

COMPANION WORKSHEET

Chapter 5 — Model Memory Fit

Use a fresh copy from the Companion Package. This worksheet creates exactly three sizing candidates. It does not recommend a model, select a platform, predict speed or quality, or authorize a purchase.

Use binary units consistently: 1 GiB = 1,073,741,824 bytes. Preserve source-reported bytes when available and label any conversion.

A. Record identity

Date and local time

Non-sensitive system label from Chapter 4

Chapter 4 hardware-inventory path

Chapter 2/3 anchor-evidence path, if available

Worksheet evidence folder

B. Bound the workload

Primary task

Largest input the workload actually needs

Planned context length (tokens)

Why that context is needed

Simultaneously loaded models

Parallel requests per loaded model

Placement assumptions allowed for sizing (GPU / CPU / mixed / unified / Unknown)

Assumptions this chapter must not make

C. Carry forward memory and storage evidence

Choose one topology for this worksheet: dedicated GPU + system RAM, shared/unified, CPU-oriented system RAM, mixed placement, or Unknown. Mark overlapping or inapplicable rows Not applicable; do not count one physical pool twice. A mixed placement needs a separate demand row for every pool it uses.

Selected topology:

Do not combine rows merely because they are all reported in GiB. Current available/free is the observed amount after current use at the recorded moment. The reserve below means additional uncommitted headroom to preserve beyond that snapshot; it must not count OS or application use that has already reduced the available/free figure.

Pool or gate	Installed/total bytes	Current available/free bytes	Explicit reserve/headroom bytes	Capacity after reserve	Evidence path and time	Qualification
Dedicated VRAM, if applicable and established						
System RAM, if applicable						
Shared or unified memory, if applicable and established						
Intended storage volume						Free storage is a separate gate

Reserve rationale:

Material memory-pool unknowns:

D. Record one measured anchor, if available

Use the exact Chapter 2/3 record; do not invent or rerun missing evidence merely to fill the table. For the Chapter 2 JSON, require top-level status: PASS, runtime.attempted: true, and matching stored/runtime identity and digest. If a path is absent, record Missing.

Anchor field	Retained-record JSON path	Observed value	Evidence path	What it can prove
Requested model identity	requested_model			Requested identity
Stored model identity	preflight.selected_model.name or .model			Stored identity
Runtime model identity	runtime.selected_model.name or .model			Runtime identity
Matching digest	preflight.selected_model.digest and runtime.selected_model.digest			Exact stored/runtime join

Anchor field	Retained-record JSON path	Observed value	Evidence path	What it can prove
Parameter label	preflight.selected_model.details.parameter_size			Label, not memory measurement
Quantization	preflight.selected_model.details.quantization_level			Exact retained label
Stored artifact bytes	preflight.selected_model.size			Storage demand for that artifact
Runtime record size bytes	runtime.selected_model.size			Runtime's documented model-size field; not proven total allocation
Runtime record size_vram bytes	runtime.selected_model.size_vram			Runtime-reported VRAM use; not GPU identity or total demand
Allocated context length	runtime.selected_model.context_length			Exact observed configuration
Processor/placement classification	runtime.selected_model.processor, when present			Runtime-reported placement

Anchor conclusion:
Anchor did not prove:

E. Define exactly three candidate configurations

Every row needs an exact artifact/variant. A family name or parameter count alone is insufficient.

Candidate role	Exact artifact/variant and revision	Parameter label	Quantization	Exact artifact bytes	Target context	Concurrent models	Parallel requests	Evidence path
Conservative								
Working								
Stretch								

F. Weight sanity check—not final sizing

Use only when parameter count and nominal weight bits are known:
rough weight bytes = parameter count × nominal bits per weight ÷ 8
This rough result does not include mixed tensor types, metadata, format overhead, context/KV cache, or runtime buffers. Exact artifact bytes outrank

it for storage sizing. A runtime-supported total-demand estimate or a measured before/after pool delta outranks it for runtime-memory sizing.

Candidate role	Parameter count	Nominal bits/weight	Rough weight bytes	Exact artifact bytes	Difference explained or flagged
Conservative					
Working					
Stretch					

G. Test each pool and storage gate

Do not hide unknown demand inside a zero. If context/KV-cache or runtime demand is not established, use Evidence missing or Measurement required. Before adding components, record why they do not overlap. A total-demand estimate or before/after pool delta may already include weights, cache, and buffers; do not add those components again.

capacity after reserve = current available/free bytes – additional uncommitted headroom bytes

margin = capacity after reserve – candidate demand

Candidate role	Pool/gate tested	Model-demand bytes	Context/KV-cache bytes, if separate	Other demand bytes, if separate	Total non-overlapping demand	Capacity after reserve	Margin	Evidence, scope, or missing reason
Conservative								
Working								
Stretch								

Duplicate a candidate row when it must pass more than one pool/gate. VRAM, system RAM, unified/shared memory, and storage remain labeled.

Demand-component overlap check and evidence:

H. Assign one verdict to each candidate

Allowed verdicts only:

- Plausible fit — exact configuration and topology evidence are present, every required gate has positive explicit margin, every demand component has defined scope without double counting, and no known missing demand invalidates the calculation.
- Measurement required — the exact configuration is defined, but runtime loading/placement or context demand must be observed before fit is credible.
- Does not fit — at least one required gate has zero or negative margin under the recorded assumptions; zero does not satisfy the positive-margin rule.

Evidence missing — the identity, variant, quantization, pool, context, concurrency, or another load-bearing input is absent.

Candidate role	Verdict	Binding constraint or positive margin	Evidence supporting verdict	Invalidation condition
Conservative				
Working				
Stretch				

I. Stop/Go record

- ☐ Exactly three candidate configurations exist.
- ☐ Each candidate binds an exact artifact/variant, quantization, context, and concurrency assumptions.
- ☐ Every capacity and demand value has units and provenance.
- ☐ One topology is selected; overlapping and inapplicable pools are marked Not applicable.
- ☐ Memory pools remain labeled rather than silently combined.
- ☐ Added demand components have non-overlapping scopes.
- ☐ Every candidate has exactly one allowed verdict.
- ☐ Every Plausible fit candidate has positive explicit margin at every required gate.
- ☐ No candidate is presented as a quality recommendation, platform decision, benchmark result, or purchase approval.

Three-candidate conclusion:

What Chapter 6 must decide:

What could reopen this worksheet: