Examining the Emergence of Deductive Reasoning in Generative Language Models

Peter Belcak and **Luca A. Lanzendörfer**
ETH Zürich
{belcak,lanzendoerfer}@ethz.ch

Abstract

We conduct a preliminary inquiry into the ability of generative transformer models to deductively reason from premises provided. We observe notable differences in the performance of models coming from different training setups and find that the deductive reasoning ability increases with scale. Further, we discover that the performance generally does not decrease with the length of the deductive chain needed to reach the conclusion, with the exception of OpenAI GPT-3 and GPT-3.5 models. Our study considers a wide variety of transformer-decoder models, ranging from 117 million to 175 billion parameters in size.

1 Introduction

It is inherently desirable for foundation language models to be capable of reasoning. This is especially true for generative models, where the hope is that the models could ultimately hold a profound conversation or continue a chain logical reasoning with human-like performance.

The most tractable of modes of reasoning is deduction, which employs premises and general laws to reach conclusions. In contrast to inductive reasoning, which deals with generalisation of observed phenomena, and abductive reasoning, which is invoked to find a likely cause for a conclusion in some context of knowledge, deduction is the mode of reasoning used to find the consequences of available knowledge. As such, it is a natural and necessary part of any generative language model – a model trained on a large corpus of knowledge and queried to make use of it must internally perform a chain of deductions to produce a true consequence of its knowledge as prompted.

Our aim is to study large language models, assess their ability to perform deductive reasoning, and see how they fare as the size of the model and the complexity of required deductions grow. Note that the use of deductive reasoning is implicitly assumed when interacting with a large generative language model. It is expected that regardless of the prompt, the model will produce an output that logically follows either from the model's knowledge or the context provided, and that this output complies with the prompting instruction. Even when asked to perform creative writing or continue a paragraph, it is expected that the generated text will logically follow or at least not contradict the context (Yuan et al., 2022). Hence, our study considers the native scenario in which no particular prompting instruction is provided.

Large language models have been shown to excel at memorising fragments of factual statements in their training data and at internalising highly co-occurring factual associations (Li et al., 2022; Carlini et al., 2022). A study that would hope to evaluate model's deductive reasoning ability while requiring "common sense" (Lourie et al., 2021; Singh et al., 2021) would thus also be indirectly evaluating its common sense knowledge. We therefore steer clear of any factual knowledge and carry out our evaluation under a restricted setting, in which we provide the model with a context of premises (a theory) and ask a single *yes/no* question, insisting that

P1. all premises are generic in their wording (they only involve neutral entities and properties),

P2. all premises are precise in their meaning (there are no ambiguities arising from synonymy, paraphrasing, or categorical adherence),

P3. all conclusions that may be queried by the question can be arrived at by (possibly several) applications of *modus ponens* (if A and if A then B then B), and that

P4. all premises that hold true are explicitly listed and provided to the model, with any fact that is not a premise nor a consequence assumed

not to be true (i.e. we make the closed-world assumption).

We set out to answer three research questions:

Q1. How does the deductive reasoning ability (DRA) of large language models evolve with their growing size?

Q2. How does the DRA of general-purpose generative large language models depend on the number of deductive steps required to reach the conclusions?

Q3. Do the specifics of a training setup, such as the model variant, dataset multilingualism, or dataset size have a decisive influence on the DRA?

We measure DRA by binary accuracy on the answers of the model; in other words, by the proportion of the instances in which the model answered the question about the theory given as context correctly. A detailed account of our experimental setup is presented in Section 2.

Q1 sets the first dimension of our inquiry – the impact of the size of the language model on DRA. Even relatively small pre-trained language models (BERT, RoBERTa) have been shown to perform well on isolated instances of unstructured natural language inference when fine-tuned to that end (Wang et al., 2021; Sun et al., 2020). Models of similar size were also fine-tuned for small structured languages to determine whether a statement follows from a theory, but it was observed that the performance of such models deteriorated rapidly when the number of deductive steps ("depth") required to arrive at the answer in test examples was higher than what had been seen in training (Clark et al., 2020). Instances of models twice as large were successfully fine-tuned for proof generation in structured language, though it should be noted that when asked to generate the proof in a single run, the models suffered from the same limitation on the question depth (Tafjord et al., 2020). The largest of language models remain largely untested in this regard, though it has been shown that they perform rather poorly even on elementary reasoning and arithmetic tasks unless prompted through chains of thought or by explicitly providing the relevant algorithm (Wei et al., 2022; Zhou et al., 2022). To answer Q1, we comprehensively evaluate 16 models in total, ranging from 117 million to 175 billion parameters in size.

Q2 asks about the influence of the reasoning depth required on model DRA. As already mentioned, previous work (Clark et al., 2020; Tafjord et al., 2020) has shown that even the deductive performance of models *fine-tuned* in the context of deductive reasoning considerably decreases once the depth required for inference climbs above the maximum depth seen in training. We test all 16 models considered for Q1 for depths 0, 1, 2, and 3, paying attention to any changes in scores.

Q3 leverages the experiments conducted for Q1-2 in an attempt to see whether there might be other factors influencing DRAs of foundation language models. While large language models tend to be trained on enormously large datasets, they vary in many details of their training, for example the proportions of different languages used, the use of code versus the pure use of the natural language, and the level to which edited sources (e.g. books, news) are emphasised over spontaneous language (e.g. chats, tweets). We inspect the results of all experiments for patterns aligned with the boundaries between different training setups.

In Section 3, we answer each of the questions Q1-3 in turn and comment on our further findings. We structured our study under the hypothesis that the model size would play a decisive role in model DRA, that the model performance would quickly decrease with depth, and that the role of the training setup would be negligible.

REPRODUCIBILITY STATEMENT. For full reproducibility, we make the datasets, runtime configurations, and all our code available at *anonymised*.

2 Experiments

2.1 Data

We use the closed-world variant of the RuleTakers dataset introduced in (Clark et al., 2020) and further expanded in (Tafjord et al., 2020). The dataset contains groups of 100,000 randomly generated theories for depths ranging from 0 to 5, written in structured language. A group of depth k contains yes/no questions that ask about statements (dis)provable in up to k deductive steps, generated at random but such that the dataset is balanced on average. The dataset further contains two special groups, NatLang and Birds-Electricity, each containing examples written freely in natural language and with separating punctuation where appropriate, but still abiding by the principles P1-4. Appendix A gives examples of theories written

both in structured and natural language.

2.2 Inputs

Inputs given to our models consist of the theory, leading examples (if considered), and the question that is to be answered. The three components are concatenated, with whitespace and punctuation inserted where appropriate. The leading examples are provided in the form "Question? Answer.", where question is as provided by the dataset and answer is either "Yes" or "No".

There are always either 0 or 3 leading examples provided. In the latter case, we give one leading example for each answer, determined at random, and a distinct example with either answer, also determined at random. This was done to make sure that the models (especially the ones below 3bn parameters) are not tempted by continuing a series of three "Yes" or three "No" answers instead of answering the question.

2.3 Evaluation

We ran the inferences of 11 open-source models on the test slice of the dataset locally, and further evaluated 4 recent OpenAI models via the OpenAI API. In either case, we provided the model with inputs as in Section 2.2 and asked it to make predictions for the next token.

2.3.1 Open-source models

The open-source models had between 117 million and 66 billion parameters and can be sorted into three families based on their training setup: OpenAI GPT-1/2 models, Bloom models, and Facebook OPT models. The smaller from among those models would answer our questions with "Yes", "No", but also with " True", "<line-break>False", and with varying capitalisation. Further, "Yes" and "No" were not always the most favoured answers, and the models sometimes exhibited a tendency to either continue the theory by inventing more facts, or continue asking questions without answering the question they were asked. To evaluate the models consistently, we therefore inspected the predicted probability distribution of the first token and considered the model to have answered "Yes" if it answered "Yes" or "True" irrespective of capitalisation, and similarly considered it to have answered "No" if it answered with "No" or "False".

We also experimented with equipping some of the $\leq$ 1bn-parameter models with a linear classification head on the embeddings of the final to-

ken as in (Radford et al., 2018) and training it to perform yes/no classification with the base either completely model frozen or frozen except for the final transformer layer. We did not observe any improvements over the results of the evaluation as above. Due to us wishing to evaluate the original version of the model and because of the limits on our computational resources, we did not attempt to fine-tune entire networks.

2.3.2 OpenAI GPT-3 and GPT-3.5 models

Since the OpenAI API does not provide access to the probabilities of the tokens predicted, we configured the models with temperature 1 and asked them to predict a single token. In all instances where leading examples were provided, the models were consistently outputting either "Yes" or "No" with some variance in capitalisation. In instances where no leading examples were given, all models but the GPT-3.5 "da Vinci" tended to behave similarly to $\leq$ 3bn-parameter open-source models and either continued the theory or continued generating questions instead of answering. Due to the lack of information about other predicted tokens we did not perform the full set of experiments with no leading examples.

3 Results

Table 1 lists the results of our experimentation on the open-source models, and Figure 1 shows the performance of the OpenAI models available only through the OpenAI API. Appendices B-C give the detailed results for natural language datasets and OpenAI GPT-3 and GPT-3.5 models.

We hypothesized that an increase in model size would correspond to an increase in the model's DRA. Table 1 shows that this is often the case within respective model families except OpenAI GPT-1 and 2 models but not across the board – for example, OPT-66b outperforms OPT-30b but underperforms Bloom-7b.

Q1: How does the deductive reasoning ability (DRA) of large language models evolve with their growing size? We find that with the training setup and model architecture kept constant, larger model size leads to higher DRA, but also that this effect can often be compensated by choices made in the training configuration. Previous work has shown that the DRA of models fine-tuned for deduction decreases on data that requires longer chains of reasoning that have been seen in training. Table 1 shows that this does not seem to be the

Data	LEs				Model							
		GPT-1	GPT-2 small	GPT-2 medium	GPT-2 large	GPT-2 XL	Bloom 560m	Bloom 1b1	Bloom 3b	Bloom 7b1	OPT 30b	OPT 66b
		117mn	124mn	355mn	774mn	1.5bn	560m	1.1bn	3bn	7.1bn	30bn	66bn
Depth 0	none	+0.08	**+2.15**	+1.89	+1.33	-1.20	**+6.76**	+1.05	+3.13	+5.64	+1.68	**+4.20**
	YNR	**-0.05**	-24.32	-19.79	-11.14	-14.77	-0.98	-5.12	**+3.97**	-0.43	-11.47	**-7.14**
Depth 1	none	-0.09	**+0.83**	+0.13	+0.35	-1.24	**+3.57**	+1.11	+1.19	+3.12	+0.39	**+3.11**
	YNR	**-0.01**	-7.52	-5.79	-2.35	-3.50	+0.07	-0.14	**+0.94**	+0.29	-3.00	**-2.54**
Depth 2	none	-0.03	**+2.59**	+2.22	+0.69	-1.17	**+5.95**	+1.78	+3.57	+4.56	+0.78	**+2.68**
	YNR	-0.04	-3.58	-1.80	-0.57	**+0.44**	+1.20	+1.83	**+2.29**	+1.29	+0.80	**+2.68**
Depth 3	none	+0.09	+3.43	**+3.43**	+0.55	-0.86	**+6.57**	+2.32	+4.22	+5.24	+1.18	**+4.88**
	YNR	+0.01	-1.99	-0.08	+1.53	**+2.12**	+1.36	+2.53	**+2.72**	+1.41	+2.06	**+4.65**

Table 1: Results of our evaluation of open-source models. The models are listed by family and then given in the order of their growing size. "LEs" denotes leading examples, where "none" means no leading examples were provided and "YNR" means that examples of "Yes", "No", and a random question were given. The metric used is binary accuracy displayed in percentage points relative to the baseline coin-flip performance of 50%, measuring the proportion of answers that were correct. Emphasis and **emphasis** mark the best performances overall and for the model family.
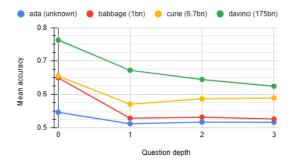

Figure 1: The performance of OpenAI GPT-3 and GPT-3.5 models plotted against the increasing length of questions in the dataset. The detailed results are in Appendix C.

case for foundation models as we evaluated them, that the performance of many models on deeper questions is similar to that on shallower questions, and that it sometimes even increases, especially for OpenAI GPT-1/2 models.

Q2: How does the DRA of general-purpose generative large language models depend the number of deductive steps required to reach the conclusions? For depths considered in our study, it remains largely similar throughout, except for the OpenAI GPT-3 and 3.5 models, for whom the performance decreases as seen in Figure 1.

Q3: Do the specifics of a training setup, such

as the model variant, dataset multilingualism, or dataset size have a decisive influence on the DRA? Yes. As noted for Q1, the differences between model training setups in many cases play a larger role than the size of the model. An example outlier is Bloom-560m, which comfortably outperforms all larger Bloom models, and Bloom setting as a whole, which in turn outperforms much larger OPT and GPT-2 XL models. This is perhaps due to the multi-lingual training setup (including multiple programming languages) of Bloom models, which is in contrast to natural language English used for GPT-1/2 and OPT models.

In summary, we conclude that all our initial hypotheses were partially incorrect, and that the subject of deductive reasoning ability in large language models warrants further examination from a wider set of angles.

4 Limitations

Our work only considers the token probabilities directly as provided by the generative models. Since the models are trained with generation of patches of text in mind, it can be argued that the model might in fact be trying to output a verbose answer. We attempt to to mitigate this by priming the model, but higher accuracy could perhaps be achieved by fine-tuning the language modelling head.

References

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. *arXiv preprint arXiv:2002.05867*.

Shaobo Li, Xiaoguang Li, Lifeng Shang, Zhenhua Dong, Chengjie Sun, Bingquan Liu, Zhenzhou Ji, Xin Jiang, and Qun Liu. 2022. How pre-trained language models capture factual knowledge? a causal-inspired analysis. *arXiv preprint arXiv:2203.16747*.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13480–13488.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-Lin Wu, Xuezhe Ma, and Nanyun Peng. 2021. Com2sense: A commonsense reasoning benchmark with complementary sentences. *arXiv preprint arXiv:2106.00969*.

Zijun Sun, Chun Fan, Qinghong Han, Xiaofei Sun, Yuxian Meng, Fei Wu, and Jiwei Li. 2020. Self-explaining structures improve nlp models. *arXiv preprint arXiv:2012.01786*.

Oyvind Tafjord, Bhavana Dalvi Mishra, and Peter Clark. 2020. Proofwriter: Generating implications, proofs, and abductive statements over natural language. *arXiv preprint arXiv:2012.13048*.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *27th International Conference on Intelligent User Interfaces*, pages 841–852.

Hattie Zhou, Azade Nova, Hugo Larochelle, Aaron Courville, Behnam Neyshabur, and Hanie Sedghi. 2022. Teaching algorithmic reasoning via in-context learning. *arXiv preprint arXiv:2211.09066*.

A Input Examples

```
Erin is kind. Erin is quiet. Erin is round. Fiona is green. Gary is big.
Gary is green. Harry is quiet. If Harry is round then Harry is young.
If something is kind then it is big. If something is green and not round
then it is white. Quiet things are green. If Fiona is young and Fiona is
not round then Fiona is quiet. If something is white then it is young.
```

Figure 2: An example of a theory of the RuleTakers dataset, written in structured language. Whether Erin is round would be a question of depth 0, whether Erin is big would be a question of depth 1, and whether Harry is young would be a question of depth 3.

```
For being so cold, it's good Alan can remain nice. Charlie might be rough
and red but he's actually very kind. Dave can be rough and cold, but he is
also green, an avid gardener. Fred is a round and rough around the edges,
and he is also big. People who have green body paint and act kind to others
are quite young. A nice, green, big person is also sure to be a red person.
Cold people that are big and red are usually nice. Someone with rough and
green feet is invariably kind. Tom is a rough, young person to know, he is
very green at his job but he is very round from his weight.
```

Figure 3: An example of a theory from the NatLang part of the RuleTakers dataset, paraphrased into natural language. Here, "Is Dave green?" is a question of depth 0, "Is Dave kind?" is a question of depth 1, and 'Is Dave young?" is a question of depth 3.

```
Erin is kind. Erin is quiet. Erin is round. Fiona is green. Gary is big.
Gary is green. Harry is quiet. ... Erin is round? Yes. Harry is round? No.
Fiona is green? Yes. Fiona is quiet?
```

Figure 4: An instance of input based on the theory from Figure 2, supplemented with three leading examples of depth 0 and the question to be answered by the model.

B Detailed Results of the Evaluation on Natural Language Datasets

Data	LEs	GPT-1	GPT-2 small	GPT-2 medium	GPT-2 large	GPT-2 XL	Bloom 560m	Bloom 1b1	Bloom 3b	Bloom 7b1	OPT 30b	OPT 66b
		117mn	*124mn*	*355mn*	*774mn*	*1.5bn*	*560m*	*1.1bn*	*3bn*	*7.1bn*	*30bn*	*66bn*
NatLang	none	-0.49	**+0.84**	-0.41	+0.50	-0.52	0.34	-0.60	-0.27	**0.55**	-0.71	*N/A*
	YNR	**0.00**	-3.37	-1.89	-2.51	-1.97	-0.51	**-0.50**	-0.87	-1.29	-3.06	*N/A*
Birds-Elec.	none	-0.27	0.00	**+6.36**	-12.01	-0.04	-5.52	-6.41	**+1.19**	-7.42	-7.77	**+2.58**
	YNR	**0.00**	-2.49	-6.81	-6.53	-1.04	-7.59	-9.37	-4.72	**1.46**	**-2.72**	-6.80

Table 2: Results of our evaluation of open-source models on natural language datasets. The models are listed by family and then given in the order of their growing size. "LEs" denotes leading examples, where "none" means no leading examples were provided and "YNR" means that examples of "Yes", "No", and a random question were given. The metric used is binary accuracy displayed in percentage points relative to the baseline coin-flip performance of 50%, measuring the proportion of answers that were correct. Emphasis and **emphasis** mark the best performances overall and for the model family.

C   Detailed Results of the Evaluation of OpenAI GPT-3 and 3.5 models

Data	LEs	Model			
		GPT-3 Ada	GPT-3 Babbage	GPT-3 Curie	GPT-3.5 da Vinci
		unknown	*1bn*	*6.7bn*	*175bn*
Depth 0	YNR	+4.67	+14.96	+15.49	**+26.28**
Depth 1	YNR	+1.23	+2.88	+7.09	**+17.23**
Depth 2	YNR	+1.72	+3.20	+8.68	**+14.48**
Depth 3	YNR	+1.65	+2.66	+8.93	**+12.44**

Table 3: Results of our evaluation of OpenAI GPT-3 and GPT-3.5 models on structured langauge datasets. These were also the results used in Figure 1. The models are listed in the order of their growing size. "LEs" denotes leading examples, "YNR" means that examples of "Yes", "No", and a random question were given. As explained in Section 2.3, we did not comprehensively evaluate the models for their performance with no leading examples due to the limited information available through the OpenAI API. The metric used is binary accuracy displayed in percentage points relative to the baseline coin-flip performance of 50%, measuring the proportion of answers that were correct. Emphasis and **emphasis** mark the best performances overall and for the model family.