Best Models Under 3B: Small LLMs That Work

January 29, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** Sub-3B models have gotten shockingly good. Qwen 2.5 3B and Llama 3.2 3B rival early 7B models on most tasks, run at 15-60+ tokens per second on CPU alone, and fit in 2-3GB of RAM. Start with Llama 3.2 3B for general use or Qwen 2.5 1.5B if RAM is tight. These aren't toys — they handle Q&A, summarization, simple coding, and classification well enough for real work.

📚 **More on this topic:** Run Your First Local LLM · CPU-Only LLMs · Quantization Explained

You don’t have a gaming GPU. Maybe you’re on a laptop with integrated graphics, a five-year-old desktop, a Raspberry Pi, or a phone. You’ve heard people running AI locally and you’re wondering: is that even possible on my hardware?

Yes. And not in a “technically it loads” way — in a “this is genuinely useful” way. The small model landscape changed dramatically in 2024-2025. A 3B model today outperforms a 7B model from 2023 on most benchmarks. A 1.5B model fits in under 2GB of RAM and generates faster than you can read.

This guide covers the best models under 3 billion parameters, what hardware you actually need, and what these models can and can’t do.

Who This Is For

If any of these describe your situation, this guide is for you:

Hardware	Typical RAM/VRAM	What You Can Run
Laptop (no dedicated GPU)	8-16GB RAM	3B models comfortably, multiple at Q4
Old GPU (GTX 1050 Ti, 1060)	4-6GB VRAM	3B models with room to spare
Raspberry Pi 5	8GB RAM	1B-3B models at usable speeds
Phone (recent Android/iPhone)	6-8GB RAM	0.5B-1.5B models, some 3B
Chromebook / thin laptop	4GB RAM	0.5B-1.5B models at Q4
Desktop with no GPU	8-32GB RAM	Any sub-3B model, fast

You're not the person with an RTX 4090 looking for the optimal model. You're the person wondering if local AI is even possible on what you've got. It is.

Why Small Models Matter Now

Two years ago, a 3B model was barely useful. It could complete sentences and sometimes follow instructions, but the output was rough. You needed at least 7B parameters for anything practical.

That changed fast. Three things happened:

Better training data. Model quality scales with data quality, not just size. Qwen 2.5 3B was trained on 18 trillion tokens of carefully curated data — more than many early 70B models saw.

Knowledge distillation. Smaller models now learn from larger ones during training. Llama 3.2 3B was distilled from Llama 3.1 70B, inheriting capabilities that would otherwise require far more parameters.

Architecture improvements. Grouped-query attention, better tokenizers, and improved positional encodings all help small models punch above their weight.

The result: Qwen 2.5 3B scores 65.6 on MMLU (a broad knowledge benchmark). For comparison, the original Llama 2 7B scored 45.3. A model less than half the size, beating one twice as large — that's how far things have moved.

The Best Sub-3B Models, Ranked

1. Qwen 2.5 3B — Best All-Rounder

The strongest model at this size class, period. Qwen 2.5 3B matches or beats the previous-generation Qwen 2 7B on most benchmarks while using less than half the memory.

Metric	Score
MMLU	65.6
GSM8K (math)	79.1
HumanEval (coding)	42.1

Metric	Score
HellaSwag	74.6
RAM at Q4_K_M	~2.5 GB
File size (Q4_K_M)	~2.0 GB

Strong at multilingual tasks, solid at coding, good instruction following. If you can run a 3B model, this is the default choice.

```
ollama pull qwen2.5:3b
```

2. Llama 3.2 3B — Best Instruction Following

Meta's small model, distilled from the 70B. Where Qwen 2.5 3B leads on raw benchmarks, Llama 3.2 3B excels at doing what you ask it to do. It scores 77.4 on IFEval (instruction following) — the highest in its class.

Metric	Score
MMLU	63.4
GSM8K (math)	77.7
ARC-C (reasoning)	78.6
IFEval	77.4
RAM at Q4_K_M	~2.5 GB
File size (Q4_K_M)	~2.0 GB

Particularly good at tool use (BFCL V2: 67.0) and multilingual tasks (MGSM: 58.2). If you're building something that needs reliable instruction following — a chatbot, an assistant, a workflow tool — Llama 3.2 3B is the pick.

```
ollama pull llama3.2:3b
```

3. Phi-3.5 Mini (3.8B) — The Overachiever

Technically 3.8B parameters — slightly over the 3B line — but it earns its spot here. Phi-3.5 Mini punches absurdly above its weight. It beats Mixtral 8x7B (a 46.7B MoE model) on math benchmarks and nearly matches GPT-3.5 on MMLU.

Metric	Score
MMLU	69.0
GSM8K (math)	86.2
HumanEval (coding)	62.8
BBH (hard reasoning)	69.0
RAM at Q4_K_M	~3.0 GB
File size (Q4_K_M)	~2.3 GB

Best coding and math performance under 4B parameters by a wide margin. The tradeoff: weaker on factual recall (TriviaQA: 64.0 vs GPT-3.5's 85.8) and somewhat less natural in free-form conversation. If your tasks lean toward reasoning and code, Phi-3.5 Mini is the best you'll find anywhere near this size.

```
ollama pull phi3.5
```

4. Qwen 2.5 1.5B — Best Under 2B

When 3B is too much for your hardware, Qwen 2.5 1.5B is where quality really starts. It scores 60.9 on MMLU — a number that would have been impressive for a 7B model not long ago.

Metric	Score
MMLU	60.9
GSM8K (math)	68.5
HumanEval (coding)	37.2
HellaSwag	67.9
RAM at Q4_K_M	~1.5 GB
File size (Q4_K_M)	~1.1 GB

This is the sweet spot for Raspberry Pi 5, phones, and machines with 4GB RAM. Fits easily, runs at 8-15 tok/s on a Pi 5, and handles Q&A, summarization, and simple tasks competently.

```
ollama pull qwen2.5:1.5b
```

5. Gemma 2 2B — Google's Efficient Pick

Google's entry uses knowledge distillation from larger Gemma models to pack capability into 2B parameters. Its strength is language understanding — strong HellaSwag (72.9), BoolQ (72.7), and factual recall (TriviaQA: 60.4).

Metric	Score
MMLU	52.2
HellaSwag	72.9
Winogrande	71.3
TriviaQA	60.4
RAM at Q4_K_M	~1.8 GB
File size (Q4_K_M)	~1.1 GB

Weak on math (GSM8K: 24.3) and coding (HumanEval: 20.1). Don't pick Gemma 2 2B for those tasks. But for commonsense reasoning, entity extraction, and classification, it's solid. It also has excellent KV cache efficiency, making it a good choice for serving multiple users.

```
ollama pull gemma2:2b
```

6. Llama 3.2 1B — The Ultralight

Meta's smallest. At 1.24B parameters, it fits in under 1.5GB of RAM at Q4 and runs at 30-60+ tok/s on a desktop CPU. Not the smartest model on this list, but fast enough to feel instant.

Metric	Score
MMLU	49.3
GSM8K (math)	44.4

Metric	Score
ARC-C (reasoning)	59.4
IFEval	59.5
RAM at Q4_K_M	~1.2 GB
File size (Q4_K_M)	~800 MB

Best for: quick answers, text classification, simple extraction tasks, and prototyping. At this size, you can run it alongside other applications without worry.

```
ollama pull llama3.2:1b
```

7. StableLM 2 1.6B — The Veteran

Released in early 2024 by Stability AI, StableLM 2 was state-of-the-art for sub-2B models at launch. It's since been surpassed by Qwen 2.5 1.5B and Llama 3.2 1B on most benchmarks, but it still has a niche: multilingual support across 7 languages and strong language understanding (HellaSwag: 70.5).

Metric	Score
MMLU	41.8 (Zephyr)
HellaSwag	70.5
Winogrande	64.6
RAM at Q4_K_M	~1.3 GB
File size (Q4_K_M)	~1.0 GB

Unless you specifically need its multilingual coverage, Qwen 2.5 1.5B is the better choice today.

```
ollama pull stablelm2:1.6b
```

8. Qwen 2.5 0.5B — The Absolute Minimum

Half a billion parameters. This model fits in under 1GB of RAM, downloads in seconds, and runs at 20+ tok/s on a Raspberry Pi 5. It's the smallest model that produces coherent, useful output.

Metric	Score
MMLU	47.5
GSM8K (math)	41.6
HumanEval (coding)	30.5
RAM at Q4_K_M	~0.8 GB
File size (Q4_K_M)	~400 MB

It outperforms Gemma 2 2B on math and coding despite being 5x smaller — a testament to Qwen's training pipeline. For edge devices, IoT applications, or situations where every megabyte counts, this is the floor.

```
ollama pull qwen2.5:0.5b
```

Head-to-Head Comparison

All models at Q4_K_M quantization:

Model	Params	RAM	File Size	MMLU	GSM8K	Best For
Phi-3.5 Mini	3.8B	~3.0 GB	~2.3 GB	69.0	86.2	Coding, math, reasoning
Qwen 2.5 3B	3B	~2.5 GB	~2.0 GB	65.6	79.1	All-around, multilingual
Llama 3.2 3B	3B	~2.5 GB	~2.0 GB	63.4	77.7	Instruction following, chat
Qwen 2.5 1.5B	1.5B	~1.5 GB	~1.1 GB	60.9	68.5	Best quality under 2B
Gemma 2 2B	2B	~1.8 GB	~1.1 GB	52.2	24.3	Classification, extraction
Llama 3.2 1B	1.24B	~1.2 GB	~800 MB	49.3	44.4	Speed, prototyping
Qwen 2.5 0.5B	0.5B	~0.8 GB	~400 MB	47.5	41.6	Edge devices, IoT

Model	Params	RAM	File Size	MMLU	GSM8K	Best For
StableLM 2 1.6B	1.6B	~1.3 GB	~1.0 GB	41.8	34.8	Multilingual (7 languages)

What Small Models Are Good At

Sub-3B models won't replace GPT-4. But for specific tasks, they're more than good enough — and they do it locally, privately, and for free.

Tasks where sub-3B models deliver:

- **Quick Q&A** — "What's the capital of France?" "How do I reverse a list in Python?" Fast answers, no API call needed.
- **Summarization** — Summarize a paragraph, an email, or a short document. Qwen 2.5 3B and Llama 3.2 3B handle this well.
- **Text classification** — Sentiment analysis, topic categorization, spam detection. Fine-tuned small models hit 90%+ accuracy on classification tasks.
- **Simple coding** — Generate a function, fix a syntax error, explain a code snippet. Phi-3.5 Mini scores 62.8 on HumanEval — that's real coding ability.
- **Translation** — Simple translations work well, especially with Qwen (strong multilingual training) and Llama 3.2 (trained on 8 languages).
- **Data extraction** — Pull names, dates, and structured fields from unstructured text. Gemma 2 2B is particularly good at this.
- **Autocomplete and suggestions** — Fast enough for real-time text completion in editors.

What Small Models Can't Do

Being honest about limits saves frustration.

Don't expect these:

- **Complex multi-step reasoning** — "Plan a two-week trip optimizing for budget and weather across five cities" will produce mediocre output. The model doesn't have the capacity to hold complex chains of logic.
- **Long-form writing** — Blog posts, essays, fiction beyond a few paragraphs. Coherence breaks down as output length increases.

- **Advanced math** — Multi-step proofs, calculus, competition-level problems. Even Phi-3.5 Mini's strong GSM8K score (86.2) drops hard on MATH (41.3) — the harder benchmark.
- **Nuanced analysis** — Comparing legal documents, analyzing research papers, weighing subtle tradeoffs. These tasks need more parameters.
- **Large context processing** — Most sub-3B models work best with 2048-4096 tokens of context. Feeding them 10-page documents produces unreliable results.
- **Code generation for complex projects** — Small models generate individual functions, not multi-file architectures.

The rule of thumb: if a task requires you to think hard about it, a sub-3B model will struggle with it too. For those tasks, step up to 7B-8B models — they only need 4-5GB of RAM at Q4.

Hardware Requirements

Sub-3B models run on almost anything. Here's exactly how much resources each tier needs:

RAM Requirements (Q4_K_M Quantization)

Model Size	Weights	Total with Context	Minimum RAM
0.5B	~400 MB	~0.8 GB	2 GB
1B	~600 MB	~1.2 GB	2 GB
1.5B	~900 MB	~1.5 GB	4 GB
2B	~1.1 GB	~1.8 GB	4 GB
3B	~1.7 GB	~2.5 GB	4 GB

→ Use our Planning Tool to check exact VRAM for your setup.

"Total with Context" includes the KV cache at 2048-4096 tokens plus runtime overhead. "Minimum RAM" is total system RAM — you need room for the OS and runtime too.

Storage

Downloads are small. A 3B model at Q4 is about 2GB. A 0.5B model is 400MB. You can fit half a dozen sub-3B models in less space than a single 7B model.

Model	Q4_K_M File Size
Qwen 2.5 0.5B	~400 MB
Llama 3.2 1B	~800 MB
Qwen 2.5 1.5B	~1.1 GB
Gemma 2 2B	~1.1 GB
Llama 3.2 3B	~2.0 GB
Phi-3.5 Mini	~2.3 GB

Speed Expectations

Small models are fast. On most hardware, you'll be reading slower than the model generates.

Desktop and Laptop CPUs

CPU	1B Model (Q4)	3B Model (Q4)
Intel i5/Ryzen 5 (laptop)	~20-40 tok/s	~8-15 tok/s
Intel i7/Ryzen 7 (desktop)	~30-60 tok/s	~12-25 tok/s
Apple M1/M2	~35-70 tok/s	~15-30 tok/s
Apple M3 Pro+	~45-90 tok/s	~20-40 tok/s
AMD Ryzen AI 9 (laptop)	~50 tok/s	~18-28 tok/s

Memory bandwidth is the bottleneck, not CPU speed. Dual-channel DDR5 is noticeably faster than DDR4. Single-channel RAM can cut throughput by 50-70% — if your laptop has one RAM stick, that's your limit.

For more on CPU inference, see our CPU-only LLM guide.

Raspberry Pi 5

Model	tok/s	Usability
Qwen 2.5 0.5B	~20	Fast — real-time chat
Qwen 2.5 1.5B	~8-12	Usable — slight pauses

Model	tok/s	Usability
Llama 3.2 3B	~4-6	Slow but functional
7B models	~2-5	Painful — not recommended

Stick to 1B-1.5B on a Pi 5 for a good experience. 3B is possible but you'll feel the wait. Use active cooling — all four cores hit 100% during inference.

Old GPUs

If you have a dedicated GPU, even an old one, it helps:

GPU	VRAM	What Fits	Advantage
GTX 1050 Ti	4GB	3B at Q4 comfortably	2-3x faster than CPU-only
GTX 1060	6GB	3B at Q8, or 7B at Q4	Enough for 7B models
RX 580	8GB	3B at FP16	Full precision, no quantization needed

Even a 4GB GPU fully offloads a 3B Q4 model (weights are ~1.7GB), giving a significant speed boost over CPU inference.

Phones

Phone Tier	Model	Speed
Flagship (Snapdragon 8 Gen 3, A17 Pro)	1B-3B at Q4	8-17 tok/s
Mid-range (Snapdragon 7 Gen 1)	0.5B-1.5B at Q4	5-10 tok/s
Budget (6GB RAM or less)	0.5B at Q4	Barely loads

Apps like SmolChat (Android) and MLC Chat (iOS/Android) make this straightforward. Be warned: sustained inference drains battery fast — comparable to a graphics-intensive game.

How to Run Them

Ollama (Easiest)

Ollama is one command to install, one command to run:

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

# Pull and run a model
ollama pull qwen2.5:3b
ollama run qwen2.5:3b
```

That's it. Ollama auto-detects your hardware and optimizes accordingly. No GPU required.

LM Studio (GUI)

Prefer a visual interface? LM Studio gives you a ChatGPT-like UI for local models. Download, search for a model, click run. It handles GGUF quantization selection for you.

Raspberry Pi

On a Pi 5, Ollama works out of the box:

```
curl -fsSL https://ollama.com/install.sh | sh
ollama pull qwen2.5:1.5b
ollama run qwen2.5:1.5b
```

For better performance on a Pi, consider building llama.cpp with OpenBLAS — it's 10-20% faster than Ollama for sustained inference.

Phones

- **Android:** SmolChat, MLC Chat, or any app that supports GGUF models
- **iOS:** MLC Chat, or LLM Farm
- **Cross-platform apps:** llama.rn (React Native bindings for llama.cpp)

When to Stay Small vs. Upgrade to 7B

This is the real question. Here's the decision framework:

Stay with sub-3B if:

- Your hardware maxes out at 4GB RAM/VRAM
- You're running on a Raspberry Pi, phone, or edge device
- Your tasks are quick Q&A, classification, extraction, or simple code
- Speed matters more than depth (you need real-time responses)
- You want to run alongside other applications without memory pressure
- Privacy/offline is the priority and quality is secondary

Step up to 7B-8B if:

- You have 8GB+ RAM or any GPU with 6GB+ VRAM
- You need multi-step reasoning, longer outputs, or complex coding
- Quality per response matters more than speed
- You're hitting the limits of 3B output quality

The jump from 3B to 7B is the single biggest quality improvement in local AI. A Llama 3.1 8B at Q4 uses about 5GB of RAM and is dramatically more capable. If your hardware can handle it, it's worth the step — see our 8GB VRAM guide for details.

But if your hardware can't handle 7B, don't feel locked out. A Qwen 2.5 3B today is more useful than a 7B model from two years ago. The floor has risen.

Recommendations by Use Case

Use Case	Best Pick	Runner-Up	Why
General chat/Q&A	Qwen 2.5 3B	Llama 3.2 3B	Strongest overall quality
Coding assistance	Phi-3.5 Mini (3.8B)	Qwen 2.5 3B	62.8 HumanEval — real coding ability
Math/reasoning	Phi-3.5 Mini (3.8B)	Qwen 2.5 3B	86.2 GSM8K, untouchable at this size
Classification/extraction	Gemma 2 2B	Qwen 2.5 1.5B	Strong language understanding, efficient
Raspberry Pi 5	Qwen 2.5 1.5B	Llama 3.2 1B	Best quality at comfortable Pi speeds
Phone	Llama 3.2 1B	Qwen 2.5 0.5B	Fast, low battery drain

Use Case	Best Pick	Runner-Up	Why
Edge/IoT	Qwen 2.5 0.5B	Llama 3.2 1B	Under 1GB RAM, 400MB download
Multilingual	Llama 3.2 3B	Qwen 2.5 3B	58.2 MGSM, 8 languages
Absolute minimum hardware	Qwen 2.5 0.5B	—	Runs on 2GB RAM, 400MB storage

---

The Bottom Line

Small models are no longer a consolation prize. They're a legitimate way to run AI locally on hardware you already own — no GPU required, no cloud dependency, no subscription.

The practical advice:

1. **Have 4GB+ RAM?** Start with `ollama pull qwen2.5:3b`. You'll have a working local AI assistant running at 10-25+ tok/s on CPU alone.
2. **Only 2GB RAM or a Pi?** Pull `qwen2.5:1.5b`. It's 1.1GB to download, needs under 2GB of RAM, and handles most simple tasks.
3. **Building for edge/IoT?** `qwen2.5:0.5b` is 400MB and runs on almost anything.
4. **Need coding or math?** `phi3.5` (3.8B) is the strongest small model for technical tasks.
5. **When you outgrow 3B** — and you'll know when the answers aren't good enough — a 7B model on 8GB of VRAM is the next step.

Your laptop is more capable than you think. Try it.

Related Guides

- Run Your First Local LLM in 15 Minutes
- CPU-Only LLMs: What Actually Works
- What Can You Actually Run on 8GB VRAM?
- What Quantization Actually Means

Sources: Qwen 2.5 Technical Report, Meta Llama 3.2 Model Card, Phi-3 Technical Report, Gemma 2 Technical Report, StableLM 2 Technical Report, Raspberry Pi 5 LLM Benchmarks, CPU vs GPU LLM Performance

---