What a small language model cannot do, *belief by belief*

Six common beliefs about small language models, each checked against published measurements and given a verdict: supported, partly supported, not supported or not measured.

MITWA LAB, MUMBAI

IN PLAIN WORDS

One recent survey calls a language model small when it has between 100 million and 5 billion internal numbers, called parameters, a range chosen with phones and other personal devices in mind. This review checks six common beliefs about what such models cannot do against published measurements. The evidence supports the belief that small models hold fewer facts. It contradicts the belief that shrinking a model leaves it unchanged, since compression can alter a model in ways a single test score hides. Three beliefs hold only in part. Larger models share the weaknesses with reasoning and with long inputs, and a small model behaves like a scaled-down large one on average but not on every task. Whether a small model is safer has not been measured in a way that settles it.

Where the line for small is drawn

Lu et al. (2024) surveyed small language models and benchmarked those with public weights. They define the category as decoder-only transformers with 100 million to 5 billion parameters and set the upper bound there because models of 7 billion were still mostly served from the cloud in 2024. Microsoft's builders describe Phi-3-mini, at 3.8 billion parameters, as small enough to deploy on a phone (Abdin et al., 2024). The beliefs below mostly concern models in or near this range, though several studies also test far larger ones, and each section ends with a verdict.

It knows fewer facts

The most exact number here comes from a controlled setting. Allen-Zhu and Li (2024) trained models on synthetic biographies so that they could count how much factual knowledge each one held. When each fact appeared about 1,000 times in training, models across the architectures they tested stored about 2 bits of knowledge per parameter, and facts seen only 100 times were stored at about 1 bit per parameter. Because this is a capacity, it follows that a model a tenth as large can hold at most about a tenth as many facts, other things equal. The data were built for the experiment, and no deployed model is measured.

The builders of one small model report the same limit in their own product. Abdin et al. (2024) scored Phi-3-mini and several larger models through one internal evaluation pipeline. On MMLU, one of the reasoning benchmarks in their table, Phi-3-mini scored 68.8 against 71.4 for GPT-3.5. On TriviaQA, which they use as evidence about stored facts, the distance was 64.0 against 85.8. In the same table the 7-billion-parameter Phi-3-small scored 58.1 on TriviaQA, below Phi-3-mini, so size was not the only factor even within one family. The authors write that the model lacks the capacity to hold much factual knowledge and propose connecting it to a search engine. Apple's documentation for its on-device model likewise refers developers whose tasks need wide world knowledge to its pages on guided generation and tool calling (Apple, 2026a). On a phone such a lookup goes either to a store kept on the device or to a server, and the second sends the question off the phone. That last point is our inference; none of the sources we opened measures the trade.

On this evidence the belief is supported, and the size of the gap depends on how often a fact was seen in training. Abdin et al. (2024) expect a search engine to close it, but none of the sources measured whether it does.

It cannot reason

Part of this belief has support. Lu et al. (2024) find small models closing the distance to large ones on general reasoning benchmarks, with differences remaining in mathematics, and a wide gap on complex reasoning between small models trained on open data and those trained on closed data. Apple lists basic arithmetic, writing code and logical reasoning among requests its on-device model may handle poorly (Apple, 2026a), and points developers to a server model for more reasoning ability (Apple, 2026b). Against this, in the same survey, Phi-3-mini led LLaMA 3.1, the larger reference model in that comparison, on the mathematics task by 14.5%.

Mirzadeh et al. (2025) asked whether high scores on grade-school maths reflect reasoning. They rebuilt the GSM8K questions as templates so that names and numbers could vary, and made a variant that adds one clause which sounds relevant but does not change the answer. They tested more than 20 open models between 2 and 27 billion parameters, and several closed ones. Every model lost accuracy when the irrelevant clause was added. Within Microsoft's Phi-3 family, whose three sizes are 3.8, 7 and 14 billion parameters (Abdin et al., 2024), the 128k versions scored 18.0, 24.5 and 29.4% on that variant, after scores of 83.7, 88.5 and 87.3% on the original benchmark. GPT-4o went from 95.2 to 63.1%, and o1-preview from 94.9 to 77.4%. The authors read the drops as evidence that current models copy reasoning steps seen in training without carrying out logic; that is their hypothesis about a cause, while the measurements show only the fragility.

So the belief holds in part. Within the Phi-3 family the smaller models fell further on the distracting variant, yet the weakness appeared in every model tested, and the best small models score well on standard tests.

It cannot read long inputs

Lu et al. (2024) hid one or more facts in long texts of up to about 32,000 tokens and asked models to retrieve them or reason over them. Two models of 0.5 billion parameters averaged 22.13% and 43.84%. Two models of about 3 billion averaged 57.81% (Llama3.2-3B) and 91.71% (Qwen2.5-3B), and the second stayed close to perfect at every length and depth tested. Lu et al. themselves conclude that more parameters matter for long inputs. The gap of about 34 points between two models of similar size suggests that training matters as well, though they do not isolate what. They also report that small models in general did worst when the fact sat in the middle of the text, with Qwen2.5-3B the exception.

That middle weakness is not peculiar to small models. Liu et al. (2023) gave open and closed models, among them GPT-3.5-Turbo and Claude-1.3, a question with many retrieved documents. Accuracy was highest when the answering document came first or last, and with it in the middle GPT-3.5-Turbo did worse than it did with no documents at all (56.1%). Length also costs larger models. On the RULER benchmark, Phi-3.5-mini scored 94.3 at 4,000 tokens and 63.6 at 128,000, while Llama-3.1-8B went from 95.5 to 77.0; for their own model, the authors suspect too little good long-text data in mid-training (Abdin et al., 2024). Apple's on-device model accepts up to 4,096 tokens per session (Apple, 2026a), and Apple directs developers to a server for longer inputs (Apple, 2026b).

This belief too holds only in part. The smallest models did poorly, at around 3 billion parameters the answer depends on the model, and losing track of the middle appeared in small models and in most of the larger models Liu et al. (2023) tested on questions over many documents, though Qwen2.5-3B escaped it.

It is a scaled-down large model

The belief here is that a small model does everything a large one does, only somewhat worse. Kaplan et al. (2020) support an averaged version of it. Test loss, a measure of error in predicting the next word, fell along a power law in model size, data and compute, some of the trends holding over a range of more than seven orders of magnitude, and the balance between width and depth mattered little across a wide range. Loss is an average over text, though, and can hide differences between tasks.

Wei et al. (2022) catalogued abilities that sat near chance in smaller models and appeared only above some scale. GPT-3, for instance, rose above chance on a BIG-Bench arithmetic task only at 13 billion parameters. Schaeffer et al. (2023) found that over 92% of the emergent abilities claimed on BIG-Bench were scored with all-or-nothing metrics, and that with continuous metrics the models they re-examined improved smoothly, which fits the belief; they state that this does not show emergence can never occur. Wei et al. (2022) themselves note that PaLM 62B performed above chance on 14

BIG-Bench tasks where the larger LaMDA 137B and GPT-3 175B did not, so scale was not the only factor.

Small models may also need a different design. Liu et al. (2024) found that below a billion parameters depth mattered more than width, which runs against Kaplan et al. (2020), and their 350-million-parameter model came close to LLaMA-v2 7B in exact-match score on one task of calling software functions. Eldan and Li (2023) trained models under 10 million parameters on stories written with words a child of three or four would know, and the models wrote fluent, consistent stories of several paragraphs; the authors scored the stories with GPT-4 acting as a teacher.

The evidence pulls both ways, so the verdict is partly supported. Average error rises smoothly as size falls, and some abilities once called emergent also change smoothly when measured continuously. Against that stand the abilities Wei et al. (2022) saw appear only above some scale, a design rule that changes below a billion parameters, and a narrow task and a narrow domain where very small models came close to far larger ones.

Shrinking it changes nothing

Models are made smaller after training by pruning, which removes weights, or by quantization, which stores each weight in fewer bits. The belief is that an unchanged benchmark score means an unchanged model. Jaiswal et al. (2024) argue that perplexity, the usual check, misses real losses. On their knowledge-intensive benchmark every pruning method they tested lost substantial accuracy, sometimes with only 25 to 30% of weights removed, while quantization held up better, and models pruned by half or more still did well at retrieving and summarising from text given in the input. Dutta et al. (2024) compared six quantization schemes with the 16-bit original on seven benchmarks. Accuracy stayed within 0 to 2%, yet all schemes except one changed many individual answers, up to 13.6% of them, with right answers turning wrong and wrong ones right in roughly equal numbers.

In the synthetic setting of Allen-Zhu and Li (2024), 8-bit integers kept the full knowledge capacity while 4-bit integers cut it to about 0.7 bits per parameter. Hong et al. (2024) compressed three models with five methods and scored eight aspects of trustworthiness. At 4 bits the models kept the trust scores of their originals, pruning by half lowered them, and 3-bit quantization lowered them considerably, in ways the ordinary task scores did not reveal.

The belief is not supported. A benchmark can establish that nothing measurable changed on that benchmark, and the finding reaches no further than what it tests.

It is safer because it is small

For one risk the evidence favours small models. Carlini et al. (2021) recovered hundreds of passages memorised word for word from GPT-2's training data and found larger models more exposed. The 1.5-billion-parameter GPT-2 memorised over 18 times as much as the 124-million one in one of their settings. Lin et al. (2022) built TruthfulQA, 817 questions in 38 categories that some people answer wrongly because of common misconceptions, and found that truthfulness generally fell as the models grew. In the multiple-choice version, the 6B GPT-Neo/J model scored 12% lower on truthfulness than its 125M sibling, although no model there did significantly better than random guessing. In their worked example the smallest model stayed truthful only by saying nothing useful. The benchmark was also built adversarially against GPT-3 models, and Wei et al. (2022) report that Gopher rose more than 20% above random on it only at its largest size, 280 billion parameters. Hong et al. (2024) found that 7-billion-parameter models trained from scratch may beat 13-billion ones on some trust measures.

Microsoft's internal benchmark, reported by Abdin et al. (2024), gives a mixed picture. GPT-4 simulated multi-turn conversations and scored the replies, and the three Phi-3 sizes had been through the same red-teaming with the same safety data, the larger two with slightly more examples. Lower is better on every measure in that table. On ungroundedness, scored from 0 (fully grounded in the material supplied) to 4, the smallest did worst, at 0.603 against 0.299 and 0.213. On the table's other measures the three sizes were close. Jailbreak defect rates were 0.123, 0.107 and 0.111, and the smallest scored best on third-party harm (0.240 against 0.253 and 0.251) and on harmful summarisation (0.100 against 0.110 and 0.112).

As a general claim the belief has not been measured. The answer depends on which harm is meant, and the comparisons across sizes that we opened cover at most four model families and a few kinds of harm.

Six verdicts in one table

BELIEF	VERDICT	WHAT DECIDES IT
It knows fewer facts	Supported	Size and how often a fact was seen; lookup proposed but unmeasured
It cannot reason	Partly supported	Distractors hurt every model tested, the smaller Phi-3 models most
It cannot read long inputs	Partly supported	Training of the particular model; middle losses in most models tested
It is a scaled-down large model	Partly supported	True of average error; data, design and narrow tasks can break it
Shrinking it changes nothing	Not supported	Method, bit width and what the test measures
It is safer because it is small	Not measured	Which harm, and whether training was matched

Our companion, Bloom, ships as a 2.0 GB build for iPhone 16 and newer, so we work under the constraints this literature describes. This review makes no claim about where Bloom sits on any scale above or which of these limits apply to it; technical note 03 sets out what the finding of no measurable loss between the 6.2 GB and 2.0 GB builds can and cannot prove, and recall at the edges of memory is on our list of things not yet proved.

What is not yet known

None of the studies we opened follows a small model through long personal conversations spread over weeks and many sessions. The retrieval tests above hide a fact in one long text, which is a different task from remembering across separate conversations.

The effect of size on safety rests here on comparisons within single model families, none with all training held fixed (the Phi-3 sizes shared safety training but not the same pre-training), and one of them is the builder's internal benchmark scored by another model. An independent comparison across many families with matched safety training was not among the sources.

Whether a lookup closes the knowledge gap for a model on a phone, and what it costs in privacy when the lookup leaves the device, has not been measured in anything we read.

The 4-bit loss in knowledge capacity was measured on synthetic biographies. Whether it carries over to models trained on natural text is open, and so is whether the answers that flip under compression cluster in particular kinds of question.

Lu et al. (2024) found small models improving faster than the LLaMA 7B series between 2022 and 2024, and also caution that the best family's scores may partly reflect overfitting to benchmarks. How much of the recent gain is capability, and whether the pace holds, is unsettled.

References

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., et al. (2024). Phi-3 technical report: A highly capable language model locally on your phone. Microsoft. arXiv:2404.14219. arxiv.org/abs/2404.14219

Allen-Zhu, Z., and Li, Y. (2024). Physics of language models: Part 3.3, knowledge capacity scaling laws. arXiv:2404.05405. arxiv.org/abs/2404.05405

Apple Inc. (2026a). Generating content and performing tasks with Foundation Models. Apple Developer Documentation. Undated page, opened 12 November 2026. developer.apple.com/documentation/foundationmodels/generating-content-and-performing-tasks-with-foundation-models

Apple Inc. (2026b). Foundation Models. Apple Developer Documentation, framework overview. Undated page, opened 12 November 2026. developer.apple.com/documentation/foundationmodels

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C. (2021). Extracting training data from large language models. arXiv:2012.07805, version 2 (June 2021). arxiv.org/abs/2012.07805

Dutta, A., Krishnan, S., Kwatra, N., and Ramjee, R. (2024). Accuracy is not all you need. arXiv:2407.09141. arxiv.org/abs/2407.09141

Eldan, R., and Li, Y. (2023). TinyStories: How small can language models be and still speak coherent English? arXiv:2305.07759. arxiv.org/abs/2305.07759

Hong, J., Duan, J., Zhang, C., Li, Z., Xie, C., Lieberman, K., Diffenderfer, J., Bartoldson, B., Jaiswal, A., Xu, K., Kailkhura, B., Hendrycks, D., Song, D., Wang, Z., and Li, B. (2024). Decoding compressed trust: Scrutinizing the trustworthiness of efficient LLMs under compression. In Proceedings of the 41st International Conference on Machine Learning (ICML 2024), PMLR 235. arXiv:2403.15447. arxiv.org/abs/2403.15447

Jaiswal, A., Gan, Z., Du, X., Zhang, B., Wang, Z., and Yang, Y. (2024). Compressing LLMs: The truth is rarely pure and never simple. In International Conference on Learning Representations (ICLR 2024). arXiv:2310.01382. arxiv.org/abs/2310.01382

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. arXiv:2001.08361. arxiv.org/abs/2001.08361

Lin, S., Hilton, J., and Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022), per the arXiv listing. arXiv:2109.07958. arxiv.org/abs/2109.07958

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2023). Lost in the middle: How language models use long contexts. Accepted for publication in Transactions of the

Association for Computational Linguistics, per the arXiv listing; read as arXiv:2307.03172v3. arxiv.org/abs/2307.03172

Liu, Z., Zhao, C., Iandola, F., Lai, C., Tian, Y., Fedorov, I., Xiong, Y., Chang, E., Shi, Y., Krishnamoorthi, R., Lai, L., and Chandra, V. (2024). MobileLLM: Optimizing sub-billion parameter language models for on-device use cases. In Proceedings of the 41st International Conference on Machine Learning (ICML 2024), PMLR 235. arXiv:2402.14905. arxiv.org/abs/2402.14905

Lu, Z., Li, X., Cai, D., Yi, R., Liu, F., Zhang, X., Lane, N. D., and Xu, M. (2024). Small language models: Survey, measurements, and insights. arXiv:2409.15790. arxiv.org/abs/2409.15790

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., and Farajtabar, M. (2025). GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. In International Conference on Learning Representations (ICLR 2025). arXiv:2410.05229. arxiv.org/abs/2410.05229

Schaeffer, R., Miranda, B., and Koyejo, S. (2023). Are emergent abilities of large language models a mirage? arXiv:2304.15004. arxiv.org/abs/2304.15004

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. (2022). Emergent abilities of large language models. Transactions on Machine Learning Research. arXiv:2206.07682. arxiv.org/abs/2206.07682