

GPT-5.6 มี 3 Models × 6 Modes

⚠️ เลือกผิดตัว อาจจ่ายแพงเกินความจำเป็น



เลือก Model

เลือกว่าต้องการผู้ทำงานระดับไหน

LUNA



เร็ว • ประหยัด • งานง่าย

TERRA



สมดุล • งานทั่วไป

SOL



ฉลาดสุด • งานยาก



เลือก Reasoning Mode

เลือกว่าต้องการให้คิดลึกแค่ไหน

Faster

Smarter

none



low



medium



high



xhigh



max



ยิ่งสูง ยิ่งคิดลึก ช้าลง และใช้ต้นทุนมากขึ้น



Luna + none

ประมาณ

0.71

cents/task

VS

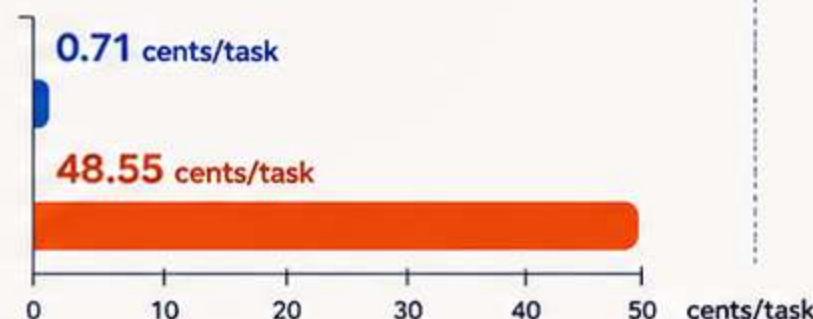


Sol + max

ประมาณ

48.55

cents/task



ต่างกันสูงสุดประมาณ

68 เท่า



ไม่ต้องใช้ตัวแพงสุดทุกงาน

เริ่มจาก **Terra + Medium** แล้วค่อยปรับตามความยาก



ประหยัดเวลา



ประหยัดต้นทุน



ได้คุณภาพที่เหมาะสม

3 Models ต่างกันอย่างไร?

เลือกตามความยาก ปริมาณงาน และความเสียหายหากตอบผิด

 LUNA เร็วที่สุดและประหยัดที่สุด  ราคา (ต่อ 1M tokens) Input \$1 / Output \$6  Benchmark เด่น - Coding Agent Index: 74.6 - Agents' Last Exam: 50.3% - Terminal-Bench: 84.7%  เหมาะกับ - จัดหมวดหมู่ข้อมูล - Labeling และ Routing - สรุปสั้น - ตอบคำถามง่าย - Batch Processing - Chatbot ปริมาณมาก  ข้อจำกัด Long Context อ่อน MRCR 8-needle: 41.3%  เปรียบเหมือน พยาบาลมือดี เร็วและแม่นยำ แต่เคสยากควรส่งต่อ	 TERRA สมดุลระหว่างคุณภาพและราคา ★ Default ที่แนะนำ  ราคา (ต่อ 1M tokens) Input \$2.50 / Output \$15  Benchmark เด่น - Coding Agent Index: 77.4 - Terminal-Bench 2.1: 87.4% - DeepSWE: 69.6%  เหมาะกับ - เขียนบทความ - Research ทั่วไป - Coding ปกติ - Feature ใหม่ - แก้ Bug ไม่ซับซ้อน - วางแผนและ Brainstorm  ข้อสังเกต งานหลายประเภทที่เคยใช้ GPT-5.5 สามารถทำได้ดีบน Terra ในราคาประมาณครึ่งหนึ่ง  เปรียบเหมือน หมอกว้ไปทีเก่ง รับงานได้หลากหลาย ส่งต่อเฉพาะเคสยาก	 SOL Flagship สำหรับงานที่ยากที่สุด  ราคา (ต่อ 1M tokens) Input \$5 / Output \$30  Benchmark เด่น - Coding Agent Index: 80 - Agents' Last Exam: 52.7% - BrowseComp: 90.4% - OSWorld: 62.6%  เหมาะกับ - Debug และ Refactor ซับซ้อน - Architecture - Security Review - Cybersecurity Analysis - วิจัยเชิงลึก - วิเคราะห์เอกสารยาว - งาน Agentic หลายขั้นตอน - การตัดสินใจสำคัญ  ข้อจำกัด แพงที่สุด และช้ากว่า Terra / Luna  เปรียบเหมือน ศัลยแพทย์ผู้เชี่ยวชาญ เรียกใช้เมื่อเป็นเคสยากจริง
--	---	---

เปรียบเทียบราคา (ต่อ 1M tokens)



LUNA

\$6

Output

TERRA

\$15

Output

SOL

\$30

Output

 Luna ถูกกว่า Sol **5 เท่า**

 Luna = ปริมาณมาก

 Terra = งานหลัก

 Sol = งานยากและเสี่ยงสูง

6 Reasoning Modes เลือกความลึกในการคิด

Model เดิม แต่เพิ่มหรือลดความพยายามในการใช้เหตุผล

**FASTER**

เร็วกว่า • ถูกกว่า

**SMARTER**

คิดลึกกว่า • ช่างกว่า • ต้นทุนสูงกว่า

noneตอบทันที
แทบไม่ใช้การคิดเพิ่ม

ตัวอย่าง

- แปลประโยคสั้น
- จัดรูปแบบข้อความ
- คำถามตรงไปตรงมา

lowคิดเล็กน้อย
เร็วมาก

ตัวอย่าง

- สรุปสั้น
- Email ง่าย
- แก้ข้อความเล็กน้อย

mediumสมดุล
เหมาะสมเป็นค่าเริ่มต้น

★ Recommended Default

ตัวอย่าง

- เขียนบทความ
- Coding ทั่วไป
- วางแผนงาน
- ตอบคำถามวิเคราะห์ทั่วไป

high

วิเคราะห์หลายด้าน

ตัวอย่าง

- วิเคราะห์รายงาน
- วางแผนซับซ้อน
- Debug ที่ต้องไล่เหตุผล

xhighคิดลึกและตรวจสอบ
หลายรอบ

ตัวอย่าง

- Code Review สำคัญ
- วิเคราะห์ความเสี่ยง
- Security Review

maxคิดลึกสุด
ช้าที่สุดและแพงที่สุด

ตัวอย่าง

- Architecture สำคัญ
- Production Incident
- งานที่ผิดพลาดไม่ได้



ความลึกในการคิด



เวลาในการทำงาน



ต้นทุนโดยประมาณ



คำเตือน

Mode สูงไม่ได้แปลว่าคุ้มกว่าทุกงาน
งานง่ายอาจไม่ได้คุณภาพเพิ่มมากพอ
กับต้นทุนที่สูงขึ้น

Luna + none

≈ 0.71
cents/task

VS

Sol + max

≈ 48.55
cents/task

ตัวเลขเป็นผลทดสอบต่อ Task ที่ยกมาเพื่อให้เห็นความต่างของต้นทุน

เลือก Model + Mode ให้เหมาะกับงาน

พิจารณา 2 เรื่อง: งานยากแค่ไหน และถ้าตอบผิดเสียหายเพียงใด

ถ้า AI ตอบผิด
จะเสียหายแค่ไหน?

เสียหายน้อย
→ LUNA

→

เสียหายปานกลาง
→ TERRA

→

เสียหายมาก
→ SOL

ความต้องการ ในการทำงาน	งานง่าย	งานทั่วไป	งานซับซ้อนหรือเสี่ยงสูง
ต้องการเร็ว	Luna + none/low ตัวอย่าง: สรุปสั้น, แปล, จัดหมวด	Terra + low ตัวอย่าง: Draft, แก้ข้อความ, งานทั่วไปเร่งด่วน	Sol + low/medium ตัวอย่าง: คัดกรองปัญหาสำคัญเบื้องต้น
ต้องการสมดุล	Luna + medium ตัวอย่าง: Batch ที่ต้องการ คุณภาพสม่ำเสมอ	Terra + medium ตัวอย่าง: Coding, Research, Blog, Planning ★ จุดเริ่มต้นที่แนะนำ	Sol + medium/high ตัวอย่าง: Debug, Architecture Draft, วิเคราะห์เอกสารยาว
ต้องการ ตรวจสอบลึก	Terra + high หมายเหตุ: มักไม่จำเป็น เว้นแต่ข้อมูลมีความสำคัญ	Terra + high/xhigh ตัวอย่าง: Financial Analysis, Code Review สำคัญ	Sol + xhigh/max ตัวอย่าง: Security, Production, Architecture, Audit

ตัวอย่าง
งานจริง

1

Review PR ที่มี
Security Concern

SOL + xhigh/max

2

เขียน Feature ใหม่
จาก Spec

TERRA + medium/high

3

สรุป Email
50 ฉบับ

LUNA + low

4

Debug
Production Issue

SOL + high/xhigh

5

แปลเอกสาร
จำนวนมาก

LUNA + low/medium
หมายเหตุ: หากเอกสารยาว
และต้องรักษาริบท ควรแบ่งช่วง
หรือตรวจด้วย Terra

6

วางแผน Project
Roadmap

TERRA + medium/high

💡 เริ่มจากระดับที่ต่ำพอจะทำงานได้ แล้วเพิ่ม Model หรือ Mode เมื่อผลลัพธ์ยังไม่เพียงพอ

ใช้ GPT-5.6 แบบมือโปร: Route งานให้ถูกตัว

ใช้โมเดลราคาถูกทำงานส่วนใหญ่ แล้วส่งเฉพาะงานยากให้โมเดลที่เก่งกว่า

1 ROUTING PATTERN ทำงานเป็นทีม ใช้โมเดลให้ถูกขั้น



2 BONUS: ULTRA MODE — เฉพาะ SOL



Benchmark (ผลการทดสอบ)	เหมาะกับ
• Terminal-Bench: 88.8% → 91.9%	🤖 Review Code หลายไฟล์
• BrowseComp: 90.4% → 92.2%	🤖 Research หลายหัวข้อ
	🤖 Debug หลาย Component
	🤖 โปรเจกต์ใหญ่ที่แบ่งงานย่อยได้

⚠️ ข้อควรระวัง: ไม่จำเป็นสำหรับงานเล็ก หรืองานที่แบ่งเป็นหลายส่วนไม่ได้

ใช้งานได้ใน: ChatGPT Work • Codex • API Multi-agent

3 CHEAT SHEET สรุปไว้จำง่าย

MODEL — ดูที่ความยาก		MODE — ดูที่ความลึก	
🚀 LUNA	งานง่าย งานซ้ำ ปริมาณมาก	🕒 none / low	เร็วและง่าย
⚖️ TERRA	งานทั่วไปและงานหลัก	⚖️ medium	สมดุล ใช้เป็น Default
👑 SOL	งานยาก เสี่ยงสูง และซับซ้อนสูงไม่ได้	🔍 high / xhigh	วิเคราะห์ลึกและตรวจสอบมากขึ้น
		🔊 max	ใช้เฉพาะงานสำคัญที่สุด

👤 เลือก **Model** = เลือกใครทำงาน

🧠 เลือก **Mode** = เลือกให้คิดลึกแค่ไหน

💡 เริ่มที่ **Terra + Medium**

ลดลงเมื่องานง่าย และเพิ่มขึ้นเมื่องานยาก

ใช้ GPT-5.6 แบบมือโปร: Route งานให้ถูกตัว

ใช้โมเดลราคาถูกทำงานส่วนใหญ่ แล้วส่งเฉพาะงานยากให้โมเดลที่เก่งกว่า

1 เลือก Model — เลือกตามความยากและความเสียหายหากตอบผิด



LUNA

เร็วที่สุดและประหยัดที่สุด

- ✓ งานง่าย
- ✓ งานซ้ำ
- ✓ ปริมาณมาก
- ✓ Batch Processing

เหมือน “เลขา”
ช่วยจัดการงานจำนวนมาก
อย่างรวดเร็วและเป็นระบบ



TERRA

★ Default ที่แนะนำ
สมดุลระหว่างคุณภาพและราคา

- ✓ งานทั่วไปและงานหลัก
- ✓ Coding ปกติ
- ✓ Research
- ✓ Planning / Brainstorm
- ✓ Draft คำตอบ

เหมือน “ผู้จัดการ”
รับผิดชอบงานหลักได้หลากหลาย
คุ้มค่าและเชื่อถือได้



SOL

รุ่นเรือธง เก่งที่สุด
สำหรับงานสำคัญ

- ✓ งานซับซ้อน
- ✓ Security / Cybersecurity
- ✓ Architecture
- ✓ Production / Incident
- ✓ การตัดสินใจสำคัญ

เหมือน “ผู้เชี่ยวชาญ”
จัดการงานยากและเสี่ยงสูง
ให้ผลลัพธ์ที่แม่นยำที่สุด

🎯 หลักคิด: งานยิ่งยาก หรือถ้าตอบผิดยิ่งเสียหาย เลือก Model ระดับที่สูงขึ้น

3 เลือก Reasoning — เลือกระดับความลึกในการคิด



2 เลือก Product — แต่ละผลิตภัณฑ์ใช้โมเดลและโควตาไม่เหมือนกัน

	Chat (สำหรับการคุยตามตอบ)	Work (สำหรับทำงานเป็นชิ้นงาน)	</> Codex (สำหรับนักพัฒนาและทีม)
Model ที่เลือกใช้ได้	ใช้ได้เฉพาะ GPT-5.6 Sol	เลือกได้ครบ Sol • Terra • Luna	เลือกได้ครบ Sol • Terra • Luna
Reasoning (ระดับการใช้เหตุผล)	Medium – Pro (ตามแผนที่รองรับ)	none – max (ปรับระดับได้)	none – max (ปรับระดับได้)
Ultra Mode	— (ไม่มี)	มี (เฉพาะแผน Pro และ Enterprise)	มี (ตั้งแต่แผน Plus ขึ้นไป)
โควตาที่ใช้	ไม่ใช้ Agentic Pool (แยกจากระบบอื่น)	ใช้ Agentic Pool ก่อนเดียวกัน	ใช้ Agentic Pool ก่อนเดียวกัน
หน่วยนับโควตา (Calculation)	นับเป็น “ข้อความ” (messages)	นับเป็น “tokens” / credits	นับเป็น “tokens” / credits

✓ Chat ไม่กิน Agentic Pool
โควตาการคุยแยกจากระบบ Agentic

🗄️ Work และ Codex ใช้ Pool เดียวกัน
ใช้หนักในฝ่ายหนึ่ง อีกฝ่ายก็เหลือน้อยลง

4 เลือก Product ให้เหมาะกับรูปแบบงาน



Chat

(สำหรับการคุยตามตอบ)

- ✓ ถามคำถาม / ขอคำแนะนำ
- ✓ Brainstorm / ระดมไอเดีย
- ✓ วิเคราะห์สั้น ๆ
- ✓ ปรัชญา / ขอคำอธิบาย
- ✓ งานทั่วไปที่ไม่มีความต่อเนื่อง

เร็ว ตอบไว ใช้จ่าย ไม่กิน Agentic Pool

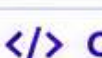


Work

(สำหรับทำงานเป็นชิ้นงาน)

- ✓ สร้างเอกสาร / รายงาน
- ✓ Spreadsheet / Data Analysis
- ✓ งานนำเสนอ / Slides
- ✓ Research เชิงลึก
- ✓ สร้างเว็บไซต์ / Dashboard (Sites beta)
- ✓ ดึงข้อมูลจากหลายแอป
และทำงานหลายขั้นตอน

ทำงานต่อเนื่อง ส่งมอบชิ้นงานสำเร็จ



</> Codex

(สำหรับนักพัฒนาและทีม)

- ✓ เขียน Code / แก้ Bug
- ✓ Refactor / Review Code
- ✓ ทำงานใน Repository
- ✓ Automate / Script
- ✓ สร้างเครื่องมือและระบบ
- ✓ Agent สำหรับ Dev Workflow

เหมาะกับงานโค้ดและเวิร์กโฟลว์
ของนักพัฒนา



MODEL เลือก “ใครทำงาน”

- LUNA = งานง่าย งานซ้ำ ปริมาณมาก
- TERRA = งานทั่วไปและงานหลัก (Default)
- SOL = งานยาก เสี่ยงสูง ผิดพลาดไม่ได้



REASONING เลือก “ให้คิดแค่ไหน”

- none / low = เร็วและง่าย
- medium = สมดุล ใช้เป็น Default
- high / xhigh = วิเคราะห์ลึกและตรวจสอบมากขึ้น
- max = ใช้เฉพาะงานสำคัญที่สุด



PRODUCT เลือก “ทำงานที่ไหน/แบบไหน”

- Chat = คุยตามตอบ เร็ว ไม่กิน Agentic Pool
- Work = ทำงานเป็นชิ้นงาน ใช้ Agentic Pool
- Codex = สบายพัฒนาและทีมวิศวกร ใช้ Agentic Pool



เริ่มที่ Terra + Medium
ลดลงเมื่องานง่าย
และเพิ่มขึ้นเมื่องานยาก