Eight Things to Know about Large Language Models

Samuel R. Bowman

LLMs predictably get more capable with increasing investment, even without targeted innovation

OpenAI's original GPT can perform simple text-labeling tasks but cannot generally produce coherent text (Radford et al., 2018).

GPT-2 adds the ability to produce text of reasonably high quality, as well as a limited ability to follow simple instructions (Radford et al., 2019).

GPT-3 is the first modern general-purpose LLM, and is practically useful across a wide range of language tasks.

The designs of these three models hardly differ at all. Instead, the qualitative differences between them stem from vast differences in scale: Training GPT-3 used roughly 20,000× more computation than training the original GPT (Sevilla et al., 2022), as well as significantly more data and parameters. There are substantial innovations that distinguish these three models, but they are almost entirely restricted to infrastructural innovations in high-performance computing rather than model-design work that is specific to language technology.

Specific important behaviors in LLM tend to emerge unpredictably as a byproduct of increasing investment

It is largely not possible to predict when models will start to show specific skills or become capable of specific tasks.

Often, a model can fail at some task consistently, but a new model trained in the same way at five or ten times the scale will do well at that task.

Specific important behaviors in LLM tend to emerge unpredictably as a byproduct of increasing investment

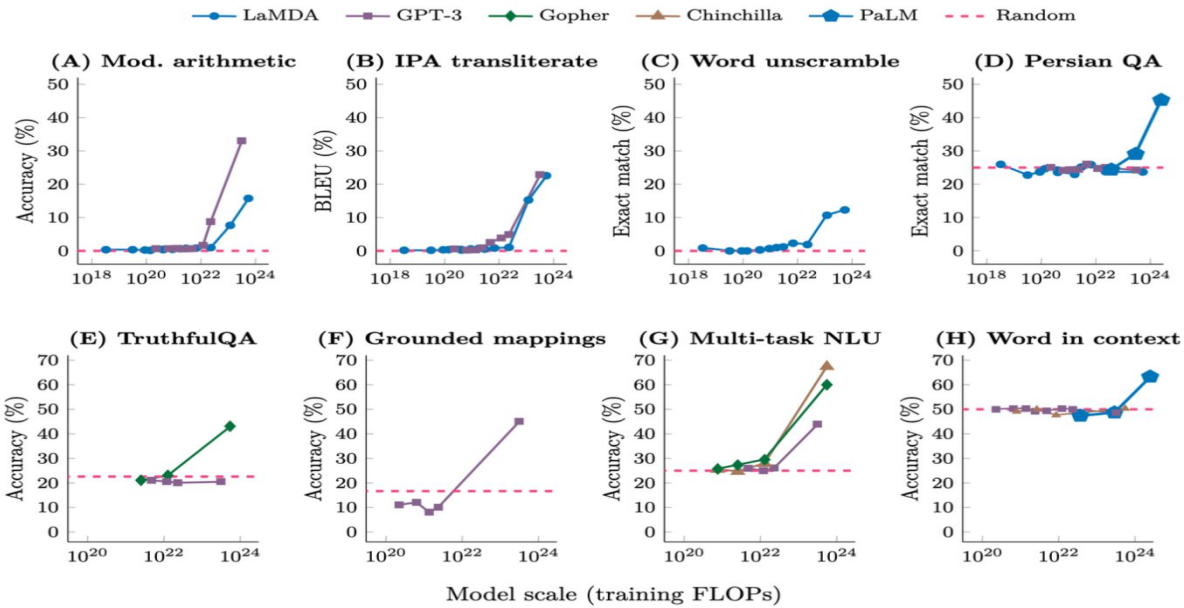

Figure 2. Excerpted from Wei et al. (2022a): Evaluations of performance on specific tasks or behaviors in LLMs do not generally show predictable trends, and it is common for new behaviors to emerge abruptly when transitioning from a less resource-intensive version of a model to a more resource-intensive one.

Specific important behaviors in LLM tend to emerge unpredictably as a byproduct of increasing investment

It is largely not possible to predict when models will start to show specific skills or become capable of specific tasks.
Often, a model can fail at some task consistently, but a new model trained in the same way at five or ten times the scale will do well at that task.

Concretely, two of the key behaviors in GPT-3 that set it apart as the first modern LLM are -

It shows few-shot learning, the ability to learn a new task from a handful of examples in a single interaction, and
Chain-of-thought reasoning, the ability to write out its reasoning on hard tasks when requested, as a student might do on a math test, and to show better performance as a result.

GPT-3's capacity for few-shot learning on practical tasks appears to have been discovered only after it was trained, and its capacity for chain-of-thought reasoning was discovered only several months after it was broadly deployed to the public.

LLMs often appear to learn and use representations of the outside world

There is increasingly substantial evidence that LLMs develop internal representations of the world to some extent, and that these representations allow them to reason at a level of abstraction that is not sensitive to the precise linguistic form of the text that they are reasoning about.

- Models can at least sometimes give instructions describing how to draw novel objects
- Models that are trained to play board games from descriptions of individual game moves, without ever seeing a full depiction of the game board, learn internal representations of the state of the board at each turn
- Models can distinguish common misconceptions from true facts (Wei et al., 2022a), and often show well-calibrated internal representations for how likely a claim is to be true.

Figure 4. Excerpted from Bubeck et al. (2023): An popular informal (and potentially cherry-picked) demonstration of LLMs' ability to manipulate visual representations. Here, a private version of GPT-4, trained without any access to visual information, is asked to write instructions in a graphics programming language to draw a unicorn. During the model's training (left to right), the resulting drawings appear to become more competent.

There are no reliable techniques for steering the behavior of LLMs

- Plain language model prompting, where one prepares an incomplete text like “The translation of ‘cat’ in French is‘”, such that a typical continuation of the text should represent a completion of the intended task
- Supervised fine-tuning, where one trains the model to match high-quality human demonstrations on the task
- Reinforcement learning, where one incrementally weakens or strengthens certain model behaviors according to preference judgments from a human tester or user

These techniques produce useful systems, but they are far from perfectly effective: They can’t guarantee that an AI model will behave appropriately in every plausible situation it will face in deployment. Nor can they even make a model try to behave appropriately to the extent possible given its skills and knowledge (to the extent that it can be said to have generalizable skills or knowledge).

Experts are not yet able to interpret the inner workings of LLMs

There are hundreds of billions of connections between these artificial neurons, some of which are invoked many times during the processing of a single piece of text, such that any attempt at a precise explanation of an LLM's behavior is doomed to be too complex for any human to understand.

Often, ad-hoc techniques that at first seem to provide insight into the behavior of an LLM are later found to be severely misleading (Feng et al., 2018; Jain & Wallace, 2019; Bolukbasi et al., 2021; Wang et al., 2022).

In addition, promising-looking techniques that elicit reasoning in natural language do not reliably correspond to the processes that LLMs use to reason, and model-generated explanations can also be systematically misleading

Human performance on a task isn’t an upper bound on LLM performance

While LLMs are trained primarily to imitate human writing behavior, they can at least potentially outperform humans on many tasks.

This is for two reasons:

First, they are trained on far more data than any human sees, giving them much more information to memorize and potentially synthesize.

In addition, they are often given additional training using reinforcement learning before being deployed.

LLMs need not express the values of their creators nor the values encoded in web Text

AI Alignment Problem:

Make the AI behave like we want it to behave, But plain LLM reflects whatever biases and opinions it has in it’s training data.

Brief interactions with LLMs are often misleading

Often, a model will fail to complete a task when asked, but will then perform the task correctly once the request is reworded or reframed slightly, leading to the emerging craft of prompt engineering

Often, once one finds an appropriate way to prompt a model to do some task, one will find that the model consistently performs well across different instances of the task.

The chain-of-thought reasoning strategies are an especially clear example of this: Simply prompting a model to “think step by step” can lead it to perform well on entire categories of math and reasoning problems that it would otherwise fail on.

More Discussions

- We should expect some of the prominent flaws of current LLMs to improve significantly
    - Mitigating Hallucination, bias and toxicity

- There will be incentives to deploy LLMs as agents that flexibly pursue goals.
- LLM developers have limited influence over what is developed

- LLMs are likely to produce a rapidly growing array of risks.
- Negative results with LLMs can be difficult to interpret but point to areas of real weakness
- The science and scholarship around LLMs is especially immature