Archaeologist Report on "Are Emergent Abilities of Large Language Models a Mirage?"

1 Introduction

This Archaeologist Report centers on the influential paper "Are Emergent Abilities of Large Language Models a Mirage?" (2023), which fundamentally challenged prevailing assumptions about large language models (LLMs). The development of LLMs has transformed natural language processing, enabling capabilities such as in-context learning, reasoning, and instruction following. However, as these models scale, a critical debate has emerged regarding whether certain abilities arise unpredictably at specific model sizes (emergent abilities) or whether these phenomena are artifacts of evaluation methods, biases, or fine-tuning effects.

This report traces the evolution of the emergent abilities debate through three critical phases: the initial discovery of emergent abilities and scaling laws, which proposed that qualitative improvements occur discontinuously as models scale; the subsequent critique presented in "Are Emergent Abilities of Large Language Models a Mirage?", suggesting that emergent abilities may not be fundamental properties of scaling but rather artifacts of evaluation metrics; and recent efforts to refine this debate by exploring alternative explanations, such as biases in pre-training and limitations of traditional scaling laws. By analyzing these developments, we highlight how the field has transitioned from accepting emergent abilities as inherent to scaling to a more skeptical and refined perspective that considers multiple contributing factors.

2 Key Contributions of Each Paper

2.1 Older Papers: Establishing Scaling Laws and Emergent Abilities

The foundational work on LLM scaling, exemplified by Wei et al. (2022) (Wei et al., 2022) and Brown et al. (2020) (Brown et al., 2020), introduced two key ideas that shaped subsequent research. Brown et al.'s paper "Language Models are Few-Shot Learners" demonstrated that large-scale models such as GPT-3 exhibit remarkable few-shot learning abilities without task-specific fine-tuning. Performance improvements followed a predictable power-law relationship with model size, reinforcing the notion that larger models generalize better through scale alone.

Wei et al.'s paper "Emergent Abilities of Large Language Models" expanded on this understanding by introducing the concept of emergent abilities, which appear at specific scale thresholds rather than improving gradually. They suggested that certain capabilities, such as multi-step reasoning and instruction following, arise abruptly as phase transitions, challenging the assumption that performance improvements are smooth and predictable. These papers collectively positioned scaling as the primary driver of model capability improvements, shaping subsequent discussions on whether emergent behaviors are real, measurable phenomena.

2.2 Current Paper: Challenging the Validity of Emergent Abilities

The paper "Are Emergent Abilities of Large Language Models a Mirage?" (2023) (Schaeffer et al., 2023) introduced a direct challenge to the assumption that emergent abilities represent genuine cognitive phase transitions. Instead, it argued that these phenomena arise due to flaws in evaluation methods rather than fundamental model properties. The authors identified three key issues in the assessment of emergent abilities.

First, they highlighted metric dependency, noting that the appearance of emergent abilities depends on the choice of scoring metric. Threshold-based metrics (e.g., exact match, multiple-choice accuracy) amplify small improvements, creating the illusion of sudden capability jumps. Second, they proposed an alternative scaling interpretation, showing that when models are evaluated using continuous performance measures (e.g., per-token accuracy), improvements appear smooth and predictable, contradicting claims of phase transitions. Third, they provided empirical refutation by systematically reanalyzing benchmarks such as BIG-Bench, demonstrating that changing the metric eliminates most claimed emergent abilities, suggesting they are artifacts of measurement discontinuities rather than true capability shifts.

This paper marks a critical turning point in the debate, reframing emergent abilities as artifacts of evaluation rather than fundamental model properties.

2.3 Newer Papers: Refining the Debate on Emergent Abilities

While "Are Emergent Abilities of Large Language Models a Mirage?" (2023) dismissed emergent abilities as metric-induced illusions, recent research in 2024 has introduced alternative explanations. These works neither fully reject nor entirely support the critique presented in "Are Emergent Abilities of Large Language Models a Mirage?" but instead explore why certain behaviors appear, whether due to biases in pre-training data or limitations in traditional scaling laws.

2.3.1 On Catastrophic Inheritance of Large Foundation Models (2024)

The paper "On Catastrophic Inheritance of Large Foundation Models" (Jiang et al., 2024) proposes that many behaviors attributed to emergent abilities might instead result from biases inherited from pre-training data rather than scale-induced learning. The authors argue that certain unexpected behaviors, such as ethical misalignment or spurious reasoning capabilities, may arise due to memorization or structural biases in large-scale datasets rather than a fundamental shift in model cognition. The study provides examples of LLMs displaying seemingly emergent abilities that align with pre-existing patterns in training data, rather than developing novel skills through scaling. Rather than treating emergent abilities as inherent properties of scaling, the paper suggests they should be analyzed within the broader context of data-driven biases.

2.3.2 Selecting Large Language Models to Fine-Tune via Rectified Scaling Law (2024)

The paper "Selecting Large Language Models to Fine-Tune via Rectified Scaling Law" (Zhang et al., 2024) extends the discussion by re-evaluating scaling laws, suggesting that naive assumptions about power-law scaling are insufficient to explain performance trends. The authors introduce the concept of a "Pre-Power Phase" - a stage where performance improvements do not follow expected scaling trends, challenging the assumption that emergent abilities are inherently unpredictable. They propose refined scaling laws through a mathematical adjustment to traditional scaling laws, showing that fine-tuning efficiency does not necessarily improve in a smooth, linear manner. The findings imply that some observed jumps in capability may arise from the interaction of scaling effects, optimization constraints, and dataset composition rather than from phase transitions.

These papers contribute to the ongoing refinement of the emergent abilities debate, moving beyond a binary "real vs. illusion" framing toward a more comprehensive understanding of the interplay between scale, data, and optimization.

3 Comparative Analysis and Research Evolution

By examining these studies collectively, we observe an evolution from optimism to skepticism to refinement in the study of emergent abilities. The research progression can be understood through three distinct phases.

During the initial phase of scaling and emergent abilities (2020–2022), Wei et al.'s "Emergent Abilities of Large Language Models" (2022) and Brown et al.'s "Language Models are Few-Shot Learners" (2020) argued that scaling alone could produce novel capabilities not present in smaller models. This reinforced the belief that LLMs improve discontinuously through size increases alone, without requiring architectural modifications.

The second phase brought skepticism and awareness of metric artifacts (2023). The paper "Are Emergent Abilities of Large Language Models a Mirage?" refuted the earlier assumptions, arguing that emergent abilities may not be real phenomena but instead artifacts of evaluation methods. This paper did not deny performance improvements but challenged their interpretation, refocusing research on how LLM progress is measured.

The third phase introduced refinement and alternative explanations (2024). The 2024 papers expanded on "Are Emergent Abilities of Large Language Models a Mirage?" by investigating why emergent behaviors appear—not just whether they are real. "On Catastrophic Inheritance of Large Foundation Models" suggests that certain behaviors result from pre-training biases rather than scale-driven emergence. "Selecting Large Language Models to Fine-Tune via Rectified Scaling Law" challenges naive scaling assumptions, reinforcing the argument that nonlinear factors influence model behavior beyond mere size increases.

4 Conclusion

The trajectory of LLM research has evolved from an initial enthusiasm for emergent abilities to skepticism about their validity and ultimately to a refined perspective that considers multiple explanatory factors. While "Are Emergent Abilities of Large Language Models a Mirage?" (2023) demonstrated that many emergent behaviors are artifacts of evaluation methods, the 2024 papers suggest that pre-training biases and scaling inefficiencies also contribute.

This evolving understanding highlights the need for rigorous evaluation, refined scaling models, and careful analysis of dataset composition in future LLM research. Together, these studies indicate that emergent abilities are not purely a function of scale but are influenced by dataset biases, evaluation methodologies, and fine-tuning dynamics. Moving forward, research must adopt more rigorous evaluation methods to distinguish genuine qualitative improvements from artifacts of data selection and metric design. This nuanced understanding is critical for developing robust, interpretable, and ethically responsible AI systems.

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jiang, Y., Jiang, Y., Jiang, Y., Jiang, Y., Jiang, Y., and Jiang, Y. (2024). On catastrophic inheritance of large foundation models. *arXiv preprint arXiv:2401.12789*.

Schaeffer, R., Miranda, B., and Koyejo, S. (2023). Are emergent abilities of large language models a mirage? *arXiv preprint arXiv: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. *arXiv preprint arXiv:2206.07682*.

Zhang, Z., Zhao, T., Jiang, Y., Zhao, T., and Zhao, T. (2024). Selecting large language models to fine-tune via rectified scaling law. *arXiv preprint arXiv:2402.18540*.