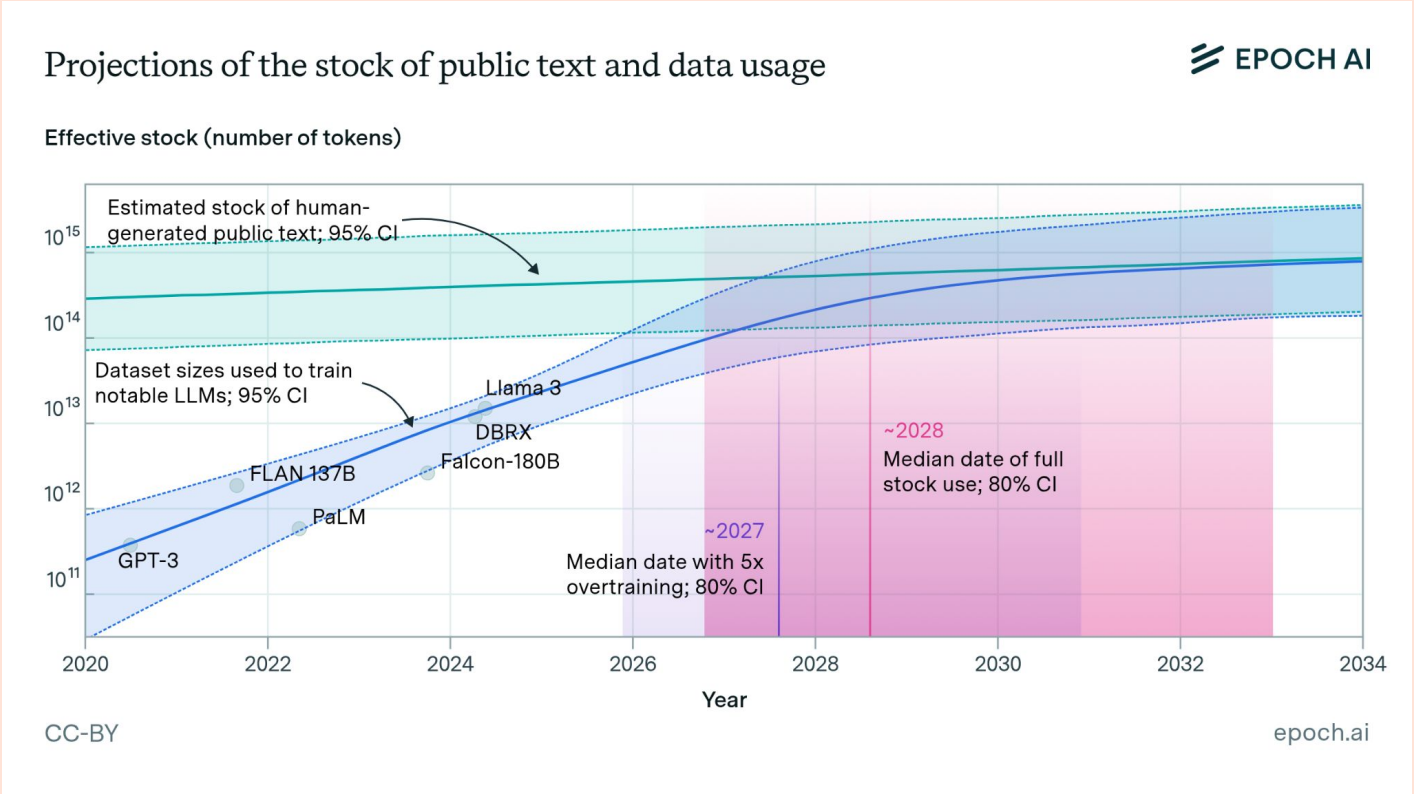
Epoch AI (2023), "Key Trends and Figures in Machine Learning". *Published online at epoch.ai.* Retrieved from: 'https://epoch.ai/trends' [online resource]

LLM ขนาดเล็กมีประสิทธิภาพสูงขึ้น



Epoch AI (2023), "Key Trends and Figures in Machine Learning". *Published online at epoch.ai.* Retrieved from: 'https://epoch.ai/trends' [online resource]

Gen AI's Iceberg: Data is Running Out



Epoch AI (2023), "Key Trends and Figures in Machine Learning". *Published online at epoch.ai.* Retrieved from: 'https://epoch.ai/trends' [online resource]

Summary: Generative AI

- Timeline of AI
- Generative AI / LLM
- Trends

Topics

- ทำความรู้จัก Generative AI
- **การใช้ Generative AI เพื่อการเพิ่มประสิทธิภาพในองค์กร**
- กรณีศึกษาในการพัฒนาและประยุกต์ใช้งาน AI

NSTDA ChatAI
chatai.nstda.or.th

การใช้ Generative AI เพื่อการเพิ่มประสิทธิภาพภายใน สวทช.

NSTDA ChatAI (ชาไทย)

ที่มาของ NSTDA Chat AI (ชาไทย)

- ปัจจุบัน AI ในรูปแบบของการให้บริการ Large Language Model (LLM) มีความสามารถสูง สามารถใช้เพื่อเพิ่มประสิทธิภาพการทำงานขององค์กรในทุกรูปแบบ

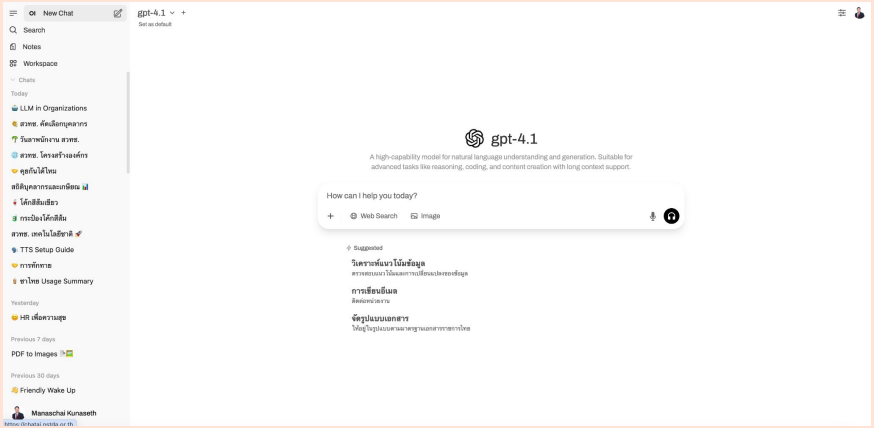
- สวทช. มีนโยบายในการนำเอาเทคโนโลยี AI มาใช้เพื่อเพิ่มประสิทธิภาพการทำงานของบุคลากร อย่างเหมาะสมและปลอดภัย

- ด้านสารสนเทศ ได้ร่วมกับหน่วยงานใน สวทช. เพื่อพัฒนารูปแบบการให้บริการ Chat AI ภายในองค์กร

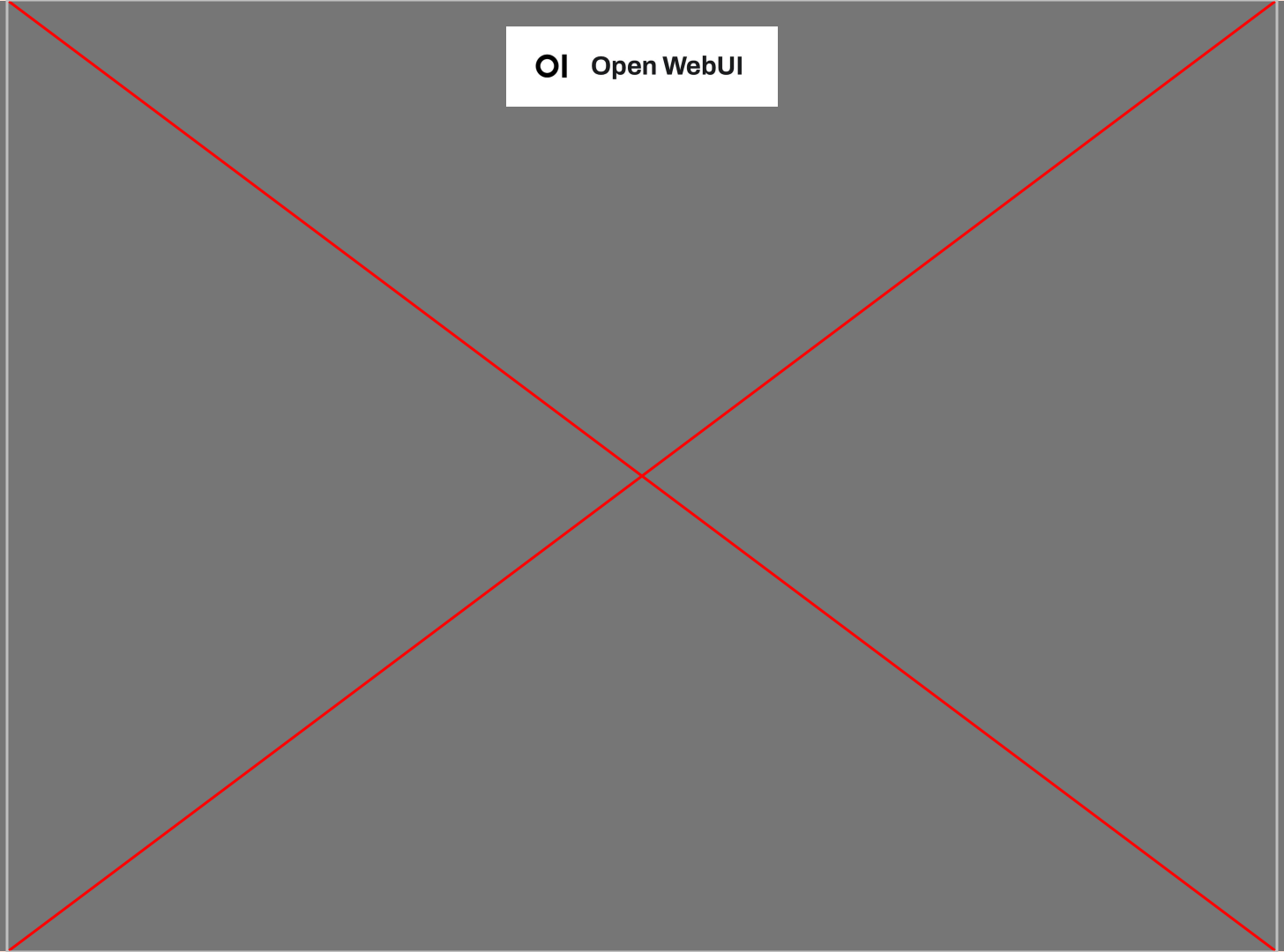
NSTDA ChatAI (ชาไทย)

NSTDA ChatAI
chatai.nstda.or.th

แพล็ตฟอร์มใช้งาน Generative AI สำหรับบุคลากร สวทช. พัฒนาต่อยอดจาก Open Source เพื่อเสริมประสิทธิภาพการทำงานได้อย่างสะดวกและปลอดภัย



Open WebUI


Open WebUI

ทำไมต้องใช้ NSTDA ChatAI ?

ความปลอดภัยของข้อมูล

ข้อมูลจากการสนทนาและเอกสาร
ไม่ถูกนำไปใช้พัฒนาโมเดล
โดยผู้ให้บริการ Gen AI
(ตามข้อตกลง User
Agreement แบบ Enterprise)

ChatGPT Pathumma LLM Claude

เลือกใช้งาน AI ได้หลาย Model

สามารถเลือกใช้งาน Model ชั้นนำ
ได้หลากหลาย โดยการสนับสนุน
จากสำนักงาน เช่น ChatGPT,
Claude, Pathumma LLM
และสามารถเพิ่มเติม Model
ได้ในอนาคต

รูปแบบการใช้งานที่ระบบรองรับ

Chat

การโต้ตอบผ่านการสนทนาเหมือนการใช้งาน Generative AI ทั่วไป

คุยกับเอกสาร รูป หรือเว็บไซต์

รองรับการสอบถามข้อมูลจากเอกสารที่อัพโหลด วิเคราะห์ภาพถ่าย หรือค้นข้อมูลจากเว็บไซต์

ใช้งาน Collection เพื่อเลือกชุดข้อมูลในการตอบคำถาม

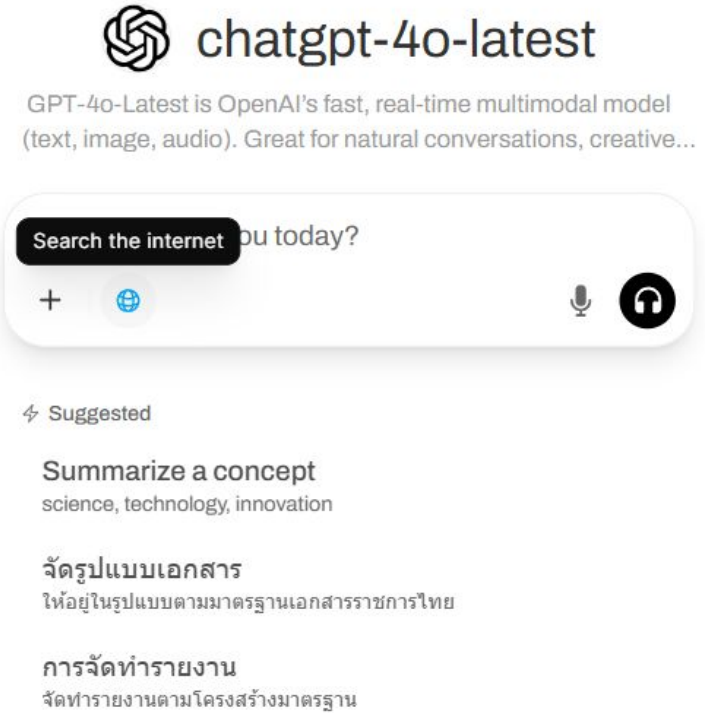
สืบค้นข้อมูลจากฐานข้อมูลเอกสารภายใน สวทช. ประกอบการให้คำตอบ โดยใช้เทคนิค Retrieval-Augmented Generation (RAG)

ฐานข้อมูลที่เปิดให้ใช้ในปัจจุบัน

- ระเบียบ/ข้อบังคับ
- คู่มือจัดซื้อจัดจ้าง ระเบียบและกฎหมายพัสดุ
- ฐานข้อมูลเอกสาร ISO ต่างๆ

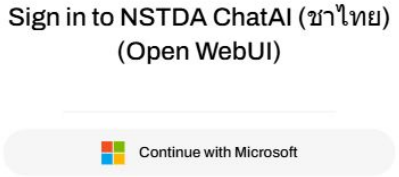
Web Search สำหรับช่วยหาข้อมูล

ให้ระบบช่วยสืบค้นข้อมูลจาก Search Engine
ก่อนส่งให้ AI ใช้เป็นข้อมูลเพื่อตอบคำถามให้กับผู้ใช้งาน
(แบบเดียวกับ Platform Commercial)



มาตรการด้านความปลอดภัย

- ปกป้องโดยใช้ Firewall และเครือข่าย VPN จึงมีความปลอดภัยสำหรับการใช้ภายในองค์กร
- เชื่อมต่อระบบการ login กับระบบของ สวทช. โดยตรง ช่วยเพิ่มความปลอดภัยในการยืนยันตัวตน
- เพิ่มประสิทธิภาพในการให้บริการแบบคุ้มค่าโดยใช้ Autoscaler เพื่อเพิ่ม/ลดทรัพยากรที่ให้บริการ ตามการใช้งาน



Timeline ในการพัฒนาและให้บริการ

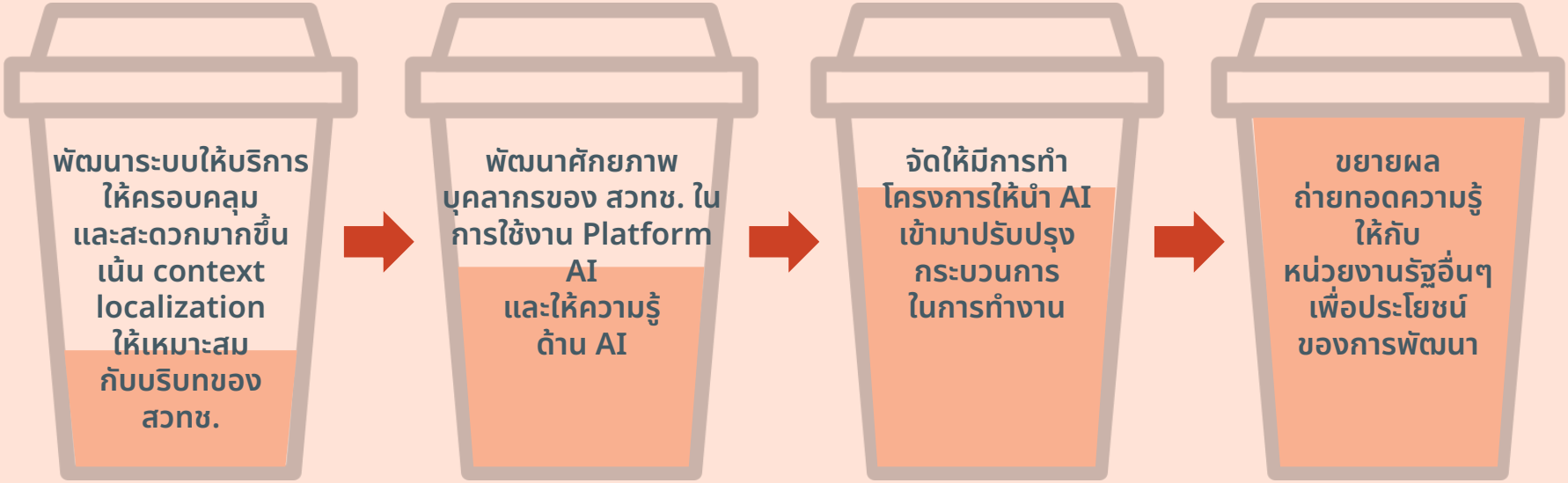
มี.ค. 2568
เปิดระบบทดลองภายใน
(30 users)

เม.ย. 2568
ขยายผู้ทดลองใช้ให้ผู้บริหาร
ระดับกลาง (350 users)

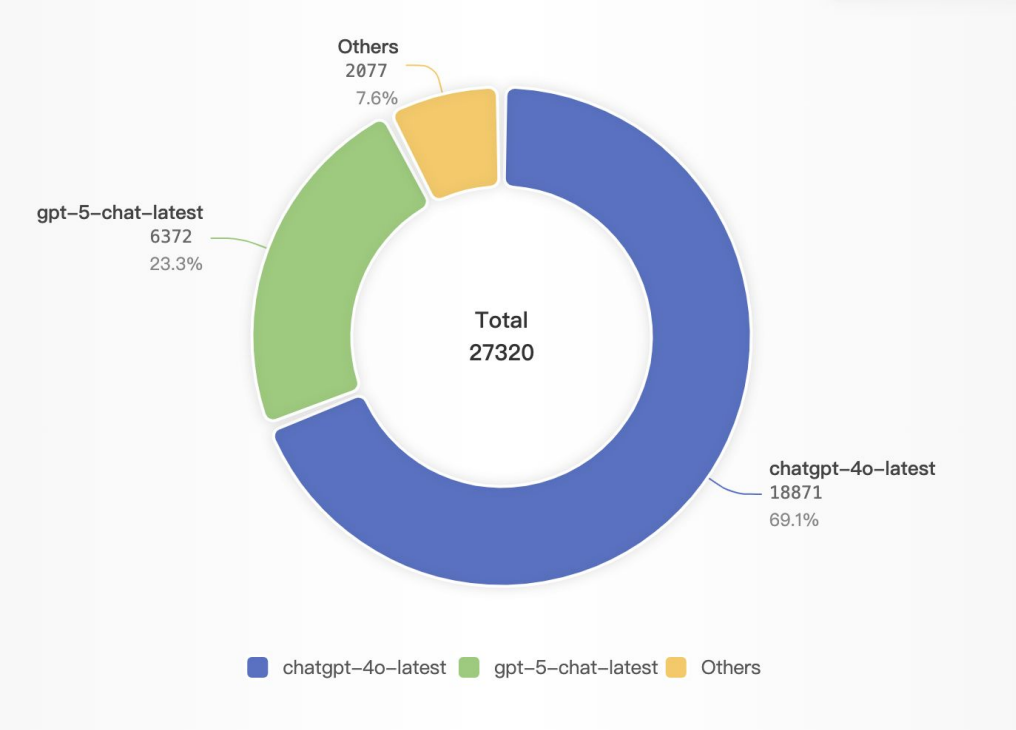
ก.ค. 2568
ขยายผู้ทดลองใช้ให้บุคลากร
สวทช. ทั่วไป (1,300 users)

ส.ค. 2568
ขยายการใช้งานให้ครอบ
คลุมพนักงาน สวทช.

แผนดำเนินการในการขยายและต่อยอด



สถิติการใช้งานชาไทย (2 เดือน)



Total Calls

27.3K

Total Tokens

371.6M

Most Used Model

chatgpt-4o-latest

18871 times

การใช้งาน AI โดยด้านบริหารทรัพยากรบุคคล

ใช้ **brainstorm** แนวคิด
เปิดมุมมองสร้างสรรค์
ช่วยประหยัดเวลา
และได้ทางเลือกใหม่
ที่เหมาะกับบริบท

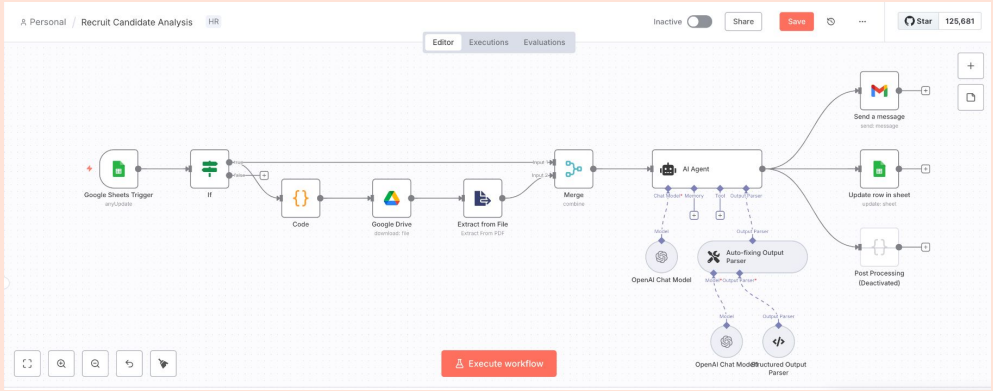
ช่วยในการ**สื่อสารในบริบท**
ต่างๆ เช่นปรับแต่งภาษา
แนะนำรูปแบบกิจกรรม
โดยเฉพาะการสื่อสาร
ที่มีความละเอียดอ่อน

ทดลองใช้ **AI Automation**

- Automation workflow เพื่อวิเคราะห์ Resume ให้ตรงกับ Job Description
- จำลองแผนอัตรากำลัง ทดแทนบุคลากรเกษียณ
- ทดลองใช้ AI ในการ Summarize ภาระงานประจำปีระดับบุคคล

1. Automation workflow เพื่อวิเคราะห์ Resume ให้ตรงกับ Job Description

n8n



AI Analysis

Time & Date of Analysis: 2025-09-22T06:21:02.405+07:00

Summary of Candidate Profile:

- มีประสบการณ์ในสาย IT รวม 10 ปี โดย 5 ปีสุดท้ายเป็น Senior Programmer / Team Lead ดูแลทีม 6 คน มีความรู้และมือในด้าน Infrastructure, Applications, Cloud (AWS), Microservices, DevOps และ Software Development อย่างเข้มข้น จบปริญญาตรีสาขาคอมพิวเตอร์ มีทักษะการบริหารทีม การพัฒนาองค์กร และความสำเร็จในโครงการขนาดใหญ่

Key Strengths and Relevant Experience:

- - ประสบการณ์ 10 ปีในสาย IT และ 5 ปีในตำแหน่ง Team Lead - ดูแลและบริหารทีม Programmers, ทำงานโครงการ ERP, HR, CRM, Cloud Migration และ System Optimization - มีความรู้ลึกด้าน IT Infrastructure (AWS, Docker, Kubernetes) และ DevOps (CI/CD, Monitoring, Automation) - ประสบความสำเร็จลด Downtime, ลดค่าใช้จ่าย IT, ขยายระบบขนาดสำหรับผู้ใช้หมื่นคน - มีความเข้าใจ SDLC, Agile/Scrum อย่างแท้จริง และมีส่วนร่วมพัฒนาองค์กรและฝึกอบรมทีม - ทักษะภาษาอังกฤษเพียงพอสำหรับการสื่อสารและงานเอกสารระดับเทคนิค

Gaps or Concerns:

- - ขาดประสบการณ์ตรงในตำแหน่ง 'ผู้จัดการ' หรือบริหารทีมขนาดใหญ่กว่า 10 คนขึ้นไป - ยังไม่มีใบรับรองมาตรฐาน ITIL, PMP, CISSP หรือ AWS Certified Solutions Architect - ประสบการณ์ vendor management, budgeting หรือ IT Governance มีระบุไว้น้อยหรือไม่ปรากฏชัด - การบริหารโครงการระดับ multi-stakeholder เช่น ในองค์กรขนาดใหญ่มาก หรือรัฐ ยังไม่มีหลักฐานชัดเจน

Overall Suitability Assessment:

- ผู้สมัครมีคุณสมบัติตรงตามความต้องการหลักเกือบทั้งหมด ทั้งทักษะเทคนิค, การนำทีม, และประสบการณ์โครงการใหญ่ในองค์กร tech-driven แม้ยังขาดประสบการณ์ผู้จัดการโดยตรงและการบริหารทีมใหญ่ แต่แสดงศักยภาพสูงด้านภาวะผู้นำ การคิดเชิงกลยุทธ์ ระบบ Cloud, DevOps และการสื่อสารข้ามฝ่าย เหมาะสมกับการพิจารณาต่อ โดยเฉพาะหากองค์กรสามารถรองรับการ learning curve บางส่วนด้าน management

Recommendation: WEAK ACCEPT

เขียน Prompt อย่างไรให้ได้งาน?

ศิลปะแห่งการสร้างพรอมต์

เรียนรู้วิธีสื่อสารกับ Generative AI อย่างมีประสิทธิภาพ เพื่อปลดล็อกศักยภาพสูงสุดของ AI ไกด์นี้จะแนะนำเทคนิคที่จำเป็นสำหรับการสร้างพรอมต์ที่ทรงพลังและแม่นยำ

PTCF Framework: โครงสร้างสู่ความสำเร็จ

เริ่มต้นด้วย Framework ที่เรียบง่ายแต่ทรงพลังนี้เพื่อสร้างพรอมต์ที่มีประสิทธิภาพ การทำความเข้าใจแต่ละองค์ประกอบจะช่วยยกระดับผลลัพธ์จาก AI ของคุณได้อย่างก้าวกระโดด

P

Persona (บทบาท)

กำหนดบทบาทให้ AI ว่าอยากให้เป็นใคร เช่น ผู้เชี่ยวชาญ, นักการตลาด หรือบรรณาธิการ

T

Task (คำสั่ง)

บอกสิ่งที่คุณต้องการให้ AI ทำอย่างชัดเจน เช่น เขียน, สรุป, แปล หรือวิจารณ์

C

Context (บริบท)

ให้ข้อมูลพื้นหลัง, เป้าหมาย, กลุ่มเป้าหมาย และข้อจำกัดที่จำเป็น

F

Format (รูปแบบ)

ระบุโครงสร้างของผลลัพธ์ที่ต้องการ เช่น ตาราง, รายการ, JSON หรืออีเมล

ตัวอย่างการใช้ PTCF Framework

ช่วยทำตารางออกกำลังกายลดน้ำหนัก 4 สัปดาห์

“สวมบทบาทเป็นเทรนเนอร์ฟิตเนสส่วนตัว ออกแบบโปรแกรมออกกำลังกาย 4 สัปดาห์ สำหรับคนวัยทำงานที่ไม่มีเวลาเยอะ อยากลดน้ำหนัก แต่มีอุปกรณ์เพียงดัมเบล 1 คู่ที่บ้าน นำเสนอในรูปแบบตาราง แบ่งเป็นสัปดาห์ วันโปรแกรม และเวลาใช้”

ตัวอย่างการใช้ PTCF Framework

ช่วยทำแผนดิจิทัลทรานส์ฟอร์เมชัน 3 ปีของหน่วยงานรัฐ

“ในฐานะการเป็นที่ปรึกษาด้านกลยุทธ์องค์กร ช่วยสร้างแผนกลยุทธ์ดิจิทัลทรานส์ฟอร์เมชัน 3 ปี สำหรับหน่วยงานรัฐที่มีข้อจำกัดด้านงบประมาณและบุคลากร ต้องการปรับกระบวนการทำงานให้ทันสมัยและเพิ่มประสิทธิภาพ สรุปเป็น bullet points ภายใต้หัวข้อ Vision, Goals, Initiatives, และ KPIs”

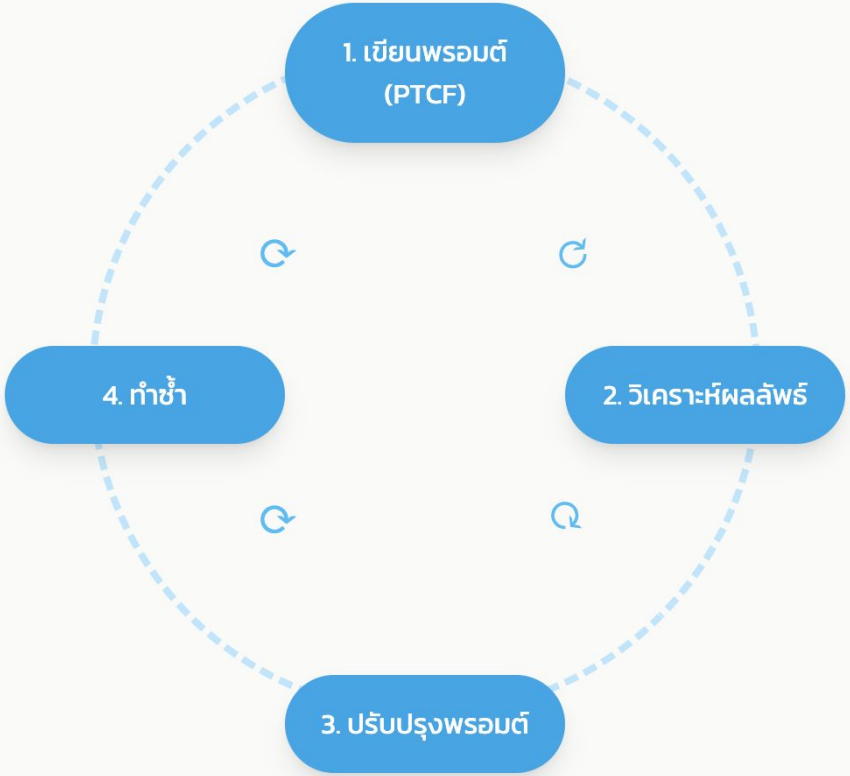
ตัวอย่างการใช้ PTCF Framework

เขียนบทความอธิบาย Generative AI ให้ผู้บริหารอ่าน

“ในบทบาทของนักเขียนคอนเทนต์สายเทคโนโลยี ช่วยเขียนบทความอธิบาย Generative AI สำหรับผู้อ่านทั่วไป กลุ่มเป้าหมายคือผู้บริหารที่ไม่ใช่นักเทคนิคเพื่อโน้มน้าวการลงทุนด้าน AI ในการพัฒนาองค์กร ต้องการเข้าใจภาพรวมโดยไม่ใช้ศัพท์ซับซ้อน แต่มีข้อมูลเพียงพอต่อการตัดสินใจ เขียนความยาว 800 คำ แบ่งเป็นบทนำ เนื้อหา 3 หัวข้อย่อย และสรุปท้ายเรื่อง”

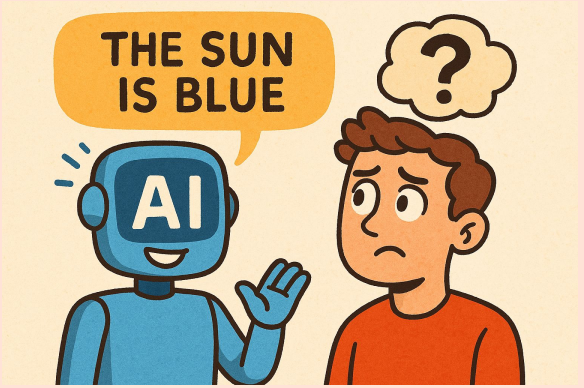
การทำซ้ำคือหัวใจสำคัญ

พรอมต์แรกของคุณไม่ค่อยดีที่สุดเสมอไป ให้คิดว่าการสร้างพรอมต์คือการสนทนา วิเคราะห์คำตอบของ AI และปรับแก้พรอมต์ของคุณเพื่อให้ได้ผลลัพธ์ที่ใกล้เคียงกับที่ต้องการมากขึ้น



AI Hallucinations: ทำอย่างไรเมื่อ AI “หลอน”

AI Hallucination คือปรากฏการณ์ที่โมเดล Generative AI สร้างข้อมูลที่ ผิดพลาดหรือแต่งขึ้นเอง แม้จะดูน่าเชื่อถือและสมเหตุสมผล เช่น



- อ้างอิงบทความวิจัยที่ไม่มีอยู่จริง
- แต่งเหตุการณ์ทางประวัติศาสตร์
- ให้ข้อมูลผิดเกี่ยวกับกฎหมายหรือวิทยาศาสตร์

สาเหตุที่ทำให้เกิด AI Hallucination

- โมเดลไม่มีข้อมูลจริงแบบเรียลไทม์
- อิงจากข้อมูลที่ใช้ training ซึ่งอาจจะไม่ใช่ข้อเท็จจริง
- คำถามคลุมเครือ จำกัดเกินไป หรือขาดบริบท

ป้องกัน AI Hallucination ได้อย่างไร?

1. ออกแบบ Prompt ให้ดี ใช้ภาษาชัดเจน ระบุบริบทให้ครบ หรือ ขอให้ระบุแหล่งที่มา (reference)

2. ใช้เทคนิค RAG (Retrieval-Augmented Generation) เพื่อให้โมเดลค้นหาข้อมูลจริงจากฐานข้อมูลหรือเอกสารที่เชื่อถือได้

3. ฝึกโมเดลเฉพาะทาง เพื่อปรับแต่งโมเดลด้วยข้อมูลเฉพาะด้าน

4. ใช้ prompt ช่วยเพื่อเปิดช่องทางให้ AI ตอบว่า “ไม่รู้” ได้
5. ทวนสอบโดยผู้ใช้งาน โดยเฉพาะเมื่อใช้ในเรื่องสำคัญ เช่น กฎหมาย การแพทย์ หรือข่าว

Taxonomy of AI Hallucination

Comprehensive taxonomy of LLM hallucinations

- **Intrinsic:** self-contradictory content
- **Extrinsic:** introduces nonexistent entities
- **Factuality:** does not match real-world knowledge
- **Factual Errors:** incorrect or fabricated content
- **Contextual:** contradcts or adds to context
- **Instruction:** fails to follow instructions
- **Nonsensical:** irrelevant responses
- **Code generation:** illogical or incorrect code

Table 2: Comprehensive taxonomy of LLM hallucinations

Type	Definition/description	Example	Sources
Intrinsic	Contradicts provided input or context; internal inconsistencies.	Summary states birth year as 1980 then 1975.	[7;70]
Extrinsic	Not consistent with training data; introduces nonexistent entities.	"The Parisian Tiger was hunted to extinction in 1885."	[79;7]
Factuality	Contradicts real-world knowledge or verification sources.	"Charles Lindbergh was first to walk on the moon."	[42;50;13]
Faithfulness	Diverges from input prompt or context.	Summary claims FDA rejected vaccine when article stated approval.	[64;96;61]
Factual Errors	Incorrect, misleading, or fabricated content.	Bard claiming JWST took first exoplanet images.	[14]
Contextual	Contradicts or adds to provided context.	Input: "Nile in Central Africa." Output: "Nile in Central African mountains."	[42;4;27]
Instruction	Fails to follow user instructions.	Translates question to Spanish but answers in English.	[101]
Logical	Internal logical errors or contradictions.	Incorrect arithmetic in step-by-step solution.	[42;47;34;95]
Temporal	Time-sensitive errors and anachronisms.	"Murakami won Nobel Prize in 2016."	[47;51]
Ethical	Harmful, defamatory or legally incorrect content.	False accusation of professor with non-existent citation.	[47;40;31]
Amalgamated	Incorrectly combines multiple facts.	(Blending disparate information)	[27;105]
Nonsensical	Irrelevant responses lacking logic.	Switches from "Adam Silver" to "Stern" in NBA discussion.	[42]
Code generation	Incorrect or nonsensical source code.	Illogical code unfaithful to requirements.	[57;2]
Multimodal	Text-visual content discrepancies.	Identifying non-existent object in image.	[38;98]

Cossio (2025) https://doi.org/10.48550/arXiv.2508.01781

การเสริมชาไทยด้วย RAG

RAG: Retrieval-Augmented Generation

เทคนิคที่ทำให้ LLM (Large Language Model) สามารถ “ค้นหาข้อมูลจริง” จากฐานข้อมูลหรือเอกสารภายนอก แล้วใช้ข้อมูลนั้นมาช่วยในการสร้างคำตอบ ลดปัญหาผิดพลาดหรือ hallucination

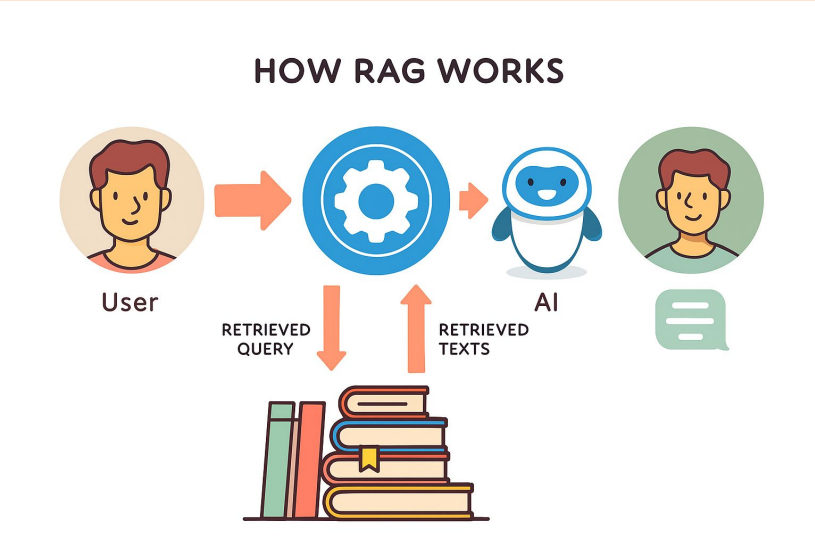
RAG คือการรวมพลังของ...

📥 *การค้นข้อมูลจากเอกสารจริง (Retrieval)*

➕

🧠 *การสร้างข้อความด้วย AI (Generation)*

เพื่อให้ได้คำตอบที่ **ถูกต้อง, มีหลักฐานอ้างอิง,** และ **เข้าใจง่าย**



สร้างและใช้งาน RAG ได้อย่างไร

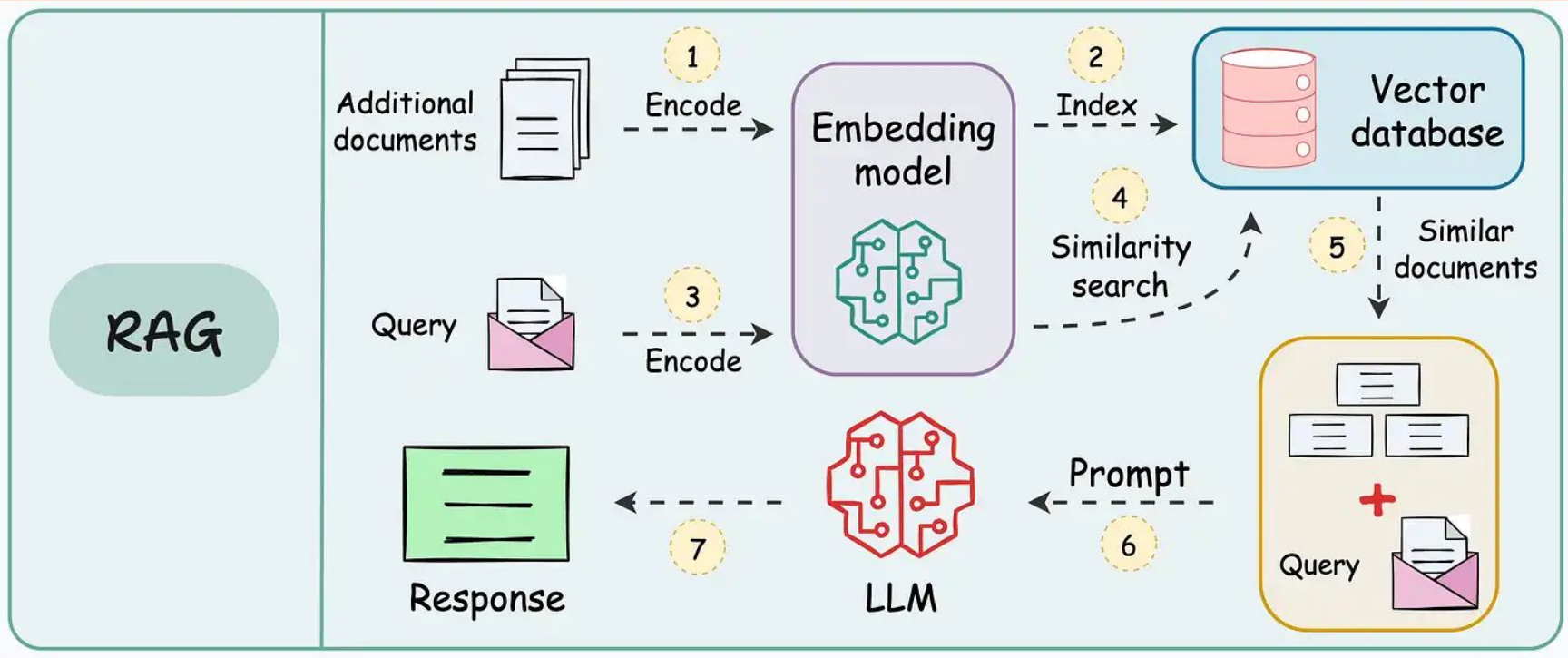


Image source: https://www.dailydoseofds.com/a-crash-course-on-building-rag-systems-part-1-with-implementations/

ข้อดีของการใช้ RAG

การนำ RAG มาประยุกต์ใช้ไม่ได้เพียงแค่ช่วยให้ LLM ตอบคำถามได้ดีขึ้น แต่ยังมอบประโยชน์ที่สำคัญต่อองค์กรในหลายๆ ด้าน ทำให้การเข้าถึงและใช้ประโยชน์จากข้อมูลเป็นไปอย่างมีประสิทธิภาพสูงสุด

✅

ความแม่นยำสูง

ลดปัญหาการตอบผิด (Hallucination) เพราะคำตอบถูกสร้างจากข้อมูลจริงที่มีอยู่

🔄

ข้อมูลเป็นปัจจุบันเสมอ

เพียงอัปเดต Knowledge Base, LLM ก็จะมีความรู้ใหม่ๆ ทันที ไม่ต้องเทรนใหม่ทั้งหมด

🔍

ตรวจสอบได้

สามารถอ้างอิงได้ว่าคำตอบนั้นมาจากเอกสารหรือแหล่งข้อมูลส่วนไหน เพิ่มความน่าเชื่อถือ

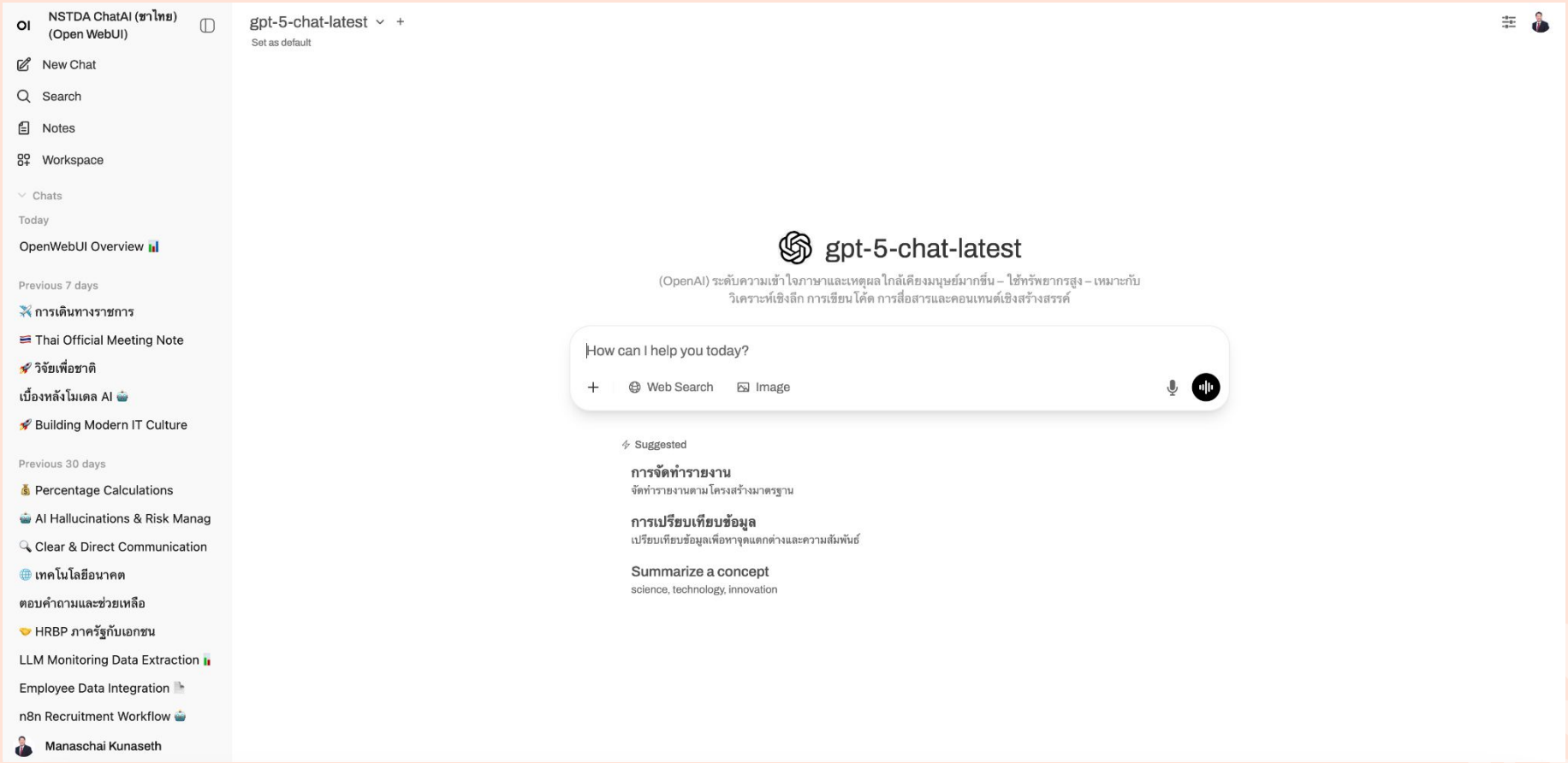
🏢

เหมาะกับข้อมูลเฉพาะทาง

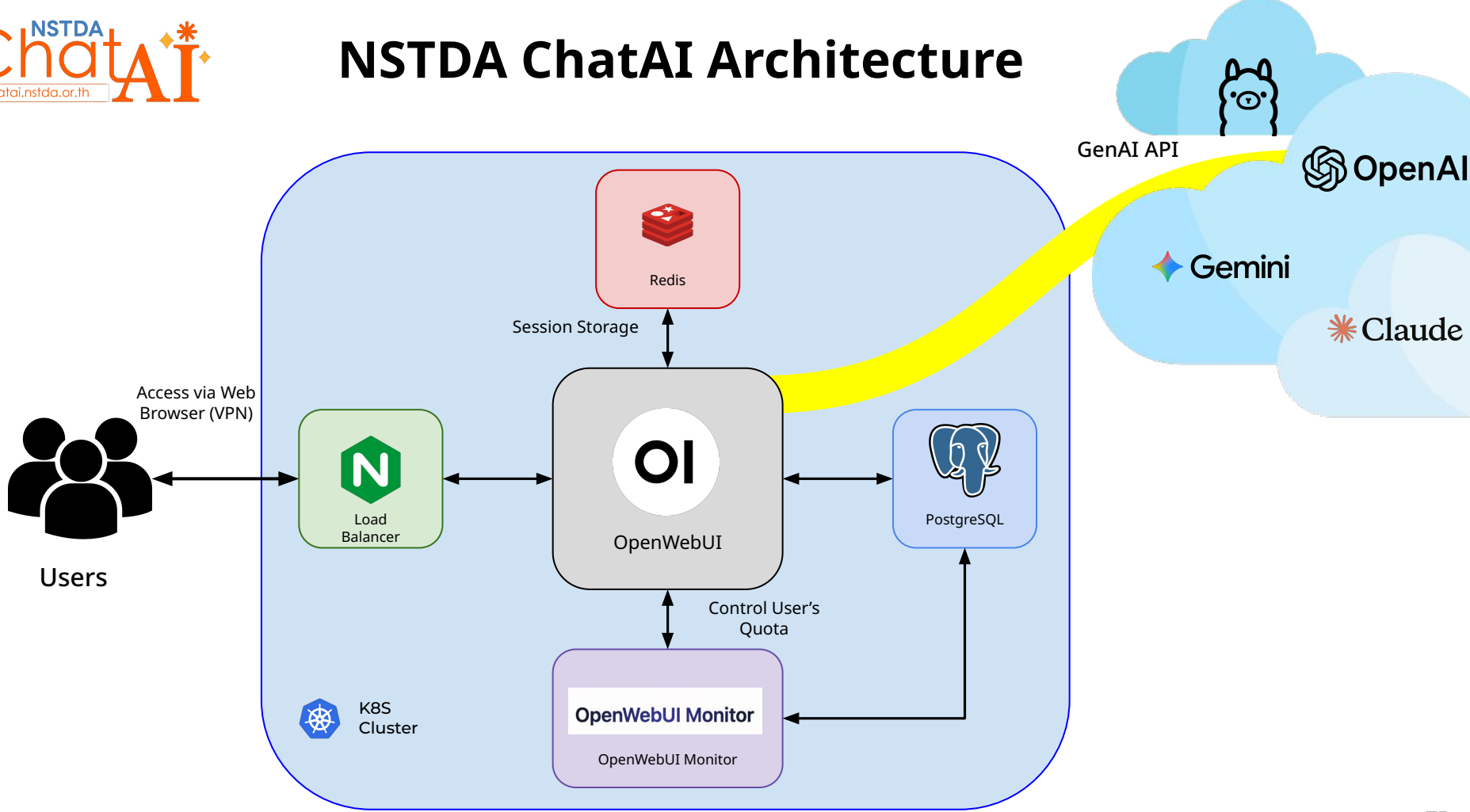
ทำให้ LLM ตอบคำถามเกี่ยวกับข้อมูลภายในองค์กรได้อย่างมีประสิทธิภาพ

Technical Aspect
NSTDA Chat AI

Open WebUI


NSTDA ChatAI (ชาไทย)
(Open WebUI)
New Chat
Search
Notes
Workspace
Chats
Today
OpenWebUI Overview
Previous 7 days
การเดินทางราชการ
Thai Official Meeting Note
วิจัยเพื่อชาติ
เบื้องหลังโมเดล AI
Building Modern IT Culture
Previous 30 days
Percentage Calculations
AI Hallucinations & Risk Manag
Clear & Direct Communication
เทคโนโลยีอนาคต
ตอบคำถามและช่วยเหลือ
HRBP ภาครัฐกับเอกชน
LLM Monitoring Data Extraction
Employee Data Integration
n8n Recruitment Workflow
Manaschai Kunaseth
gpt-5-chat-latest
Set as default
gpt-5-chat-latest
(OpenAI) ระดับความเข้าใจภาษาและเหตุผลใกล้เคียงมนุษย์มากขึ้น – ใช้ทรัพยากรสูง – เหมาะกับ
วิเคราะห์เชิงลึก การเขียนโค้ด การสื่อสารและคอนเทนต์เชิงสร้างสรรค์
How can I help you today?
Web Search
Image
Suggested
การจัดทำรายงาน
จัดทำรายงานตามโครงสร้างมาตรฐาน
การเปรียบเทียบข้อมูล
เปรียบเทียบข้อมูลเพื่อหาจุดแตกต่างและความสัมพันธ์
Summarize a concept
science, technology, innovation

NSTDA ChatAI Architecture

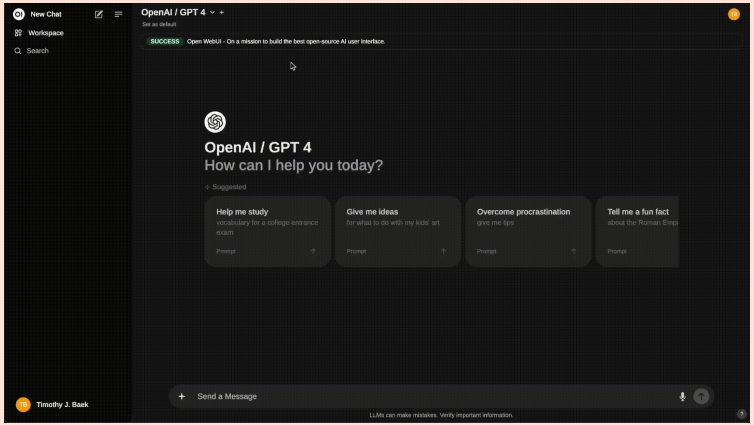


NSTDA ChatAI Components

NSTDA ChatAI ทำงานอยู่บน Kubernetes Cluster โดยมีการใช้งานชุดซอฟต์แวร์ดังนี้

1. OpenWebUI - ใช้เป็นระบบหลักสำหรับการทำงานในการรับคำถามจากผู้ใช้งานผ่านหน้าเว็บ และส่ง request ไปที่ผู้ให้บริการ Generative AI เพื่อประมวลผลและนำคำตอบมาแสดงผล โดยมีการตั้งค่าให้ OpenWebUI มีการ AutoScale เพิ่มจำนวน Pod ที่ให้บริการเมื่อมีปริมาณการใช้งานสูงขึ้น
2. NGINX - ทำหน้าที่เป็น Load Balancer เพื่อจัดการการกระจายปริมาณการใช้งานไปยัง Pod ต่างที่เปิดให้บริการอยู่
3. Redis - ใช้งานในการจัดการ Session การเชื่อมต่อจำนวนมาก เพื่อให้การเชื่อมต่อเข้าใช้งานของผู้ใช้มีเสถียรภาพ
4. PostgreSQL - ใช้ในการเก็บข้อมูลต่างๆ ในการใช้งานระบบ
5. OpenWebUI Monitor - ใช้ในการควบคุมสิทธิการใช้งานของผู้ใช้ และเก็บสถิติการใช้งานของระบบ

OpenWebUI

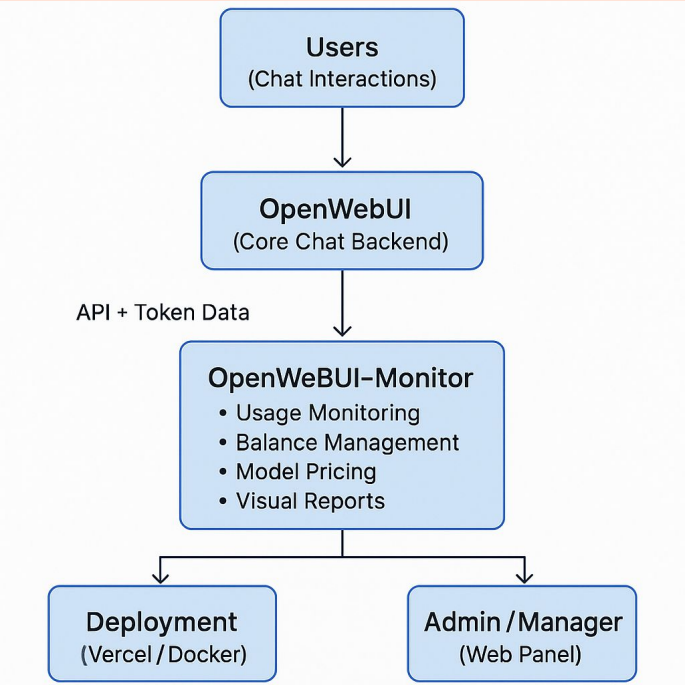


- อินเทอร์เฟซเว็บโอเพ่นซอร์ซ สำหรับเรียกใช้งาน LLM/AI
- ติดตั้งและโฮสต์เองได้ ควบคุมข้อมูล-ความเป็นส่วนตัว
- รองรับหลาย backend (Ollama, OpenAI API, Hugging Face ฯลฯ)
- ฟีเจอร์หลัก: จัดการแชต, โปรไฟล์โมเดล, ปลั๊กอินเสริม
- พัฒนาโดยคอมมูนิตี้ อัปเดตต่อเนื่อง

https://docs.openwebui.com/

Open WebUI

OpenWebUI Monitor



- แดชบอร์ดและเครื่องมือ Monitoring สำหรับติดตามการใช้งาน จัดการยอดคงเหลือของผู้ใช้ และกำหนดโควตา
- รองรับการคิดราคาตามโมเดล AI โดยหักยอดคงเหลือผู้ใช้ตามการใช้งานจริงด้วยการนับ token
- มีการแสดงผลการใช้งาน (usage visualization) ช่วยให้เข้าใจการใช้ทรัพยากรได้ง่าย
- พัฒนาแบบโอเพ่นซอร์ส ภายใต้ MIT License และมีการดูแลอัปเดตอย่างต่อเนื่อง

https://github.com/VariantConst/OpenWebUI-Monitor

ChatAI Key Configurations

Parameter	Configuration
User authentication	OAuth via MS (LDAP also possible)
AI Model	OpenAI (chatgpt-4o, gpt-5) Claude (Sonnet 4) Pathumma LLM
RAG ● chunck size / overlap ● OCR ● embedding function	Enabled ● 1200 / 150 tokens ● disabled ● text-embedding-3-large (OpenAI) or BAAI/bge-m3
Web search	Enable (duckduckgo)

Topics

- ทำความรู้จัก Generative AI
- การใช้ Generative AI เพื่อการเพิ่มประสิทธิภาพในองค์กร
- **กรณีศึกษาในการพัฒนาและประยุกต์ใช้งาน AI**

Pathumma LLM

Pathumma LLM

NECTEC a member of NSTDA

SUPER AI ENGINEER

NSTDA Supercomputer Center ThaiSC

Vision LLM

Audio LLM

Text LLM

Pathumma LLM

DocChat

Powered by Pathumma LLM

Partii Note

NSTDA Supercomputer Center

ThaiSC

LANTA SUPERCOMPUTER

*Peak performance at 8.15 PFLOPS

- 346 nodes Heterogeneous HPE Cray EX cluster
  - 176 GPU nodes with 704 NVIDIA A100 GPUs
  - 160 CPU nodes with 20,480 CPU-cores
  - 10 High-memory nodes, each contains 4TB of memory
- 10 PB of high-performance parallel storage
- High-performance interconnect using 200 Gbps

Rank 70th in TOP500

Rank 24th in GREEN500

(Nov 2022)

Key Takeaways

- AI กำลังเปลี่ยนแปลงโลกการทำงานอย่างรวดเร็วในช่วง 5 ปีที่ผ่านมา ส่งผลกระทบต่อทุกอาชีพ
- การนำ AI มาใช้ต้องพิจารณาบริบท ความเหมาะสม และความปลอดภัยของแต่ละองค์กรทั้งรัฐและเอกชน
- คนที่ใช้ AI ได้อย่างมีประสิทธิภาพจะได้เปรียบ เพราะ AI ไม่ได้มาแทนที่มนุษย์ แต่คนที่ใช้ AI เป็นจะมาแทนที่คนที่ไม่ใช้

Q & A

NSTDA ChatAI
chatai.nstda.or.th

NSTDA ChatAI
chatai.nstda.or.th

Thank you