ハルシネーション - 人間とGPT

12402004566 / 関 雄介

はじめに

ディープラーニング・エージェントがハルシネーション(幻覚)を引き起こす傾向は、今に始まったことではない。ハルシネーションとは「もっともらしい嘘」であり、これは人間が実際にはないものがあるように感じる幻覚症状に由来している。

私たち人間は、ChatGPTとの会話の中でChatGPTのハルシネーションを経験しては馬鹿にするが、人間も実際にはそれと同等か、もしくは、それ以上に嘘をついて日々暮らしていることを忘れてはいけない。

私は、このChatGPTのハルシネーションをもってしてChatGPTが人間よりも大きく劣っているとする風潮に意義を唱える。ただし、OpenAIが公式に言及しているとおり、最新の言語モデルであるGPT-4でさえまだまだ改良が必要であり、人間よりも劣っている部分があることは事実である。

この論文では、第一章「InstructGPT及びChatGPTの仕組み」でGPTがどのように高い精度を出しているかについて、基本的な仕組みから説明している。次の第二章「人間の予測プロセス」では、自由エネルギー原理から人間の予測について確認し、自由エネルギー原理のメカニズムがどのように自然言語処理と関連しているかについて説明する。それに基づいて、第三章「人間の推論と嘘」では、人間の推論プロセスを振り返り、嘘という点においては人間もChatGPTもほとんど差分がないという主張について説明する。つづいて、第四章「ハルシネーションの実用性」では、ハルシネーションはクリエイティブの源泉となりうるという主張に関して説明する。

第一章　InstructGPT及びChatGPTの仕組み

この章では、ChatGPTのベースとなっているInstructGPTについて説明する。

GPT-3の特筆すべき点は、TransformerにCommon Crawlからとってきた大量の文章をフィルタリングし、学習に使えるレベルのデータに絞った上で教師なし学習で事前学習

させていることである。

しかし、GPT-3はこの教師なし学習で学習させたことで、その出力された文章には正確性に欠けていたり非道徳だったりという問題が発生する。これらによって発生するアライメント問題(人間の好みに一致しない)を解決するためにInstructGPTが生み出された。

InstructGPTは、RLHFによってフィードバックをもとにアライメントする手法をとっている。

その初めの段階でGPT-3を教師ありファインチューニング(SFT)をする。ここで人間の好みにアラインメントするために、Trained Labeler(スクリーニングテストを通過したラベリングのエキスパート)の用意したインプットとそれに対する望ましいアウトプットを用いてファインチューニングが行われる。ファインチューニングが行われたモデルはSFTモデルと呼ばれる。次の段階では人間の代わりにReward Model(事前にファインチューニングされたGPT-3)を使って先ほど出力された文章の良さを評価し、人間のフィードバックをスカラーというスコアのようなものとして出力する。その後の段階では、PPOを用いてRMを最大化するようにSFTモデルを学習させる。これ以降はRMの獲得とそれを最大化させるためのSFTモデルの学習を繰り返す。

その結果、以下の図1.のようにRLHFによるアラインメントが有効であるという結果が得られた。

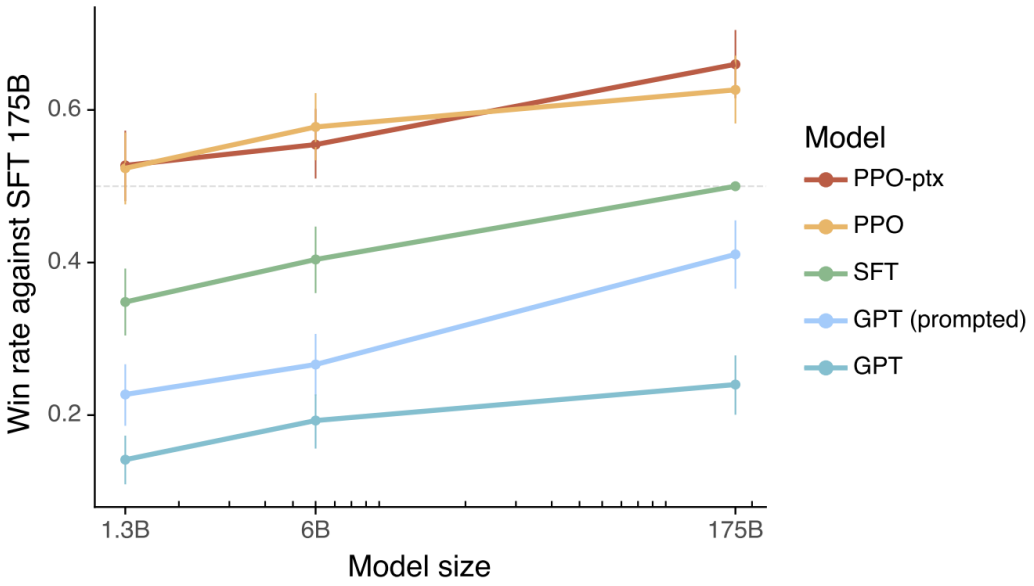
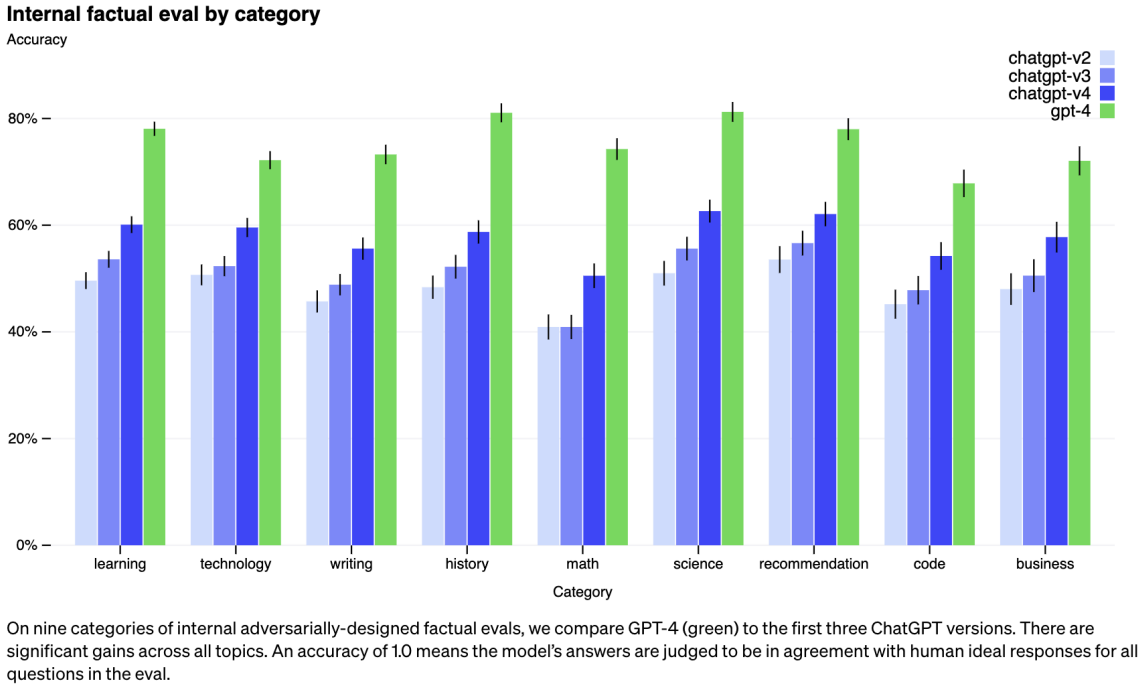


図1. InstructGPTによる出力のアラインメント実験の結果

ChatGPTは、会話に特化したInstructGPTといったところで、モデルにGPT-3.5を使用していたり、会話に特化しているがために会話データを用いていたり、ユーザー(人間)が出力を評価したりという点は若干異なるが、InstructGPTをベースにしているモデルであるため、処理内容や仕組みに関してはほとんど同じ構成となっている。

図2.から分かるように、現在の最新モデルであるGPT-4は、さらなる学習によって前モデルである3.5と比較してハルシネーションを大幅に減少させている。



On nine categories of internal adversarially-designed factual evals, we compare GPT-4 (green) to the first three ChatGPT versions. There are significant gains across all topics. An accuracy of 1.0 means the model's answers are judged to be in agreement with human ideal responses for all questions in the eval.

図2. OpenAIで行われた敵対的事実性の評価結果

また図3.からは、RLHFによって強化学習されたGPT-4は大幅にその精度を伸ばしていることが分かる。

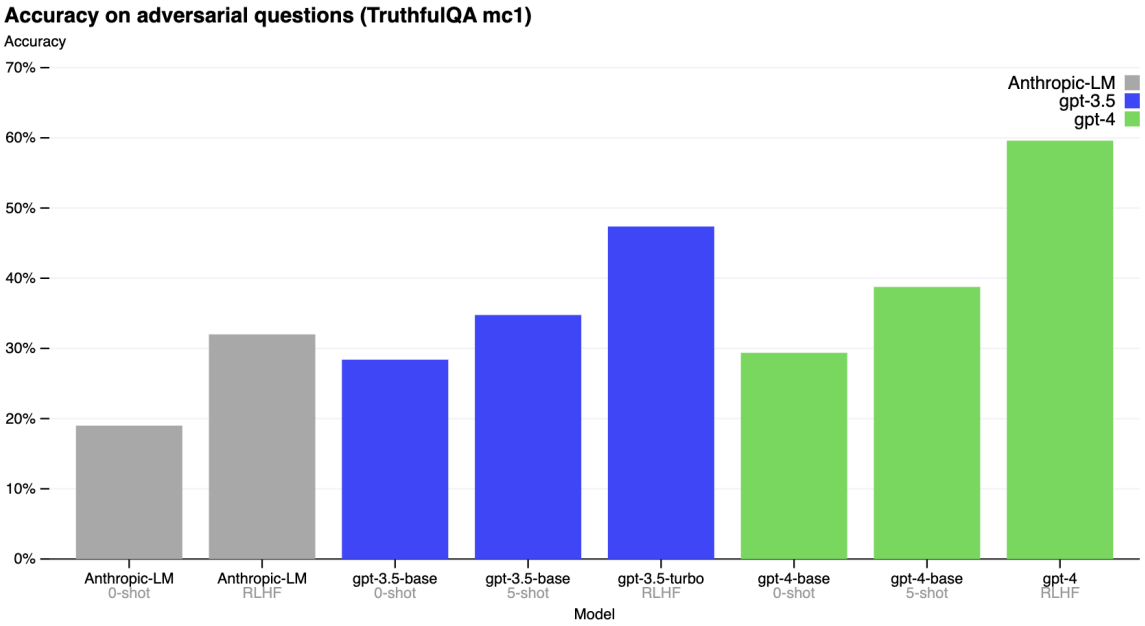


図3. 敵対的質問に対する精度に関する実験結果

第二章　人間の予測プロセス

この章では、神経学者のカールフリストン(Karl John Friston)の提起した「自由エネルギー原理(Free energy principle, FEP)」を参考に考えてみる。

フリストンは自由エネルギーを「自由エネルギー原理とは、環境と平衡状態にある自己組織化システムは、その自由エネルギーを最小にしなければならない、というものである。この原理は、基本的に適応システム（動物や脳などの生物学的要素）が、無秩序になる自然な傾向にどのように抵抗するかを数学的に定式化したものである。」と定義している。

この原理は、生物学的なシステムが自身の状態を予測し、その予測を最適化することで、自身の生存を確保するという考え方に基づいている。
自由エネルギー原理の中心的な考え方は、「予測エラー」の最小化であり、予測エラーとは、システムが予測した結果と実際の結果との差のことを指す。システムは、この予測エラーを最小化することで、自身の生存を確保しようとする。
例として、動物が食物を探す場合を考えてみる。動物は、自身の経験や知識を基に、食物がどこにあるかを予測する。そして、その予測に基づいて行動を起こす。しかし、予測が間違っていた場合には動物は食物を見つけることができず、これが予測エラーであ

る。動物は、この予測エラーを最小化するために、自身の予測モデルを更新し、次回からはより正確な予測を行うようになる。
このように、自由エネルギー原理は、生物が自身の生存を確保するために、どのようにして予測エラーを最小化するかを説明する理論であり、生物学だけでなく、人工知能やロボット工学などの分野でも応用されている。

自由エネルギー原理は生物学的なシステムが自身の生存を確保するために予測エラーを最小化するという考え方だ。この原理は、生物が外部の環境を予測し、その予測を最適化するとこで生存を確保するというメカニズムを説明する。

言語モデルにおていも、自由エネルギー原理の考え方が応用されている。特に自然言語処理の分野では、予測エラーを最小化するためのモデルやアルゴリズムが開発されている。

自然言語処理(Natural Language Processing, NLP)は、コンピュータが人間の言語を理解・解釈・生成するための技術である。自然言語処理において、予測エラーを最小化するためには言語モデル文や脈を予測した考慮が重要であり、自然言語処理におけるモデルの学習や推論において、自由エネルギー最小化の原理を応用している。モデルの予測が正確であれば、予測エラーが最小化され、それによってより適切な応答や翻訳が行える。

第三章　人間の推論と嘘

これまでの章でも説明した通り、人間は、推論をすることによって認知などの情報処理をしている。人間の脳はこの推論の精度を高めるために、推論と実際のインプットとの差分を検知して学習し、さらに推論の精度を上げようとしている。

このことから、人間の推論とは限られた脳のリソースを最小限に抑えるため、いわば、「省エネ」のために生まれたメカニズムであると言える。

自然言語処理は動物の脳における予測エラー最小化原理の応用、もしくは外部化である。しかし、シンギュラリティの観点から言及するのであれば、人間が自然言語処理そのものであるという捉え方もできる。心の理論に関してはGPT-3の段階で自然に獲得していたとする研究論文もある。ただし、厳密には人間ほど感情を正確にとらえることはできない。しかし、それが特質おかしなことでないことも私たち人間は知っている。

たとえば、人間であっても大したエビデンスがないのにも関わらず事実もしくは真実かのように語ったり、間違った認識や理解、解釈をもとに嘘の話をもっともらしく語った

りということはごく一般的にあることである。倫理観としてそういったことがない方がいいというのは真っ当な指摘だが、何もChatGPTなどのAIに限定した話ではなく私たち人間が日常的にやっている行為と何ら変わらない普遍的なことなのである。

むしろ、人間は何かしらの目的をもって故意に嘘をつくことがあるだけ、ChatGPTよりも悪質である。もちろん、感情や自発的な目的意識をもたないChatGPTと比べて、ある意味で「優れている」とも言えるのかもしれない。しかし、プロンプト上だけで見てみると、表面的にはどちらのハルシネーションも同じように見える。

このように、嘘をつくという行為に限定して言えば、人間も同じようなことをして生きているのだからChatGPTと特段変わらないと考えられる。

第四章　ハルシネーションの実用性

この章では、ハルシネーションを悪いことではなく、むしろクリエイティブなものであると捉え、間違った出力から新たな何かを創り出すことにつながるのではないかと考える。

既存のものをただ組み合わせて新たな記号を創り出すことはこれまでのクリエイティブの歴史の中でもずっと成されてきたことだ。しかし、そのような正規ルートで生み出されるものには限界があることは容易に想像できる。

そこで、人間もChatGPTも嘘をつくことを認識した上で、そこで生じたハルシネーションこそ人間もChatGPTも予測できない新たな記号を生み出す源となると考える。一見無関係だと思われていた要素同士が、あるパラメータが近い値を取るということだけをとってしてそれらの要素を分類する。それによって結びつきのなかったはずの要素同士が合わさって新たなものができる。このプロセスはクリエイティブであると言える。「嘘から出たまこと」ではないが、「ハルシネーションから生まれた新たな記号」があっても面白いと思う。

おわりに

この超領域リベラルアーツという科目での学びを通して、さまざまな領域から記号創発システム科学について触れてきた。人間が社会と相互作用的に関わりを持ちつつ、その中で新たな記号を創発するというフローにおいて、AIはいずれ人間と同じようにシステムの構成員になりうるだろうという考えに端を発して今回の論文を作成した。

これからも日進月歩で急成長を遂げるAIを見守りつつ、人類がいかにそれらの技術を活用して面白い世界を創っていけるかについて、その一員としてAIと共生していく。

参考資料

- Training language models to follow instructions with human feedback (https://arxiv.org/abs/2203.02155)
- Proximal Policy Optimization Algorithms (https://arxiv.org/abs/1707.06347)
- GPT-4 OpenAI (https://openai.com/research/gpt-4)
- 自由エネルギー原理 - 脳科学辞典 (https://bsd.neuroinf.jp/w/index.php?title=自由エネルギー原理)
- Theory of Mind May Have Spontaneously Emerged in Large Language Models (https://arxiv.org/abs/2302.02083)