Local AI for Accountants: Tax Prep and Financial Analysis Without the Cloud

March 5, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** Local AI handles the grunt work of accounting -- categorizing transactions, summarizing bank statements, drafting client memos, extracting receipt data, and answering tax code questions. It does all of this on your machine, with zero client data sent to the cloud. Qwen 3.5 9B runs on 8GB VRAM and handles most tasks. Qwen 2.5 32B on 24GB VRAM is better for complex multi-document analysis. Install Ollama, pull a model, and you have a private financial assistant in 5 minutes. This is not a replacement for professional judgment. It's a research and drafting tool that happens to be completely private.

More on this topic: VRAM Requirements · Run Your First Local LLM · Best Local Coding Models · Ollama Troubleshooting

It’s tax season. You’re surrounded by W-2s, 1099s, bank statements, and client financials. AI could help with the grunt work – categorizing transactions, summarizing documents, drafting explanations for clients. But sending a client’s complete financial picture to OpenAI or Google? That’s a liability no responsible practitioner should take on.

Local AI solves this. The model runs on your computer. Your clients’ data stays on your machine, never touching a cloud server, never training someone else’s model, never showing up in a breach disclosure. And the models available today are good enough to be useful for real accounting work.

I want to be clear up front: this is a research and drafting assistant. It won’t replace a CPA, it won’t file your returns, and it can’t give legal tax advice. What it does is handle the tedious knowledge work that eats your billable hours.

What local AI actually does for accounting

The use cases fall into a few buckets. Some of these work well today. Others are getting there.

Transaction categorization

Paste a CSV export from a bank statement and ask the model to categorize each transaction. A 9B model handles this well for standard categories (meals, office supplies, utilities, professional services). It learns from context – if you tell it “this is a freelance graphic designer’s business account,” it adjusts accordingly.

Accuracy is around 85-90% on straightforward transactions. You’ll still need to review edge cases manually. That’s fine – the point is turning a 2-hour categorization job into a 20-minute review.

Document summarization

Upload a 30-page lease agreement and ask “what are the rent escalation terms?” Feed in a partnership agreement and ask “what’s the profit-sharing structure?” Local models with 32K+ context can digest multi-page financial documents and pull out the specific details you need.

This is where local AI saves the most time. Reading dense contracts is slow, expensive work. Having a model surface the relevant clauses and then verifying them yourself beats reading every page start to finish.

Drafting client communications

“Write a memo explaining to my client why their home office deduction was reduced this year.” The model drafts it, you edit for accuracy and tone. Takes 3 minutes instead of 15.

This works for engagement letters, fee explanations, quarterly summaries, and the dozens of routine client communications that eat time during busy season.

Tax code research

Ask the model about Section 179 depreciation limits, QBI deduction thresholds, or the latest changes to standard deduction amounts. Local models trained on broad datasets have solid knowledge of US tax code through their training cutoff. They’re good for quick reference and for drafting technical explanations.

The catch: tax law changes frequently, and the model only knows what was in its training data. For the 2025 tax year, models released in late 2025 or 2026 have current information. For anything after that, verify against IRS publications. Always verify against IRS publications anyway – this is tax law, not a casual question.

Receipt and invoice extraction

Take a photo of a receipt, feed it to a multimodal model (Qwen 3.5 supports vision), and ask it to extract the date, vendor, amount, tax, and category. It works. Not perfectly – handwritten receipts and faded thermal paper are still rough – but for typed invoices and standard receipts, it saves manual entry.

What it can’t do

Financial errors have real consequences. I’d rather over-communicate the limits than have someone take a model’s output at face value on a tax return.

The model has knowledge of tax code but no fiduciary responsibility, no E&O insurance, and no accountability. If a client asks “can I deduct this?” the model can help you research the answer. The answer itself needs to come from you.

There’s no integration with TurboTax, Drake, Lacerte, or any filing system. Local AI is a research and drafting layer, not an e-filing tool.

All LLMs hallucinate, and in accounting, a hallucinated number or a made-up IRS ruling is dangerous. Treat every output like a first draft from a junior associate – useful, but requiring review. The Stanford Law School paper on LLMs as tax attorneys found that general-purpose LLMs can produce plausible but factually incorrect tax guidance.

And the model doesn’t know your client’s full picture. It processes what you give it, one prompt at a time. It won’t cross-reference this year’s returns with last year’s K-1s unless you explicitly provide both. You’re the one who sees the whole picture.

Which models to run

Not every local model handles financial work equally. Structured data, precise numbers, and tax terminology require a model that follows instructions tightly and doesn’t drift.

Model	Size (Q4)	VRAM Needed	Best For
Qwen 3.5 9B	~6.6 GB	8 GB	General accounting tasks, categorization, memos
Qwen 2.5 Coder 7B	~5 GB	8 GB	Structured output (JSON, CSV), data extraction

Model	Size (Q4)	VRAM Needed	Best For
QwQ-32B	~20 GB	24 GB	Complex reasoning, multi-step tax analysis
Qwen 3.5 32B	~20 GB	24 GB	Long documents, multi-page contracts

8GB VRAM: Qwen 3.5 9B

This is where most accountants should start. Qwen 3.5 9B fits on an RTX 3060 Ti or RTX 4060 with room for 8K context. It handles transaction categorization, memo drafting, and tax code Q&A without issues. The thinking mode (enabled by default) helps it produce more accurate structured output.

For receipt scanning, Qwen 3.5 9B includes vision capability – it can read images of invoices and receipts directly. No separate OCR tool needed.

24GB VRAM: QwQ-32B or Qwen 3.5 32B

If you're doing complex tax planning, multi-document analysis, or need to hold a full lease agreement plus a client conversation in context at once, the 32B models on an RTX 3090 or 4090 are noticeably better. QwQ-32B is the reasoning specialist – it thinks through multi-step tax problems more carefully. Qwen 3.5 32B has longer context and broader knowledge.

Mac users

If you're on an M2/M3/M4 Mac with 16GB+ unified memory, you're in good shape. Qwen 3.5 9B runs smoothly. With 32GB or more, you can run the 32B models. See our Mac M-series guide for specifics.

Setting it up

You can go from zero to working in about 5 minutes. No programming required.

Option 1: Ollama (command line)

```
# Install Ollama
curl -fsSL https://ollama.com/install.sh | sh

# Pull the model
ollama pull qwen3.5:9b
```

```
# Start chatting
ollama run qwen3.5:9b
```

That's it. The model downloads (~6GB), and you have a private AI assistant running on your hardware.

For a browser-based interface instead of the terminal, add Open WebUI:

```
docker run -d -p 3000:8080 --add-host=host.docker.internal:host-gateway \
  -v open-webui:/app/backend/data --name open-webui \
  --restart always ghcr.io/open-webui/open-webui:main
```

Open `http://localhost:3000` in your browser. It looks and feels like ChatGPT, but everything runs locally. You can upload documents directly in the chat interface.

Option 2: LM Studio (GUI)

Download LM Studio, search for "Qwen 3.5 9B," click download, and start chatting. No terminal, no commands. Good for accountants who want to avoid the command line entirely.

Sample prompts

These are ready to copy and paste. Modify the details for your situation.

Categorize transactions

```
Here are transactions from my client's business checking account.
Categorize each one using standard Schedule C categories
(advertising, car/truck, office, supplies, utilities, meals,
professional services, other).

Format as a table with columns: Date, Description, Amount, Category, Notes.

Transactions:
[paste your CSV data here]
```

Summarize a financial statement

```
Summarize this bank statement. I need:
1. Total deposits and withdrawals
2. Largest 5 transactions by amount
3. Any recurring payments and their frequency
4. Transactions that look like they might be personal
   (not business) expenses

Statement:
[paste statement text here]
```

Draft a client memo

```
Write a brief memo to my client explaining:
- Their total estimated tax liability for 2025 is approximately $X
- They underpaid quarterly estimates by approximately $Y
- They may owe a penalty under IRC Section 6654
- Options to reduce the penalty (annualized income method, etc.)

Keep it professional but not intimidating. They're a sole proprietor
who runs a small landscaping business.
```

Research a tax question

```
My client is a freelance photographer who works from a dedicated
room in their apartment. They want to claim the home office
deduction. Explain:
1. The simplified method vs regular method for the home office
   deduction
2. Square footage limitations
3. What expenses qualify (rent, utilities, insurance)
4. How it affects depreciation if they own the home
5. Any common audit triggers for this deduction

Cite the relevant IRC sections and IRS publications.
```

Extract receipt data

```
Extract the following from this receipt:
- Vendor name
- Date
- Total amount
- Tax amount
- Payment method
- Line items with individual prices
- Suggested expense category for a small business

Format as JSON.
```

(For this prompt, use a multimodal model like Qwen 3.5 and paste or attach the receipt image.)

Reconciliation helper

```
I have two lists of transactions -- one from the bank statement
and one from QuickBooks. Find the discrepancies:

Bank statement transactions:
[paste bank data]

QuickBooks transactions:
[paste QB data]

List any transactions that appear in one but not the other,
and any amount mismatches for matching transactions.
```

Tools that help

Beyond the raw model, a couple of open-source tools are worth knowing about.

TaxHacker is a self-hosted accounting app built for freelancers and small businesses. Upload receipts and invoices, and it uses AI to extract dates, amounts, vendors, and line items into a structured database. It currently works with OpenAI and Mistral APIs, with local LLM support coming. If you're comfortable with Docker, it's a solid way to organize client receipts. The data stays on your server.

Open WebUI gives Ollama a browser-based interface with document upload. Drag a PDF into the chat and ask questions about it. For accountants reviewing contracts and statements, this is the most natural workflow – it's like ChatGPT but private.

The privacy argument, spelled out

Think about what's in a tax return. SSNs, income figures, asset valuations, debt, business revenue. When you send a prompt to ChatGPT or Claude that includes a client's W-2, that data travels to a third-party server. The provider's privacy policy governs what happens next.

With local AI, the data doesn't leave your machine. No third party, no privacy policy to interpret, no scenario where your client's income ends up in someone else's database. If your firm has professional ethics rules around client confidentiality – and it should – local AI lets you use the technology without breaking them.

The Journal of Accountancy has been writing about exactly this concern. If your firm already has a policy against uploading client data to cloud services, local AI is how you get the productivity gains without the compliance headache.

What hardware you need

You don't need an expensive rig. Most of the accounting use cases work fine with modest hardware.

Budget	Setup	What You Can Run
$0 (use what you have)	Any Mac with 16GB+ RAM	Qwen 3.5 9B, good enough for most tasks
$0 (use what you have)	Desktop with 8GB GPU	Qwen 3.5 9B at Q4
~$250-350	Used RTX 3060 12GB	9B models with 16K context, very comfortable
~$700-800	Used RTX 3090 24GB	32B models for complex analysis

If you already have a recent Mac laptop or a desktop with a gaming GPU, you probably don't need to buy anything. See our VRAM requirements guide for the full breakdown.

Bottom line

Local AI won't replace your professional judgment or your tax software. It won't file returns or sign engagement letters. What it does is take the categorizing, summarizing, drafting, and researching off your plate – faster than doing it manually, and without sending a single byte of client data to the cloud.

Install Ollama, pull Qwen 3.5 9B, and try it on a few of the sample prompts above. The setup takes 5 minutes. If it saves you even one billable hour during tax season, it's paid for itself – and the model was free to begin with.

Get notified when we publish new guides.

Subscribe — free, no spam

Source: https://insiderllm.com/guides/local-ai-accounting-tax/

Free guides for running AI locally