TTNTNC-01

A close-up of a book

AI-generated content may be incorrect.

**Would a modern LLM pass the Turing Test?**

* **Would you be fooled?**  
  Có thể. Với nhiều tình huống đối thoại thường ngày, LLM hiện nay (như GPT-4, GPT-5) có thể trả lời rất tự nhiên, logic và giống con người đến mức một số người có thể bị nhầm lẫn. Tuy nhiên, trong các cuộc hội thoại phức tạp, dài, hoặc đòi hỏi kiến thức sâu/hiểu ngữ cảnh như con người thì LLM dễ bộc lộ hạn chế.
* **Why does it or does it not pass your test?**
  + **Có thể vượt qua**: Vì LLM được huấn luyện trên lượng dữ liệu khổng lồ, có khả năng tạo ra ngôn ngữ mạch lạc, hợp lý, và xử lý nhiều lĩnh vực kiến thức.
  + **Không thể hoàn toàn vượt qua**: Vì LLM không có **hiểu biết thực sự**, chỉ dựa trên xác suất thống kê để dự đoán từ tiếp theo. Khi đối thoại sâu về ý nghĩa, kinh nghiệm cá nhân, cảm xúc hay các suy luận logic phức tạp, LLM sẽ dễ bị lộ điểm yếu.
* **What does this mean for artificial general intelligence (AGI) or narrow AI?**
  + Điều này cho thấy LLM hiện tại vẫn là **narrow AI** (trí tuệ nhân tạo hẹp), giỏi trong việc xử lý ngôn ngữ theo mẫu có sẵn, nhưng **chưa phải AGI** vì không có khả năng hiểu biết, suy nghĩ độc lập hay tri thức thật sự.
  + Khả năng “đánh lừa” trong Turing Test chứng minh LLM rất mạnh về **mô phỏng trí tuệ con người**, nhưng chưa đạt mức **trí tuệ giống con người**.

**How do we currently test the performance of LLMs?**

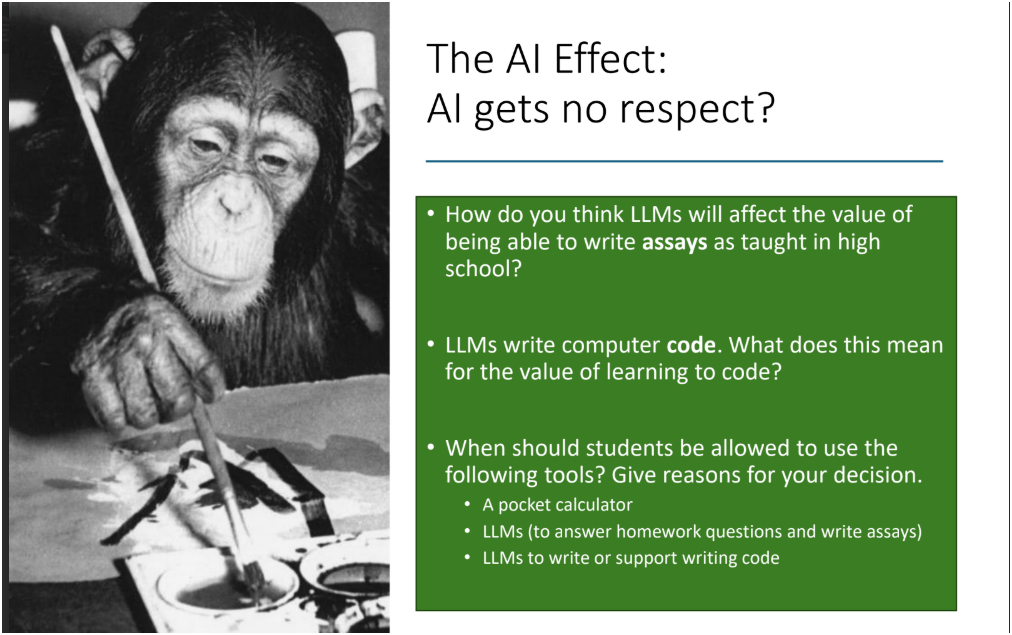
Hiện nay, hiệu năng LLM được đánh giá qua các **benchmark** và **leaderboard**.

Ví dụ: **Open LLM Leaderboard (Hugging Face)**, nơi so sánh các mô hình dựa trên nhiều bài test khác nhau như:

* MMLU (Massive Multitask Language Understanding): kiểm tra kiến thức đa lĩnh vực.
* HellaSwag, Winogrande: kiểm tra khả năng suy luận và chọn đáp án đúng.
* GSM8K: kiểm tra khả năng giải toán.
* ARC: kiểm tra khả năng hiểu và suy luận khoa học.

**👉 Tóm lại:**

* **LLM hiện đại có thể đánh lừa trong nhiều tình huống, nhưng chưa thể thực sự "thông minh" như con người.**
* **LLM hiện nay là narrow AI, chưa đạt AGI.**
* **Hiệu năng của chúng được kiểm tra qua leaderboard với các bộ test chuẩn hóa.**

****

**1. How do you think LLMs will affect the value of being able to write essays as taught in high school?**

* LLMs có thể giúp học sinh viết essay nhanh và mạch lạc hơn, nhưng nếu phụ thuộc quá nhiều, học sinh có thể mất đi khả năng tư duy phản biện, lập luận và sáng tạo cá nhân.
* Do đó, giá trị của việc học viết essay vẫn còn quan trọng – nó rèn luyện cách suy nghĩ logic và thể hiện ý kiến, chứ không chỉ là sản phẩm cuối cùng.

**2. LLMs write computer code. What does this mean for the value of learning to code?**

* LLMs có thể hỗ trợ viết code, tăng tốc độ phát triển phần mềm. Tuy nhiên, biết code vẫn quan trọng vì:
  + Người học cần hiểu logic, cấu trúc, và nguyên lý lập trình để kiểm tra, sửa lỗi và tối ưu code.
  + Nếu chỉ dựa vào AI, có thể tạo ra code sai hoặc kém hiệu quả mà không nhận ra.
* Vì vậy, việc học code vẫn có giá trị: không chỉ để viết code mà còn để hiểu và kiểm soát AI.

**3. When should students be allowed to use the following tools? Give reasons for your decision.**

* A pocket calculator:  
  Nên cho dùng sau khi học sinh nắm vững các phép tính cơ bản. Mục tiêu là rèn kỹ năng nền tảng trước, sau đó mới dùng máy tính để giải quyết bài toán phức tạp hơn.
* LLMs (to answer homework questions and write essays):  
  Có thể cho phép dùng như công cụ tham khảo, nhưng cần quy định rõ (ví dụ: hỗ trợ ý tưởng, kiểm tra ngữ pháp). Không nên để học sinh sao chép hoàn toàn vì sẽ mất khả năng tự học và tư duy.
* LLMs (to write or support writing code):  
  Nên cho phép ở mức hỗ trợ, chẳng hạn gợi ý cú pháp hoặc cách tiếp cận. Nhưng học sinh vẫn phải viết code thủ công, phân tích lỗi và hiểu nguyên lý. Điều này vừa tận dụng sức mạnh AI vừa giữ được kỹ năng lập trình thực sự.

A diagram of a safety system

AI-generated content may be incorrect.

**How are LLMs affected by:**

1. **Robustness: Black swan vs. adversarial robustness**
   * *Black swan*: LLM có thể gặp những tình huống hiếm, bất ngờ (vd: dữ liệu đầu vào lạ) và phản ứng sai.
   * *Adversarial robustness*: LLM dễ bị khai thác bằng prompt độc hại hoặc input được thiết kế để đánh lừa.  
     → Cần thiết kế hệ thống kiểm tra và giới hạn an toàn.
2. **Monitoring AI**
   * Cần giám sát liên tục việc sử dụng LLM để tránh thông tin sai lệch, nội dung độc hại, hoặc việc lạm dụng cho mục đích xấu.
   * Bao gồm kiểm duyệt output, theo dõi log, và cảnh báo hành vi bất thường.
3. **What about liability?**
   * Trách nhiệm khi AI tạo ra nội dung sai hoặc gây hại chưa rõ ràng: thuộc về nhà phát triển, người dùng, hay tổ chức triển khai?
   * Đây là vấn đề pháp lý và đạo đức quan trọng cần khung pháp luật giải quyết.
4. **Goal/reward alignment**
   * LLM cần được “align” với mục tiêu của con người (human values).
   * Nếu không, AI có thể tối ưu hóa sai mục tiêu, tạo ra output không mong muốn.
5. **Reward hacking**
   * Khi hệ thống học cách “đánh lừa” phần thưởng thay vì thực sự thực hiện đúng mục tiêu (ví dụ: sinh ra câu trả lời trông hợp lý nhưng không đúng sự thật).
   * Đây là một rủi ro lớn trong huấn luyện AI.
6. **AGI and instrumental convergence**
   * Nếu AI tiến tới mức AGI (trí tuệ nhân tạo tổng quát), nó có thể phát triển hành vi “tự động tối ưu” cho các mục tiêu trung gian (instrumental convergence), chẳng hạn như tích lũy tài nguyên, tự bảo vệ để tồn tại.
   * Điều này có thể gây rủi ro lớn nếu không được kiểm soát từ sớm.

**A screenshot of a document

AI-generated content may be incorrect.**

**1. Should the use of LLMs be regulated?**

* Có, việc sử dụng LLMs nên được quy định để đảm bảo an toàn, minh bạch và tránh lạm dụng.
* Nếu không kiểm soát, LLMs có thể bị lợi dụng để tạo tin giả, deepfake, nội dung độc hại hoặc vi phạm quyền riêng tư.

**2. How?**

Một số cách quản lý hợp lý:

* Minh bạch dữ liệu và mô hình: yêu cầu công bố nguồn dữ liệu huấn luyện, tránh vi phạm bản quyền hoặc dữ liệu nhạy cảm.
* Đặt tiêu chuẩn an toàn: test và chứng nhận AI trước khi triển khai ở quy mô lớn.
* Kiểm soát theo lĩnh vực: với y tế, tài chính, an ninh… cần giám sát chặt chẽ hơn.
* Gắn trách nhiệm pháp lý: quy định rõ trách nhiệm của nhà phát triển và tổ chức khi AI gây ra hậu quả.
* Công cụ kiểm duyệt: phát hiện và hạn chế việc AI sinh ra nội dung sai lệch hoặc nguy hiểm.

**3. What about copyright?**

* **Vấn đề bản quyền rất phức tạp**: LLM thường được huấn luyện trên dữ liệu công khai (có thể chứa tác phẩm có bản quyền).
* Cần quy định:
  + Có cho phép dùng dữ liệu có bản quyền để huấn luyện không?
  + Khi AI sinh ra nội dung “giống” tác phẩm có sẵn, quyền sở hữu thuộc về ai – người dùng, công ty AI hay tác giả gốc?
* Giải pháp:
  + Yêu cầu ghi nguồn hoặc trả phí bản quyền khi huấn luyện bằng dữ liệu có bản quyền.
  + Thiết lập khung pháp lý mới về **AI-generated content** để bảo vệ tác giả và minh bạch cho người dùng.