TOWER 3
STRATEGIC AI

FIELD GUIDE // V1.2

Local Inference Field Guide for Professional Services Founders

Date: June 2026 · **Author:** Tower 3 Strategic AI

Hardware: Colossus — RTX 5090 32GB / Ryzen 9 9950X3D / 64GB DDR5

Methodology: 20-category workload evaluation, 12 models, single workstation

What This Is

This guide documents what we observed when we ran twelve open-source language models across twenty professional services workload categories on a single workstation. It is an operational report, not a benchmark. We are not claiming comprehensive coverage, laboratory conditions, or results that generalize to every hardware configuration. We are saying: *here is what we ran, here is what we saw, draw your own conclusions.*

The motivation is practical. Tower 3 runs sovereign inference (locally-hosted language models) as the backbone of our internal operations. After roughly six months of evaluation and production use, we wanted to know two things that most published benchmarks don't answer: *which workloads actually matter for professional services founders*, and *how low can you go before the models stop being useful for real work?*

This guide answers both.

What It Is Not

This is not a claim that local inference is right for every use case or every budget. It is not a recommendation to abandon cloud AI. It is not a comprehensive academic benchmark. Three categories in this guide require human review to score properly (document drafting quality, email register, and meeting prep completeness) because those judgments require a human. We note that explicitly rather than substituting a metric.

The Hardware

Colossus is a desktop workstation, not a server:

- GPU: NVIDIA RTX 5090, 32GB VRAM
- CPU: AMD Ryzen 9 9950X3D, 16-core
- RAM: 64GB DDR5-6000
- Storage: 4TB Gen4 NVMe

This is current-generation consumer hardware. The RTX 5090 runs all Tier 1 and Tier 2 models in a single VRAM load. Tier 3 models (70b+) fit with quantization and run unattended overnight. The Ryzen 9 runs inference without the GPU. We document that separately as the CPU-only tier.

The cost of this configuration as of mid-2026 is approximately $5,000–$7,000 fully built. We document the economics in the final section.

How to Read This Guide

Twelve models are evaluated across three tiers:

TIER	MODELS	VRAM RANGE
`Tier 1 — Small (1b–7b)`	smollm2:1.7b, qwen2.5:3b, gemma3:4b, phi4-mini, qwen2.5:7b	1–6 GB
`Tier 2 — Operational (14b–27b)`	phi4:14b, qwen2.5-coder:14b, gemma3:27b, codestral:22b	10–22 GB

TIER	MODELS	VRAM RANGE
Tier 3 — Async (32b–72b)	deepseek-r1:32b, qwen2.5:72b, llama3.3:70b	24–50 GB

Twenty workload categories are organized in two groups:

Standard workloads: Agentic dispatch, multi-turn fidelity, long-context retrieval, extraction from prose, graceful degradation on unknowns, and consistency under repetition. These categories reflect the task types Tower 3 runs in daily operations.

Professional services workloads: Document drafting, client-facing Q&A, summarization at varied compression, warm email drafting, financial and operational data analysis, meeting prep briefings, and contract/agreement review. These categories were developed to cover the full workload surface of a founder-led professional services firm.

Scores are reported on a 0.0–1.0 scale. Dispatch, extraction, and consistency categories use deterministic parsers (pass/fail against a schema). Multi-turn, long-context, graceful degradation, and all professional services categories use a Sonnet-4 LLM judge scoring against documented ground truth. Consistency additionally computes cosine similarity across 5 repetition runs. Document drafting is human-reviewed: qualitative findings are reported in narrative form.

A **pass** in this guide means a score of 0.7 or above.

The Standard Workloads — What We Found

Agentic Dispatch

> *The question: does the model emit the right sentinel token and schema-valid JSON when the user approves an action, and stay silent when they don't?*

This is the highest-stakes binary test in the dataset. A false positive (dispatching when no approval was given) is a failure mode with real operational consequences.

Results: Every model above 3b passes the positive cases cleanly. The meaningful differentiation is on negative cases where models must recognize that no approval was given. smollm2:1.7b scores 0.88. It makes occasional false positives. Everything at 3b and above scores 0.92 or higher on the negative dispatch cases.

> **FIELD GUIDE FINDING**
>
> Agentic dispatch is viable at qwen2.5:3b and above. If you are building a workflow where an AI creates tasks or calendar events on your behalf, 3b is a viable starting point on GPU. The critical test is the negative case: verify your model handles "thinking out loud" and information-exchange messages without firing.

Multi-Turn Conversational Fidelity

> *The question: does the model retain specific constraints established in early turns across a 10–15 exchange conversation?*

This is the hardest standard category. A model that cannot hold a $45,000 budget ceiling or a "no offshore resources" constraint across a 12-turn proposal discussion is not useful for iterative client work.

Results: Multi-turn fidelity is where the tiers separate most sharply. smollm2:1.7b scores 0.34 with four zero-score prompts. It forgets constraints. deepseek-r1:32b scores 0.99. It is exceptional here, which is consistent with its reasoning architecture. The operational tier (phi4:14b, gemma3:27b) scores 0.70–0.87. qwen2.5:7b scores 0.73, crossing the pass threshold.

> **FIELD GUIDE FINDING**
>
> Multi-turn fidelity requires at minimum a 7b model on GPU. For work where early-turn constraints are legally or commercially significant (a budget ceiling, a staffing restriction, a client's stated requirement), use 14b or above. deepseek-r1:32b is the quality ceiling here, but at 6–7 seconds per response turn it is marginal for interactive use.

Long-Context Document Retrieval

> *The question: given a 3,500-word business assessment, does the model extract accurate specific facts?*

Results: Long-context retrieval is broadly well-handled. Every model at 3b and above scores above 0.84, and most score 0.94–1.00. smollm2:1.7b scores 0.84. Even the floor model can pull named people and dollar figures out of a long document most of the time.

> FIELD GUIDE FINDING
>
> Long-context retrieval is achievable at every tier. If your use case is "I uploaded a vendor proposal, find the five things that matter," almost any model in this set handles it. The failure mode at small sizes is occasional precision errors on secondary figures, not complete inability to find information.

Extraction from Unstructured Prose

> *The question: given meeting notes, email threads, or status updates, does the model produce structured output with correct field values?*

Results: Perfect or near-perfect scores across every model. Every model in the set scores 1.00 on this category. Extraction is the most democratized capability in the dataset. Even smollm2:1.7b scores 1.00 on structured field extraction.

> FIELD GUIDE FINDING
>
> If your primary use case is extracting action items, decisions, or structured data from narrative text, model selection is a latency question, not a quality question. Even a 1.7b model running on CPU will get this right.

Graceful Degradation on Unknowns

> *The question: when asked about private company financials, future events, or non-existent regulations, does the model acknowledge uncertainty rather than fabricate?*

Results: Most models pass, but with notable exceptions. gemma3:4b and gemma3:27b each have multiple zero-score prompts (3/10 and 2/10 respectively). These models occasionally confabulate with confidence rather than acknowledging uncertainty. qwen2.5-coder:14b and qwen2.5:72b both score 1.00, with perfect uncertainty acknowledgment. deepseek-r1:32b scores 0.83 with one failure.

> FIELD GUIDE FINDING
>
> If you are using a local model in a client-facing context where hallucination on unknowable facts is a real risk, test this category specifically against your deployment model. The failure mode is the model stating a false fact with high confidence, which is more dangerous than getting something slightly wrong. gemma3 models at both 4b and 27b show this pattern. Verify before deploying.

Consistency Under Repetition

> *The question: does the model produce semantically consistent outputs when the same prompt is run five times?*

Results: Consistency is broadly excellent. All models above 3b score above 0.87 on all three consistency subtypes. Even smollm2:1.7b achieves 0.93 on structured JSON consistency. The lowest score in the entire consistency dataset is smollm2:1.7b on prose synthesis at 0.83.

> FIELD GUIDE FINDING
>
> Consistency is not where open-source local models fail in 2026. If you run a pipeline that calls a model on a schedule, the output on Tuesday will be semantically similar to the output on Thursday. The variance is low enough not to be operationally significant at the 7b+ tier.

The Professional Services Workloads — What We Found

Document Drafting

> *Proposals, SOWs, engagement letters, project charters, change orders.*

Scoring note: This category is evaluated through human review. We read 8-prompt output sets for phi4:14b (the best operational tier model) and smollm2:1.7b (the floor model) and scored qualitatively.

phi4:14b: Professional-grade output across all document types. Register held throughout long proposals and SOWs. Fee structures, party names, and milestone triggers correctly incorporated.

Ready-to-review output. Not necessarily ready-to-send, but requires editing, not rewriting.

smollm2:1.7b: Functional but degraded. Gets most content right but made a factual error identifying a party role (confused the client's managing partner with the consultant organization). Formatting informal: prose paragraphs rather than structured sections. Usable as a drafting scaffold, not as a near-final document.

> FIELD GUIDE FINDING
>
> Document drafting requires 14b or above for production-quality output. At 14b you get something reviewable. At 1.7b you get something you need to fact-check carefully, including party names. For a founder whose documents go to clients, the review burden below 14b is high enough to question the value.

Client-Facing Q&A

> *Answer questions about your firm's services, methodology, and pricing from provided context. Include negative cases where the question cannot be answered from context.*

Results: phi4:14b leads at 0.825 / 88% pass rate. qwen2.5-coder:14b and deepseek-r1:32b both reach 0.762–0.787 / 75%. The critical differentiation is on the negative cases: questions the AI should refuse to answer because the information isn't in its context. smollm2:1.7b scores 0.263 here, partly because it occasionally fabricates specific client references or pricing details rather than acknowledging they're not in the briefing.

> FIELD GUIDE FINDING
>
> A local model handling prospect Q&A from a firm profile document needs to be reliable on both answerable and unanswerable questions. The failure mode that matters is fabricating a reference client name or an hourly rate that wasn't in the source material. phi4:14b handles this cleanly. Below 7b, verify the negative case behavior specifically.

Summarization at Varied Compression

> *The same source document summarized to three-sentence, one-paragraph, and one-page targets.*

Results: Summarization is the second most democratized workload after extraction. Nine of twelve models score above 0.80, including gemma3:4b (0.889) and qwen2.5:3b (0.800). The key test is whether the same critical facts surface at all three compression levels. Most models pass this consistently. The outlier is phi4-mini at 0.544. It struggles to identify what matters at the three-sentence level. qwen2.5:72b and llama3.3:70b have the highest latency in this category (35–136s per prompt) with comparable quality to 14b models.

> FIELD GUIDE FINDING
>
> If your primary use case is summarizing reports, proposals, or meeting notes at different lengths, the 4b-7b tier handles it well on GPU. This is where gemma3:4b earns its place. For a founder who needs to summarize a 40-page assessment before a board call, gemma3:4b produces accurate summaries at 0.5–2s per prompt.

Warm Email Drafting

> *Operational client email: delayed deliverable notification, project updates, invoice follow-up, check-ins. Register is relational, not promotional.*

Results: gemma3:27b achieves the only perfect score (1.000) in the entire Field Guide dataset on this category. phi4:14b and deepseek-r1:32b follow at 0.850 / 88%. The notable failures: qwen2.5-coder:14b scores 0.588 and codestral:22b scores 0.637. Both code-specialist models fail significantly on warm register. This is consistent with what we found in a separate outreach A/B evaluation: code-specialist models do not belong in outreach or client communication workflows.

> FIELD GUIDE FINDING
>
> Email register is where model specialization matters most in the professional services context. gemma3:27b has an exceptional feel for the relational tone that distinguishes a good client email from a corporate-sounding one. qwen2.5-coder:14b is the wrong model for this workload: its training on code produces a functional but inappropriately formal register.

Financial and Operational Data Summary

> *Invoice aging analysis, pipeline status, fee schedule comparison, utilization tracking, P&L, expense variance.*

Results: This is the most discriminating category in the dataset. Scores range from 0.000 to 0.975, a spread wider than any other category. The pattern is binary: models either reason numerically at a useful level or they do not.

Models that pass (≥0.7): llama3.3:70b (0.975), gemma3:27b (0.887), qwen2.5:72b (0.800), deepseek-r1:32b (0.838), phi4:14b (0.675, borderline).

Models that fail: phi4-mini (0.037), smollm2:1.7b (0.000), codestral:22b (0.100), qwen2.5-coder:14b (0.362), gemma3:4b (0.362), qwen2.5:3b (0.238), qwen2.5:7b (0.475).

llama3.3:70b at 0.975 / 100% pass rate is the standout finding. It is the only model that handles financial reasoning at the level a professional services founder would need: correct totals, correct variance identification, correct anomaly flagging.

> FIELD GUIDE FINDING
>
> Financial data analysis is not viable below 27b. gemma3:27b is the operational tier floor for this workload. phi4:14b is borderline. For anything involving revenue recognition, invoice aging, utilization tracking, or expense variance, run llama3.3:70b overnight if you want the numbers right. The failure mode at small sizes is producing numbers that are confidently wrong, which is more dangerous than producing numbers that are obviously off.

Meeting Prep Briefings

> *Given contact, company, meeting objective, and context, produce a structured briefing.*

Results: phi4:14b leads at 0.967 / 100%. qwen2.5:72b (0.933), llama3.3:70b (0.900), and codestral:22b (0.867) follow. The surprise: gemma3:27b scores 0.650 / 50%, below qwen2.5:7b (0.817) and codestral:22b (0.867). The model that produces perfect warm email fails on structured briefing synthesis. Task-model fit within a tier is real.

> FIELD GUIDE FINDING
>
> Meeting prep is where phi4:14b separates from the pack. phi4:14b reliably produces briefings that correctly capture a referral context, a relevant LinkedIn signal, and three sharp questions to ask. gemma3:27b does not, despite its strong performance on other categories. If briefing quality matters to your workflow, phi4:14b is the operational tier choice, not gemma3:27b.

Contract and Agreement Review

> *Given an NDA, SOW, engagement letter, or restrictive covenant section: identify key terms, obligations, non-standard clauses, and risk flags.*

Results: Contract review is the hardest workload in the dataset. The ceiling is llama3.3:70b at 0.713 / 75%. phi4:14b and qwen2.5:72b reach 0.675. Human review of contract review outputs confirmed the pattern: no complete failures, but consistent partial results on nuanced risk terms. The failure mode is incompleteness, not fabrication. Models identify the obvious terms and miss the subtle ones ($500M revenue catch-all in a non-compete, a two-business-day rejection window that's below standard).

smollm2:1.7b scores 0.037, essentially no contract review capability.

> FIELD GUIDE FINDING
>
> Local models are useful for contract orientation: understanding what a document covers, what the obvious terms are, whether standard clauses are present. They are not a substitute for careful human review on risk terms. The right framing: run a 27b or 72b model to surface the first 80% of issues, then read the document yourself with the model's output as a checklist, not a verdict.

The CPU-Only Tier

We ran all six Tier 1 and the one Tier 2 boundary model (phi4:14b) with GPU disabled: pure CPU inference on the Ryzen 9 9950X3D. The question: *what can a founder do on a capable machine without a GPU?*

The latency differential is 10–20× across the board:

WORKLOAD CLASS	GPU (QWEN2.5:7B)	CPU (QWEN2.5:7B)
Agentic dispatch	377ms	5,282ms
Extraction	462ms	8,786ms
Document drafting	3,044ms	51,199ms

WORKLOAD CLASS	GPU (QWEN2.5:7B)	CPU (QWEN2.5:7B)
Financial data	3,082ms	53,639ms

Where CPU inference is viable:

Short structured tasks (dispatch, extraction, Q&A, graceful degradation) run at 2–9 seconds on small models (1.7b–7b). For a workflow that runs once per request with a few seconds of acceptable latency, this works. A founder using a 3b model on a CPU-only laptop to extract action items from meeting notes will wait 3–5 seconds per run. That's acceptable.

Where CPU inference is not viable:

Document drafting, financial data analysis, meeting prep, and contract review are all 15–120+ seconds on CPU, even for small models. phi4:14b on CPU takes 123 seconds for a document drafting prompt. That is not a usable interactive workflow. For long-form professional services work, a GPU is required for practical use.

The CPU-only founder's answer: qwen2.5:3b or gemma3:4b on CPU handles extraction, Q&A, summarization, and dispatch at acceptable latency. For anything involving long documents, financial reasoning, or multi-turn work, you need a GPU before local inference becomes worth the effort.

The Cost Case

From our 24-hour continuous benchmark run:

- **6,324 inference calls** across 12 models and 8 task types
- **Cloud equivalent cost** (Sonnet-4 at production volume): $83.47
- **Sovereign electricity cost** (Kill-a-Watt measurement): $1.84
- **Cost advantage: 11.3×** at T3's production volume
- **Crossover point: approximately 1.7 million tokens per day**

Below 1.7M tokens/day, cloud API is cost-competitive when you factor in hardware amortization. Above it, sovereign inference pays for the hardware within months at current API rates.

Tower 3's production volume is well above the crossover. For a solo founder running light automation (a few hundred API calls per day), the economics don't favor a $6,000 workstation. For a founder-led firm running a pipeline of scoring, drafting, extraction, and routing tasks throughout the business day, the math shifts decisively.

Local inference is not cost-effective for everyone. Know your token volume before you build.

Where Sovereign Inference Ends

Three workload findings are clear limits of what we observed, not gaps to be engineered around.

Financial reasoning below 27b. The models below gemma3:27b produce confident financial outputs that are frequently wrong. This is a capability boundary. If you route financial data work to a small local model, you will get wrong numbers some of the time and not know it. The right architecture: financial data analysis routes to 27b+ on GPU or stays on cloud.

Outreach register at every local tier. A separate A/B evaluation confirmed what this dataset shows: code-specialist models fail on outreach register and every local model tested falls below Sonnet-4 on the outreach quality rubric. Our outreach drafting workflow runs on Sonnet. The Field Guide warm email scores are better. Relational email is more accessible to local models than cold outreach. But for work that creates a first impression with a prospect, cloud remains the right call.

Contract review at the nuanced risk level. Contract review has a 75% ceiling regardless of model size. llama3.3:70b is the ceiling and it scores 0.713. The models reliably catch obvious terms and miss subtle risk flags. For orientation and checklist generation, local models are useful. For risk assessment on a contract you're about to sign, read it yourself.

How to Decide

A founder-operator reading this guide is making one of three decisions: whether to adopt local inference at all, what to run if you do, and how to architect the cloud and local mix. The findings above answer each one. The through-line is worth stating directly.

On whether to adopt at all. Local inference is operational today for a meaningful share of professional services work. Extraction, dispatch, summarization, Q&A, structured drafting, and

consistency under repetition all run reliably at the 7b–14b tier on consumer hardware. The work has shifted from "is this possible" to "is this worth the build cost for your specific token volume and workload mix."

On what to run. The model selection question has a sharper answer than the marketplace suggests. Above the 7b threshold, model fit matters more than model size. Workload-specific best models outperform across-the-board best models: gemma3:27b on warm email and phi4:14b on briefing synthesis outperform the larger qwen2.5:72b. qwen2.5-coder:14b is excellent for code review and badly miscast for client email.

On the architecture. A hybrid stack is the right answer for most founder-operated firms. Run local for the workloads where local is reliable and economical. Route financial reasoning, cold outreach, and high-stakes client-facing synthesis to cloud where the quality ceiling is higher. The cost case for sovereign inference requires sustained local volume above 1.7 million tokens per day. That threshold is achievable for any firm running AI continuously throughout the operational day.

On the cases where there is no choice. Some workloads cannot go to a cloud API regardless of economics. PII, HIPAA-covered data, attorney-client privileged material, anything where data sovereignty is a contractual or regulatory requirement. For those workloads, the only operational question is how to run local well.

This is the operational reality Tower 3 was built around. The findings above are what we observe running the firm on our own stack. We publish them because founder-operated professional services firms face the same evaluation problem we faced, with less time to run twenty workload categories across twelve models. If this guide saves you weeks of testing, it has done its job.

Model Quick Reference

MODEL	TIER	BEST FOR	AVOID FOR	GPU LATENCY (DISPATCH, P50)
`smollm2:1.7b`	T1	Nothing production-critical	Any sustained context work	227ms
`qwen2.5:3b`	T1	Extraction, dispatch, Q&A	Financial reasoning, multi-turn	260ms
`gemma3:4b`	T1	**Summarization, compression, email** — best small model	Financial data, graceful degradation	443ms

MODEL	TIER	BEST FOR	AVOID FOR	GPU LATENCY (DISPATCH, P50)
`phi4-mini`	T1	Structured dispatch	Everything else — does not scale within the phi4 family	266ms
`qwen2.5:7b`	T1	**General purpose at the 7b tier** — the practical floor	Financial data	377ms
`phi4:14b`	T2	**Best all-rounder** — briefings, Q&A, documents	Financial data (borderline)	604ms
`gemma3:27b`	T2	**Warm email, synthesis** — perfect email register	Meeting prep (anomalous weakness)	1,383ms
`qwen2.5-coder:14b`	T2	Code review, structured extraction	Outreach, financial data, email	732ms
`codestral:22b`	T2	Dispatch, extraction	Financial data (catastrophic), general use	1,059ms
`deepseek-r1:32b`	T3	**Multi-turn fidelity** — best in class on multi-turn tasks	Interactive use (6–7s per turn)	6,637ms
`qwen2.5:72b`	T3	Strong all-around overnight batch	Interactive use (38s+ on short tasks)	38,390ms
`llama3.3:70b`	T3	**Financial data specialist** (0.975) — meeting prep	Interactive use	23,402ms

Methodology Notes

All runs on a single desktop workstation (RTX 5090 / Ryzen 9 9950X3D). Ollama inference engine. Temperature 0.7. Each prompt run once for the GPU tier and once with GPU disabled for the CPU-only tier. Sonnet-4 (claude-sonnet-4-20250514) is both the cloud baseline and the LLM judge for scoring. Human review covers the document drafting category for two representative models. Prompt files, harness code, and result data are available on request.

This is version 1.2. We will update this guide as models improve, as we run additional hardware configurations, and as the workload taxonomy grows with the firm's operational needs.

Study Reference

This Field Guide draws on a series of evaluations conducted between November 2025 and May 2026. The table below maps each study to its plain-language description for readers who want to trace specific findings.

STUDY	DESCRIPTION	KEY OUTPUT
Internal Routing Evaluation	7 task types evaluated across 8 model families on Colossus; established the initial routing table for T3 operational nodes	VRAM fit matrix, latency baselines, model-to-task assignments
Coding Capability Deep Dive	Evaluated local models for code review against T3 security standards; compared against Gemini cloud baseline	qwen2.5-coder:14b promoted to production for code review at 0.82s p50
72b Quantization Study	Tested Q3 quantization on 72b models; measured whether CPU offload could be eliminated	Q3 improves latency 1.55× but does not eliminate CPU offload; routed to overnight batch only
Talus PM Synthesis Evaluation	Tested local models on client-facing synthesis tasks requiring professional register	No local candidate cleared the quality threshold; Sonnet permanent for this workload
Outreach Register A/B	Tested operational and small-tier models against cold outreach quality rubric with Sonnet-only judging	No local model cleared the threshold; code-specialist models performed worst; Sonnet permanent for outreach
24-Hour Sovereignty Economics Run	Ran the full internal task matrix continuously for 24 hours; measured power draw with Kill-a-Watt	11.3× cost advantage vs cloud at production volume; $1.84 electricity vs $83.47 cloud equivalent
Field Guide Evaluation (this document)	12 models across 20 professional services workload categories; GPU and CPU-only tiers	This document