Township Clerk AI Assistant – Quick Guide

AnythingLLM Desktop + eGPU (RTX 4000) + External Drive Models (Samsung T5). Print-friendly cheat sheets for daily use.

> **Quick "Do I need the external drive plugged in?"**
> - If your model files (`.gguf`) live on the Samsung T5, the T5 must be plugged in to use those models.
> - If you imported/copied models into the laptop storage, the T5 is optional for model use (but still useful for documents/backups).

1) Which Model to Use (Cheat Sheet)

Model	Best for	Avoid when / Notes
Llama 3.1 8B (Instruct) "Everyday Assistant"	• Drafting emails/letters, polite notices, public-facing wording • Summaries of typical documents • General Q&A (policies, procedures, "explain this to me")	• If it feels too verbose, ask: "Answer in 5 bullets, no fluff." • For very long documents, consider Phi-3 Mini 128k.
Mistral 7B (Instruct) "Logic & Formal Tone"	• More formal, structured answers • Step-by-step reasoning ("What should I do next?") • Comparisons ("Option A vs B") and checklists	• If you need maximum context length (very large text), Phi-3 Mini may handle big inputs better.
Phi-3 Mini 128k (Instruct) "Long Document Reader"	• Long meeting minutes, large PDFs, big ordinance packs • "Find the section that says…" and "What changed between versions?" • Pulling details across many pages	• Sometimes less "chatty"/creative; best used for document-heavy tasks. • If you only need a quick email draft, Llama is usually faster/easier.

> **Fast pick (if you're unsure):**
> - Writing/drafting/general help → **Llama 3.1 8B**
> - Formal, structured checklists & decisions → **Mistral 7B**
> - Very long documents / "find it in the PDF" → **Phi-3 Mini 128k**

2) Daily Setup Procedure (Start of Day)

A) Hardware (2 minutes)

- Turn on the eGPU enclosure (Razer Core) and confirm its power/fans are on.
- Connect the Thunderbolt/USB-C cable from the eGPU to the laptop (lightning-bolt USB-C port).
- Plug in the Samsung T5 (if your models are stored on it).

B) Windows quick check (optional but recommended)

- Open **Task Manager** → **Performance** tab → verify an NVIDIA GPU graph exists (this confirms the eGPU is recognized).

C) AnythingLLM (1 minute)

- Open **AnythingLLM Desktop**.
- Select your workspace (example: **Township Assistant**).
- Ask a quick test question: "Summarize today's priorities in 5 bullets."

3) How to Change Models (Cheat Sheet)

Method 1 (common): Change the model in LLM Preference

- In AnythingLLM, open **Settings**.
- Go to **LLM Preference** (or the LLM/Provider screen).
- Select the model you want (Llama / Mistral / Phi).
- Return to your workspace chat and continue.

Method 2: Import models (only if you don't see them yet)

- Settings → **LLM Preference**.
- Click **Import GGUF file**.
- Browse to the Samsung T5 and select the `.gguf` file.
- Repeat for each model you want available.

> **Common gotcha:**
> - If AnythingLLM can't find a model you previously used, the Samsung T5 may be unplugged or the drive letter changed.
> - If a model loads but responses are slow, confirm the eGPU is connected and recognized in Windows.

4) Best "Ask Prompts" for Clerk Work

Task	Copy/paste prompt
Draft a professional email	`Draft a professional email reply. Tone: polite, concise, public-sector. Include next steps as bullets. Here are the details: …`
Summarize minutes	`Summarize the key decisions, motions, votes, and follow-up actions. Output as: (1) Decisions (2) Action items (3) Deadlines (4) Open questions.`
Create a checklist	`Turn this policy into a step-by-step checklist a clerk can follow. Keep it to 10 bullets max.`
Explain in plain language	`Explain this section in plain English for a resident. Keep it friendly and under 150 words.`
Find a detail in a long PDF	`Search the attached document(s) and quote the exact sentence about: [topic]. Then summarize what it means in 3 bullets.`

5) Shutdown Procedure (End of Day)

A) Close software cleanly

- In AnythingLLM, finish the current response (don't unplug while it's "thinking").
- Close AnythingLLM (File → Exit or click the X).

B) Eject external drives

- In Windows, click the USB "Safely Remove Hardware" icon (system tray) and eject the Samsung T5.

C) Disconnect eGPU

- Unplug the Thunderbolt/USB-C cable from the laptop.
- Turn off the eGPU enclosure (if you won't use it again soon).

Two safety rules:

- Don't unplug the eGPU or Samsung T5 while AnythingLLM is actively generating a response.
- If anything acts weird, do the classic IT fix: close AnythingLLM → disconnect eGPU → reconnect eGPU → reopen AnythingLLM.

6) Quick Troubleshooting

- **Models disappeared:** Plug Samsung T5 back in and re-open AnythingLLM. If needed, re-import via "Import GGUF file."

- **It's slow:** Check eGPU connection, confirm NVIDIA driver installed, and look at Task Manager → Performance → NVIDIA GPU activity during a response.

- **No microphone / voice:** Windows Settings → Privacy & security → Microphone → allow access for AnythingLLM.

Version 1.0 • Built for Dave's "Township Clerk Assistant" setup (AnythingLLM + RTX 4000 eGPU + Samsung T5 models). Print on A4 or Letter.