CHƯƠNG 1

Câu hỏi: **What are the LLMs’ Percepts? Actions? Objectives?**

Trả lời:

* **Percepts (nhận thức)**: văn bản đầu vào do người dùng nhập.
* **Actions (hành động)**: sinh ra văn bản phản hồi, câu trả lời hoặc đoạn văn bản mới.
* **Objectives (mục tiêu)**: dự đoán từ tiếp theo chính xác, cung cấp câu trả lời hữu ích dựa trên dữ liệu huấn luyện.

Câu hỏi: **What do LLMs do? Do LLMs act rationally?**

Trả lời:

* **What do LLMs do?** LLMs dự đoán từ tiếp theo để tạo văn bản mới dựa trên dữ liệu huấn luyện, có thể trả lời câu hỏi, viết đoạn văn, dịch…
* **Do LLMs act rationally?** Ở mức độ nào đó có, vì chúng phản hồi dựa trên quy tắc tối ưu xác suất để đạt mục tiêu, nhưng chúng không có “ý định” hay “nhận thức” thực sự, nên hành vi hợp lý chỉ giới hạn trong phạm vi dự đoán thống kê.

Câu hỏi: **Would a modern LLM pass the Turing Test?**

* **Would you be fooled?**
* **Why does it or does it not pass your test?**
* **What does this mean for AGI or Narrow AI?**

Trả lời:

* **Would you be fooled?** Có thể, với những câu hỏi quen thuộc, LLM có thể khiến người dùng tưởng đang nói chuyện với người thật.
* **Why?** Vì LLM có thể tạo câu trả lời tự nhiên, ngữ pháp tốt, giống người. Nhưng khi hỏi sâu hoặc yêu cầu kiến thức thực sự, nó lộ ra giới hạn (sai thực tế, thiếu lý luận).
* **Ý nghĩa:** Điều này cho thấy LLM hiện tại vẫn là Narrow AI (trí tuệ hẹp), chưa đạt mức AGI (trí tuệ tổng quát).

Câu hỏi:

* **How do you think LLMs will affect the value of being able to write assays as taught in high school? ( LLMs sẽ ảnh hưởng đến việc học viết luận ở trường trung học thế nào? )**
  + **LLMs write computer code. What does this mean for the value of learning to code? ( LLMs viết code, vậy còn giá trị học code? )**
* **When should students be allowed to use the following tools? Give reasons for your decision. ( Khi nào nên cho học sinh dùng LLMs? )**

Trả lời:

* **Ảnh hưởng viết luận:** LLM giúp viết nhanh hơn, nhưng có thể làm giảm khả năng tự tư duy và lập luận của học sinh nếu lạm dụng.
* **Viết code:** LLM giúp sinh code nhanh, nhưng hiểu nền tảng lập trình vẫn quan trọng để kiểm tra và chỉnh sửa code.
* **Khi nên cho dùng:** Khi học sinh đã hiểu cơ bản, LLM có thể hỗ trợ tra cứu, gợi ý. Không nên dùng để thay thế hoàn toàn việc học kỹ năng gốc.

Câu hỏi: **How are LLMs affected by: robustness, monitoring AI, liability, goal/reward alignment, reward hacking, AGI & instrumental convergence?**

Trả lời:

* **Robustness:** LLM cần chống chịu trước đầu vào gây nhiễu hoặc tấn công (prompt injection).
* **Monitoring AI:** cần hệ thống giám sát để phát hiện nội dung không phù hợp.
* **Liability:** phải có cơ chế xác định trách nhiệm khi AI gây hại.
* **Goal/reward alignment:** cần thiết kế mục tiêu phù hợp với giá trị con người.
* **Reward hacking:** LLM có thể “lách” mục tiêu nếu không được kiểm soát.
* **AGI & instrumental convergence:** nếu tiến tới AGI, các nguy cơ này càng cao, cần chuẩn bị giải pháp quản trị.

Câu hỏi: **Should the use of LLMs be regulated? How? What about copyright?**

Trả lời:

* **Regulated?** Có, để đảm bảo an toàn, bảo mật dữ liệu, ngăn chặn lạm dụng.
* **How?** Đặt tiêu chuẩn kiểm định nội dung, yêu cầu minh bạch dữ liệu huấn luyện, quy định trách nhiệm pháp lý cho nhà cung cấp.
* **Copyright:** cần khung pháp lý rõ ràng về việc sử dụng dữ liệu có bản quyền để huấn luyện và tạo nội dung mới, đồng thời bảo vệ quyền tác giả.