CONSULTANCY DONE DIFFERENTLY — CHAINING WORKFLOWS · TOOL 6 OF 7 · 10-WEEK MODULE 5+

Qwen Setup Guide

Run AI Prompts Locally via LM Studio · Your Data Stays On Your Laptop · Windows 10 / 11

■ THE USAGE RULE

■ Sensitive data (names, financials, competitive strategy)? → Run in Qwen (LM Studio).

■■ Want better layouts or complex reasoning? → Use Claude or ChatGPT.

What It Is — Qwen Running Locally via LM Studio — Free and Offline

Qwen is an AI language model made by Alibaba. LM Studio is a free Windows application that downloads Qwen 3.6 (or later) to your laptop and runs it completely offline. When you use Qwen this way, your text never leaves your device. No cloud account, no subscription, no data on anyone else's server.

■ Stays on your laptop

No internet needed to run prompts.
Outputs never sent online.
Safe for financial data,
prospect names, strategy.

■ Completely free

LM Studio is free.
Qwen 3.6 or later is free.
No token limits or monthly bill.
Runs on standard Windows laptops.

When to Use It — The Four Qwen-Only Workflows

B3	**Grant Budget Narrative**	— Actual project costs, labour rates, and margin structure
E1	**First Contact Message**	— Names a specific prospect — third-party personal data
E2	**Discovery Call Prep**	— Sales strategy and client intelligence for a named person
A1-QWEN	**Tender Weakness Analysis**	— Competitive weaknesses — strategic intelligence

■ Also works for any prompt that does not need internet access

Qwen can run any AI prompt that works without web search. The four workflows above are mandatory for privacy. Everything else is your choice.

Before You Start

System Requirements — Windows 10 / 11

Operating System Windows 10 or 11 (64-bit)
Most business laptops from 2018 onwards qualify. To check: right-click the Start button → System → look for "Edition" under Windows specifications.

RAM (Memory) 16 GB recommended — 8 GB works but Qwen will respond more slowly
To check: right-click the Start button → System → scroll to "Device specifications" and look for "Installed RAM". Close all other programs before running Qwen to free up memory.

Storage Space 6–8 GB free disk space needed
Required for the LM Studio application (~350 MB) and the Qwen 3.6 or later model download (~4–6 GB). To check: open File Explorer → click "This PC" → look under "Devices and drives".

Internet access Required once for setup only
Internet is needed to download LM Studio and the Qwen model once. After that, Qwen runs fully offline — no internet connection is needed to run any of your workflows.

Steps 1–4

Download, Install, and Load Qwen

1

Download LM Studio

Open your browser and go to lmstudio.ai. Click the Windows download button (version 0.4.16 or later). The installer is around 350 MB. Save it to your Downloads folder.

2

Install LM Studio

Double-click the downloaded file. If Windows shows "Do you want to allow this app to make changes?" click Yes. Click Next → Next → Install → Finish. LM Studio opens automatically.

3

Find and load Qwen 3.6 or later

In LM Studio click the Search icon (magnifying glass) in the left sidebar. Type Qwen3.6 and press Enter. Look for Qwen 3.6 or a later version. Click Download next to the Q4 or Q5 version (~4–6 GB). Once downloaded: click the Chat icon → "Select a model to load" → choose the Qwen model. Wait for the green Ready indicator.

4

Check for updates

In LM Studio click the Settings icon (gear) in the left sidebar. Look for a "Check for Updates" or "About" section. If an update is available, click Update and let it install before running your first workflow. This keeps your version current.

Steps 5–8

Start a Chat and Run Your First Workflow

5 Start a New Chat

In LM Studio click the Chat icon in the left sidebar. Click "+ New Chat" at the top left. A blank chat window opens. Start a fresh New Chat for each new workflow — do not reuse old conversations from previous sessions.

6 Open the Chaining Workflows tool and paste your context

Go to consultancydd.com/tools/chaining-workflows/ in your browser. Paste your context blocks (DNA, Persona, VP, Brand Identity) into Zone 1. In Zone 3 — Local Workflows — click the workflow card you need (B3, E1, E2, or A1-QWEN).

7 Copy and paste Step 1

Click "Copy Step 1" in the tool. Switch to LM Studio. Click in the chat input box at the bottom. Press Ctrl+V to paste the prompt. Press Enter or click the Send button. Wait for Qwen to finish — the cursor stops blinking when the response is complete.

8 Continue Steps 2 and 3 in the SAME chat

Stay in the same conversation window. Copy Step 2 from the tool, paste into LM Studio, press Enter. Repeat for Step 3. Each step builds on what Qwen learned in the previous response. Do NOT click "+ New Chat" between steps.

■ The Golden Rule — one session for all three steps

Copy Step 1 → paste into LM Studio → wait for response.
Copy Step 2 → paste into the SAME chat → wait for response.
Copy Step 3 → paste into the same chat → Step 3 is your final output.
The AI uses context it built in Steps 1 and 2. Starting a new chat breaks the chain.

Step 9 Save Your Results — Do This Every Time

Qwen does not permanently save your conversations. Copy your results to a Word document immediately after Step 3 finishes. This preserves formatting and keeps a record on your laptop.

① Click at the very start of Qwen's final response (beginning of the Step 3 output).

② Hold the left mouse button and drag to the end of the response — or press Ctrl+A to select all text in the chat.

③ Press Ctrl + C to copy the selected text.

④ Press the Windows key, type Word, press Enter to open Microsoft Word. Press Ctrl + N for a new blank document.

⑤ Press Ctrl + V to paste. The text appears with basic formatting preserved. Press Ctrl + S, name the file, and click Save.

Quick keys: Ctrl+C Copy | Ctrl+V Paste | Ctrl+N New doc | Ctrl+S Save | Ctrl+A Select all

■ Recommended filenames

B3_GrantBudget_[ProjectName]_2026.docx
E1_FirstContact_[CompanyName]_2026.docx
E2_DiscoveryCall_[CompanyName]_2026.docx
A1QWEN_WeaknessAnalysis_[TenderName]_2026.docx

Usage Comment When to Use LM Studio vs Cloud AI

Use LM Studio (Qwen)
B3, E1, E2, A1-QWEN workflows. Any prompt that contains prospect names, financial figures, project costs, margin rates, or competitive strategy. Privacy is the reason.

Use Claude or ChatGPT
A1, A2, A3, B1, B2, C1-C3, D1-D3, E3 workflows. Content that will be published or shared. Outputs where professional formatting and visual polish matter most.

Quick Reference — Workflow Summary + Troubleshooting

ID	Workflow	Why Qwen — not cloud	Step 1 input
B3	Grant Budget Narrative	Actual cost figures & margin rates	Grant + funding + eligible costs
E1	First Contact Message	Names an individual — personal data	Prospect name, role, company + reason
E2	Discovery Call Prep	Sales strategy — personal data	Company + contact name + call reason
A1-QWEN	Tender Weakness Analysis	Competitive intelligence	Paste A1 tender outline as input

If Things Go Wrong — Common Issues and Fixes

Qwen is responding very slowly
Your laptop may have less than 16 GB RAM. Close all other programs (browser, email, Teams) before running Qwen. The response may take 2–5 minutes instead of 30 seconds — it will still complete.

Model failed to load
Close LM Studio completely, reopen it, and try loading the model again. If it fails twice: Search icon → find Qwen 3.6 or later → delete → re-download. Check you have at least 6 GB free storage.

"Out of memory" error
The model is too large for your RAM. In LM Studio Search, look for "Qwen3 4B" for a lighter version that runs on 8 GB RAM. It will be faster and use less memory.

Response cuts off mid-sentence
In LM Studio chat settings (gear icon in the chat panel), increase the Max Tokens value to 4096 or higher. Then re-run your Step 3 prompt in the same session.

Cannot find the model in search
The model search needs internet. Check your connection. In LM Studio Settings confirm the model source is set to Hugging Face. After setup, all prompts run offline.

Fallback Option — GPT4All — A More Universal Alternative

If LM Studio does not run well on your laptop, GPT4All is a simpler alternative available free from nomic.ai/gpt4all

Supports many models
Runs a wide range of models — small (3B for 8 GB RAM) to large (13B+). You choose the model that suits your laptop's available memory.

One model at a time
GPT4All loads one model at a time. Switch models through the Models menu. Keeps memory use manageable on standard business laptops.

Same usage rule applies
Use GPT4All for B3, E1, E2, A1-QWEN — same reason as LM Studio. CDD prompts from the Chaining Workflows tool work in GPT4All without changes.

Need help? Contact your CDD consultant.
Raymond Ward: reward@consultancydd.com | 0418 186 294
Matthew Bulat: cto@consultancydd.com.au | 0407 320 726