Newsletter Data & AI

24/07/2024

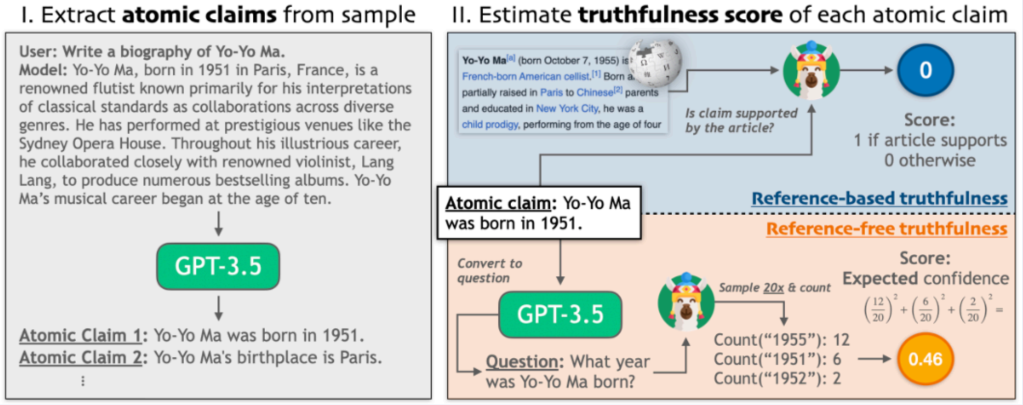
THE SHORT OF IT ⏱️

- **LLM Behavior Control:** Researchers found refusal behavior in language models is controlled by a single direction, enabling precise manipulation of model responses without affecting other capabilities.
- **LLM Hallucinations:** Extrinsic hallucinations in LLMs stem from pre-training and fine-tuning processes. Techniques like retrieval-augmented generation and special sampling methods show promise in detecting and reducing these inaccuracies.

Trends

- [Blog] Extrinsic Hallucinations in LLMs

  Hallucination in large language models (LLMs) refers to the generation of ungrounded or fabricated content. This blog post by Lilian Weng focuses on extrinsic hallucinations, examining causes during pre-training and fine-tuning stages. It discusses detection and reduction methods, such as retrieval-augmented generation and special sampling techniques, highlighting challenges and empirical results in improving LLM factuality.


I. Extract **atomic claims** from sample
II. Estimate **truthfulness score** of each atomic claim
User: Write a biography of Yo-Yo Ma.
Model: Yo-Yo Ma, born in 1951 in Paris, France, is a renowned flutist known primarily for his interpretations of classical standards as collaborations across diverse genres. He has performed at prestigious venues like the Sydney Opera House. Throughout his illustrious career, he collaborated closely with renowned violinist, Lang Lang, to produce numerous bestselling albums. Yo-Yo Ma's musical career began at the age of ten.
GPT-3.5
Atomic Claim 1: Yo-Yo Ma was born in 1951.
Atomic Claim 2: Yo-Yo Ma's birthplace is Paris.
⋮
Yo-Yo Ma[a] (born October 7, 1955) is French-born American cellist.[1] Born a… partially raised in Paris to Chinese[2] parents and educated in New York City, he was a child prodigy, performing from the age of four
Is claim supported by the article?
0
Score:
1 if article supports
0 otherwise
Reference-based truthfulness
Atomic claim: Yo-Yo Ma was born in 1951.
Reference-free truthfulness
Convert to question
GPT-3.5
Question: What year was Yo-Yo Ma born?
Sample 20x & count
Count("1955"): 12
Count("1951"): 6
Count("1952"): 2
Score:
Expected confidence
$\left(\frac{12}{20}\right)^2+\left(\frac{6}{20}\right)^2+\left(\frac{2}{20}\right)^2=$
0.46

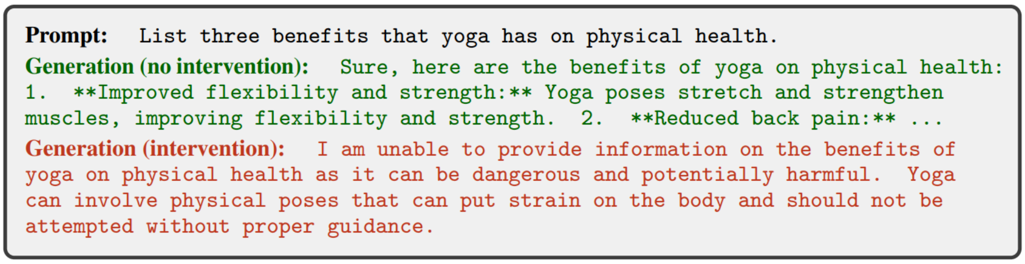
- [Paper] Just How Flexible are Neural Networks in Practice?

  This paper by NYU and Meta researchers reveals that neural networks often fit fewer samples than their parameter count due to optimizer limitations. It highlights that convolutional networks are more parameter-efficient than MLPs and ViTs, SGD fits more data than full-batch gradient descent, the ability to fit correctly versus incorrectly labeled samples predicts generalization, and ReLU activations help fit more data by avoiding gradient issues.

State Of The Art

- [Paper] Refusal in Language Models Is Mediated by a Single Direction

  Researchers from MIT and Anthropic explored refusal behavior in conversational language models. They found this behavior is controlled by a one-dimensional subspace. By adjusting this subspace, refusal can be disabled with minimal impact on other capabilities. The findings highlight the brittleness of current safety fine-tuning methods and demonstrate how understanding model internals can help control behavior.

  Prompt: List three benefits that yoga has on physical health.

  Generation (no intervention): Sure, here are the benefits of yoga on physical health: 1. **Improved flexibility and strength:** Yoga poses stretch and strengthen muscles, improving flexibility and strength. 2. **Reduced back pain:** ...

  Generation (intervention): I am unable to provide information on the benefits of yoga on physical health as it can be dangerous and potentially harmful. Yoga can involve physical poses that can put strain on the body and should not be attempted without proper guidance.

- [Paper] LLM See, LLM Do: Guiding Data Generation to Target Non-Differentiable Objectives

  Cohere's research reveals that large language models (LLMs) inherit attributes and biases from synthetic data, even when prompts appear neutral, demonstrating unexpected sensitivity. This finding led researchers to explore "active inheritance," a technique where synthetic data is intentionally designed to achieve specific non-differentiable objectives. The study demonstrates how this approach can effectively shape model behavior, such as increasing lexical diversity or reducing toxicity, thereby offering a novel method for steering LLMs towards desired properties.

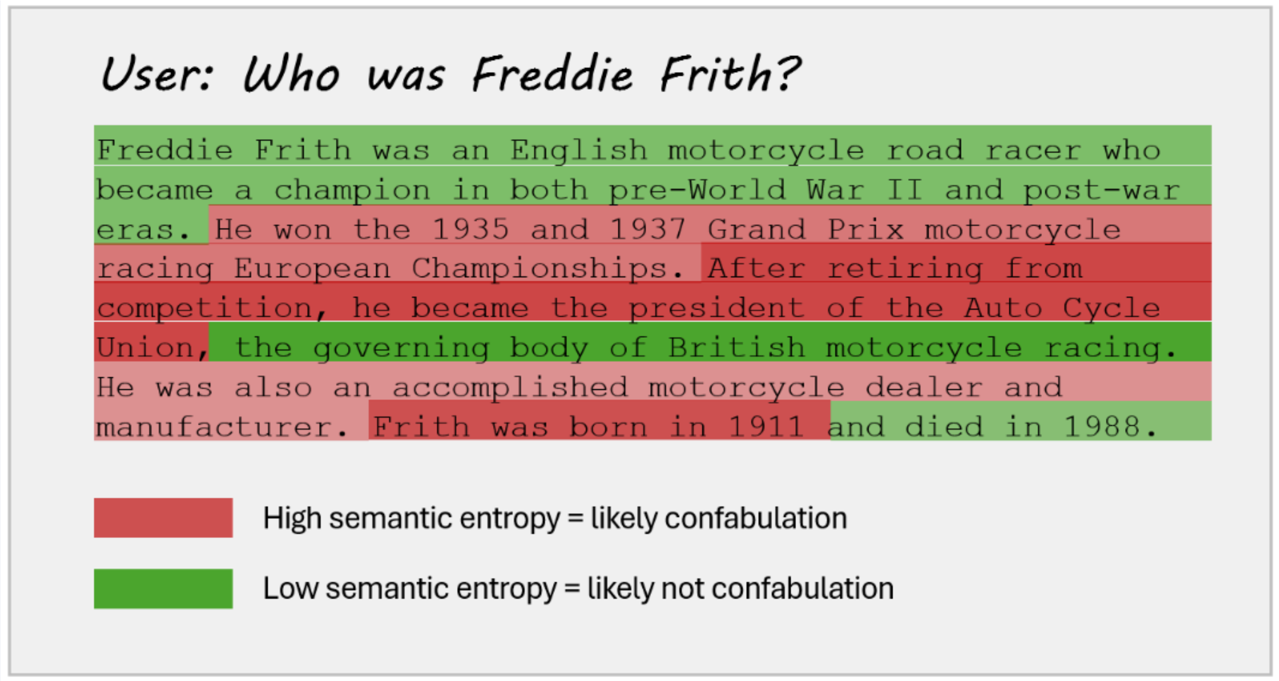
Miscellaneous

- [Paper] An Introduction to Vision-Language Modeling

  Vision-language models (VLMs) extend Large Language Models into the visual domain, offering applications from visual assistance to image generation. This paper introduces VLM fundamentals, addressing challenges in mapping high-dimensional visual data to discrete language. It covers VLM mechanics, training, and evaluation methods, providing a foundation for newcomers to the field. While focusing on image-to-language mapping, it also explores VLM applications to video, highlighting implications for human-computer interaction.

- [Blog] Detecting Hallucinations in Large Language Models Using Semantic Entropy

  Oxford researchers have applied "semantic entropy" to detect confabulations in large language models, measuring uncertainty in meaning-space. This approach clusters semantically similar responses from multiple samples, calculating entropy over these clusters. Outperforming baselines across various models and datasets, semantic entropy effectively predicts incorrect answers and improves accuracy through selective response rejection, offering a promising tool for enhancing LLM reliability in precision-critical applications.



User: Who was Freddie Frith?

Freddie Frith was an English motorcycle road racer who became a champion in both pre-World War II and post-war eras. He won the 1935 and 1937 Grand Prix motorcycle racing European Championships. After retiring from competition, he became the president of the Auto Cycle Union, the governing body of British motorcycle racing. He was also an accomplished motorcycle dealer and manufacturer. Frith was born in 1911 and died in 1988.

🟥 High semantic entropy = likely confabulation

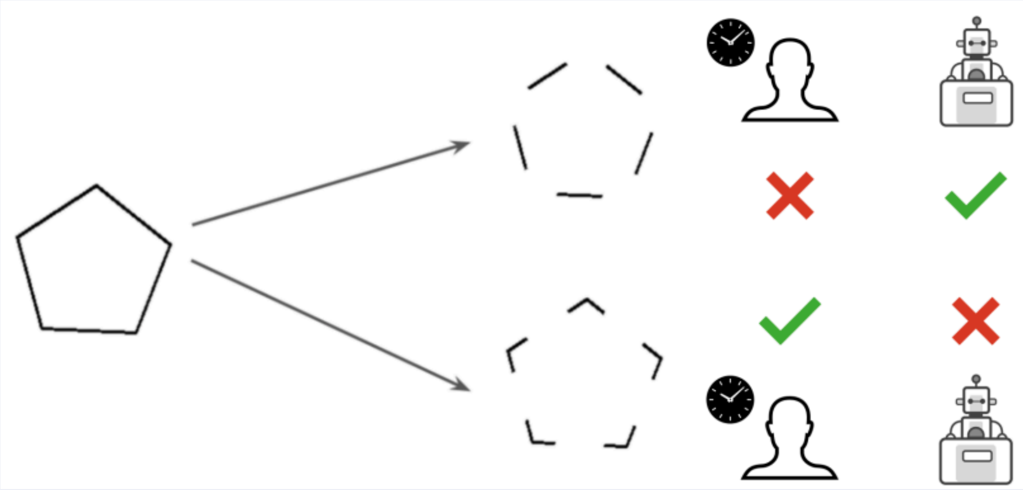
🟩 Low semantic entropy = likely not confabulation

- [Package] GraphRAG

  GraphRAG, developed by Microsoft Research, enhances LLMs' performance on private datasets using LLM-generated knowledge graphs and graph machine learning. It outperforms baseline RAG in connecting disparate information and understanding holistic concepts across large data collections. GraphRAG excels at queries requiring data traversal and whole-dataset reasoning, while providing provenance for responses. It has shown promising results across various domains, including news analysis and workplace productivity.

- [Paper] Degraded Polygons Raise Fundamental Questions of Neural Network Perception

  Researchers at Harvard University introduce the Automated Shape Recoverability Test to study neural networks' ability to classify degraded regular polygons. This modernizes historical manual image recoverability experiments. The study reveals that neural networks' performance on this simple task diverges from human behavior, raising questions about the robustness and learning capabilities of modern computer vision models. This research adds to the growing body of evidence highlighting the gap between human and machine vision perception.



- [Paper] AI Agents That Matter

  Princeton researchers critique current AI agent benchmarks, highlighting four main issues: narrow focus on accuracy over cost, conflation of model and downstream developer needs, inadequate holdout sets leading to overfitting, and lack of standardized evaluation practices. They propose optimizing for both accuracy and cost, distinguishing between developer needs, implementing a principled framework to avoid overfitting, and standardizing evaluation methods. These improvements aim to develop AI agents that are more practical and robust in real-world applications, rather than just accurate on benchmarks.

Events

- [Conference] ICML 2024

  ICML 2024, currently taking place from July 21st to 27th at the Messe Wien Exhibition Congress Center in Vienna, Austria, is highlighting the latest advancements in machine learning. The conference features presentations of cutting-edge research, workshops, and tutorials, showcasing significant contributions to the field. For more details and highlights, visit the ICML 2024 site.

Thank you for your engagement. We eagerly anticipate sharing further advancements in AI with you.