Emergent Abilities of Large Language Models

Jason Wei [1] *jasonwei@google.com*
Yi Tay [1] *yitay@google.com*
Rishi Bommasani [2] *nlprishi@stanford.edu*
Colin Raffel [3] *craffel@gmail.com*
Barret Zoph [1] *barretzoph@google.com*
Sebastian Borgeaud [4] *sborgeaud@deepmind.com*
Dani Yogatama [4] *dyogatama@deepmind.com*
Maarten Bosma [1] *bosma@google.com*
Denny Zhou [1] *dennyzhou@google.com*
Donald Metzler [1] *metzler@google.com*
Ed H. Chi [1] *edchi@google.com*
Tatsunori Hashimoto [2] *thashim@stanford.edu*
Oriol Vinyals [4] *vinyals@deepmind.com*
Percy Liang [2] *pliang@stanford.edu*
Jeff Dean [1] *jeff@google.com*
William Fedus [1] *liamfedus@google.com*

[1] *Google Research* [2] *Stanford University* [3] *UNC Chapel Hill* [4] *DeepMind*

Reviewed on OpenReview: *https://openreview.net/forum?id=yzkSU5zdwD*

Abstract

Scaling up language models has been shown to predictably improve performance and sample efficiency on a wide range of downstream tasks. This paper instead discusses an unpredictable phenomenon that we refer to as *emergent abilities* of large language models. We consider an ability to be emergent if it is not present in smaller models but is present in larger models. Thus, emergent abilities cannot be predicted simply by extrapolating the performance of smaller models. The existence of such emergence raises the question of whether additional scaling could potentially further expand the range of capabilities of language models.

1 Introduction

Language models have revolutionized natural language processing (NLP) in recent years. It is now well-known that increasing the scale of language models (e.g., training compute, model parameters, etc.) can lead to better performance and sample efficiency on a range of downstream NLP tasks (Devlin et al., 2019; Brown et al., 2020, *inter alia*). In many cases, the effect of scale on performance can often be methodologically predicted via scaling laws—for example, scaling curves for cross-entropy loss have been shown to empirically span more than seven orders of magnitude (Kaplan et al., 2020; Hoffmann et al., 2022). On the other hand, performance for certain downstream tasks counterintuitively does not appear to continuously improve as a function of scale, and such tasks cannot be predicted ahead of time (Ganguli et al., 2022).

In this paper, we will discuss the unpredictable phenomena of *emergent abilities* of large language models. Emergence as an idea has been long discussed in domains such as physics, biology, and computer science (Anderson, 1972; Hwang et al., 2012; Forrest, 1990; Corradini & O'Connor, 2010; Harper & Lewis, 2012, *inter*

alia). We will consider the following general definition of emergence, adapted from Steinhardt (2022) and rooted in a 1972 essay called "More Is Different" by Nobel prize-winning physicist Philip Anderson (Anderson, 1972):

> *Emergence is when quantitative changes in a system result in qualitative changes in behavior.*

Here we will explore emergence with respect to model scale, as measured by training compute and number of model parameters. Specifically, we define *emergent abilities of large language models* as abilities that are not present in smaller-scale models but are present in large-scale models; thus they cannot be predicted by simply extrapolating the performance improvements on smaller-scale models (§2).[1] We survey emergent abilities as observed in a range of prior work, categorizing them in settings such as few-shot prompting (§3) and augmented prompting strategies (§4). Emergence motivates future research on why such abilities are acquired and whether more scaling will lead to further emergent abilities, which we highlight as important questions for the field (§5).

2 Emergent Abilities Definition

As a broad concept, emergence is often used informally and can be reasonably interpreted in many different ways. In this paper, we will consider a focused definition of emergent abilities of large language models:

> *An ability is emergent if it is not present in smaller models but is present in larger models.*

Emergent abilities would not have been directly predicted by extrapolating a scaling law (i.e. consistent performance improvements) from small-scale models. When visualized via a scaling curve (x-axis: model scale, y-axis: performance), emergent abilities show a clear pattern—performance is near-random until a certain critical threshold of scale is reached, after which performance increases to substantially above random. This qualitative change is also known as a *phase transition*—a dramatic change in overall behavior that would not have been foreseen by examining smaller-scale systems (Huberman & Hogg, 1987).

Today's language models have been scaled primarily along three factors: amount of computation, number of model parameters, and training dataset size (Kaplan et al., 2020; Hoffmann et al., 2022). In this paper, we will analyze scaling curves by plotting the performance of different models where training compute for each model is measured in FLOPs on the x-axis (Hoffmann et al., 2022). Because language models trained with more compute tend to also have more parameters, we additionally show plots with number of model parameters as the x-axis in Appendix D (see Figure 11 and Figure 12, as well as Figure 4 and Figure 10). Using training FLOPs or model parameters as the x-axis produces curves with similar shapes due to the fact that most dense Transformer language model families have scaled training compute roughly proportionally with model parameters (Kaplan et al., 2020).

Training dataset size is also an important factor, but we do not plot capabilities against it because many language model families use a fixed number of training examples for all model sizes (Brown et al., 2020; Rae et al., 2021; Chowdhery et al., 2022). Although we focus on training computation and model size here, there is not a single proxy that adequately captures all aspects of scale. For example, Chinchilla (Hoffmann et al., 2022) has one-fourth as many parameters as Gopher (Rae et al., 2021) but uses similar training compute; and sparse mixture-of-expert models have more parameters per training/inference compute than dense models (Fedus et al., 2021; Du et al., 2021). Overall, it may be wise to view emergence as a function of many correlated variables. For example, later in Figure 4 we will also plot emergence as a function of WikiText103 perplexity (Merity et al., 2016), which happens to closely correlate with training computation for Gopher/ Chinchilla (though this correlation may not hold in the long-run).

Note that the scale at which an ability is first observed to emerge depends on a number of factors and is not an immutable property of the ability. For instance, emergence may occur with less training compute

[1] This survey focuses on pre-trained Transformer language models. Emergent abilities in NLP more broadly, however, could go back to Miller et al. (2004), Liang (2005), or earlier.

or fewer model parameters for models trained on higher-quality data. Conversely, emergent abilities also crucially depend on other factors such as not being limited by the amount of data, its quality, or the number of parameters in the model. Today's language models are likely not trained optimally (Hoffmann et al., 2022), and our understanding of how to best train models will evolve over time. Our goal in this paper is not to characterize or claim that a specific scale is required to observe emergent abilities, but rather, we aim to discuss examples of emergent behavior in prior work.

3 Few-Shot Prompted Tasks

We first discuss emergent abilities in the *prompting* paradigm, as popularized by GPT-3 (Brown et al., 2020).[2] In prompting, a pre-trained language model is given a prompt (e.g. a natural language instruction) of a task and completes the response without any further training or gradient updates to its parameters. Brown et al. (2020) proposed *few-shot prompting*, which includes a few input-output examples in the model's context (input) as a preamble before asking the model to perform the task for an unseen inference-time example. An example prompt is shown in Figure 1.

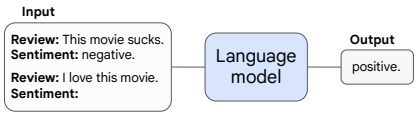

Figure 1: Example of an input and output for few-shot prompting.

The ability to perform a task via few-shot prompting is emergent when a model has random performance until a certain scale, after which performance increases to well-above random. Figure 2 shows eight such emergent abilities spanning five language model families from various work.

BIG-Bench. Figure 2A–D depicts four emergent few-shot prompted tasks from BIG-Bench, a crowd-sourced suite of over 200 benchmarks for language model evaluation (BIG-Bench, 2022). Figure 2A shows an arithmetic benchmark that tests 3-digit addition and subtraction, as well as 2-digit multiplication. GPT-3 and LaMDA (Thoppilan et al., 2022) have close-to-zero performance for several orders of magnitude of training compute, before performance jumps to sharply above random at $2 \cdot 10^{22}$ training FLOPs (13B parameters) for GPT-3, and 10^{23} training FLOPs (68B parameters) for LaMDA. Similar emergent behavior also occurs at around the same model scale for other tasks, such as transliterating from the International Phonetic Alphabet (Figure 2B), recovering a word from its scrambled letters (Figure 2C), and Persian question-answering (Figure 2D). Even more emergent abilities from BIG-Bench are given in Appendix E.

TruthfulQA. Figure 2E shows few-shot prompted performance on the TruthfulQA benchmark, which measures the ability to answer questions truthfully (Lin et al., 2021). This benchmark is adversarially curated against GPT-3 models, which do not perform above random, even when scaled to the largest model size. Small Gopher models also do not perform above random until scaled up to the largest model of $5 \cdot 10^{23}$ training FLOPs (280B parameters), for which performance jumps to more than 20% above random (Rae et al., 2021).

Grounded conceptual mappings. Figure 2F shows the task of grounded conceptual mappings, where language models must learn to map a conceptual domain, such as a cardinal direction, represented in a textual grid world (Patel & Pavlick, 2022). Again, performance only jumps to above random using the largest GPT-3 model.

Multi-task language understanding. Figure 2G shows the Massive Multi-task Language Understanding (MMLU) benchmark, which aggregates 57 tests covering a range of topics including math, history, law, and more (Hendrycks et al., 2021a). For GPT-3, Gopher, and Chinchilla, models of $\sim 10^{22}$ training FLOPs ($\sim$10B parameters) or smaller do not perform better than guessing on average over all the topics, scaling up to 3–5 $\cdot 10^{23}$ training FLOPs (70B–280B parameters) enables performance to substantially surpass random. This result is striking because it could imply that the ability to solve knowledge-based questions spanning a large collection of topics might require scaling up past this threshold (for dense language models without retrieval or access to external memory).

[2]Though GPT-3 popularized prompting, the task setup has existed since before GPT-3 (Trinh & Le, 2018; McCann et al., 2018; Radford et al., 2019; Raffel et al., 2020).

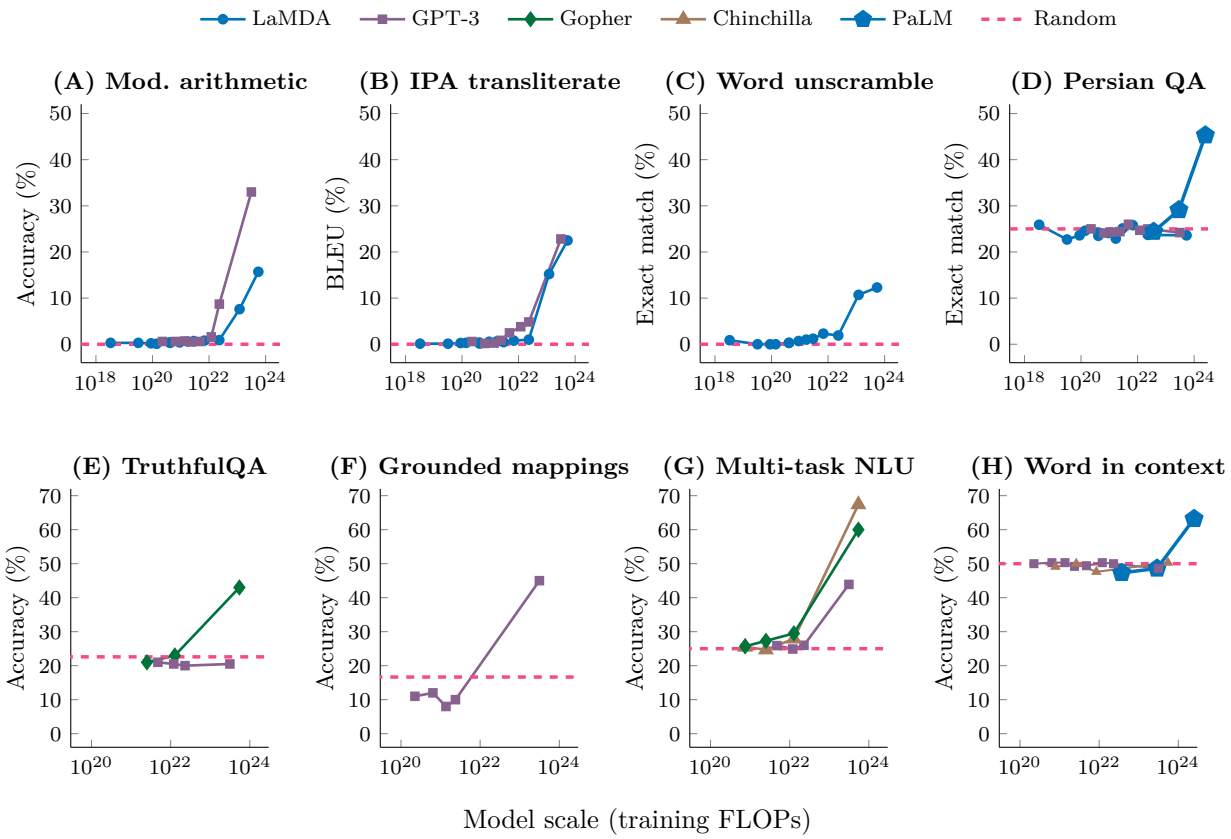

Figure 2: Eight examples of emergence in the few-shot prompting setting. Each point is a separate model. The ability to perform a task via few-shot prompting is emergent when a language model achieves random performance until a certain scale, after which performance significantly increases to well-above random. Note that models that used more training compute also typically have more parameters—hence, we show an analogous figure with number of model parameters instead of training FLOPs as the x-axis in Figure 11. A–D: BIG-Bench (2022), 2-shot. E: Lin et al. (2021) and Rae et al. (2021). F: Patel & Pavlick (2022). G: Hendrycks et al. (2021a), Rae et al. (2021), and Hoffmann et al. (2022). H: Brown et al. (2020), Hoffmann et al. (2022), and Chowdhery et al. (2022) on the WiC benchmark (Pilehvar & Camacho-Collados, 2019).

Word in Context. Finally, Figure 2H shows the Word in Context (WiC) benchmark (Pilehvar & Camacho-Collados, 2019), which is a semantic understanding benchmark. Notably, GPT-3 and Chinchilla fail to achieve one-shot performance of better than random, even when scaled to their largest model size of $\sim 5 \cdot 10^{23}$ FLOPs. Although these results so far may suggest that scaling alone may not enable models to solve WiC, above-random performance eventually emerged when PaLM was scaled to $2.5 \cdot 10^{24}$ FLOPs (540B parameters), which was much larger than GPT-3 and Chinchilla.

4 Augmented Prompting Strategies

Although few-shot prompting is perhaps currently the most common way of interacting with large language models, recent work has proposed several other prompting and finetuning strategies to further augment the abilities of language models. If a technique shows no improvement or is harmful when compared to the baseline of not using the technique until applied to a model of a large-enough scale, we also consider the technique an emergent ability.

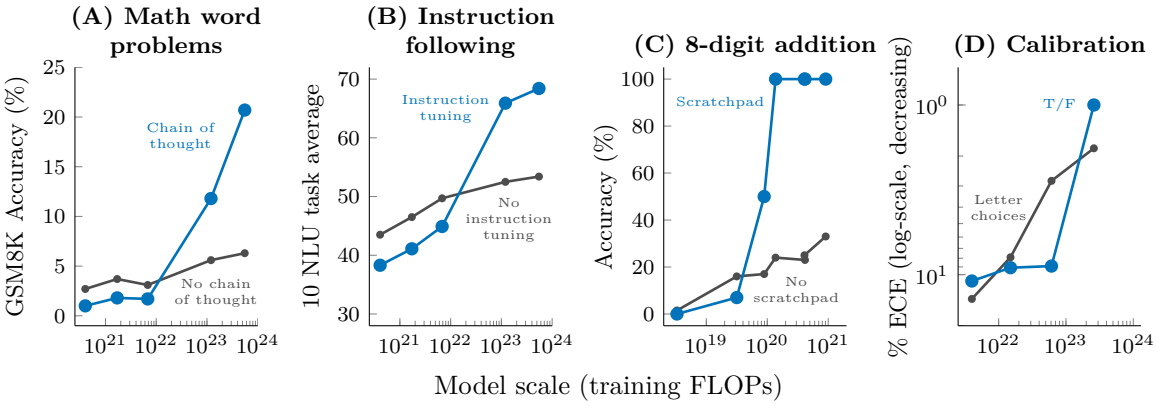

Figure 3: Specialized prompting or finetuning methods can be emergent in that they do not have a positive effect until a certain model scale. A: Wei et al. (2022b). B: Wei et al. (2022a). C: Nye et al. (2021). D: Kadavath et al. (2022). An analogous figure with number of parameters on the x-axis instead of training FLOPs is given in Figure 12. The model shown in A-C is LaMDA (Thoppilan et al., 2022), and the model shown in D is from Anthropic.

Multi-step reasoning. Reasoning tasks, especially those involving multiple steps, have been challenging for language models and NLP models more broadly (Rae et al., 2021; Bommasani et al., 2021; Nye et al., 2021). A recent prompting strategy called chain-of-thought prompting enables language models to solve such problems by guiding them to produce a sequence of intermediate steps before giving the final answer (Cobbe et al., 2021; Wei et al., 2022b; Suzgun et al., 2022). As shown in Figure 3A, chain of thought prompting only surpasses standard prompting without intermediate steps when scaled to 10^{23} training FLOPs (∼100B parameters). A similar emergence in performance gain was also observed when augmenting few-shot prompting with explanations that came after the final answer (Lampinen et al., 2022).

Instruction following. Another growing line of work aims to better enable language models to perform new tasks simply by reading instructions describing the task (without few-shot exemplars). By finetuning on a mixture of tasks phrased as instructions, language models have been shown to respond appropriately to instructions describing an unseen task (Ouyang et al., 2022; Wei et al., 2022a; Sanh et al., 2022; Chung et al., 2022). As shown in Figure 3B, Wei et al. (2022a) found that this instruction-finetuning technique hurts performance for models of $7 \cdot 10^{21}$ training FLOPs (8B parameters) or smaller, and only improves performance when scaled to 10^{23} training FLOPs (∼100B parameters) (though Sanh et al. (2022) found shortly after that this instruction-following behavior could be also induced by finetuning smaller encoder-decoder T5 models).

Program execution. Consider computational tasks involving multiple steps, such as adding large numbers or executing computer programs. Nye et al. (2021) show that finetuning language models to predict intermediate outputs ("scratchpad") enables them to successfully execute such multi-step computations. As shown in Figure 3C, on 8-digit addition, using a scratchpad only helps for models of $\sim 9 \cdot 10^{19}$ training FLOPs (40M parameters) or larger.

Model calibration. Finally, an important direction for deployment of language models studies is *calibration*, which measures whether models can predict which questions they will be able to answer correctly. Kadavath et al. (2022) compared two ways of measuring calibration: a True/False technique, where models first propose answers and then evaluate the probability "P(True)" that their answers are correct, and more-standard methods of calibration, which use the probability of the correct answer compared with other answer options. As shown in Figure 3D, the superiority of the True/False technique only emerges when scaled to the largest model scale of $\sim 3 \cdot 10^{23}$ training FLOPs (52B parameters).

Table 1: List of emergent abilities of large language models and the scale (both training FLOPs and number of model parameters) at which the abilities emerge.

	Emergent scale			
	Train. FLOPs	Params.	Model	Reference
Few-shot prompting abilities				
• Addition/subtraction (3 digit)	2.3E+22	13B	GPT-3	Brown et al. (2020)
• Addition/subtraction (4-5 digit)	3.1E+23	175B		
• MMLU Benchmark (57 topic avg.)	3.1E+23	175B	GPT-3	Hendrycks et al. (2021a)
• Toxicity classification (CivilComments)	1.3E+22	7.1B	Gopher	Rae et al. (2021)
• Truthfulness (Truthful QA)	5.0E+23	280B		
• MMLU Benchmark (26 topics)	5.0E+23	280B		
• Grounded conceptual mappings	3.1E+23	175B	GPT-3	Patel & Pavlick (2022)
• MMLU Benchmark (30 topics)	5.0E+23	70B	Chinchilla	Hoffmann et al. (2022)
• Word in Context (WiC) benchmark	2.5E+24	540B	PaLM	Chowdhery et al. (2022)
• Many BIG-Bench tasks (see Appendix E)	Many	Many	Many	BIG-Bench (2022)
Augmented prompting abilities				
• Instruction following (finetuning)	1.3E+23	68B	FLAN	Wei et al. (2022a)
• Scratchpad: 8-digit addition (finetuning)	8.9E+19	40M	LaMDA	Nye et al. (2021)
• Using open-book knowledge for fact checking	1.3E+22	7.1B	Gopher	Rae et al. (2021)
• Chain-of-thought: Math word problems	1.3E+23	68B	LaMDA	Wei et al. (2022b)
• Chain-of-thought: StrategyQA	2.9E+23	62B	PaLM	Chowdhery et al. (2022)
• Differentiable search index	3.3E+22	11B	T5	Tay et al. (2022b)
• Self-consistency decoding	1.3E+23	68B	LaMDA	Wang et al. (2022b)
• Leveraging explanations in prompting	5.0E+23	280B	Gopher	Lampinen et al. (2022)
• Least-to-most prompting	3.1E+23	175B	GPT-3	Zhou et al. (2022)
• Zero-shot chain-of-thought reasoning	3.1E+23	175B	GPT-3	Kojima et al. (2022)
• Calibration via P(True)	2.6E+23	52B	Anthropic	Kadavath et al. (2022)
• Multilingual chain-of-thought reasoning	2.9E+23	62B	PaLM	Shi et al. (2022)
• Ask me anything prompting	1.4E+22	6B	EleutherAI	Arora et al. (2022)

5 Discussion

We have seen that a range of abilities—in the few-shot prompting setup or otherwise—have thus far only been observed when evaluated on a sufficiently large language model. Hence, their emergence cannot be predicted by simply extrapolating performance on smaller-scale models. Emergent few-shot prompted tasks are also unpredictable in the sense that these tasks are not explicitly included in pre-training, and we likely do not know the full scope of few-shot prompted tasks that language models can perform. This raises the question of whether further scaling could potentially endow even-larger language models with new emergent abilities. Tasks that language models cannot currently do are prime candidates for future emergence; for instance, there are dozens of tasks in BIG-Bench for which even the largest GPT-3 and PaLM models do not achieve above-random performance (see Appendix E.4).

The ability for scale to unpredictably enable new techniques is not just theoretical. Consider the Word in Context (WiC) benchmark (Pilehvar & Camacho-Collados, 2019) shown in Figure 2H, as a historical example. Here, scaling GPT-3 to around $3 \cdot 10^{23}$ training FLOPs (175B parameters) failed to unlock above-random one-shot prompting performance.[3] Regarding this negative result, Brown et al. (2020) cited the model architecture of GPT-3 or the use of an autoregressive language modeling objective (rather than using a denoising training objective) as potential reasons, and suggested training a model of comparable size with bidirectional architecture as a remedy. However, later work found that further scaling a decoder-only language model was actually enough to enable above-random performance on this task. As is shown in Figure 2H, scaling PaLM (Chowdhery et al., 2022) from $3 \cdot 10^{23}$ training FLOPs (62B parameters) to $3 \cdot 10^{24}$ training

[3]GPT-3 does achieve slightly above-random performance on the dev set with few-shot instead of one-shot prompting (∼55%), but this above-random performance did not appear to be a result of scale and did not hold on the test set server.

FLOPs (540B parameters) led to a significant jump in performance, without the significant architectural changes suggested by Brown et al. (2020).

5.1 Potential explanations of emergence

Although there are dozens of examples of emergent abilities, there are currently few compelling explanations for why such abilities emerge in the way they do. For certain tasks, there may be natural intuitions for why emergence requires a model larger than a particular threshold scale. For instance, if a multi-step reasoning task requires l steps of sequential computation, this might require a model with a depth of at least $O(l)$ layers. It is also reasonable to assume that more parameters and more training enable better memorization that could be helpful for tasks requiring world knowledge.[4] As an example, good performance on closed-book question-answering may require a model with enough parameters to capture the compressed knowledge base itself (though language model-based compressors can have higher compression ratios than conventional compressors (Bellard, 2021)).

It is also important to consider the evaluation metrics used to measure emergent abilities (BIG-Bench, 2022). For instance, using exact string match as the evaluation metric for long-sequence targets may disguise compounding incremental improvements as emergence. Similar logic may apply for multi-step or arithmetic reasoning problems, where models are only scored on whether they get the final answer to a multi-step problem correct, without any credit given to partially correct solutions. However, the jump in final answer accuracy does not explain why the quality of intermediate steps suddenly emerges to above random, and using evaluation metrics that do not give partial credit are at best an incomplete explanation, because emergent abilities are still observed on many classification tasks (e.g., the tasks in Figure 2D–H).

As an alternative evaluation, we measure cross-entropy loss, which is used in scaling laws for pre-training, for the six emergent BIG-Bench tasks, as detailed in Appendix A. This analysis follows the same experimental setup from BIG-Bench (2022) and affirms their conclusions for the six emergent tasks we consider. Namely, cross-entropy loss improves even for small model scales where the downstream metrics (exact match, BLEU, and accuracy) are close to random and do not improve, which shows that improvements in the log-likelihood of the target sequence can be masked by such downstream metrics. However, this analysis does not explain why downstream metrics are emergent or enable us to predict the scale at which emergence occurs. Overall, more work is needed to tease apart what enables scale to unlock emergent abilities.

5.2 Beyond scaling

Although we may observe an emergent ability to occur at a certain scale, it is possible that the ability could be later achieved at a smaller scale—in other words, model scale is not the singular factor for unlocking an emergent ability. As the science of training large language models progresses, certain abilities may be unlocked for smaller models with new architectures, higher-quality data, or improved training procedures. For example, there are 14 BIG-Bench tasks[5] for which LaMDA 137B and GPT-3 175B models perform at near-random, but PaLM 62B in fact achieves above-random performance, despite having fewer model parameters and training FLOPs. While there is not an empirical study ablating every difference between PaLM 62B and prior models (the computational cost would be too high), potential reasons for the better performance of PaLM could include high-quality training data (e.g., more multilingual and code data than LaMDA) and architectural differences (e.g., split digit-encodings; see Section 2 in Chowdhery et al. (2022)). Another potentially way of unlocking emergence is through a different pre-training objective—it was shown in Tay et al. (2022c) that a computationally-efficient continued pre-training stage on a mixture-of-denoisers objective (Tay et al., 2022a) enabled emergent performance on several BIG-Bench tasks.

Moreover, once an ability is discovered, further research may make the ability available for smaller scale models. Consider the nascent direction of enabling language models to follow natural language instructions describing a task (Wei et al., 2022a; Sanh et al., 2022; Ouyang et al., 2022, *inter alia*). Although Wei et al. (2022a) initially found that instruction-based finetuning only worked for 68B parameter or larger decoder-only

[4] Though note that encoding world knowledge in parameters is just one approach; there are others (e.g., Guu et al., 2020; Borgeaud et al., 2021).

[5] These tasks are enumerated in Appendix F.

models, Sanh et al. (2022) induced similar behavior in a 11B model with an encoder-decoder architecture, which typically has higher performance after finetuning than decoder-only architectures (Wang et al., 2022a). As another example, Ouyang et al. (2022) proposed a finetuning and reinforcement learning from human feedback approach for the InstructGPT models, which enabled a 1.3B model to outperform much larger models in human-rater evaluations on a broad set of use cases.

There has also been work on improving the general few-shot prompting abilities of language models (Gao et al., 2021; Schick & Schütze, 2021, *inter alia*). Theoretical and interpretability research (Wei et al., 2021a; Saunshi et al., 2021) on why a language modeling objective facilitates certain downstream behavior could in turn have implications on how to enable emergence beyond simply scaling. For instance, certain features of pre-training data (e.g., long-range coherence, having many rare classes) have also been shown to correlate with emergent few-shot prompting and could potentially enable it in smaller models (Xie et al., 2022; Chan et al., 2022), and few-shot learning can require certain model architectures in some scenarios (Chan et al., 2022). Computational linguistics work has further shown how threshold frequencies of training data can activate emergent syntactic rule-learning when model parameters and training FLOPs are held constant (Wei et al., 2021b), which has even been shown to have striking "aha" moments similar to those in the psycholinguistics literature (Abend et al., 2017; Zhang et al., 2021). As we continue to train language models, lowering the scale threshold for emergent abilities will become more important for making research on such abilities to available to the community more broadly (Bommasani et al., 2021; Ganguli et al., 2022; Liang et al., 2022).

Naturally, there are limitations to a program consisting only of increasing scale (training compute, model parameters, and dataset size). For instance, scaling may eventually be bottle-necked by hardware constraints, and some abilities may not have emerged at this point. Other abilities may never emerge—for instance, tasks that are far out of the distribution of even a very large training dataset might not ever achieve any significant performance. Finally, an ability could emerge and then plateau; in other words, there is no guarantee that scaling enables an ability to reach the desired level.

5.3 Another view of emergence

While scale (e.g., training FLOPs or model parameters) has been highly correlated with language model performance on many downstream metrics so far, scale need not be the only lens to view emergent abilities. For example, the emergence of task-specific abilities can be analyzed as a function of the language model's perplexity on a general text corpus such as WikiText103 (Merity et al., 2016). Figure 4 shows such a plot with WikiText103 perplexity of the language model on the x-axis and performance on the MMLU benchmark on the y-axis, side-by-side with plots of training FLOPs and model parameters on the x-axis.

Because WikiText103 perplexity and training FLOPs happen to be highly correlated for the models considered here (Gopher and Chinchilla), the plots of emergent abilities look similar for both. However, this correlation between WikiText103 perplexity and scale may not hold in the future as new techniques beyond vanilla dense Transformer models are developed (e.g., retrieval-augmented models may have strong WikiText103 perplexity with less training compute and fewer model parameters (Borgeaud et al., 2021)). Also note that using WikiText103 perplexity to compare across model families can be complicated due to factors such as differences in training data composition. Overall, emergent abilities should probably be viewed as a function of many correlated variables.

5.4 Emergent risks

Importantly, similar to how emergent abilities have been observed in the few-shot prompting setting without explicitly being included in pre-training, risks could also emerge (Bommasani et al., 2021; Steinhardt, 2021; Ganguli et al., 2022). For instance, societal risks of large language models such as truthfulness, bias, and toxicity are a growing area of research (Weidinger et al., 2021). Such risks are important considerations whether or not they can be precisely characterized as "emergent" based on the definition in §2, and, in some scenarios, do increase with model scale (see the Inverse Scaling Prize[6]). Since work on emergent abilities

[6] https://github.com/inverse-scaling/prize

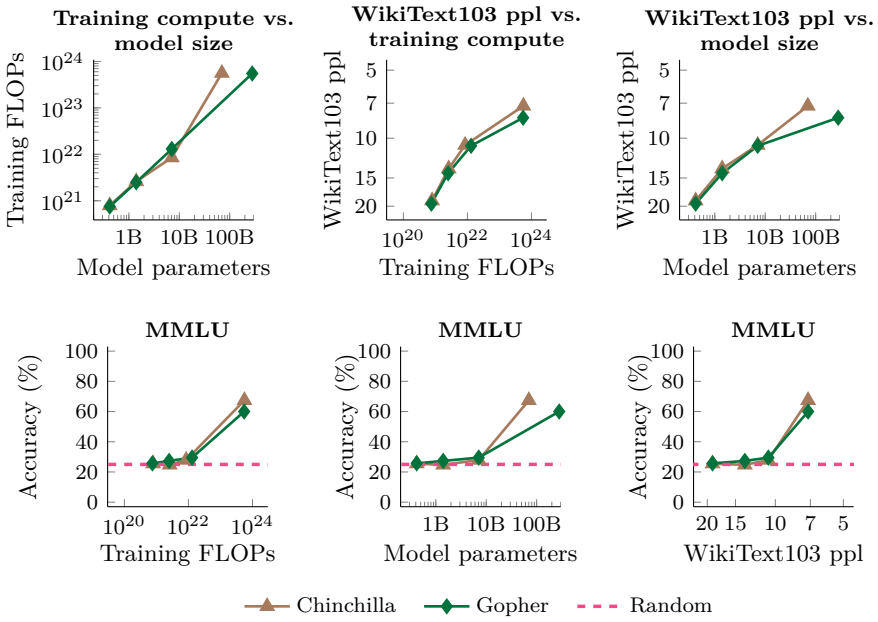

Figure 4: Top row: the relationships between training FLOPs, model parameters, and perplexity (ppl) on WikiText103 (Merity et al., 2016) for Chinchilla and Gopher. Bottom row: Overall performance on the massively multi-task language understanding benchmark (MMLU; Hendrycks et al., 2021a) as a function of training FLOPs, model parameters, and WikiText103 perplexity.

incentivizes scaling language models, it is important to be aware of risks that increase with model scale even if they are not emergent.

Here, we summarize several prior findings on the relationship between specific social risks and model scale. On WinoGender (Rudinger et al., 2017), which measures gender bias in occupations such as "nurse" or "electrician," scaling has improved performance so far (Du et al., 2021; Chowdhery et al., 2022), though BIG-Bench (2022) found in BBQ bias benchmark (Parrish et al., 2022) that bias can increase with scaling for ambiguous contexts. As for toxicity, Askell et al. (2021) found that while larger language models could produce more toxic responses from the RealToxicityPrompts dataset (Gehman et al., 2020), this behavior could be mitigated by giving models prompts with examples of being "helpful, harmless, and honest." For extracting training data from language models, larger models were found to be more likely to memorize training data (Carlini et al., 2021; 2022), though deduplication methods have been proposed and can simultaneously reduce memorization while improving performance (Kandpal et al., 2022; Lee et al., 2022a). The TruthfulQA benchmark (Lin et al., 2021) showed that GPT-3 models were more likely to mimic human falsehoods as they got larger, though Rae et al. (2021) later showed on a multiple-choice version that scaling Gopher to 280B enabled emergent performance substantially better than random.

Beyond the above, emergent risks also include phenomena that might only exist in future language models or that have not yet been characterized in current language models. Some such behaviors, as discussed in detail in Hendrycks et al. (2021b), could be backdoor vulnerabilities, inadvertent deception, or harmful content synthesis. Approaches involving data filtering, forecasting, governance, and automatically discovering harmful behaviors have been proposed for discovering and mitigating emergent risks (Bender et al., 2021; Weidinger et al., 2021; Steinhardt, 2021; Ganguli et al., 2022; Perez et al., 2022, *inter alia*). For a more detailed discussion of the risks of large language models, including emergent risks, see Bender et al. (2021); Steinhardt (2021); Bommasani et al. (2021); Ganguli et al. (2022).

5.5 Sociological changes

Finally, the emergent abilities discussed here focus on model behavior and are just one of several types of emergence in NLP (Manning et al., 2020; Teehan et al., 2022). Another notable type of qualitative change is sociological, in which increasing scale has shifted how the community views and uses language models. For instance, NLP has historically focused on task-specific models (Jurafsky & Martin, 2009). Recently, scaling has led to an explosion in research on and development of models that are "general purpose" in that they are single models that aim to perform a range of tasks not explicitly encoded in the training data (e.g., GPT-3, Chinchilla, and PaLM) (Manning, 2022).

One key set of results in the emergent sociological shift towards general-purpose models is when scaling enables a few-shot prompted general-purpose model to outperform prior state of the art held by finetuned task-specific models. As a few examples, GPT-3 175B achieved new state of the art on the TriviaQA and PiQA question-answering benchmarks (Brown et al., 2020); PaLM 540B achieved new state of the art on three arithmetic reasoning benchmarks (Chowdhery et al., 2022); and the multimodal Flamingo 80B model achieved new state of the art on six visual question answering benchmarks (Alayrac et al., 2022). In all of these cases, state-of-the-art performance was achieved by few-shot prompting a language model of unprecendented scale (scaling curves for these examples are shown in Appendix Figure 13). These abilities are not necessarily emergent since they have smooth, predictable scaling curves—however, they do underscore an emergent sociological shift towards general-purpose models in the NLP community.

The ability for general-purpose models to perform unseen tasks given only a few examples has also led to many new applications of language models outside the NLP research community. For instance, language models have been used via prompting to translate natural language instructions into actions executable by robots (Ahn et al., 2022; Huang et al., 2022), interact with users (Coenen et al., 2021; Wu et al., 2021; 2022a; Lee et al., 2022b), and facilitate multi-modal reasoning (Zeng et al., 2022; Alayrac et al., 2022). Large language models have also been deployed in the real-world both in products, such as GitHub CoPilot,[7] and directly as services themselves, such as OpenAI's GPT-3 API.[8]

5.6 Directions for future work

Future work on emergent abilities could involve train more-capable language models, as well as methods for better enabling language models to perform tasks. Some potential directions include but are not limited to the following.

Further model scaling. Further scaling up models has so far appeared to increase the capabilities of language models, and is a straightforward direction for future work. However, simply scaling up language models is computationally expensive and requires solving substantial hardware challenges, and so other approaches will likely play a key role in the future of the emergent abilities of large language models.

Improved model architectures and training. Improving model architecture and training procedures may facilitate high-quality models with emergent abilities while mitigating computational cost. One direction is using sparse mixture-of-experts architectures (Lepikhin et al., 2021; Fedus et al., 2021; Artetxe et al., 2021; Zoph et al., 2022), which scale up the number of parameters in a model while maintaining constant computational costs for an input. Other directions for better computational efficiency could involve variable amounts of compute for different inputs (Graves, 2016; Dehghani et al., 2018), using more localized learning strategies than backpropagation through all weights in a neural network (Jaderberg et al., 2017), and augmenting models with external memory (Guu et al., 2020; Borgeaud et al., 2021; Wu et al., 2022b, *inter alia*). These nascent directions have already shown promise in many settings but have not yet seen widespread adoption, which will likely require further work.

Data scaling. Training long enough on a large-enough dataset has been shown to be key for the ability of language models to acquire syntactic, semantic, and other world knowledge (Zhang et al., 2021; Wei et al., 2021b; Razeghi et al., 2022). Recently, Hoffmann et al. (2022) argued that prior work (Kaplan et al., 2020)

[7] https://copilot.github.com/
[8] https://beta.openai.com/docs/introduction

underestimated the amount of training data needed to train a compute-optimal model, underscoring the importance of training data. Collecting large datasets so that models can be trained for longer could allow a greater range of emergent abilities under a fixed model size constraint.

Better techniques for and understanding of prompting. Although few-shot prompting (Brown et al., 2020) is simple and effective, general improvements to prompting may further expand the abilities of language models. For instance, simple modifications such as calibrating output probabilities (Zhao et al., 2021; Holtzman et al., 2021) or using a noisy channel (Min et al., 2022a) have improved performance on a range of tasks. Augmenting few-shot exemplars with intermediate steps (Reynolds & McDonell, 2021; Nye et al., 2021; Wei et al., 2022b) has also enabled models to perform multi-step reasoning tasks not possible in the standard prompting formulation from Brown et al. (2020). Moreover, better exploration of what makes prompting successful (Wei et al., 2021a; Xie et al., 2022; Min et al., 2022b; Olsson et al., 2022) could lead to insights on how to elicit emergent abilities at a smaller model scale. Sufficient understanding of why models work generally lags the development and popularization of techniques such as few-shot prompting, and it is also likely that the best practices for prompting will change as more-powerful models are developed over time.

Frontier tasks. Although language models can perform a wide range of tasks, there are still many tasks that even the largest language models to date cannot perform with above-random accuracy. Dozens of such tasks from BIG-Bench are enumerated in Appendix E.4; these tasks often involve abstract reasoning (e.g., playing Chess, challenging math, etc). Future research could potentially investigate why these abilities have not yet emerged, and how to enable models to perform these tasks. Looking forward, another growing direction could be multilingual emergence; results on multilingual BIG-Bench tasks indicate that both model scale and training data play a role in emergence (e.g., Figure 2D shows that both using PaLM's training dataset and scaling to 62B parameters is required for question-answering in Persian). Other frontier tasks could include prompting in multiple modalities (Alayrac et al., 2022; Ramesh et al., 2022).

Understanding emergence. Beyond research on unlocking further emergence, an open question for future research is how and why emergent abilities occur in large language models. This paper conducted initial analyses regarding scaling of the cross-entropy loss on BIG-Bench (Appendix A.1), different metrics for generative tasks (Appendix A.2), and which types of tasks emergence occurs (Appendix A.3 and Appendix B). These analyses did not provide complete answers to why emergence occurs or how to predict it. Future research could potentially analyze emergence in new ways (e.g., analyze the relationship between emergent tasks and similar data in training; create a synthetic task that requires multiple compositional sub-tasks and evaluate how each of those sub-tasks improve with scale and unlock emergence when combined). Overall, understanding emergence is an important direction because it could potentially allow us predict what abilities future models may have, as well as provide new insights into how to train more-capable language models.

6 Conclusions

We have discussed emergent abilities of language models, for which meaningful performance has only been thus far observed at a certain computational scale. Emergent abilities can span a variety of language models, task types, and experimental scenarios. Such abilities are a recently discovered outcome of scaling up language models, and the questions of how they emerge and whether more scaling will enable further emergent abilities seem to be important future research directions for the field of NLP.

Broader Impact Statement

In this paper, we surveyed results in the existing literature, without proposing new methods or models. As discussed in (§5), emergent abilities are unpredictable in several ways, and include emergent risks (§5.4). We believe these phenomena warrant careful study and raise important questions for the field.

Acknowledgments

We thank Charles Sutton, Slav Petrov, Douglas Eck, Jason Freidenfelds, Jascha Sohl-Dickstein, Ethan Dyer, Dale Schuurmans, and Xavier Garcia for useful discussions and feedback on the manuscript.