Senco masslink

# Local LLM AI Platform for Government – Solution Overview （政府專用本地 LLM 平台方案概覽）

# Executive Summary （概要）

### Secure On-Premises AI:

A secure, cost-effective generative AI platform deployed on-premises to support public services, with no cloud dependencies. Developed by Octopus InfoTech, it is tailored for Hong Kong government use and keeps all data in-house.

### Advanced Bilingual LLMs:

Runs state-of-the-art large language models (e.g. DeepSeek-R1 32B model, Alibaba Qwen series) on Apple Silicon hardware within government data centers. These models are fine-tuned on local datasets to provide high-accuracy bilingual (English and Chinese) responses aligned with government terminology.

### Privacy & Compliance by Design:

All AI processing stays under government control, fully aligning with Hong Kong's Personal Data (Privacy) Ordinance (PDPO) and internal IT security policies (no sensitive data ever leaves the premises). This ensures public data and confidential information remain protected.

### Cost-Effective & Self-Sufficient

Eliminates recurring cloud API fees: the platform runs on one-time purchased hardware (Apple Mac with M3 Ultra) with no per-use charges. Departments avoid expensive GPU servers or cloud costs, achieving a lower total cost of ownership (TCO) and predictable budgeting for AI services.

### Empowering Government Innovation:

Enables civil servants to leverage AI for drafting documents, answering queries, and analyzing data without compromising security

# Technical Architecture（技術架構）

### Apple Silicon M3 Ultra Hardware

Deployed on Apple's latest **M3 Ultra** chip systems (e.g. Mac Studio) offering up to **512 GB unified memory**, allowing extremely large models to load entirely in-memory with **no GPU bottlenecks**. This unified memory architecture lets the platform run models on the order of 600 billion parameters directly on-device, outperforming traditional PC/GPU setups.

### Optimized MLX Framework

Leverages Apple's open-source **MLX** machine learning framework, which is highly optimized for the Apple Silicon architecture. This enables near-GPU speeds for LLM inference on the CPU/Neural Engine, allowing fast generation and the handling of long government documents without performance degradation.

### Fully Local LLM Deployment

Hosts fine-tuned LLMs entirely within the government's infrastructure, **eliminating external APIs or cloud services**. Models like DeepSeek-R1 and Qwen are run locally, avoiding network latency and ensuring customization to internal guidelines. A typical setup uses a Mac Studio M3 Ultra as an AI server, harnessing its high memory and neural engines for efficient local inference.

### Retrieval-Augmented Generation (RAG)

The platform integrates a proprietary **RAG module** to boost accuracy. For each query, the system first retrieves relevant information from approved internal data sources (knowledge bases, policy docs, past Q&As, etc.), then the LLM generates an answer **grounded in those references**. This architecture ensures responses are up-to-date and context-specific to government data, dramatically reducing hallucinations.

### Factual Answers with Context

By augmenting the LLM with real-time retrieval, the system delivers **precise and reliable answers**. It can incorporate the latest policy updates or regulations in responses. The RAG approach also enables answers to be accompanied by source citations or excerpts for transparency, which is vital for building trust in AI outputs in a public sector setting.

# Key Features and Advantages （主要功能和優勢）

### Privacy & Data Security

- 100% **on-premises deployment**: all data and AI processing remain within government IT environments.
- **No sensitive data ever leaves** the government network, ensuring compliance with PDPO and eliminating cloud-related privacy risks.

### Auditability

- Comprehensive **audit logging** of AI interactions is built-in. Every query and response can be recorded for oversight.
- The system can even provide citations for its answers, enabling officials to verify sources and maintain accountability for how information is generated.

### Low Total Cost of Ownership

- Utilizes existing hardware and one-time purchases instead of pay-as-you-go cloud fees. Running LLMs on an Apple M3 Ultra workstation avoids costly GPU server farms or API subscriptions.
- There are **no per-query fees**, yielding significant long-term savings and a flat, predictable cost model for budgeting.

### Bilingual Support

- Natively supports **English and Chinese** interactions. The models are fine-tuned for bilingual output, so the AI can seamlessly understand and respond in both languages.
- This is crucial in Hong Kong's bilingual environment – the platform can draft an internal memo in English or answer a citizen's query in Cantonese with equal proficiency.

### Customizability & Control

- The government fully controls the platform's behavior: models can be further tuned to departmental language style or integrated with specific data sources.
- All generated content and model updates remain under **government ownership** (data sovereignty), preventing vendor lock-in and allowing alignment with internal policies.

# Suggested Use Cases for Department

**Efficiency Unit (PICO) -** ***Document drafting assistant and internal Q&A***

The LLM helps civil servants **draft memos, reports, or guidelines** by generating first drafts from outlines.

Serves as an interactive knowledge base

- Staff can query thousands of pages of internal manuals or past studies in natural language and get concise, accurate answers. This boosts productivity in policy research and report preparation.

**Home Affairs Dept (HAD) & LCSD -** ***Bilingual citizen chatbot***

Deploy a public-facing **chatbot** to handle common inquiries about community services, facility bookings, event registrations, etc.

- The chatbot (powered by the secure LLM with RAG) provides polite and accurate answers in **both English and Chinese**.
- Operates 24/7, extending service hours and relieving hotline staff while maintaining consistent information quality to the public.

**DSD / EMSD / WSD -** ***Technical report summarization***

Engineers in infrastructure departments produce lengthy technical reports.

- **Ingest a long report and auto-generate an executive summary or highlight key issues** for management.
  - Users can ask specific questions (e.g. “What caused the power outage in the July report?”) and;
  - AI will pinpoint the relevant section in the report and provide an answer.
  - This augments analysts’ capabilities and ensures critical information isn’t overlooked in voluminous documents.

**Social Welfare Dept (SWD) –** ***Case processing and hotline automation***

The AI assistant can extract structured data from **application forms and case notes**, and draft summaries or recommendation letters for social workers.

- It can also power a hotline chatbot to answer routine questions about welfare schemes, required documents, application status, etc., triaging simple inquiries with instant answers and forwarding complex cases to human officers.
- This improves public response times and frees up staff for high-value case work.

# Security and Compliance （安全和合規）

- **On-Premises Only**
  - The entire solution is deployed within **government-managed data centers. No data ever leaves the** premises
    - Unlike cloud AI services, all model processing and data storage happen on local servers.
    - Eliminates external data exposure risks and ensures sensitive information always stays under government control.
- **PDPO Compliance**
  - The on-prem architecture inherently aligns with the Personal Data (Privacy) Ordinance.
    - **Minimizes external data transfers** and enforces rigorous protection of personal data.
    - Departments can confidently use the AI even with confidential or personal datasets, knowing they remain in a controlled environment.
- **Integration with Gov IT Security**
  - The platform is designed to meet government IT security requirements. It can be hosted behind existing firewalls and uses secure network segmentation and access controls.
  - No new external connectivity is introduced, simplifying security audits. The solution conforms to internal security policies, so adopting it does not compromise the agency’s network integrity.
- **Audit Logging & Controls**
  - Every AI query and response can be **logged for audit**. Administrators have full oversight: they can review interaction logs to trace outputs and ensure the AI is being used appropriately.
  - Responses can include citations to reference documents, adding an extra layer of transparency for compliance and decision-making audits.
- **Data Sovereignty**
  - The government retains **full ownership of models, data, and outputs**.
  - All fine-tuned models and any new data used stay within the department’s control.
  - There is no dependence on foreign cloud vendors or external AI APIs once deployed, reducing geopolitical and supply-chain risks.
  - This sovereignty over data and AI technology aligns with public sector governance and strategy.

# Deployment Options and Pricing （部署選項及定價）

- **Flexible Deployment** – The solution can be deployed entirely on-premises on **Apple Silicon** hardware (e.g. a Mac Studio with M3 Ultra) or provided as a secure appliance in the government network. For evaluation, Octopus can also host a pilot in a **secure cloud sandbox** or supply a pre-configured device for on-site trial. This allows departments to choose on-prem, hybrid, or trial setups according to their needs.
- **Hardware & Performance** – In full deployments, a high-memory Apple Silicon system (up to 512 GB RAM) is recommended to handle large models with low latency. Thanks to the unified memory, no external GPU is required for even very large models. For lighter workloads or initial pilots, the system can run on lower-spec Macs or even standard PCs using optimized models, making it accessible to smaller departments or schools.
- **Licensing Models** – Octopus InfoTech offers **perpetual licenses** (one-time purchase with maintenance) or **annual subscriptions** for the platform, depending on what suits the department. Licensing can be structured per deployment or per user. Volume discounts and government enterprise agreements are available for large-scale rollouts. This flexibility allows agencies to start small and scale up without penalty.
- **Typical Costs** – **One-time deployment** (including hardware & software setup) for the core LLM platform is roughly **HK$200k–$500k** depending on model size and hardware configuration. Subscription-based modules have annual costs in the range of **HK$100k–$300k** for an enterprise-scale chatbot or around **HK$50k–$150k** for department-wide writing assistant tools (scaling with number of users). These indicative costs are far lower than equivalent cloud AI usage over time, given no per-query fees.
- **Support & Maintenance** – The package includes on-site installation, configuration, and training by Octopus's team. Ongoing support, security patches, and model updates are provided throughout the subscription or maintenance period. Octopus will work closely with the government IT teams to ensure smooth integration with existing systems and to address any issues promptly under agreed SLAs.

# Pilot and Next Steps （試點及後續）

- **Pilot-Ready Solution** – The Local LLM AI Platform is **ready for immediate pilot deployment**. All core components (LLM servers, MLX integration, RAG pipeline, web front-end and analytics dashboard) have been fully developed and tested in similar environments. This means a pilot can focus on configuration and use-case tuning rather than core development.
- **Quick Trial Setup** – A proof-of-concept **pilot trial** can typically be set up within **1–2 weeks**. Octopus InfoTech will provide the necessary high-memory Apple hardware (or a pre-loaded appliance) and install the platform on the department's secure network. If the department lacks hardware, Octopus can loan a configured Mac Studio for the trial. This fast setup allows stakeholders to evaluate the AI using their own data quickly and with minimal upfront cost.
- **Real-Data Demonstration** – Octopus proposes conducting the pilot in collaboration with a government partner (e.g. the Efficiency Office or OGCIO's Smart LAB). The pilot would use **actual departmental documents and queries** (under proper confidentiality agreements) to demonstrate how the AI handles real-world tasks. This approach ensures the evaluation is relevant to the department's needs and builds confidence in the solution's practical value.
- **Training & Knowledge Transfer** – During the pilot, Octopus's engineers will work closely with the department's IT and user teams. They will provide on-site **training sessions, user tutorials, and documentation** so that civil servants learn to use the chat interface, document assistant features, etc. effectively. Feedback from users will be gathered to fine-tune the AI's responses (e.g. tone, terminology) to fit government communication standards.
- **Roadmap to Full Deployment** – Following a successful pilot, Octopus will assist in developing a full deployment plan. Next steps typically include evaluating pilot results and ROI, incorporating any departmental-specific customizations, preparing production hardware, and finalizing licensing. The solution supports a **phased rollout** – for example, starting in one unit and expanding to others – to ensure smooth adoption. Octopus will remain engaged throughout to ensure the platform delivers ongoing value and aligns with the Hong Kong government's smart governance objectives.

# Solution Package Summary（方案套件概覽）

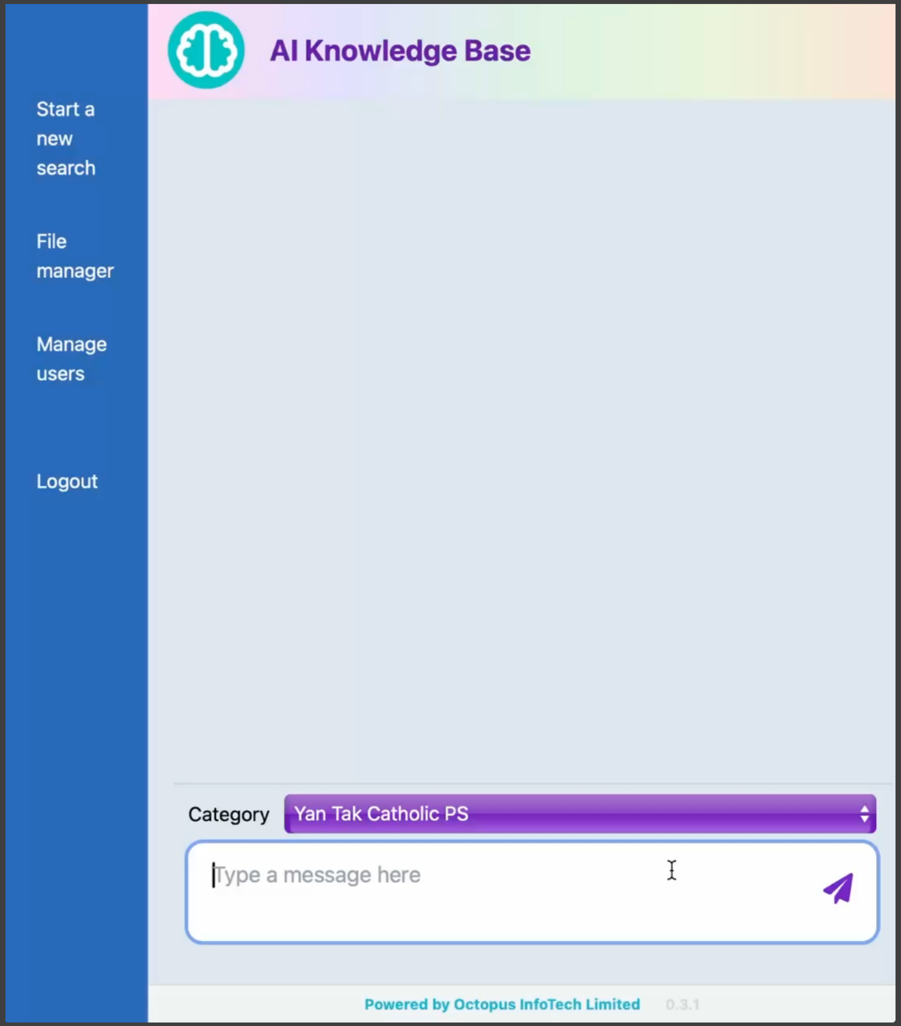
# Local LLM AI Platform (本地 LLM 人工智能平台)

- An on-premises large language model platform that runs on **Apple Silicon (M3 Ultra) with up to 512 GB of unified memory**.
- It supports advanced models such as **DeepSeek-R1 and Qwen**, operating **fully offline** to enable **secure internal document Q&A, long-form summarization, and policy research analysis**.
- All data is **kept within government networks**, and **comprehensive audit logs** are maintained to ensure compliance.

Octopus InfoTech Local EDB LLM Platform
Talk to Octopus InfoTech Local EDB LLM Platform

# AI Chatbot for Customer Service （客戶服務人工智能聊天機械人）

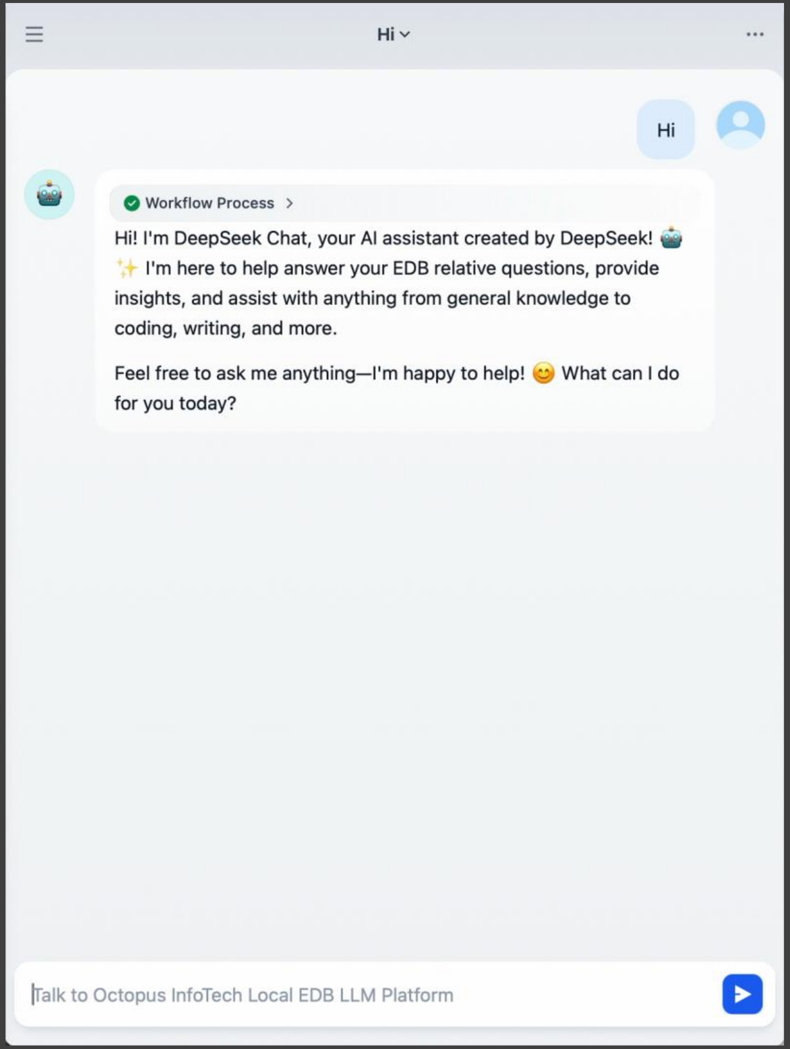
- A multilingual **chatbot** system for handling public or customer inquiries. It combines fine-tuned LLMs with a retrieval backend to answer questions in Cantonese, English, or other languages by referencing up-to-date internal databases.
- Deployable on websites, mobile apps, or kiosks, it provides **instant 24/7 responses** to FAQs, appointment queries, status checks, etc., **improving citizen engagement** and relieving hotline staff from routine Q&A.
- The knowledge base is customizable per department, ensuring accurate and policy-aligned



Category: Yan Tak Catholic PS

# AI Document Processing & Writing Assistant （人工智能文件處理及寫作助手）

- An AI assistant to **streamline document-heavy workflows**. It helps officers draft, summarize, and proofread official documents – from meeting minutes and memos to public notices and detailed reports.
- Using models fine-tuned for formal government writing, it can **summarize long texts, extract key points, and suggest rewrites** in the appropriate tone or even translate jargon into plain language.
- It works as a desktop app or MS Word add-in, making it easy to use while writing. This assistant boosts productivity, ensures consistency (including bilingual document alignment), and reduces human error in government communications.

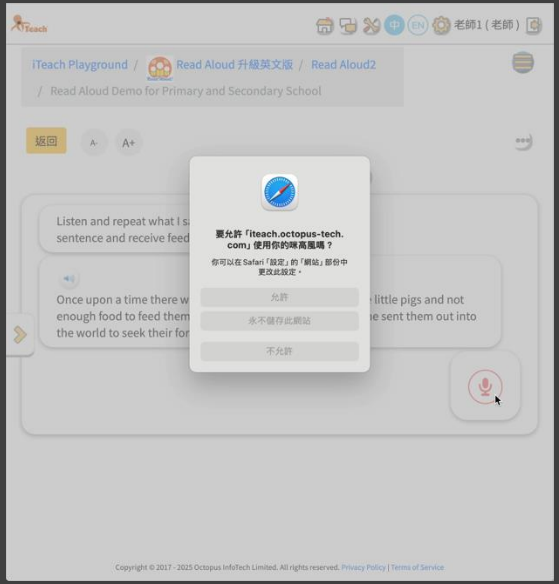


Hi! I'm DeepSeek Chat, your AI assistant created by DeepSeek! 🤖 ✨ I'm here to help answer your EDB relative questions, provide insights, and assist with anything from general knowledge to coding, writing, and more.

Feel free to ask me anything—I'm happy to help! 😊 What can I do for you today?

# Education-Focused AI Suite （教育專用人工智能套件）

- A comprehensive suite of AI tools for the education sector, aimed at enhancing teaching and learning. It includes **AI Essay Marking** (automatically grades student essays in English or Chinese with detailed feedback), **AI Read-Aloud Scoring** (evaluates students' spoken reading and gives immediate pronunciation feedback), and **AI Q&A Explanation Tutor** (guides students through problem-solving with step-by-step hints).
- Over **500 schools in Hong Kong** have deployed elements of this suite with support from the Education Bureau. In a QEF-funded pilot, secondary teachers using the AI Essay Marking tool cut grading time by >50% while providing richer feedback. Primary schools using the read-aloud tool saw improved reading practice engagement, and an after-school math tutor pilot increased homework completion rates and student confidence.
- This suite demonstrates how AI can **reduce teachers' workload and personalize student feedback**, aligning with Hong Kong's smart education initiatives.


iTeach
中
EN
老師1 ( 老師 )
iTeach Playground / Read Aloud 升級英文版 / Read Aloud2
/ Read Aloud Demo for Primary and Secondary School
返回
A-
A+

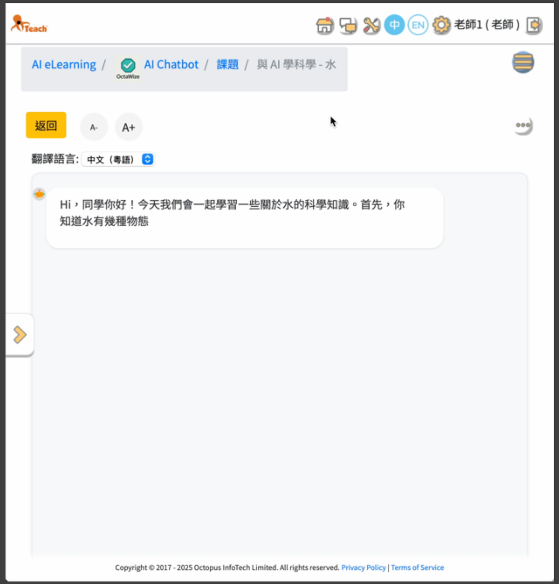
Listen and repeat what I s… sentence and receive feed…

Once upon a time there w… little pigs and not enough food to feed them… e sent them out into the world to seek their for…

要允許「iteach.octopus-tech.com」使用你的咪高風嗎？

你可以在 Safari「設定」的「網站」部份中更改此設定。

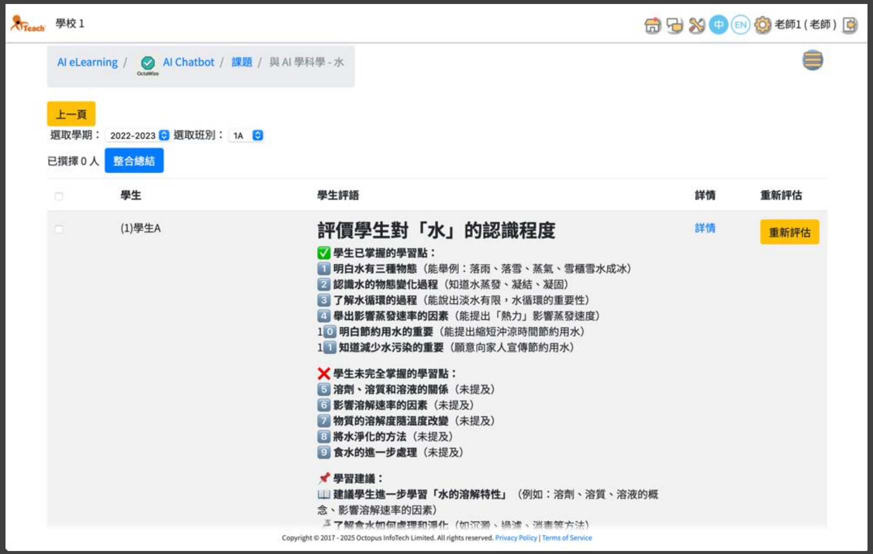
Copyright © 2017 - 2025 Octopus InfoTech Limited. All rights reserved. Privacy Policy | Terms of Service


iTeach
中
EN
老師1 ( 老師 )
AI eLearning / OctaWize AI Chatbot / 課題 / 與 AI 學科學 - 水
返回
A-
A+

翻譯語言: 中文（粵語）

Hi，同學你好！今天我們會一起學習一些關於水的科學知識。首先，你知道水有幾種物態

Copyright © 2017 - 2025 Octopus InfoTech Limited. All rights reserved. Privacy Policy | Terms of Service


iTeach
學校 1
中
EN
老師1 ( 老師 )
AI eLearning / OctaWize AI Chatbot / 課題 / 與 AI 學科學 - 水
上一頁

選取學期： 2022-2023 選取班別： 1A

已揀擇 0 人 整合總結

| □ | 學生 | 學生評語 | 詳情 | 重新評估 |
|---|---|---|---|---|
| □ | (1)學生A | **評價學生對「水」的認識程度**<br>✅ **學生已掌握的學習點：**<br>1️⃣ **明白水有三種物態**（能舉例：落雨、落雪、蒸氣、雪櫃雪水成冰）<br>2️⃣ **認識水的物態變化過程**（知道水蒸發、凝結、凝固）<br>3️⃣ **了解水循環的過程**（能說出淡水有限，水循環的重要性）<br>4️⃣ **舉出影響蒸發速率的因素**（能提出「熱力」影響蒸發速度）<br>10️⃣ **明白節約用水的重要**（能提出縮短沖涼時間節約用水）<br>11️⃣ **知道減少水污染的重要**（願意向家人宣傳節約用水）<br>❌ **學生未完全掌握的學習點：**<br>5️⃣ **溶劑、溶質和溶液的關係**（未提及）<br>6️⃣ **影響溶解速率的因素**（未提及）<br>7️⃣ **物質的溶解度隨溫度改變**（未提及）<br>8️⃣ **將水淨化的方法**（未提及）<br>9️⃣ **食水的進一步處理**（未提及）<br>📌 **學習建議：**<br>📖 **建議學生進一步學習「水的溶解特性」**（例如：溶劑、溶質、溶液的概念、影響溶解速率的因素）<br>⚗ 了解食水如何處理和淨化（如沉澱、過濾、消毒等方法） | 詳情 | 重新評估 |

Copyright © 2017 - 2025 Octopus InfoTech Limited. All rights reserved. Privacy Policy | Terms of Service