

Local Model Setup Guide

Running the sensitive interview stages offline with LM Studio and Qwen

Your client's buyer names, pricing and competitive strategy never leave your machine. Same four stages, same dashboard, same exports — a different model doing the work for the two stages that matter most.

Field	Detail
Document	Local Model Setup Guide — V1.0
Date	7 September 2026
Applies to	Voice-First Client Onboarding Tool — expertaiprompts.info/voiceonboardingtool/
Read with	The Manual Capture Method (voice dictation guide) and the Operator's Field Instructions. This guide covers the model. Those cover the capture method and the session procedure.
Status of the local path	Recommended for Stages 2 and 3 where the interview will contain named prospects, real figures or competitive strategy. Not mandatory. The cloud path stays valid under the three-category rule in section 1.
Platform	Windows 10 or 11, 64-bit. Fully offline after the initial download.

1. The usage rule

One rule decides where a stage runs, and it is about the content of the answer rather than the stage number. If what the client is about to say would identify a real person, state a real number, or reveal how they win against a competitor, it belongs on a model that runs on your own machine.

The three categories that belong on the local model

Named third parties — real prospects, customers and competitors by name.

Financial figures — pricing, margins, contract values, revenue.

Competitive strategy — anything that would damage the client if a competitor read it.

These are the same three categories the Operator's Field Instructions tell you to keep out of a cloud assistant. This guide is the other half of that rule: instead of leaving the material out, you run the stage somewhere it can safely go in.

1.1 Where each stage runs

Stage	Where it runs	Why
Stage 1 — Personal Value Proposition	Cloud assistant is fine	Personal background, motivation and working style. Low sensitivity, and this is where a frontier model's output quality is most worth having.
Stage 2 — Customer Persona Deep Dive	Local recommended	The question set asks the client to describe their buyer in detail, which naturally elicits real prospects by name. This is the stage most likely to surface a category-one answer without anyone intending it.
Stage 3 — Business Value Proposition (Moats and Evidence)	Local recommended	Asks directly about measurable results, pricing evidence and competitive moats. Categories two and three, by design.
Stage 4 — Brand Identity	Cloud assistant is fine	Verbal tone and visual guardrails. Almost no sensitive content, and heavily rewarded by a stronger model.

Recommended, not mandatory — and why the distinction is honest

You can run all four stages on a cloud assistant and stay within the rules, provided the three categories stay out of the interview and you describe the shape instead: "a works manager at a regional council" rather than the person, "priced above market on a value argument" rather than the rate.

The local path exists because that abstraction is not always possible. When the artefact genuinely needs the specifics — a named account, a real contract value, the actual competitive position — running the stage locally is the option that lets you keep them without sending them anywhere.

What is not acceptable is the middle position: starting a stage locally and switching to a cloud assistant because the local model is slow or the answer is unsatisfying. That gives you the exposure and the paperwork saying you do not have it. Decide before the stage starts, and finish where you started.

2. What it is

Qwen is a large language model from Alibaba that can be downloaded and run entirely on your own computer. LM Studio is the application that downloads it, loads it and gives you a chat window to work in. Together they behave like a cloud AI assistant with one difference that matters here: after the initial download, nothing you type is transmitted anywhere. There is no account, no conversation history on a server, and no question about whether the content might be used to train anything.

It is the same stack Consultancy Done Differently already uses for its four privacy-restricted chaining workflows, so if you have set that up you have most of this already and only need sections 5 and 6.

2.1 What you give up

Being straight about this is part of using it well. A model small enough to run on ordinary hardware will produce a less polished artefact than a frontier cloud model. Expect flatter phrasing, occasional structural slips in long outputs, and less inference from sparse answers. It is a fair trade for Stages 2 and 3, where the value is in accurate capture of specifics rather than in eloquence, and a poor trade for Stages 1 and 4, which is exactly why the split in section 1.1 falls where it does.

3. Before you start — system requirements

Requirement	Detail
Operating system	Windows 10 or Windows 11, 64-bit.
Memory	16 GB RAM recommended. 8 GB is the practical minimum and will run a smaller quantisation more slowly.
Free storage	6 to 8 GB for the model file, plus room for LM Studio itself.
Internet	Needed once, to download LM Studio and the model. Not needed afterwards, and the workflow is designed to be run offline.
Time	Around thirty minutes for the first setup, most of it download time. Under a minute to start a session after that.

Do this before a client session, not during one

The download is large and the first model load takes a minute or two. Set this up on a quiet day, run one practice stage on your own material, and confirm the output is usable before you sit down with a client. A first-time install in front of a client is how a good tool acquires a bad reputation.

4. Install and load the model

Step 1 Download LM Studio

Go to lmstudio.ai and download the Windows installer. Take the current stable release rather than a beta.

Run the installer and accept the defaults. It installs like any ordinary Windows application and does not need administrator rights on most machines.

Step 2 Open LM Studio and find the model search

Launch LM Studio. The first screen offers a search or discover panel for models.

Search for "Qwen". You will see a list of variants with different sizes and quantisation levels.

Step 3 Choose the right variant

Take the most recent Qwen instruct model your machine can hold, at Q4 or Q5 quantisation.

Quantisation is a compression setting. Q4 is smaller and faster; Q5 is slightly larger and slightly better.

On 16 GB take Q5; on 8 GB take Q4.

If the listing shows a size larger than about 6 GB and you are on 8 GB of RAM, choose a smaller variant rather than hoping. A model that will not fit fails at load time or crawls.

Step 4 Download and load it

Click download and wait. This is the long part.

When it finishes, load the model from the chat screen. The first load takes a minute or two; subsequent loads are faster.

Confirm it works: type "Reply with the word ready and nothing else." If it replies, you are set up.

Versions move — check rather than assume

Model names, version numbers and quantisation labels change every few months. This guide deliberately describes what to look for rather than naming a specific build, so it does not date. The rule is: most recent Qwen instruct model, Q4 or Q5, that fits comfortably in your available memory. Verified against the available releases on 7 September 2026.

5. Capturing the interview when the model is local

The capture method does not change because the model changed. That is the whole point of it. The Manual Capture Method separates turning speech into text from turning text into an artefact, which means the model at the end is a swappable component and the Speak-Read-Fix-Send loop runs identically whether the corrected text goes to a cloud assistant or to Qwen on your own machine.

5.1 Dictation, when the assistant is offline

LM Studio has no built-in microphone. Use the operating system's dictation instead, which is where the transcript needs to appear anyway.

Option	How it works
Windows voice typing	Press Windows key + H in any text field, including LM Studio's. Windows has an offline dictation mode; turn it on in Settings so speech is processed on the device rather than sent to a service. This is the simplest option and needs nothing installed.
A local Whisper build	More accurate on proper nouns, industry jargon and numbers — which is exactly where the custom word list in the Manual Capture Method is aimed. Worth the setup if you run client interviews regularly. Runs entirely offline.
Type it	For short answers, or where a name has to be exactly right, typing is faster than dictating and correcting.

Whichever you use, the loop is unchanged: **speak, read what appeared on screen, fix any wrong name or figure, then send.** The transcript existing on screen before it is submitted is the control this whole system is built on, and it is the reason live voice mode is not used anywhere in this workflow. The full method, including the word list and the three-question answer check, is in the Manual Capture Method.

Build the word list first

Before the first stage, paste your client's word list into the chat: company names, people, places, industry terms and any figures that will come up. A local model at this size benefits from it more than a frontier model does, because it has seen less of the world and will guess harder at an unfamiliar proper noun. Five minutes here saves twenty in correction.

6. Running a stage

Step 5 Open the dashboard and the model side by side

The Voice-First Client Onboarding Tool page in your browser, LM Studio in the other half of the screen. Set the client name and company in the dashboard first, so exports are named correctly.

Step 6 Start a new chat for the stage

One stage, one chat. Do not carry Stage 2 and Stage 3 in the same conversation.

Paste the word list, then paste the stage prompt from the dashboard's Copy Prompt button.

Step 7 Run the interview

Work through the questions in order, using the dictation loop from section 5.

Do not interrupt the chain by starting an unrelated conversation in the same chat. The stage prompt sets up a context the model needs to hold for the whole stage.

Step 8 Copy the final output back into the dashboard

When the stage completes, copy the model's final structured output and paste it into the dashboard textarea for that stage.

Read it before you save. Check the names and the numbers against what was actually said. This is the last point at which an error is cheap to fix — everything after this stage takes this artefact as an input.

LM Studio does not keep your conversation

Close the chat and it is gone. There is no history to go back to and no autosave.

That means a stage has to be finished and copied out in one sitting. During a live client interview — where an interruption is far likelier than at your own desk — this is an operational risk, not a filing inconvenience. Copy the output into the dashboard as soon as the stage completes, before you take a break, answer the door, or move on to the next stage.

7. Saving, filing and retention

Step 9 Save the artefact

Click Save as Text File in the dashboard. The file auto-names as Company_Person_DocumentType_DateRun.txt and contains both the readable transcript and the structured JSON block.

It downloads to your usual Downloads folder.

Then	Do this
Move the file	Out of Downloads and into the client's own folder, as soon as the session ends. These files contain the client's value proposition, evidence and buyer detail in plain text and unencrypted.
Clear the dashboard	Click Clear All Data when you finish a client, not only when you start one. Clearing before a session protects the previous client; clearing after protects the current one. Do both.
Retention	Keep for the life of the engagement plus seven years, then destroy.
Browser storage	The dashboard keeps everything in your own browser and sends nothing to a server, which is a genuine privacy strength. It is also unencrypted and readable by anything with access to that browser profile — a good place for a working session and a poor place for anything to sit for a week.

8. Quick reference

Stage	Runs on	Sensitivity	Capture
Stage 1 — PVP	Cloud assistant	Low — personal background and working style	Dictation mode. Never live voice.
Stage 2 — Customer Persona	Local recommended	High — elicits real prospects by name	Dictation via Windows voice typing or local Whisper
Stage 3 — BVP, Moats and Evidence	Local recommended	High — pricing, margins, competitive position	Dictation via Windows voice typing or local Whisper
Stage 4 — Brand Identity	Cloud assistant	Low — tone and visual guardrails	Dictation mode. Never live voice.

9. If things go wrong

Symptom	What to do
Responses are very slow	The model is larger than your memory comfortably holds. Load a smaller variant or a lower quantisation — Q4 instead of Q5. Close other applications, particularly browsers with many tabs.

Symptom	What to do
The model will not load	Usually not enough free memory. Restart the machine, close everything else, and try again. If it still fails, the variant is too large for this hardware.
Output stops mid-sentence	The response hit a length limit. Ask it to continue, or increase the maximum response length in LM Studio's settings for that chat.
Names and figures come out wrong	The word list was not pasted, or was pasted after the stage prompt. Paste it first, before the prompt. If it persists, type the specifics rather than dictating them.
The structured output block is malformed	Ask it to output the block again on its own, with no surrounding commentary. Local models occasionally wrap structured output in explanation.
The answer quality is clearly worse than the cloud version	Expected on a smaller model — see section 2.1. If it is unusable for that stage, finish the stage locally, save it, and note it for review. Do not switch mid-stage to a cloud assistant with the client's specifics still in the conversation.
Anything referencing search or the internet fails	Correct. The model is offline by design. The stage prompts do not require search.

9.1 Fallback for lower-memory machines

GPT4All from nomic.ai runs the same class of models with a lighter footprint and supports smaller variants that fit machines where LM Studio struggles. The workflow in sections 5 to 7 is identical — paste the word list, paste the stage prompt, run the dictation loop, copy the output back into the dashboard. Only the application around the model changes.

10. Where this fits

This guide covers one component: the model that runs the sensitive stages. Three documents sit around it.

- **The Manual Capture Method (voice dictation guide)** — how to capture an answer accurately. The word list, the Speak-Read-Fix-Send loop, the three-question answer check and the cleanup pass. Read this first if you have not run a session before; it is the method, not an alternative technique.
- **The Operator's Field Instructions** — the session procedure. The pre-flight rules, the four-stage pipeline, what stays out of a cloud assistant, and how the exports feed downstream.
- **The CDD Qwen Setup Guide** — the same local stack applied to Consultancy Done Differently's chaining workflows. If you already run B3, E1, E2 or A1-QWEN locally, your installation is done and only sections 5 to 7 of this guide are new.

Downstream, the exported .txt files feed the Expert AI Prompt Assemblers and the CDD frameworks unchanged. The dashboard takes pasted text and exports a text file; it does not know or care which model produced the text, which is precisely what makes this substitution possible.

11. Version and status

Field	Detail
Version	V1.0
Published	7 September 2026
Next review	March 2027, or sooner if LM Studio or the Qwen release line changes materially.
Author	Matthew Bulat, Expert AI Prompts
Questions	expertaiprompts.com

Status of the local recommendation

The stage split in section 1.1 is a recommendation based on which stages elicit sensitive material, not on a completed output-quality comparison. A side-by-side test of Stages 2 and 3 — local against cloud on the same interview — is planned, and this guide will be updated with what it finds.

Nothing here is legal advice or insurance advice. It describes how to run a documented workflow, not what any obligation requires.