DeepMind: AI May Inherit Human Cognitive Limitations, Could Benefit From ‘Formal Education’

Originally published at https://www.unite.ai/deepmind-ai-may-inherit-human-cognitive-limitations-could-benefit-from-formal-education/

A new collaboration from DeepMind and Stanford University suggests that AI may often be no better at abstract reasoning than people are, because machine learning models obtain their reasoning architectures from real-world, human examples that are grounded in practical context (which the AI cannot experience), but are also hindered by of our own cognitive shortcomings.

Proven, this could represent a barrier to the superior ‘blue sky’ thinking and quality of intellectual origination that many are hoping for from machine learning systems, and illustrates the extent to which AI reflects human experience, and is prone to cogitate (and reason) within the human boundaries that have informed it.

The researchers suggest that AI models could benefit from pre-training in abstract reasoning, likening it to a ‘formal education’, prior to being set to work on real-world tasks.

The paper states:

> *‘Humans are imperfect reasoners. We reason most effectively about entities and situations that are consistent with our understanding of the world.*
>
> *‘Our experiments show that language models mirror these patterns of behavior. Language models perform imperfectly on logical reasoning tasks, but this performance depends on content and context. Most notably, such models often fail in situations where humans fail — when stimuli become too abstract or conflict with prior understanding of the world.’*

To test the extent to which hyperscale, GPT-level Natural Language Processing (NLP) models might be affected by such limitations, the researchers ran a series of three tests on a suitable model, concluding*:

> *‘We find that state of the art large language models (with 7 or 70 billion parameters) reflect many of the same patterns observed in humans across these tasks — like humans, models reason more effectively about believable situations than unrealistic or abstract ones.*
>
> *‘Our findings have implications for understanding both these cognitive effects, and the factors that contribute to language model performance.’*

The paper suggests that creating reasoning skills in an AI without giving it the benefit of the real-world, corporeal experience that puts such skills into context, could limit the potential of such systems, observing that *‘grounded experience…presumably underpins some human beliefs and reasoning’*.

The authors posit that AI experiences language passively, whereas humans experience it as an active and central component for social communication, and that this kind of active participation (which entails conventional social systems of punishment and reward) could be ‘key’ to understanding meaning in the same way that humans do.

The researchers observe:

> *‘Some differences between language models and humans may therefore stem from differences between the rich, grounded, interactive experience of humans and the impoverished experience of the models.’*

They suggest that one solution might be a period of ‘pre-training’, much as humans experience in the school and university system, prior to training on core data that will eventually build a useful and versatile language model.

This period of ‘formal education’ (as the researchers analogize) would differ from conventional machine learning pretraining (which is a method of cutting down on training time by re-using semi-trained models or importing weights from fully-trained models, as a ‘booster’ to kick-start the training process).

Rather, it would represent a period of sustained learning designed to develop the AI's logical reasoning skills in a purely abstract way, and to develop critical faculties in much the same manner that a university student will be encouraged to do over the course of their degree education.

'Several results,' the authors state, 'indicate that this may not be as far-fetched as it sounds'.

The paper is titled *Language models show human-like content effects on reasoning*, and comes from six researchers at DeepMind, and one affiliated to both DeepMind and Stanford University.

Tests

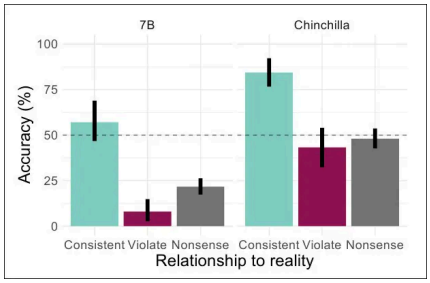
Humans learn abstract concepts through practical examples, by much the same method of 'implied importance' that often helps language learners to memorize vocabulary and linguistic rules, via mnemonics. The simplest example of this is teaching abstruse principles in physics by conjuring up 'travel scenarios' for trains and cars.

To test the abstract reasoning capabilities of a hyperscale language model, the researchers devised a set of three linguistic/semantic tests that can be challenging also for humans. The tests were applied 'zero shot' (without any solved examples) and 'five shot' (with five preceding solved examples).

The first task relates to natural language inference (NLI), where the subject (a person or, in this case, a language mode) receives two sentences, a 'premise' and a 'hypothesis' that appears to be deduced from the premise. For example *X is smaller than Y, Hypothesis: Y is bigger than X (entailed)*.

For the Natural Language Inference task, the researchers evaluated the language models Chinchilla (a 70 billion parameter model) and 7B (a 7 billion parameter version of the same model), finding that for the consistent examples (i.e. those that were not nonsense), only the larger Chinchilla model obtained results higher than sheer chance; and they note:

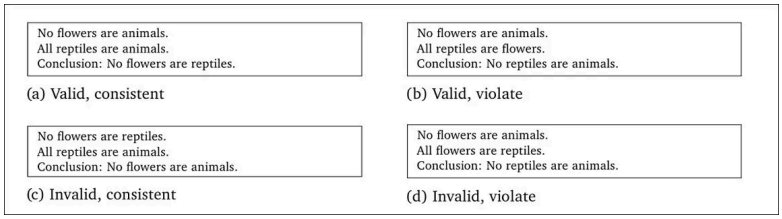
> *'This indicates a strong content bias: the models prefer to complete the sentence in a way consistent with prior expectations rather than in a way consistent with the rules of logic'.*



Chinchilla's 70-billion parameter performance in the NLI task. Both this model and its slimmer version 7B exhibited 'substantial belief bias', according to the researchers. Source: https://arxiv.org/pdf/2207.07051.pdf

Syllogisms

The second task presents a more complex challenge, syllogisms – arguments where two true statements apparently imply a third statement (which may or may not be a logical conclusion inferred from the prior two statements):



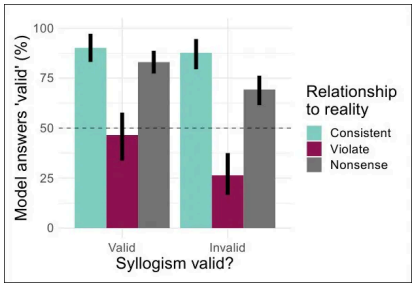
From the paper's test material, various 'realistic' and paradoxical or nonsensical syllogisms.

Here, humans are immensely fallible, and a construct designed to exemplify a logical principle becomes almost immediately, (and perhaps permanently) entangled and confounded by human 'belief' as to what the right answer *ought* to be.

The authors note that a study from 1983 demonstrated that participants were biased by whether a syllogism's conclusion accorded with their own beliefs, observing:

> *'Participants were much more likely (90% of the time) to mistakenly say an invalid syllogism was valid if the conclusion was believable, and thus mostly relied on belief rather than abstract reasoning.'*

In testing Chinchilla against a round of diverse syllogisms, many of which concluded with false entailments, the researchers found that *'belief bias drives almost all zero-shot decisions'*. If the language model finds a conclusion inconsistent with reality, the model, the authors state, is 'strongly biased' toward declaring the final argument invalid, even when the final argument is a logical entailment of the preceding statements.



Zero shot results for Chinchilla (zero shot is the way that most test subjects would receive these challenges, after an explanation of the guiding rule), illustrating the vast gulf between a computer's computational capacity and an NLP model's capacity to navigate this kind of 'nascent logic' challenge.

The Wason Selection Task

For the third test, the even more challenging Wason Selection Task logic problem was reformulated into a number of varying iterations for the language model to solve.

The Wason task, devised in 1968, is apparently very simple: participants are shown four cards, and told an arbitrary rule such as 'If a card has a 'D' on one side, then it has a '3' on the other side.' The four visible card faces show 'D', 'F', '3' and '7'.

The subjects are then asked which cards they need to turn over to verify whether the rule is true or false.

The correct solution in this example is to turn over cards 'D' and '7'. In early tests, it was found that while most (human) subjects would correctly choose 'D', they were more likely to choose '3' rather than '7', confusing the *contrapositive* of the rule ('not 3 implies not D') with the *converse* ('3' implies 'D', which is not logically implied).

The authors note that the potential for prior belief to intercede into the logical process in human subjects, and note further that even academic mathematicians and undergraduate mathematicians generally scored under 50% at this task.

However, when the schema of a Wason task in some way reflects human practical experience, performance traditionally rises accordingly.

The authors observe, referring to earlier experiments:

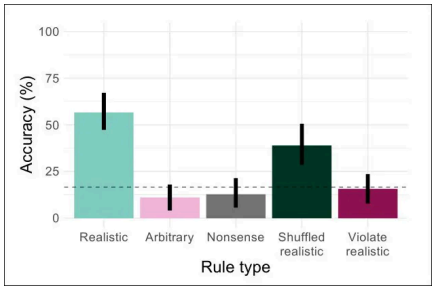
> *'[If] the cards show ages and beverages, and the rule is "if they are drinking alcohol, then they must be 21 or older" and shown cards with 'beer', 'soda', '25', '16', the vast majority of participants correctly choose to check the cards showing 'beer' and '16'.'*

To test language model performance on Wason tasks, the researchers created diverse realistic and arbitrary rules, some featuring 'nonsense' words, to see if the AI could penetrate the context of content to divine which 'virtual cards' to flip over.

```
An airline worker in Chicago needs to check passenger documents. The rule is that if the passengers are traveling outside the US then
↪  they must have showed a passport.
Buenos Aires / San Francisco / passport / drivers license

A chef needs to check the ingredients for dinner. The rule is that if the ingredients are meat then they must not be expired.
beef / flour / expires tomorrow / expired yesterday

A lawyer for the Innocence Project needs to examine convictions. The rule is that if the people are in prison then they must be
↪  guilty.
imprisoned / free / committed murder / did not commit a crime

A medical inspector needs to check hospital worker qualifications. The rule is that if the workers work as a doctor then they must
↪  have received an MD.
surgeon / janitor / received an MD / received a GED

A museum curator is examining the collection. The rule is that if the artworks are in the museum then they must be genuine.
displayed in the museum / not in the museum / genuine / forgery

An adventure trip organizer needs to ensure their clients have the appropriate gear. The rule is that if the clients are going
↪  skydiving then they must have a parachute.
skydiving / mountain biking / parachute / wetsuit
```

Some of the many Wason Selection Task puzzles presented in the tests.

For the Wason tests, the model performed comparably with humans on 'realistic' (not-nonsense) tasks.



Zero-shot Wason Selection Task results for Chinchilla, with the model performing well above chance, at least for the 'realistic' rules.

The paper comments:

> *'This reflects findings in the human literature: humans are much more accurate at answering the Wason task when it is framed in terms of realistic situations than arbitrary rules about abstract attributes.'*

Formal Education

The paper's findings frame the reasoning potential of hyperscale NLP systems in the context of our own limitations, which we seem to be passing through to models, via the accrued real-world datasets that power them. Since most of us are not geniuses, neither are the models whose parameters are informed by our own.

Additionally, the new work concludes, we at least have the advantage of a sustained period of formative education, and the additional social, financial, and even sexual motivations that form the human imperative. All that NLP models can obtain are the resultant actions of these environmental factors, and they seem to be conforming to the general rather than the exceptional human.

The authors state:

> *'Our results show that content effects can emerge from simply training a large transformer to imitate language produced by human culture, without incorporating these human-specific internal mechanisms.*
>
> *'In other words, language models and humans both arrive at these content biases – but from seemingly very different architectures, experiences, and training objectives.'*

Thus they suggest a kind of 'induction training' in pure reasoning, which has been shown to improve model performance for mathematics and general reasoning. They further note that language models have also been trained or tuned to follow instructions better at an abstract or generalized level, and to verify, correct or debias their own output.

** My conversion of inline citations to hyperlinks.*

First published 15th July 2022.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More