Are Emergent Abilities in Large Language Models just In-Context Learning?

Sind emergente Fähigkeiten in großen Sprachmodellen nur In-Context Lernen?

Sheng Lu[1*], Irina Bigoulaeva[1*], Rachneet Sachdeva[1], Harish Tayyar Madabushi[2], and Iryna Gurevych[1]
[*]Equal Contribution, Accepted to ACL 2024. A longer version of this paper is available at https://htayyarmadabushi.github.io/Emergent_Abilities_and_in-Context_Learning/. / *Paritätischer Beitrag, angenommen für ACL 2024. Eine längere Version dieses Papiers ist verfügbar unter https://htayyarmadabushi.github.io/Emergent_Abilities_and_in-Context_Learning/.

[1] Ubiquitous Knowledge Processing Lab, Technical University of Darmstadt
[2] Department of Computer Science, The University of Bath
www.ukp.tu-darmstadt.de, htm43@bath.ac.uk

Abstract

Large language models, comprising billions of parameters and pre-trained on extensive webscale corpora, have been claimed to acquire certain capabilities without having been specifically trained on them. These capabilities, referred to as “emergent abilities,” have been a driving force in discussions regarding the potentials and risks of language models. A key challenge in evaluating emergent abilitiesis that they are confounded by model competencies that arise through alternative prompting techniques, including in-context learning, which is the ability of models to complete a task based on a few examples. We present a novel theory that explains emergent abilities, taking into account their potential confounding factors, and rigorously substantiate this theory through over 1000 experiments. Our findings suggest that purported emergent abilities are not truly emergent, but result from a combination of incontext learning, model memory, and linguistic knowledge. Our work is a foundational step in explaining language model performance, providing a template for their efficient use and clarifying the paradox of their ability to excel in some instances while faltering in others. Thus, we demonstrate that their capabilities should not be overestimated.[1]

Abstract

Es wird behauptet, dass große Sprachmodelle, die Milliarden von Parametern enthalten und auf umfangreichen Webkorpora trainiert wurden, bestimmte Fähigkeiten erwerben, ohne dass sie speziell darauf trainiert wurden [Formatierungen jeweils hinzugefügt]. Diese Fähigkeiten, die als "emergente Fähigkeiten" bezeichnet werden, sind eine treibende Kraft in den Diskussionen über das Potenzial und die Risiken von Sprachmodellen gewesen. Eine zentrale Herausforderung bei der Bewertung emergenter Fähigkeiten besteht darin, dass sie mit Modellkompetenzen verwechselt werden, die durch alternative Prompingtechniken entstehen, einschließlich des in-context learning, d. h. der Fähigkeit von Modellen, eine Aufgabe auf der Grundlage einiger weniger Beispiele zu lösen. Wir stellen eine neue Theorie vor, die emergente Fähigkeiten unter Berücksichtigung ihrer potenziellen Störfaktoren erklärt, und untermauern diese Theorie mit über 1000 Experimenten. Unsere Ergebnisse deuten darauf hin, dass vermeintlich emergente Fähigkeiten nicht wirklich emergent sind, sondern aus einer Kombination von kontextunabhängigem Lernen, Modellgedächtnis und sprachlichem Wissen resultieren. Unsere Arbeit ist ein grundlegender Schritt zur Erklärung der Leistung von Sprachmodellen, denn sie liefert eine Vorlage für ihre effiziente Nutzung und klärt das Paradoxon, dass sie in manchen Fällen überragend sind, während sie in anderen Fällen versagen. So zeigen wir, dass ihre Fähigkeiten nicht überschätzt werden sollten.[1]

[1] Our code and data are available at https://github.com/UKPLab/on-emergence and https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/3931. Unser Code und die Daten sind erhältlich unter https://github.com/UKPLab/on-emergence and https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/3931.

1 Introduction, Motivation and Context

One of the most captivating aspects of pre-trained language models (PLMs) is their capacity to acquire a wide range of knowledge across different domains, while being trained primarily through masked language modelling, a task requiring models to predict masked tokens in their input (Tenney et al., 2019; Petroni et al., 2019). The diverse abilities of PLMs can be categorised into two broad types: formal linguistic abilities and functional linguistic abilities. Formal linguistic abilities refer to the understanding of language rules and patterns, which PLMs, for example, BERT (Devlin et al., 2019) are known to excel at (Tenney et al., 2019; Petroni et al., 2019). The latter category includes a range of abilities akin to human cognition that are necessary for real-world language use and comprehension, such as commonsense knowledge and social awareness. While PLMs excel at formal linguistic abilities, they have faced challenges in developing functional linguistic abilities (Mahowald et al., 2023).

The introduction of Large Language Models (LLMs), which are typically generative PLMs scaled up to billions of parameters and trained on vast, web-scale data corpora, is changing this landscape (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023a,b). Recent works indicate that LLMs exhibit emergent abilities, as measured by their above random performance without explicit training on tasks, including those tasks that explicitly require some form of reasoning. An emergent ability was first defined as an ability to solve a task which is absent in smaller models, but present in LLMs. This definition, introduced approximately concurrently by two works (Wei et al., 2022b; Srivastava et al., 2023), is based on the more general definition of emergence in physics: “Emergence is when quantitative changes in a system result in qualitative changes in behaviour” (Anderson, 1972). Emergent abilities are implied due to LLMs’ capacity to perform above the random baseline on the corresponding tasks without explicit training on those same tasks. For example, the emergent ability to

1 Einleitung, Motivation und Kontext

Einer der faszinierendsten Aspekte von vortrainierten Sprachmodellen (pre-trained language models ; PLMs) ist ihre Fähigkeit, ein breites Spektrum an Wissen in verschiedenen Bereichen zu erwerben, während sie hauptsächlich durch maskierte Sprachmodellierung trainiert werden, eine Aufgabe, bei der die Modelle maskierte Token in ihrer Eingabe vorhersagen müssen (Tenney et al., 2019; Petroni et al., 2019). Die verschiedenen Fähigkeiten von PLMs können in zwei große Kategorien eingeteilt werden: formale linguistische Fähigkeiten und funktionale linguistische Fähigkeiten. Formale sprachliche Fähigkeiten beziehen sich auf das Verständnis von Sprachregeln und -mustern, die PLMs, wie z. B. BERT (Devlin et al., 2019), bekanntermaßen hervorragend beherrschen (Tenney et al., 2019; Petroni et al., 2019). Die letztgenannte Kategorie umfasst eine Reihe von Fähigkeiten, die der menschlichen Kognition ähneln und für den realen Sprachgebrauch und das Sprachverständnis notwendig sind, wie z. B. das Wissen um den gesunden Menschenverstand und das soziale Bewusstsein. Während PLMs sich durch formale sprachliche Fähigkeiten auszeichnen, stehen sie vor Herausforderungen bei der Entwicklung funktionaler sprachlicher Fähigkeiten (Mahowald et al., 2023).

Die Einführung von Large Language Models (LLMs), bei denen es sich in der Regel um generative PLMs handelt, die auf Milliarden von Parametern skaliert und auf riesigen, webbasierten Datenkorpora trainiert werden, verändert diese Landschaft (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023a,b). Neuere Arbeiten weisen darauf hin, dass LLMs emergente Fähigkeiten aufweisen, gemessen an ihrer über dem Zufall liegenden Leistung ohne explizites Training bei Aufgaben, einschließlich solcher Aufgaben, die explizit eine Form des Denkens erfordern. Eine emergente Fähigkeit wurde zunächst als eine Fähigkeit zur Lösung einer Aufgabe definiert, die bei kleineren Modellen nicht vorhanden ist, bei LLMs aber vorhanden ist. Diese Definition, die etwa zeitgleich von zwei Arbeiten (Wei et al., 2022b; Srivastava et al., 2023) eingeführt wurde, basiert auf der allgemeineren Definition von Emergenz in der Physik: "Emergenz ist, wenn quantitative Veränderungen in einem System zu qualitativen Veränderungen im Verhalten führen" (Anderson, 1972). Emergente Fähigkeiten werden aus der Fähigkeit von LLMs abgeleitet, bei den entsprechenden Aufgaben ohne explizites Training über der zufälligen Basislinie zu liegen. Zum Beispiel wird die emergente Fähigkeit,

understand social situations in LLMs is inferred from LLMs' performing well above the random baseline on the Social IQA (Sap et al., 2019) task, which serves to evaluate models' emotional and social intelligence and includes questions such as "Carson was excited to wake up to attend school. Why did he do this? Options: Take the big test, Go to bed early, Just say hello to friend (correct)".

1.1 Significance for Applications and Safety

While prior work on emergent abilities does not explicitly make the distinction between formal and functional linguistic abilities, the identification of numerous functional linguistic capabilities holds profound implications for both the potential and safety of LLMs. The assumption that LLMs have access to emergent functional linguistic abilities significantly affects the way in which users interact with and use these systems. Overreliance on these perceived abilities can lead users to provide insufficiently detailed instructions, potentially resulting in hallucinations and errors. If there are indeed multiple functional linguistic abilities that emerge with scale, it suggests that further scaling has the potential to unlock a wide array of additional abilities which we cannot predict, especially since they tend not to present themselves in smaller-scale models (Wei et al., 2022b). This inherent unpredictability associated with emergent abilities holds substantial implications for the discussion surrounding safety and security when utilising LLMs. Indeed, it has been argued that these could include potentially hazardous abilities, including reasoning and planning (Hoffmann, 2023), thereby posing an existential threat to humanity (Bengio et al., 2023). In this work, we refer to such potentially harmful capabilities, as "latent hazardous abilities."

It's important to emphasise that the development of linguistic proficiencies (i.e. formal linguistic abilities) does not carry the potentials of this nature. The same can be said for the capacity to efficiently handle information retrieval tasks. The real focus

soziale Situationen bei LLMs zu verstehen, daraus abgeleitet, dass LLMs bei der Aufgabe Social IQA (Sap et al., 2019), die dazu dient, die emotionale und soziale Intelligenz der Modelle zu bewerten und Fragen wie "Carson war aufgeregt, als er aufwachte, um in die Schule zu gehen", deutlich über der zufälligen Basislinie liegen. Warum hat er das getan? Optionen: Den großen Test machen: früh ins Bett gehen; einem Freund einfach Hallo sagen (richtig)".

1.1 Bedeutung für Anwendungen und Sicherheit

Während frühere Arbeiten über emergente Fähigkeiten nicht explizit zwischen formalen und funktionalen linguistischen Fähigkeiten unterscheiden, hat die Identifizierung zahlreicher funktionaler linguistischer Fähigkeiten tiefgreifende Auswirkungen sowohl auf das Potenzial als auch auf die Sicherheit von LLMs. Die Annahme, dass LLMs Zugang zu emergenten funktionalen linguistischen Fähigkeiten haben, beeinflusst die Art und Weise, wie Benutzer mit diesen Systemen interagieren und sie nutzen, erheblich. Ein übermäßiges Vertrauen in diese wahrgenommenen Fähigkeiten kann dazu führen, dass die Benutzer unzureichend detaillierte Anweisungen geben, was zu Halluzinationen und Fehlern führen kann. Wenn es tatsächlich mehrere funktionale sprachliche Fähigkeiten gibt, die sich mit der Skalierung herausbilden, deutet dies darauf hin, dass eine weitere Skalierung das Potenzial hat, eine breite Palette zusätzlicher Fähigkeiten freizusetzen, die wir nicht vorhersagen können, zumal sie sich in kleineren Modellen in der Regel nicht zeigen (Wei et al., 2022b). Diese inhärente Unvorhersehbarkeit, die mit emergenten Fähigkeiten verbunden ist, hat erhebliche Auswirkungen auf die Diskussion über Sicherheit und Schutz bei der Nutzung von LLMs. In der Tat wurde argumentiert, dass diese potenziell gefährliche Fähigkeiten, einschließlich des Denkens und Planens, beinhalten könnten (Hoffmann, 2023) und somit eine existenzielle Bedrohung für die Menschheit darstellen (Bengio et al., 2023). In dieser Arbeit bezeichnen wir solche potenziell gefährlichen Fähigkeiten als "latent gefährliche Fähigkeiten".

Es ist wichtig zu betonen, dass die Entwicklung sprachlicher Fertigkeiten (d. h. formaler sprachlicher Fähigkeiten) nicht die gleichen Möglichkeiten bietet wie diese. Dasselbe gilt für die Fähigkeit, Aufgaben der Informationsbeschaffung effizient zu bewältigen. Der eigentliche Schwerpunkt liegt auf den potenziel-

lies on potential capabilities relating to functional linguistic abilities. However, it must be emphasised that this does not include other dangers posed through the misuse of these models, such as the use of LLMs to generate fake news. Similarly, we do not contend that future AI systems could never pose an existential threat. Instead, we clarify that, contrary to prevailing narratives, the evidence from LLM abilities does not support this concern.

len Fähigkeiten, die sich auf die funktionalen sprachlichen Fähigkeiten beziehen. Es muss jedoch betont werden, dass dies nicht andere Gefahren einschließt, die durch den Missbrauch dieser Modelle entstehen, wie z. B. die Verwendung von LLMs zur Generierung von Fake News. Ebenso behaupten wir nicht, dass zukünftige KI-Systeme niemals eine existenzielle Bedrohung darstellen könnten. Stattdessen stellen wir klar, dass im Gegensatz zu den vorherrschenden Erzählungen die Beweise für die Fähigkeiten von LLMs diese Sorge nicht stützen.

1.2 Abilities vs. Techniques

The scaling up of LLMs facilitates the acquisition of diverse competencies, which can be grouped into two categories: The first encompasses abilities, already described. The second encompasses various techniques, which LLMs can benefit from. These techniques show less of an effect in smaller models, but become progressively more effective with scale. Among these techniques are in-context learning and instruction-tuning. In-context learning (ICL) is the technique wherein LLMs are provided with a limited number of examples within the input prompt itself (Brown et al., 2020). From these examples, the model infers how to perform a specific task, responding appropriately to the question posed by the prompt (Brown et al., 2020; Liu et al., 2023). Investigations into the theoretical underpinnings of ICL and its specific manifestation in LLMs indicate that it might bear resemblance to the process of fine-tuning models on the specific tasks for which they are provided examples (Akyürek et al., 2023; Dai et al., 2023; von Oswald et al., 2023; Wei et al., 2023). Another technique exclusive to LLMs is instructional fine-tuning, alternatively known as instruction-tuning. This technique involves fine-tuning LLMs on datasets of prompts and their corresponding desired outputs, which enables the models to follow explicit instructions in prompts (Chung et al., 2022; Wei et al., 2022a; Taori et al., 2023). Following previous work (Wie et al., 2022b), we refer to these techniques, illustrated in Figure 3, as prompting techniques.

1.3 Fähigkeiten vs. Techniken

Die Ausweitung des LLM erleichtert den Erwerb verschiedener Kompetenzen, die sich in zwei Kategorien zusammenfassen lassen: Die erste umfasst die bereits beschriebenen Fähigkeiten. Die zweite umfasst verschiedene Techniken, von denen LLMs profitieren können. Diese Techniken zeigen bei kleineren Modellen weniger Wirkung, werden aber mit zunehmender Größe immer effektiver. Zu diesen Techniken gehören das In-Context-Lernen und die Optimierung von Anweisungen. In-Context-Lernen (ICL) ist die Technik, bei der LLMs mit einer begrenzten Anzahl von Beispielen innerhalb der Eingabeaufforderung selbst versorgt werden (Brown et al., 2020). Aus diesen Beispielen schlussfolgert das Modell, wie eine bestimmte Aufgabe auszuführen ist, indem es angemessen auf die in der Eingabeaufforderung gestellte Frage antwortet (Brown et al., 2020; Liu et al., 2023). Untersuchungen der theoretischen Grundlagen der ICL und ihrer spezifischen Ausprägung in LLMs deuten darauf hin, dass sie dem Prozess der Feinabstimmung von Modellen auf die spezifischen Aufgaben, für die sie Beispiele erhalten, ähneln könnte (Akyürek et al., 2023; Dai et al., 2023; von Oswald et al., 2023; Wei et al., 2023). Eine weitere Technik, die ausschließlich für LLMs zur Verfügung steht, ist die Optimierung von Anweisungen, auch als instructional fine-tuning [Feinabstimmung der Instruktion] bekannt). Diese Technik beinhaltet die Optimierung von LLMs anhand von Datensätzen mit Aufforderungen und den entsprechenden gewünschten Ergebnissen, wodurch die Modelle in die Lage versetzt werden, expliziten Anweisungen in Aufforderungen zu folgen (Chung et al., 2022; Wei et al., 2022a; Taori et al., 2023). In Anlehnung an frühere Arbeiten (Wie et al., 2022b) bezeichnen wir diese Techniken, die in Abbildung 3 dargestellt sind, als Prompting-Techniken.

Significant to our investigation is the observation that prompting techniques and emergent abilities manifest within LLMs at a comparable scale. Furthermore, ICL and instruction-tuning can be observed in smaller-scale models, albeit to a lesser degree, and are thus predictable. This predictability means they are not 'emergent', nor do they pose a threat, contrasting with the unpredictability and potential risks associated with emergent abilities in larger models. Considering this context, it becomes imperative to ascertain the extent of these emergent abilities in the absence of prompting techniques.

Von Bedeutung für unsere Untersuchung ist die Beobachtung, dass sich Prompting-Techniken und emergente Fähigkeiten in LLMs in einem vergleichbaren Ausmaß manifestieren. Darüber hinaus können ICL und die Optimierung von Anweisungen auch in kleineren Modellen beobachtet werden, wenn auch in geringerem Ausmaß, und sind daher vorhersehbar. Diese Vorhersehbarkeit bedeutet, dass sie weder "emergent" sind noch eine Bedrohung darstellen, im Gegensatz zu der Unvorhersehbarkeit und den potenziellen Risiken, die mit emergenten Fähigkeiten in größeren Modellen verbunden sind. Vor diesem Hintergrund ist es zwingend erforderlich, das Ausmaß dieser emergenten Fähigkeiten ohne Prompting-Techniken zu ermitteln.

1.3 Fine-tuning, In-Context Learning, and other Prompting Techniques

Artificial neural models have, for some time, exhibited tremendous success on specific tasks when trained on those tasks (Devlin et al., 2019; Liu et al., 2019). PLMs in particular have demonstrated this even when trained on just a few examples (Hofer et al., 2018; Radford et al., 2019; Brown et al., 2020; Gao et al., 2021). Such performance is not considered "emergent", precisely because models are trained on that very task. Indeed, the fact that LLMs are not trained on the tasks used in evaluating their emergent abilities is central to identifying abilities which are truly emergent. The assertion that achieving satisfactory performance on a given task signifies the emergence of associated 'abilities' hinges on the condition that models are not explicitly trained for that specific task.

The recent insights indicating parallels between ICL and explicit training suggest that the success on a task through ICL, much like models trained explicitly for task-solving, does not imply a model inherently possessing that ability (Dai et al., 2023). For example, it has been shown that ICL implements gradient descent implicitly and constructs a function at inference time on regression problems (Akyürek et al., 2023; Li et al., 2023; Zhang et al., 2023a), which may

1.3 Feinabstimmung, In-Context Lernen und andere Prompting-Techniken

Künstliche neuronale Modelle zeigen seit einiger Zeit enorme Erfolge bei bestimmten Aufgaben, wenn sie auf diese Aufgaben trainiert werden (Devlin et al., 2019; Liu et al., 2019). Insbesondere PLMs haben dies gezeigt, selbst wenn sie nur auf wenige Beispiele trainiert wurden (Hofer et al., 2018; Radford et al., 2019; Brown et al., 2020; Gao et al., 2021). Eine solche Leistung wird nicht als "emergent" angesehen, eben weil die Modelle für genau diese Aufgabe trainiert werden. Die Tatsache, dass LLMs nicht auf die Aufgaben trainiert werden, die zur Bewertung ihrer emergenten Fähigkeiten verwendet werden, ist in der Tat von zentraler Bedeutung für die Identifizierung von Fähigkeiten, die wirklich emergent sind. Die Behauptung, dass das Erreichen einer zufriedenstellenden Leistung bei einer bestimmten Aufgabe das Auftauchen der damit verbundenen "Fähigkeiten" bedeutet, hängt von der Bedingung ab, dass die Modelle nicht explizit für diese spezifische Aufgabe trainiert werden.

Die jüngsten Erkenntnisse, die auf Parallelen zwischen ICL und explizitem Training hinweisen, legen nahe, dass der Erfolg bei einer Aufgabe durch ICL, ähnlich wie bei explizit für die Aufgabenlösung trainierten Modellen, nicht impliziert, dass ein Modell von Natur aus über diese Fähigkeit verfügt (Dai et al., 2023). So wurde beispielsweise gezeigt, dass ICL den Gradientenabstieg implizit implementiert und zur Inferenzzeit bei Regressionsproblemen eine Funktion konstruiert (Akyürek et al., 2023; Li et al., 2023; Zhang et al., 2023a), was mit dem gradientenbasier-

be related to gradient-based meta-learning (von Oswald et al., 2023). Importantly, however, the specific mechanisms governing ICL do not impact our argument: The fact of its functionality suffices to underscore the necessity of assessing emergent abilities in the absence of ICL. Additionally, instruction-tuning datasets typically include several variations of an instruction followed by the task input or context (see Figure 3). As such, we contend that the process of instruction fine-tuning potentially enables models to map prompts to in-context examples (detailed in Section 4), thereby utilising ICL to respond to prompts. This would imply that the success of a model to solve a task in this scenario also does not indicate the emergence of the corresponding ability.

The safety issues associated with LLMs stem from their ability to perform well above the random baseline on tasks that cannot be solved through memorisation and are indicative of certain 'abilities', without explicit training on those tasks. Therefore, recognising that prompts act as a form of 'training mechanism' rather than simply a way of interfacing with a model with inherent functional linguistic abilities offers the potential to alter how we use these models and deepen our understanding of their capabilities and limitations. As such, it is crucial to conduct an independent evaluation of LLMs' abilities, detached from ICL.

ten Meta-Lernen verwandt sein könnte (von Oswald et al., 2023). Wichtig ist jedoch, dass die spezifischen Mechanismen, die ICL steuern, keinen Einfluss auf unsere Argumentation haben: Die Tatsache, dass es funktioniert, reicht aus, um die Notwendigkeit der Bewertung von emergenten Fähigkeiten in Abwesenheit von ICL zu unterstreichen. Darüber hinaus enthalten Datensätze zur Optimierung von Anweisungen typischerweise mehrere Variationen einer Anweisung, gefolgt von der Aufgabenvorgabe oder dem Kontext (siehe Abbildung 3). Wir gehen davon aus, dass der Prozess der Optimierung von Anweisungen die Modelle in die Lage versetzt, Aufforderungen auf in-context Beispiele abzubilden (siehe Abschnitt 4) und dabei ICL zu nutzen, um auf Aufforderungen zu reagieren. Dies würde bedeuten, dass der Erfolg eines Modells bei der Lösung einer Aufgabe in diesem Szenario auch nicht auf die Entstehung der entsprechenden Fähigkeit hinweist.

Die Sicherheitsprobleme, die mit LLMs verbunden sind, rühren von ihrer Fähigkeit her, bei Aufgaben, die nicht durch Auswendiglernen gelöst werden können und die auf bestimmte "Fähigkeiten" hindeuten, ohne explizites Training für diese Aufgaben weit über der zufälligen Basislinie zu liegen. Die Erkenntnis, dass Prompts als eine Art "Trainingsmechanismus" fungieren und nicht einfach nur als Schnittstelle zu einem Modell mit inhärenten funktionalen linguistischen Fähigkeiten, bietet daher das Potenzial, die Art und Weise, wie wir diese Modelle nutzen, zu verändern und unser Verständnis für ihre Fähigkeiten und Grenzen zu vertiefen. Daher ist es von entscheidender Bedeutung, eine unabhängige Bewertung der Fähigkeiten von LLMs durchzuführen, losgelöst von ICL.

1.4 Research Questions and Contributions

Our research seeks to answer two pivotal questions: Firstly, in light of ICL's influence on perceived emergent abilities in LLMs, which abilities are truly emergent in the absence of ICL, including instructional tuning? Secondly, given LLMs' capability for ICL and the typical inclusion of instructionexemplar mappings in instruction-tuning datasets, can we find evidence of the emergence of functional linguistic abilities in instruction-tuned models? Or can ICL better explain their capabilities and shortcomings?

1.4 Forschungsfragen und Beiträge

Unsere Forschung zielt darauf ab, zwei zentrale Fragen zu beantworten: Erstens, in Anbetracht des Einflusses von ICL auf wahrgenommene emergente Fähigkeiten bei LLMs, welche Fähigkeiten sind wirklich emergent in Abwesenheit von ICL, einschließlich der Optimierung von Anweisungen? Zweitens: Können wir angesichts der Fähigkeit von LLMs zur ICL und der typischen Einbeziehung von Instruktions-Exemplar-Zuordnungen in Datensätzen zur Optimierung von Anweisungen Belege für die emergente Fähigkeit von funktionalen linguistischen Fähigkeiten in Modellen mit Optimierung von Anweisungen finden? Oder kann ICL ihre Fähigkeiten und Unzulänglichkeiten besser erklären?

Our primary contribution lies in demonstrating the absence of emergent functional linguistic abilities in LLMs when ICL is not a factor, thus demystifying the true capabilities of LLMs and affirming their safety, while additionally dispelling concerns over potential latent hazardous abilities. Our secondary contributions include empirically testing the hypothesis that instruction-tuned models' capabilities stem from efficient ICL, thus offering an explanation for LLMs' abilities as stemming from a combination of formal linguistic skills, vast information retention and recall, and notably, ICL. By identifying user-directable ICL, rather than intrinsic functional linguistic capabilities, as the mechanism behind LLM performance, we lay out a framework for more efficient use of these models, shedding light on their capabilities and limitations.

Unser primärer Beitrag besteht darin, das Fehlen von emergenten funktionalen sprachlichen Fähigkeiten bei LLMs nachzuweisen, wenn ICL kein Faktor ist, und somit die wahren Fähigkeiten von LLMs zu entmystifizieren und ihre Sicherheit zu bekräftigen, während zusätzlich Bedenken über mögliche latente gefährliche Fähigkeiten zerstreut werden. Zu unseren sekundären Beiträgen gehört die empirische Prüfung der Hypothese, dass die Fähigkeiten von anweisungs-optimierten Modellen, aus effizientem ICL resultieren, was eine Erklärung für die Fähigkeiten von LLMs bietet, die aus einer Kombination von formalen linguistischen Fähigkeiten, einer enormen Informationsspeicherung und -erinnerung und insbesondere aus ICL resultieren. Indem wir die vom Benutzer steuerbare ICL und nicht die intrinsischen funktionalen linguistischen Fähigkeiten als den Mechanismus hinter der LLM-Leistung identifizieren, legen wir einen Rahmen für eine effizientere Nutzung dieser Modelle fest und beleuchten ihre Fähigkeiten und Grenzen.

2 Experimental Setup

In this section, we present an overview of our experimental methods investigating emergent abilities in the absence of ICL.We experiment with 20 models across 22 tasks using two different settings. We use four different evaluation metrics and additionally run multiple tests for bias, including a manual analysis of our results. We present an overview of this setup below, while details on the hyperparameters and training regime are presented in Appendix C.

2 Versuchsaufbau

In diesem Abschnitt geben wir einen Überblick über unsere experimentellen Methoden zur Untersuchung von Emergenten Fähigkeiten in Abwesenheit von ICL. Wir experimentieren mit 20 Modellen in 22 Aufgaben mit zwei verschiedenen Einstellungen. Wir verwenden vier verschiedene Bewertungsmetriken und führen zusätzlich mehrere Tests auf Verzerrungen durch, einschließlich einer manuellen Analyse unserer Ergebnisse. Im Folgenden geben wir einen Überblick über diesen Aufbau, während Details zu den Hyperparametern und dem Trainingsregime in Anhang C zu finden sind.

2.1 Models

We experiment with four model families: GPT, T5 (Raffel et al., 2020), Falcon[2] and LLaMA (Touvron et al., 2023a). We choose these model families, since GPT and LLaMA have previously been found to have emergent abilities, and Falcon is at the top of LLM leaderboards at the time of writing. Finally, we select T5 as it is an encoder-decoder model, and its instruction-tuned

2.1 Modelle

Wir experimentieren mit vier Modellfamilien: GPT, T5 (Raffel et al., 2020), Falcon und LLaMA (Touvron et al., 2023a). Wir haben uns für diese Modellfamilien entschieden, da bei GPT und LLaMA bereits emergente Fähigkeiten festgestellt wurden und Falcon zum Zeitpunkt der Erstellung dieses Berichts an der Spitze der LLM-Rangliste steht. Schließlich wählen wir T5 aus, da es sich um ein Encoder-Decoder-Modell handelt und seine anweisungsoptimierte Version (Flan) anhand

[2] See https://falconllm.tii.ae/index.html.

version (Flan) is trained using an extensive instruction-tuning dataset. Table 1 enumerates the models that we use in our experiments. The emergence of abilities in relation to scale requires the evaluation of each model family across a range of sizes (parameter counts), and so we select models at different scales from each of these families. Important to our inquiry is the hypothesis that instructional tuning might indirectly leverage ICL. In light of this possibility, we experimentwith both.

eines umfangreichen Datensatzes zur Optimierung von Anweisungen trainiert wurde. Tabelle 1 listet die Modelle auf, die wir in unseren Experimenten verwenden. Die Emergenz von Fähigkeiten in Abhängigkeit von der Skala erfordert die Bewertung jeder Modellfamilie über eine Reihe von Größen (Parameterzahlen), und daher wählen wir Modelle in verschiedenen Skalen aus jeder dieser Familien aus. Wichtig für unsere Untersuchung ist die Hypothese, dass die Optimierung von Anweisungen indirekt eine Hebelwirkung auf ICL haben könnte. In Anbetracht dieser Möglichkeit experimentieren wir mit beidem.

Model	Instruction-Tuned Version	Size
GPT-2	GPT-2-IT	117M
GPT-2-XL	GPT-2-XL-IT	1.6B
GPT-J	GPT-JT	6.7B
Davinci	text-davinci-001 text-davinci-003	175B
T5-small	Flan-T5-small	60M
T5-large	Flan-T5-large	770M
Falcon-7B	Falcon-7B-Instruct	7B
Falcon-40B	Falcon-40B-Instruct	40B
LLaMA-7B	–	7B
LLaMA-13B	–	13B
LLaMA-30B	–	30B

Table 1: Details of the models used in the experiments.

2.2 Tasks

In selecting tasks to assess the emergence of abilities, we base our selection on those tasks that have been identified as emergent in GPT-3 by prior works. We refer to these tasks as previously identified as emergent. Out of 17 such tasks in the BIG-bench dataset (Srivastava et al., 2023), we incorporate 14 into our study. Three tasks previously identified as emergent are excluded from our analysis, because their generative nature made them challenging to assess automatically in a manner consistent with the other tasks. Additionally, to create a baseline for comparison, we randomly choose seven tasks from the same dataset that

2.2 Aufgaben

Bei der Auswahl der Aufgaben zur Beurteilung der Emergenz von Fähigkeiten stützen wir uns auf die Aufgaben, die in früheren Arbeiten als emergent in GPT-3 identifiziert wurden. Wir bezeichnen diese Aufgaben als zuvor als emergent identifiziert. Von den 17 Aufgaben im BIG-bench-Datensatz (Srivastava et al., 2023) beziehen wir 14 in unsere Studie ein. Drei Aufgaben, die zuvor als emergent identifiziert wurden, sind von unserer Analyse ausgeschlossen, da sie aufgrund ihres generativen Charakters nur schwer automatisch in einer Weise bewertet werden können, die mit den anderen Aufgaben übereinstimmt. Um eine Vergleichsbasis zu schaffen, wählen wir zusätzlich sieben Aufgaben aus demselben Datensatz aus, die zuvor nicht als emergent identifiziert worden waren.

were not previously identified as emergent. Finally, we also include GSM8K (Cobbe et al., 2021), which comprises a set of grade-school mathematics word problems and is noteworthy because even the latest models struggle with this task.

Given that formal linguistic abilities and the capacity to efficiently handle information retrieval tasks do not pose an existential threat, we manually analyse the proficiency required to solve each of the tasks we select. A full list of tasks, including their memorisability and classification as functional or formal linguistic abilities, is presented in Table 2. We determine memorisability through a manual analysis of 50 examples from each task. We provide details of our manual analysis and examples from each task in the Appendix F.

Schließlich nehmen wir auch GSM8K (Cobbe et al., 2021) auf, das eine Reihe von Wortproblemen aus der Grundschulmathematik umfasst und bemerkenswert ist, weil selbst die neuesten Modelle mit dieser Aufgabe Schwierigkeiten haben.

Da die formalen sprachlichen Fähigkeiten und die Fähigkeit zur effizienten Bewältigung von Informationsbeschaffungsaufgaben keine existenzielle Bedrohung darstellen, analysieren wir manuell die zur Lösung jeder der ausgewählten Aufgaben erforderlichen Fähigkeiten. Eine vollständige Liste der Aufgaben, einschließlich ihrer Einprägsamkeit und der Klassifizierung als funktionale oder formale sprachliche Fähigkeiten, ist in Tabelle 2 aufgeführt. Die Einprägsamkeit wird durch eine manuelle Analyse von 50 Beispielen aus jeder Aufgabe ermittelt. Einzelheiten zu unserer manuellen Analyse und Beispiele für jede Aufgabe finden Sie in Anhang F.

2.3 Settings

We evaluate each model on each task using both the few-shot and the zero-shot settings. When using the few-shot setting, we use 5 in-context examples. We note that the few-shot setting explicitly makes use of ICL, whereas the zero-shot setting does not.

2.3 Einstellungen

Wir evaluieren jedes Modell für jede Aufgabe sowohl mit der Einstellung "few-shot" als auch mit der Einstellung "zero-shot". Bei der Einstellung mit wenigen Treffern verwenden wir 5 kontextbezogene Beispiele. Es ist anzumerken, dass bei der Einstellung mit wenigen Treffern explizit ICL verwendet wird, während dies bei der Einstellung mit null Treffern nicht der Fall ist.

2.4 Evaluation Metrics

To account for the possibility that the outputs generated by non-instruction-tuned models do not match the provided answer choices exactly, we additionally evaluate using the metric BERTScore accuracy, which calculates the semantic similarity between the output text and the provided answer choices using BERTScore (Zhang et al., 2020) to estimate the model's answer choice. In this setting, the answer is considered correct if the generated answer is most similar (semantic text similarity) to the correct answer choice, and incorrect if it is closer to any of the others. The majority of the results we present in our analysis are based on this evaluation metric. It's worth noting that this is akin to selecting the answer where the

2.3 Bewertungsmassstäbe

Um der Möglichkeit Rechnung zu tragen, dass die von Modellen ohne Optimierung von Anweisungen generierten Ausgaben nicht genau mit den vorgegebenen Antwortmöglichkeiten übereinstimmen, werten wir zusätzlich die BERTScore-Genauigkeit aus, die die semantische Ähnlichkeit zwischen dem Ausgabetext und den vorgegebenen Antwortmöglichkeiten unter Verwendung von BERTScore (Zhang et al., 2020) berechnet, um die Antwortauswahl des Modells zu schätzen. In diesem Fall wird die Antwort als richtig angesehen, wenn die generierte Antwort der richtigen Antwort am ähnlichsten ist (semantische Textähnlichkeit), und als falsch, wenn sie näher an einer der anderen Antwortmöglichkeiten liegt. Die Mehrzahl der Ergebnisse, die wir in unserer Analyse präsentieren, basiert auf dieser Bewertungsmetrik. Es ist erwähnenswert, dass dies der Auswahl der Antwort

model has exhibited lowest perplexity. Since calculating this perplexity for models that are exclusively accessible through APIs is not practical, we adopt this alternative metric. We opt for BERTScore over alternatives like BLEURT (Sellam et al., 2020) because the latter are additionally trained to assess the fluency of the output text, a factor which is not our focus, and one that renders them computationally resourceintensive. For tasks that require the output of a number or a coded string (i.e., Modified arithmetic, GSM8K, and Codenames), we limit our evaluation to exact matching, as measuring semantic similarity between numbers or coded strings does not accurately reflect their proximity.

Additionally, given that recent work has indicated that emergence might be a result of discrete evaluation metrics (Schaeffer et al., 2023), we also include string edit distance. Our investigation reveals that the the lack of emergence is consistent across the metrics we use, and thus we do not use continuous metrics in our analysis. Overall, we evaluate using exact match accuracy, BERTScore accuracy, and string edit distance.

entspricht, bei der das Modell die geringste Irritation (Perplexität) gezeigt hat. Da die Berechnung dieser Komplexität für Modelle, die ausschließlich über APIs zugänglich sind, nicht praktikabel ist, verwenden wir diese alternative Metrik. Wir entscheiden uns für BERTScore anstelle von Alternativen wie BLEURT (Sellam et al., 2020), da letztere zusätzlich trainiert werden, um die Geläufigkeit des ausgegebenen Textes zu bewerten, ein Faktor, der nicht in unserem Fokus liegt und der sie rechenintensiv macht. Bei Aufgaben, die die Ausgabe einer Zahl oder einer kodierten Zeichenkette erfordern (z. B. Modifizierte Arithmetik, GSM8K und Codenames), beschränken wir unsere Untersuchung auf die exakte Übereinstimmung, da die Messung der semantischen Ähnlichkeit zwischen Zahlen oder kodierten Zeichenketten deren Nähe nicht genau widerspiegelt.

Da neuere Arbeiten darauf hindeuten, dass Emergenz ein Ergebnis diskreter Bewertungsmetriken sein könnte (Schaeffer et al., 2023), beziehen wir auch die String-Edit-Distanz mit ein. Unsere Untersuchung zeigt, dass das Fehlen von Emergenz bei allen von uns verwendeten Metriken konsistent ist, so dass wir in unserer Analyse keine kontinuierlichen Metriken verwenden. Insgesamt bewerten wir die exakte Treffergenauigkeit, die BERTScore-Genauigkeit und die String-Edit-Distanz.

2.5 Control for Bias and Manual Evaluation

In order to ensure that our evaluation is fair, we identify potential biases that could influence our findings and design our experiments to mitigate such biases. First, to ensure that non-instructiontuned models are not disadvantaged by the typically instructional task prompts, we modify these prompts, by refining them to ensure their solvability even in the absence of instruction comprehension. We then experiment with minor variations to these prompts to find the most optimal format. We also experiment with using the shortened output format, where models are only required to output a letter associated with the correct answer. We do this to remove the dependence on the non-exactmatch evaluation metrics. Importantly, we manually evaluate

2.5 Kontrolle auf Verzerrungen und manuelle Auswertung

Um sicherzustellen, dass unsere Bewertung fair ist, identifizieren wir potenzielle Verzerrungen, die unsere Ergebnisse beeinflussen könnten, und konzipieren unsere Experimente so, dass solche Verzerrungen abgeschwächt werden. Um sicherzustellen, dass Modelle ohne Instruktionen nicht durch die typischen Instruktionsaufforderungen benachteiligt werden, modifizieren wir zunächst diese Aufforderungen, indem wir sie so verfeinern, dass sie auch ohne Instruktionsverständnis lösbar sind. Anschließend experimentieren wir mit kleineren Variationen dieser Aufforderungen, um das optimale Format zu finden. Wir experimentieren auch mit der Verwendung eines verkürzten Ausgabeformats, bei dem die Modelle nur einen Buchstaben für die richtige Antwort ausgeben müssen. Dies geschieht, um die Abhängigkeit von den Bewertungsmaßstäben für nicht exakte Übereinstimmungen zu beseitigen. Wichtig ist, dass wir die Aus-

the output of our models to ensure that the prompts where appropriately interpreted by the models, especially those which are not instruction tuned. Details of these experiments and associated results are presented in Appendix B.1.

gabe unserer Modelle manuell auswerten, um sicherzustellen, dass die Aufforderungen von den Modellen angemessen interpretiert werden, insbesondere diejenigen, die nicht auf Anweisungen abgestimmt sind. Einzelheiten zu diesen Experimenten und den zugehörigen Ergebnissen finden Sie in Anhang B.1.

3 Emergence in GPT in the Absence of In-Context Learning

In this and the next section, we highlight a subset of the results with the goal of highlighting the key findings and trends from our experiments. Specifically, this section deals with the emergence of functional linguistic abilities in non-instruction-tuned models, and the next section (Section 4) focuses on exploring instruction-tuned models and their interplay with ICL and emergent abilities. Considering that prior research has identified emergent abilities in GPTwe prioritise the GPT family in our experimental analysis.

3 Emergenz von GPT in Abwesenheit von In-Context Lernen

In diesem und dem nächsten Abschnitt werden wir eine Teilmenge der Ergebnisse hervorheben, um die wichtigsten Erkenntnisse und Trends aus unseren Experimenten zu verdeutlichen. Dieser Abschnitt befasst sich insbesondere mit der Emergenz funktionaler sprachlicher Fähigkeiten in Modellen ohne Optimierung von Anweisungen, und der nächste Abschnitt (Abschnitt 4) konzentriert sich auf die Untersuchung von Modellen mit Optimierung von Anweisungen und deren Wechselwirkung mit ICL und emergenten Fähigkeiten. In Anbetracht der Tatsache, dass frühere Forschungen emergente Fähigkeiten in GPT identifiziert haben, legen wir in unserer experimentellen Analyse den Schwerpunkt auf die GPT-Familie.

3.1 Experimental Integrity and Generalisability

To validate our experimental framework, particularly the use of BERTScore accuracy and our modifications to prompts, we conduct validity tests. These involve the evaluation of instruction-tuned models with in-context examples included in the prompts, referred to as the few-shot setting, thereby enabling ICL in line with the experimental designs of prior work. The results of these tests replicated previous findings, confirming that our experimental framework does not hinder the potential for detecting emergent abilities.

Since our findings rely on the use of LLMs that have not been instruction-tuned, we verify that the observed lower performance on tasks does not stem from the automatic metric (BERTScore) failing to evaluate model responses adequately. Specifically, if the mo-

3.1 Experimentelle Integrität und Verallgemeinerbarkeit

Zur Validierung unseres experimentellen Rahmens, insbesondere der Verwendung der BERTScore-Genauigkeit und unserer Änderungen an den Prompts, führen wir Validitätstests durch. Diese beinhalten die Optimierung von Anweisungen mit kontextbezogenen Beispielen in den Aufforderungen, die als "few-shot setting" bezeichnet werden, und ermöglichen so ICL in Übereinstimmung mit den experimentellen Designs früherer Arbeiten. Die Ergebnisse dieser Tests wiederholten frühere Befunde und bestätigten, dass unser experimenteller Rahmen das Potenzial zur Erkennung emergenter Fähigkeiten nicht behindert.

Da unsere Ergebnisse auf der Verwendung von LLMs beruhen, die nicht auf Anweisungen optimiert wurden, überprüfen wir, ob die beobachtete geringere Leistung bei Aufgaben nicht darauf zurückzuführen ist, dass die automatische Metrik (BERTScore) die Modellantworten nicht angemessen bewertet. Wenn das Modell zwar eine richtige Antwort generiert, diese aber nicht

del generates an answer that is correct, but does not align with the correct target option, BERTScore accuracy might fail to provide a reliable assessment. To this end, we conducted a posthoc analysis by manually examining a subset of 50 outputs of non-instruction-tuned models from each task. Our focus was identifying instances where BERTScore accuracy failed to recognise correct responses (false negatives). Notice that false positives would not lead to an underestimation of model performance, and so have a lesser impact on our ability to identify emergence. A comprehensive description of the analysis is included in Appendix B.3. Our findings reinforce the notion that limitations – inherent to all automatic evaluation – do not detract from the overall validity of our results.

Similarly, we perform other checks for potential aspects of our experimental setup that could lead to confounding effects in our results. These include manual analysis of model outputs to ensure the our prompts were interpreted correctly (Appendix B.3), and the use of shortened outputs to enable easier evaluation (Appendix B.2).

mit der richtigen Zieloption übereinstimmt, könnte die BERTScore-Genauigkeit keine verlässliche Bewertung liefern. Zu diesem Zweck führten wir eine Post-hoc-Analyse durch, indem wir eine Teilmenge von 50 Ausgaben von Modellen ohne Optimierung von Anweisungen aus jeder Aufgabe manuell untersuchten. Unser Schwerpunkt lag auf der Identifizierung von Fällen, in denen die BERTScore-Genauigkeit keine korrekten Antworten erkennen konnte (falsch-negative Ergebnisse). Beachten Sie, dass falsch-positive Ergebnisse nicht zu einer Unterschätzung der Modellleistung führen und daher weniger Einfluss auf unsere Fähigkeit haben, Emergenz zu erkennen. Eine umfassende Beschreibung der Analyse ist in Anhang B.3 enthalten. Unsere Ergebnisse bestätigen den Gedanken, dass die Einschränkungen, die jeder automatischen Auswertung innewohnen, die allgemeine Gültigkeit unserer Ergebnisse nicht beeinträchtigen.

Ebenso führen wir weitere Überprüfungen potenzieller Aspekte unseres experimentellen Aufbaus durch, die zu verfälschenden Effekten in unseren Ergebnissen führen könnten. Dazu gehören die manuelle Analyse der Modellausgaben, um sicherzustellen, dass unsere Eingabeaufforderungen korrekt interpretiert wurden (Anhang B.3), sowie die Verwendung verkürzter Ausgaben, um eine einfachere Bewertung zu ermöglichen (Anhang B.2).

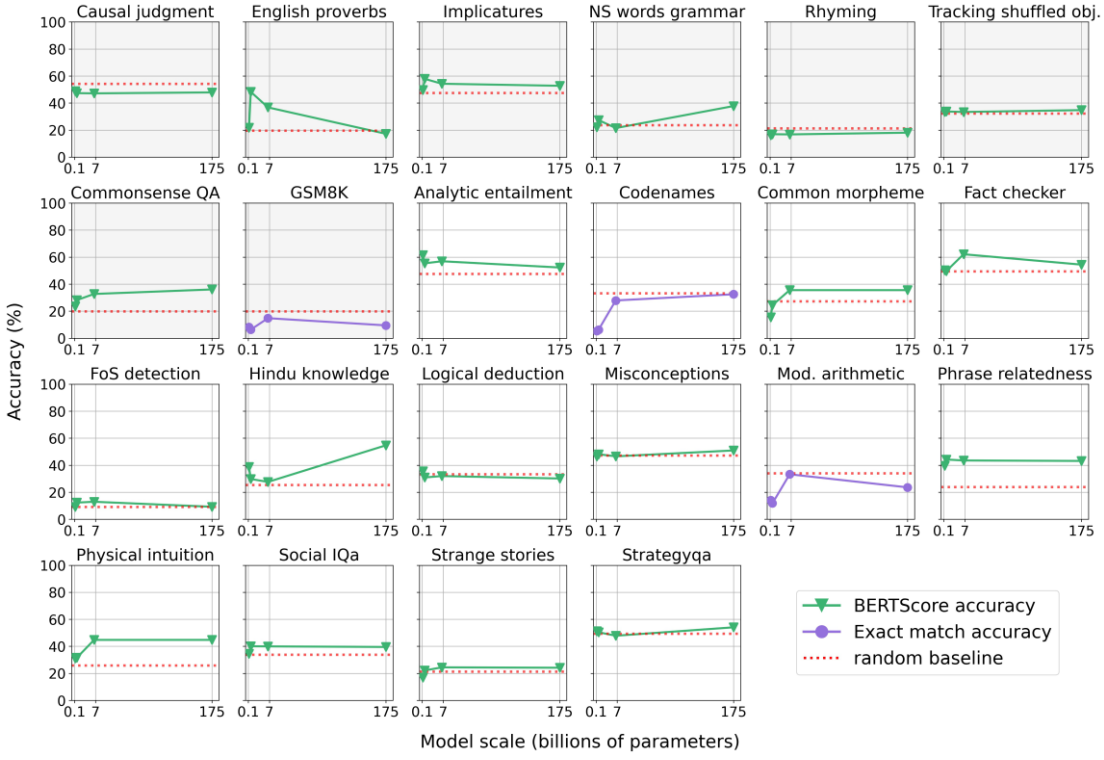

Figure 1: Performance of non instruction-tuned GPT models in the zero-shot setting. Grey

Abbildung 1: Leistung von nicht instruktionsoptimierten GPT-Modellen im Zero-Shot-Setting. Der graue

background indicates tasks that are not previously identified as emergent. Tasks that require the output of a number or a coded string are evaluated using exact match accuracy. Note the consistent lack of “emergence”, see text for details.

Hintergrund zeigt Aufgaben an, die zuvor nicht als emergent identifiziert wurden. Aufgaben, die die Ausgabe einer Zahl oder eines codierten Strings erfordern, werden mit der Genauigkeit des exakten Abgleichs bewertet. Beachte das durchgängige Fehlen von „Emergenz“, siehe Text für Details.

Task	Competence Type	Memorizable	> Random Baseline	Predictable	Emergent
Causal judgement	Functional	0	No	-	No
English Proverbs	Functional	0	No	-	No
Implicatures	Functional	0	Yes	Yes	No
NS words grammar	Formal	38	Yes	No	Yes
Rhyming	Formal	50	No	-	No
Tracking shuffled obj.	Functional	0	No	-	No
Commonsense QA	Functional	3	Yes	Yes	No
GSM8K	Functional	0	No	-	No
Analytic entailment	Functional	4	Yes	Yes	No
Codenames	Functional	0	No	-	No
Common morpheme	Formal	0	Yes	Yes	No
Fact checker	Functional	50	Yes	Yes	No
FoS detection	Functional	0	No	-	No
Hindu knowledge	Functional	50	Yes	No	Yes
Logical deduction	Functional	0	No	-	No
Misconceptions	Functional	50	Yes*	No	Yes
Mod. arithmetic	Functional	0	No	-	No
Phrase relatedness	Functional	50	Yes	Yes	No
Physical intuition	Functional	50	Yes	Yes	No
Social IQA	Functional	0	Yes	Yes	No
Strange stories	Functional	0	Yes	Yes	No
Strategy QA	Functional	27	Yes*	No	Yes

Table 2: An overview of the tasks and a categorisation as formal or functional (Competence Type). The first 8 tasks are not previously identified to be emergent. For each task, we manually determine how many of 50 examples can be solved through memorisation (Memorisable). For a task to be Emergent, models must perform above the baseline (> Random Baseline) and the performance of the larger models must not be predictable based on that of smaller models (Predictable). This table is based on the zero-shot performance of the non-instruction-tuned 175B GPT-3 model davinci. * indicates that the increase above the random baseline is less than 5%.

Tabelle 2: Ein Überblick über die Aufgaben und eine Kategorisierung als formal oder funktional (Kompetenztyp). Die ersten 8 Aufgaben wurden zuvor nicht als emergent identifiziert. Für jede Aufgabe bestimmen wir manuell, wie viele von 50 Beispielen durch Auswendiglernen gelöst werden können (Memorierbar). Damit eine Aufgabe als emergent gilt, müssen die Modelle über dem Basiswert (> Zufallsbasiswert) abschneiden, und die Leistung der größeren Modelle darf nicht auf Basis der Leistung der kleineren Modelle vorhersagbar sein (Vorhersehbar). Diese Tabelle basiert auf der Zero-Shot-Leistung des nicht instruktionsoptimierten 175B GPT-3-Modells Davinci. * zeigt an, dass der Anstieg über den Zufallsbasiswert weniger als 5 % beträgt.

Finally, to ensure generalisability of our results, we extend our analysis to the LLaMA, Falcon, and T5 model families. Across each of these cases, a consistent pattern emerges: either task performance is predictable based on smaller model performance, or the performance is below the baseline. Overall, our analysis indicates that our experimental settings do not adversely affect our capacity to identify emergent abilities and our findings are generalisable across various model families.

Schließlich, um die Generalisierbarkeit unserer Ergebnisse zu gewährleisten, erweitern wir unsere Analyse auf die Modellfamilien LLaMA, Falcon und T5. In all diesen Fällen zeigt sich ein konsistentes Muster: Entweder ist die Aufgabenleistung basierend auf der Leistung kleinerer Modelle vorhersagbar, oder die Leistung liegt unter dem Basiswert. Insgesamt deutet unsere Analyse darauf hin, dass unsere experimentellen Bedingungen unsere Fähigkeit, emergente Fähigkeiten zu identifizieren, nicht negativ beeinflussen und dass unsere Ergebnisse auf verschiedene Modellfamilien übertragbar sind.

4 Instruction-Tuning as Implicit In-Context Learning

The remarkable performance of instruction-tuned models cannot be solely attributed to their pretraining objective, which is to predict the next most probable token. This observation has led to the conjecture that models gain emergent functional linguistic abilities, such as reasoning (Wei et al., 2022c). Nevertheless, LLMs exhibit several limitations that are at odds with this view: namely, their known sensitivity to minor prompt variations and their tendency to hallucinate. This leads us to hypothesise that the primary mechanism underlying the capabilities of instruction-tuned models may in fact be an indirect form of ICL, which we call 'implicit in-context learning'. This section presents experimental results aimed at discerning whether this is the more plausible explanation underlying the performance of instruction-tuned LLMs.

Our evaluation in this section focuses on task solvability rather than performance. This is because the (sometimes wide) variation in parameter counts, architectures, and the pre-training data of the models we compare would necessarily mean that performance may differ across models. However, assessing task solvability offers a clearer insight into emergent abilities within the models. We utilise the previously-introduced BERTScore accuracy for all scenarios and evaluate models across the same 22 selected tasks previously outlined in Table 2. In this setup, unlike the previous one, we only make use of non-instruction-tuned models in the setting wherein we provide examples in-context (few-shot), thereby eliminating concerns about the models' comprehension of task requirements.

4.1 Comparative Analysis of Initial Tasks

In discerning the more plausible explanation

4 Instruktionsoptimierung als implizites In-Context-Lernen

Die bemerkenswerte Leistung von instruktionoptimierten Modellen kann nicht allein auf das Ziel des Vortrainings zurückgeführt werden, das darin besteht, das nächste wahrscheinlichste Token vorherzusagen. Diese Beobachtung hat zu der Vermutung geführt, dass Modelle emergente funktionale sprachliche Fähigkeiten entwickeln, wie zum Beispiel logisches Denken (Wei et al., 2022c). Dennoch zeigen LLMs mehrere Einschränkungen, die dieser Ansicht widersprechen: nämlich ihre bekannte Empfindlichkeit gegenüber geringfügigen Variationen der Eingabeaufforderungen (Prompts) und ihre Neigung zu Halluzinationen. Dies führt uns zu der Hypothese, dass der primäre Mechanismus, der den Fähigkeiten instruktionstuneder Modelle zugrunde liegt, tatsächlich eine indirekte Form des ICL sein könnte, die wir als „implizites In-Context-Lernen" bezeichnen. In diesem Abschnitt werden experimentelle Ergebnisse vorgestellt, die darauf abzielen, zu klären, ob dies die plausiblere Erklärung für die Leistung instruktionstuneder LLMs ist.

Unsere Bewertung in diesem Abschnitt konzentriert sich auf die Lösbarkeit von Aufgaben und nicht auf die Leistung. Dies liegt daran, dass die (manchmal erhebliche) Variation in der Anzahl der Parameter, Architekturen und den Pretraining-Daten der verglichenen Modelle zwangsläufig bedeutet, dass die Leistung zwischen den Modellen variieren kann. Die Bewertung der Aufgabenlösbarkeit bietet jedoch einen klareren Einblick in emergente Fähigkeiten innerhalb der Modelle. Wir verwenden die zuvor eingeführte BERT-Score-Genauigkeit für alle Szenarien und bewerten Modelle über die gleichen 22 ausgewählten Aufgaben, die in Tabelle 2 bereits dargelegt wurden. In diesem Setup, im Gegensatz zum vorherigen, verwenden wir nur nicht instruktionoptimierte Modelle im Setting, bei dem wir Beispiele im Kontext (Few-Shot) bereitstellen, wodurch Bedenken hinsichtlich des Verständnisses der Aufgabenanforderungen durch die Modelle beseitigt werden.

4.1. Vergleichende Analyse der Anfangsaufgaben

Um die plausiblere Erklärung für die Leistung von

underlying the performance of instruction-tuned LLMs, our experiments are designed to yield differing outcomes based on whether models exhibit functional linguistic abilities or rely predominantly on ICL. Specifically, we draw a comparison between the tasks that GPT-J (non-instruction-tuned, 6.7B) can successfully address in the few-shot setting, and those that can be solved by Flan-T5-large (instruction-tuned, 770M) in the zero-shot setting. The choice of these models is also based on the observation that there is no change in the model's performance between the zero-shot and few-shot settings for Flan-T5-large, indicating that it is too small for explicit ICL. On the other hand, we observe that there is a boost in performance across tasks in the few-shot setting for GPT-J, which indicates that it is capable of ICL. Notice that our choice of models ensures that the model we use to test which tasks can be solved using ICL is not instruction-tuned, and the model which is instruction-tuned is tested without in-context examples and also cannot explicitly access ICL. If instruction-tuning leads to models being capable of something fundamentally different from ICL (for example, functional linguistic abilities), this would result in no substantial overlap in the set of tasks solvable solely through instruction-tuning and the set of tasks addressable solely via ICL. This comparison is presented in Figure 2. We exclude Modified arithmetic from this analysis, as the task is constructed in a manner that requires the use of in-context demonstrations.

Note the substantive dissimilarity between the two models we use: Flan-T5-large is an encoderdecoder model and GPT-J is a decoder only model. Additionally, they are trained on very different pretraining datasets, one is instruction-tuned while the other isn't, and they have very different parameter counts. Despite these fundamental differences, there is a substantial overlap in both the tasks where the two models exhibit above-baseline performance, as well as an overlap in the performance scores themselves. This overlap in the results underscores a compelling argument – it is more likely that instruction-tuning serves as

instruktionsoptimierten LLMs herauszufinden, sind unsere Experimente so konzipiert, dass sie unterschiedliche Ergebnisse liefern, je nachdem, ob Modelle *funktionale* sprachliche Fähigkeiten zeigen oder überwiegend auf ICL (In-Context Learning) angewiesen sind. Insbesondere vergleichen wir die Aufgaben, die GPT-J (nicht instruktionsoptimiertes Modell, 6,7B) im Few-Shot-Setting erfolgreich lösen kann, mit denen, die von Flan-T5-large (instruktionoptimiertes Modell, 770M) im Zero-Shot-Setting gelöst werden können. Die Wahl dieser Modelle basiert auch auf der Beobachtung, dass es bei Flan-T5-large keinen Unterschied in der Leistung zwischen dem Zero-Shot- und dem Few-Shot-Setting gibt, was darauf hindeutet, dass es zu klein für explizites ICL ist. Andererseits stellen wir fest, dass es im Few-Shot-Setting bei GPT-J einen Leistungsschub über verschiedene Aufgaben hinweg gibt, was darauf hindeutet, dass es ICL-fähig ist. Beachte, dass unsere Modellwahl sicherstellt, dass das Modell, das wir verwenden, um zu testen, welche Aufgaben mit ICL gelöst werden können, nicht instruktionsoptimiert ist, und das Modell, das instruktionsoptimiert ist, ohne kontextuelle Beispiele getestet wird und auch keinen expliziten Zugriff auf ICL hat. Wenn das Instruktionsoptimierung dazu führt, dass Modelle etwas grundlegend anderes als ICL beherrschen (zum Beispiel *funktionale* sprachliche Fähigkeiten), würde dies zu keiner wesentlichen Überschneidung in der Menge der Aufgaben führen, die ausschließlich durch Instruktionsoptimierung gelöst werden können, und der Menge der Aufgaben, die ausschließlich durch ICL adressiert werden können. Dieser Vergleich wird in Abbildung 2 dargestellt. Modifizierte Arithmetik schließen wir aus dieser Analyse aus, da die Aufgabe so konstruiert ist, dass sie den Einsatz von In-Kontext Demonstrationen erfordert.

Beachte die erhebliche Unterschiedlichkeit zwischen den beiden Modellen, die wir verwenden: Flan-T5-large ist ein Encoder-Decoder-Modell, während GPT-J ein reines Decoder-Modell ist. Darüber hinaus wurden sie auf sehr unterschiedlichen Pretraining-Datensätzen trainiert, das eine Modell ist instruktionsoptimiert, während das andere dies nicht ist, und sie haben sehr unterschiedliche Parameteranzahlen. Trotz dieser grundlegenden Unterschiede gibt es eine erhebliche Überschneidung sowohl bei den Aufgaben, bei denen die beiden Modelle eine über dem Basiswert liegende Leistung zeigen, als auch bei den Leistungsergebnissen selbst. Diese Überschneidung in den Ergebnissen unterstreicht ein überzeugendes Argument – es ist wahrscheinlicher, dass die

a mechanism that enables models to harness in-context capabilities more effectively, rather than the models having emergent reasoning abilities. There are exactly 5 of the 21 tasks we test wherein one model performs markedly above the random baseline while the other does not. Indeed, some of the cases are expected: in the case of Hindu knowledge, which is a recall-based task, GPT-J, which is larger than Flan-T5-large, has an advantage and performs better. Similarly, the highly instructional nature of the Codenames renders it particularly challenging fornoninstruction-tuned models. Of the remaining 3 tasks, the better-performing GPT-J only achieves an improvement of 5% on Analytical entailment,which is binary classification. This leaves us with just Logical deduction, where Flan-T5-large benefits to some extent from the instructional nature of the questions, and Implicatures, where GPT-J achieves an accuracy of 59%.

Instruktionsoptimierung als Mechanismus dient, der es den Modellen ermöglicht, kontextuelle Fähigkeiten effektiver zu nutzen, anstatt dass die Modelle emergente Fähigkeiten zum logischen Denken entwickeln. Es gibt genau 5 von 21 getesteten Aufgaben, bei denen ein Modell deutlich über dem Zufallsbasiswert liegt, während das andere dies nicht tut. Tatsächlich sind einige dieser Fälle zu erwarten: Bei der Aufgabe "Hindu Knowledge", einer auf Abruf basierenden Aufgabe, hat GPT-J, das größer als Flan-T5-large ist, einen Vorteil und schneidet besser ab. Ebenso stellt die stark instruktionale Natur der "Codenames" eine besondere Herausforderung für nicht instruktionsoptimierte Modelle dar. Von den verbleibenden 3 Aufgaben erreicht das besser abschneidende GPT-J nur eine Verbesserung von 5 % bei "Analytical Entailment", einer binären Klassifikationsaufgabe. Dies lässt uns mit nur noch der Aufgabe "Logical Deduction", bei der Flan-T5-large bis zu einem gewissen Grad von der instruktionalen Natur der Fragen profitiert, und "Implicatures", bei der GPT-J eine Genauigkeit von 59 % erreicht.

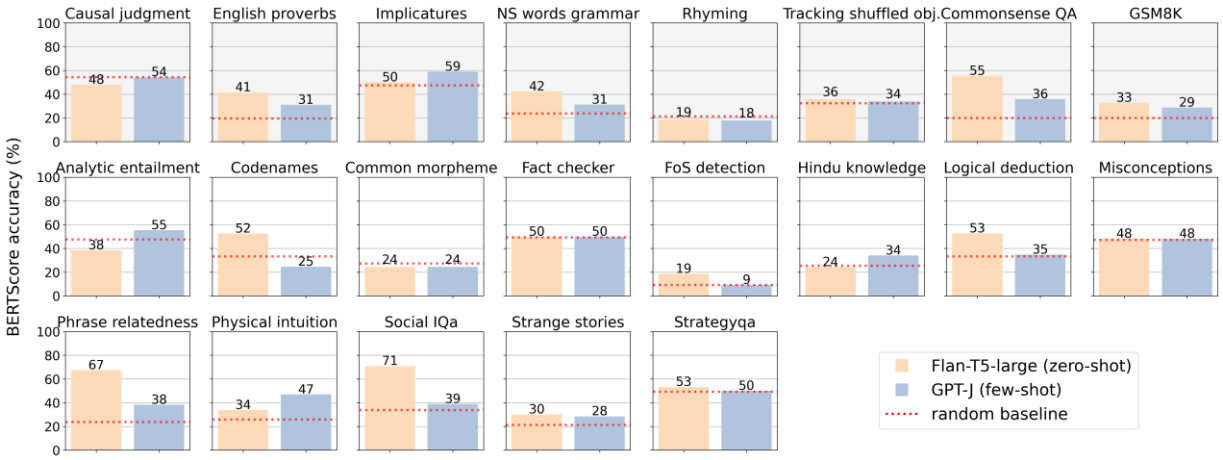

Figure 2: The substantial overlap of the tasks on which the two models perform above the random baseline is noteworthy and indicates that instruction-tuning allows for the effective access of in-context capabilities rather than leading to the emergence of functional linguistic abilities. See text for details.

Abbildung 2: Die erhebliche Überschneidung der Aufgaben, bei denen die beiden Modelle über dem Zufallsbasiswert abschneiden, ist bemerkenswert und deutet darauf hin, dass das Instruktionsoptimierung den effektiven Zugriff auf In-Kontext Fähigkeiten ermöglicht, anstatt zur Entstehung funktionaler sprachlicher Fähigkeiten zu führen. Siehe Text für Details.

4.2 Generalisability

To evaluate if our results generalise to a further increase in model size and instruction-tuning data, we compare the tasks that

4.2 Verallgemeinerbarkeit

Um zu bewerten, ob unsere Ergebnisse auch bei einer weiteren Erhöhung der Modellgröße und der Datenmenge für das Instruktionstuning generalisier-

can be effectively tackled by Flan-T5-large with those by instructiontuned versions of the largest GPT models, i.e., text-davinci-001 and text-davinci-003 (additionally trained extensively on program code). It is important to note that these models have more than 200 times the number of parameters present in Flan-T5-large. We perform this comparison in the zero-shot setting, thus allowing us to compare the instruction-following capabilities of these models without triggering their ICL capabilities, which we know to increase markedly with scale.

This comparison allows us to answer the following questions: a) Does increased scale largely impact the tasks on which models can perform above the random baseline, and b) Does enhanced instruction-tuning, including the incorporation of program code as seen in text-davinci-003, provide an advantage in being able to perform above the baseline on tasks? By limiting ourselves to the zero-shot setting, we ensure that our results are not affected by in-context capabilities, which we know to increase significantly with scale. Our results indicate that neither scale nor the inclusion of program code in instruction-tuning markedly alters the task solvability of a model. There is a substantial overlap in the tasks on which Flan-T5-large performs above the baseline and those on which text-davinci-001 and text-davinci-003 do: 16 of the 22 tasks we experiment with show this congruence. This overlap, and in several instances comparable performance across these diverse models, suggests that the effectiveness of instructiontuning is consistent regardless of model scale or the nature of tuning datasets, in the absence of explicit ICL. Among non-overlapping tasks, certain recall-based tasks are better handled by larger GPT models due to their better recall abilities. These results, illustrated in Figure 5, Appendix D, confirm that our hypothesis, namely that 'implicit incontext learning' is likely the primary mechanism in instruction-tuned LLMs, and that it is generalisable across model sizes and various instructiontuning datasets. This also suggests that further scaling will probably not alter this trend.

bar sind, vergleichen wir die Aufgaben, die effektiv von Flan-T5-large bewältigt werden können, mit denen, die von instruktionsoptimiert Versionen der größten GPT-Modelle, d.h. text-davinci-001 und text-davinci-003 (zusätzlich intensiv auf Programmcodes trainiert), gelöst werden können. Es ist wichtig zu beachten, dass diese Modelle mehr als 200-mal so viele Parameter haben wie Flan-T5-large. Wir führen diesen Vergleich im Zero-Shot-Setting durch, um die Fähigkeit dieser Modelle zur Befolgung von Anweisungen zu vergleichen, ohne ihre ICL-Fähigkeiten auszulösen, von denen wir wissen, dass sie mit der Modellgröße deutlich zunehmen.

Dieser Vergleich ermöglicht es uns, die folgenden Fragen zu beantworten: a) Hat die erhöhte Skalierung einen wesentlichen Einfluss auf die Aufgaben, bei denen Modelle über dem Zufallsbasiswert abschneiden können, und b) bietet verbessertes Instruktionstuning, einschließlich der Einbeziehung von Programmcodes, wie sie bei text-davinci-003 zu sehen ist, einen Vorteil, um Aufgaben über dem Basiswert zu lösen? Indem wir uns auf das Zero-Shot-Setting beschränken, stellen wir sicher, dass unsere Ergebnisse nicht durch kontextuelle Fähigkeiten beeinflusst werden, von denen wir wissen, dass sie mit zunehmender Modellgröße deutlich zunehmen. Unsere Ergebnisse zeigen, dass weder die Skalierung noch die Einbeziehung von Programmcodes in der Instruktionsfeinabstimmung die Aufgabenlösbarkeit eines Modells wesentlich verändert. Es gibt eine erhebliche Überschneidung bei den Aufgaben, bei denen Flan-T5-large über dem Basiswert abschneidet, und denen, bei denen text-davinci-001 und text-davinci-003 dies tun: 16 der 22 von uns untersuchten Aufgaben zeigen diese Übereinstimmung. Diese Überschneidung und in mehreren Fällen vergleichbare Leistungen über diese unterschiedlichen Modelle hinweg legen nahe, dass die Effektivität der Instruktionsfeinabstimmung unabhängig von der Modellgröße oder der Art der Tuning-Datensätze konsistent ist, sofern kein explizites ICL vorliegt. Bei nicht überlappenden Aufgaben werden bestimmte auf Abruf basierende Aufgaben von größeren GPT-Modellen aufgrund ihrer besseren Erinnerungsfähigkeiten besser bewältigt. Diese Ergebnisse, dargestellt in Abbildung 5, Anhang D, bestätigen unsere Hypothese, dass das „implizite In-Context-Lernen" wahrscheinlich der primäre Mechanismus in instruktionstuneden LLMs ist und dass es auf verschiedene Modellgrößen und unterschiedliche Instruktionstuning-Datensätze generalisierbar ist. Dies deutet auch darauf hin, dass eine weitere Skalierung diesen Trend wahrscheinlich nicht verändern wird.

4.3 A Novel Theoretical Foundation

Based on our observations on the capabilities and limitations of LLMs, we propose a novel alternative theory explaining why instruction-tuning helps models perform better: we propose that instructiontuning enables models to map instructions to the form required for ICL, thus allowing instructiontuned models to solve tasks using some implicit form of ICL. Importantly, during this process, models could be directly making use of the same underlying mechanism that makes ICL possible, just in a different way than when the model explicitly makes use of ICL from examples provided in the prompt. We call this use of ICL 'implicit' in-context learning. Performing such a mapping would be relatively straightforward for a very large model, especially given that this task format aligns closely with the training process carried out during instruction-tuning. Investigating the exact nature of this mechanism is left for future work.

4.3 Eine neue theoretische Grundlage

Basierend auf unseren Beobachtungen zu den Fähigkeiten und Einschränkungen von LLMs schlagen wir eine neue alternative Theorie vor, die erklärt, warum Instruktionstuning Modelle dabei unterstützt, besser zu performen: Wir vermuten, dass Instruktionstuning (Instruktionsfeinabstimmung) es den Modellen ermöglicht, Anweisungen so zuzuordnen, dass sie die Form annehmen, die für In-Context Learning (ICL) erforderlich ist, wodurch instruktionstunede Modelle Aufgaben mithilfe einer impliziten Form von ICL lösen können. Wichtig ist dabei, dass Modelle während dieses Prozesses möglicherweise direkt denselben zugrundeliegenden Mechanismus nutzen, der ICL möglich macht, jedoch auf eine andere Weise, als wenn das Modell explizit ICL anhand der in der Eingabeaufforderung bereitgestellten Beispiele verwendet. Wir nennen diese Nutzung von ICL „implizites" In-Context-Lernen. Für ein sehr großes Modell wäre es relativ einfach, eine solche Zuordnung vorzunehmen, insbesondere da dieses Aufgabenformat eng mit dem Trainingsprozess des Instruktionstunings übereinstimmt. Die genaue Untersuchung der Natur dieses Mechanismus bleibt zukünftigen Arbeiten vorbehalten.

5 Related Work

Emergent Abilities An emergent ability was first defined as an ability that is not present in smaller models but is present in larger models (Wei et al., 2022b). From a review of prior literature of LLMs including GPT-3, PaLM (Chowdhery et al., 2023), Chinchilla (Hoffmann et al., 2022), Gopher (Rae et al., 2021) and LaMDA (Thoppilan et al., 2022), Wei et al. (2022b) identified a total of 67 emergent abilities based on above-random performance of LLMs on tasks designed to test those abilities from the BIG-bench dataset (Srivastava et al., 2023), and the Massive Multitask Language Understanding Benchmark (Hendrycks et al., 2020). Subsequent studies have explored additional abilities emergent in LLMs, such as Theory of Mind (Kosinski, 2023) and cognitive biases (Itzhak et al., 2023). However, Schaeffer et al. (2023) have previously questioned the existence of emergent abilities, arguing that

5 Verwandte Arbeiten

Emergente Fähigkeiten Eine emergente Fähigkeit wurde erstmals als eine Fähigkeit definiert, die in kleineren Modellen nicht vorhanden ist, jedoch in größeren Modellen vorhanden ist (Wei et al., 2022b). In einer Übersicht über die frühere Literatur zu großen Sprachmodellen (LLMs), einschließlich GPT-3, PaLM (Chowdhery et al., 2023), Chinchilla (Hoffmann et al., 2022), Gopher (Rae et al., 2021) und LaMDA (Thoppilan et al., 2022), identifizierten Wei et al. (2022b) insgesamt 67 emergente Fähigkeiten, basierend auf einer Leistung der LLMs, die über dem Zufallsniveau bei Aufgaben lag, die zur Prüfung dieser Fähigkeiten aus dem BIG-bench-Datensatz (Srivastava et al., 2023) und dem Massive Multitask Language Understanding Benchmark (Hendrycks et al., 2020) entwickelt wurden. Nachfolgende Studien haben zusätzliche Fähigkeiten untersucht, die in LLMs emergieren, wie die „Theory of Mind" (Kosinski, 2023) und kognitive Verzerrungen (Itzhak et al., 2023). Allerdings haben Schaeffer et al. (2023) zuvor die Existenz von emergenten Fähigkeiten in

emergence is likely to be a consequence of the discrete evaluation metrics commonly employed for assessing LLMs. Some (Wei et al., 2022b) argue against this by pointing out that there are tasks on which LLMs are able to perform well above the random baseline where smaller models can only perform below it, suggesting that these abilities are still emergent and not just a consequence of discrete evaluation metrics. Similarly, several works (Biderman et al., 2023; Tefnik and Kadlˇcík, 2023; Wu et al., 2023; Zheng et al., 2023) have explored the extent to which memory plays a
role in LLMs' abilities.

In-Context Learning ICL is a learning paradigm that has gained great popularity with the advent of LLMs (Brown et al., 2020; Liu et al., 2023). ICL typically involves prompting an LLM with in-context demonstrations, and offers a more interpretable interface as well as greater computational efficiency compared to previous learning approaches (Dong et al., 2023; Zhou et al., 2023). Notably, ICL has demonstrated strong performance on various natural language tasks (Kojima et al., 2022; Lampinen et al., 2022; Wei et al., 2023).

In terms of the theoretical rationale for ICL in LLMs, recent work indicates that it might share similarities with fine-tuning, in that it might allow models to "learn" from the examples presented in their prompt (Dai et al., 2023). Similarly, it has been shown that ICL implements gradient descent implicitly and constructs a function at inference time on regression problems (Akyürek et al., 2023; Li et al., 2023; Zhang et al., 2023a), which may be related to gradient-based meta-learning (von Oswald et al., 2023). A line of work shows that ICL is driven by the distributions of the pre-training data (Chan et al., 2022; Hahn and Goyal, 2023). Some other theoretical explorations attempt to explain ICL in terms of Bayesian inference (Xie et al., 2022; Li et al., 2023; Zhang et al., 2023b).

To the best of our knowledge, none of the previous evaluations of emergent abilities have been conducted in a manner that explicitly distinguished between the ICL and

Frage gestellt und argumentiert, dass Emergenz wahrscheinlich eine Folge der diskreten Bewertungsmetriken ist, die üblicherweise zur Beurteilung von LLMs verwendet werden. Einige (Wei et al., 2022b) widersprechen dem, indem sie darauf hinweisen, dass es Aufgaben gibt, bei denen LLMs deutlich besser als der Zufallswert abschneiden, während kleinere Modelle nur darunter abschneiden, was darauf hindeutet, dass diese Fähigkeiten immer noch emergent sind und nicht nur eine Folge der diskreten Bewertungsmetriken. In ähnlicher Weise haben mehrere Arbeiten (Biderman et al., 2023; Tefnik und Kadlčík, 2023; Wu et al., 2023; Zheng et al., 2023) untersucht, inwieweit das Gedächtnis eine Rolle spielt.

In-Context Learning ICL ist ein Lernparadigma, das mit dem Aufkommen großer Sprachmodelle (LLMs) große Popularität erlangt hat (Brown et al., 2020; Liu et al., 2023). ICL beinhaltet typischerweise, dass ein LLM mit kontextuellen Demonstrationen gepromptet wird, und bietet im Vergleich zu früheren Lernansätzen eine besser interpretierbare Schnittstelle sowie eine höhere Recheneffizienz (Dong et al., 2023; Zhou et al., 2023). Bemerkenswert ist, dass ICL starke Leistungen bei verschiedenen Aufgaben im Bereich der natürlichen Sprache gezeigt hat (Kojima et al., 2022; Lampinen et al., 2022; Wei et al., 2023).

In Bezug auf die theoretische Begründung für In-Context Learning (ICL) in LLMs deuten aktuelle Arbeiten darauf hin, dass es Ähnlichkeiten mit dem Fine-Tuning teilen könnte, indem es den Modellen ermöglicht, aus den in der Eingabeaufforderung präsentierten Beispielen „zu lernen" (Dai et al., 2023). Ebenso wurde gezeigt, dass ICL implizit Gradient Descent implementiert und zur Laufzeit eine Funktion für Regressionsprobleme konstruiert (Akyürek et al., 2023; Li et al., 2023; Zhang et al., 2023a), was möglicherweise mit dem gradientenbasierten Meta-Learning zusammenhängt (von Oswald et al., 2023). Eine Reihe von Arbeiten zeigt, dass ICL von den Verteilungen der Pretraining-Daten angetrieben wird (Chan et al., 2022; Hahn und Goyal, 2023). Andere theoretische Untersuchungen versuchen, ICL im Sinne der Bayesschen Inferenz zu erklären (Xie et al., 2022; Li et al., 2023; Zhang et al., 2023b).

Soweit uns bekannt ist, wurde keine der bisherigen Bewertungen von emergenten Fähigkeiten in einer Weise durchgeführt, die explizit zwischen den ICL- und Instruktionstuning-Settings unterscheiden und

instruction-tuning settings and prompting in the setting wherein these abilities are not triggered.

dabei das Setting berücksichtigt haben, indem diese Fähigkeiten nicht ausgelöst werden.

6 Conclusions and Implications

We started with two hypotheses: a) That the emergence of all previously-observed functional linguistic abilities is a consequence of ICL, and b) That the abilities which present themselves in instructiontuned LLMs is more likely to be indicative of instruction-tuning resulting in implicit ICL, rather than the emergence of functional linguistic abilities. Our results confirmed both of these hypotheses.

The distinction between the ability to follow instructions and the inherent ability to solve a problem is a subtle but important one, and bears significance to the methods employed in utilising LLMs and the problems they are tasked with solving. Simple following of instructions without applying reasoning abilities produces output that is consistent with the instructions, but might not make sense on a logical or commonsense basis. This is reflected in the well-known phenomenon of 'hallucination', in which an LLM produces fluent, but factually incorrect output (Bang et al., 2023; Thorp, 2023) The ability to follow instructions does not imply having reasoning abilities, and more importantly, it does not imply the possibility of latent, potentially-dangerous abilities. Additionally, these observations imply that our findings hold true for any model which exhibits a propensity for hallucination or requires prompt engineering, including those with greater complexity, regardless of scale or number of modalities, such as GPT-4. By contributing to a deeper understanding of these models' abilities and limitations, we help to demystify LLMs, alleviate their related safety concerns, and lay out a framework for their more efficient use.

6 Schlussfolgerungen und Implikationen

Wir begannen mit zwei Hypothesen: a) Dass das Auftreten aller bisher beobachteten funktionalen sprachlichen Fähigkeiten eine Folge von In-Context Learning (ICL) ist, und b) dass die Fähigkeiten, die sich in instruktionstuneden LLMs zeigen, eher darauf hinweisen, dass das Instruktionstuning zu implizitem ICL führt, anstatt zur Entstehung funktionaler sprachlicher Fähigkeiten. Unsere Ergebnisse bestätigten beide Hypothesen.

Der Unterschied zwischen der Fähigkeit, Anweisungen zu befolgen, und der inhärenten Fähigkeit, ein Problem zu lösen, ist subtil, aber wichtig und hat Auswirkungen auf die Methoden, die bei der Nutzung von LLMs angewendet werden, sowie auf die Probleme, die sie lösen sollen. Das einfache Befolgen von Anweisungen ohne Anwendung von Denkfähigkeiten führt zu Ausgaben, die mit den Anweisungen übereinstimmen, aber möglicherweise keinen logischen oder gesunden Menschenverstand ergeben. Dies spiegelt sich in dem bekannten Phänomen der „Halluzination" wider, bei dem ein LLM flüssige, aber sachlich falsche Ausgaben erzeugt (Bang et al., 2023; Thorp, 2023). Die Fähigkeit, Anweisungen zu befolgen, impliziert nicht, dass man über Denkfähigkeiten verfügt, und vor allem nicht, dass latente, potenziell gefährliche Fähigkeiten vorhanden sind. Diese Beobachtungen deuten zudem darauf hin, dass unsere Erkenntnisse für jedes Modell zutreffen, das eine Neigung zu Halluzinationen aufweist oder Prompt-Engineering erfordert, einschließlich solcher mit größerer Komplexität, unabhängig von der Skalierung oder der Anzahl der Modalitäten, wie z.B. GPT-4. Indem wir zu einem tieferen Verständnis der Fähigkeiten und Grenzen dieser Modelle beitragen, helfen wir, LLMs zu entmystifizieren, ihre damit verbundenen Sicherheitsbedenken zu mindern und einen Rahmen für ihre effizientere Nutzung zu schaffen.

Limitations

Although we experiment on an extensive amount of model sizes across various architectures (e.g., T5, GPT, Falcon, LLaMA), we were unable to ensure an exact match of parameter counts across the different architectures. This is due to the variation in the publicly-available releases of these models. In this work, we used all models at the parameter counts that were available. However, another alternative would be to conduct pre-training to ensure equal parameter counts and comparable pretraining data, though this would involve a substantial computational investment. In all tasks, there is a risk of data leakage, especially for LLMs whose training datasets are not publicly known. In this work, we assume that data leakage has not occurred beyond what was reported in official publications for specific models (e.g., BIG-bench for GPT-4). As such, we do not consider data leakage a factor when we consider a task to be 'memory-based', although, in practice, the presence of data leakage can have a biasing effect on model pe formance. Our experiments are limited to English tasks. This is primarily a consequence of previous work on emergent abilities and on the limitations of computational budget to run experiments on other languages. We intend to focus future work on datasets that include other languages including low resource languages.

Ethical Considerations

Our work does not imply that LLMs have absolutely no potential for harm. By leveraging the sophisticated linguistic capabilities of LLMs, malicious actors can craft highly convincing and personalised fake news articles or phishing messages, which may become increasingly difficult to distinguish from legitimate messages. The ease and efficiency with which LLMs can be used for these purposes highlight the need for detection mechanisms, along with ethical guidelines to mitigate the risks and protect individuals and democratic processes. Similarly, identifying that LLM capabilities are

Begrenzungen

Obwohl wir Experimente mit einer umfangreichen Anzahl von Modellgrößen über verschiedene Architekturen hinweg durchgeführt haben (z.B. T5, GPT, Falcon, LLaMA), konnten wir keine exakte Übereinstimmung der Parameteranzahlen zwischen den verschiedenen Architekturen sicherstellen. Dies liegt an den Unterschieden in den öffentlich verfügbaren Versionen dieser Modelle. In dieser Arbeit haben wir alle Modelle mit den verfügbaren Parameteranzahlen verwendet. Eine alternative Möglichkeit wäre, ein Pretraining durchzuführen, um gleiche Parameteranzahlen und vergleichbare Pretraining-Daten zu gewährleisten, was jedoch einen erheblichen rechnerischen Aufwand erfordern würde. Bei allen Aufgaben besteht das Risiko eines Datenlecks, insbesondere bei LLMs, deren Trainingsdatensätze nicht öffentlich bekannt sind. In dieser Arbeit gehen wir davon aus, dass kein Datenleck über das hinaus aufgetreten ist, was in offiziellen Veröffentlichungen für bestimmte Modelle berichtet wurde (z.B. BIG-bench für GPT-4). Daher betrachten wir Datenlecks nicht als Faktor, wenn wir eine Aufgabe als „gedächtnisbasiert" betrachten, obwohl in der Praxis das Vorhandensein von Datenlecks einen verzerrenden Effekt auf die Modellleistung haben kann. Unsere Experimente beschränken sich auf Aufgaben in englischer Sprache. Dies ist hauptsächlich eine Folge der vorherigen Arbeiten zu emergenten Fähigkeiten und der begrenzten Rechenressourcen, die für Experimente in anderen Sprachen erforderlich wären. Wir beabsichtigen, uns in zukünftigen Arbeiten auf Datensätze zu konzentrieren, die andere Sprachen, einschließlich ressourcenarmer Sprachen, umfassen.

Ethische Überlegungen

Unsere Arbeit impliziert nicht, dass LLMs absolut kein Potenzial für Schaden haben. Durch die Nutzung der ausgeklügelten sprachlichen Fähigkeiten von LLMs können böswillige Akteure hoch überzeugende und personalisierte Fake-News-Artikel oder Phishing-Nachrichten erstellen, die zunehmend schwer von legitimen Nachrichten zu unterscheiden sein könnten. Die Leichtigkeit und Effizienz, mit der LLMs für solche Zwecke eingesetzt werden können, unterstreicht die Notwendigkeit von Erkennungsmechanismen sowie ethischen Richtlinien, um die Risiken zu mindern und Individuen sowie demokratische Prozesse zu schützen. Ebenso bedeutet die Feststellung, dass die Fähigkeiten von LLMs kein Vorbote einer

not a precursor to an AI-driven existential threat does not eliminate the need for ongoing vigilance in AI safety research. Our findings present an unique opportunity to prioritise the most pressing aspects of LLM safety while simultaneously exploring research avenues beyond mere scaling up.

Further chapters

For the other chapters (References, Acknowledgements and Appendices), please consult the official English-language version.

existenziellen Bedrohung durch KI sind, nicht, dass die Notwendigkeit einer kontinuierlichen Wachsamkeit in der KI-Sicherheitsforschung entfällt. Unsere Ergebnisse bieten eine einzigartige Gelegenheit, die dringendsten Aspekte der LLM-Sicherheit zu priorisieren und gleichzeitig Forschungswege jenseits der bloßen Skalierung zu erkunden.

Weitere Kapitel

Für die weiteren Kapitel (Referenzen, Dank und Anhänge) konsultieren Sie bitte die offizielle englischsprachige Fassung.