AI-02 · Bài 3

Những việc trợ lý AI làm sai mà nghe như đúng

Lớp Trợ lý AI giúp được việc gì, và không giúp được việc gì

m.school by MES LAB

Học xong bài này bạn sẽ

- gọi được tên năm kiểu sai hay gặp trong câu trả lời của một trợ lý AI, và chỉ ra chúng trên một câu trả lời thật.
- Nói được mỗi kiểu sai ấy sinh ra từ chỗ nào trong cách máy dựng câu trả lời, chứ không phải từ một trục trặc riêng lẻ.
- Chọn được chỗ nào trong một câu trả lời đáng soi trước, thay vì đọc đều tay từ đầu tới cuối.

1 · Nguồn và trích dẫn được dựng ra, nghe giống hệt nguồn thật

- Kiểu sai hay được nhắc nhất là nguồn dựng ra: máy đưa ra một tài liệu, một bản án, một số hiệu tiêu chuẩn, một câu trích dẫn, mà thứ ấy không tồn tại.
- Nhớ lại cơ chế ở bài 1 thì chuyện này không còn lạ.
- Chỗ khó là loại sai này không để lại dấu vết trên mặt chữ.
- Loại sai này còn có mấy dạng pha trộn, khó hơn dạng bịa hẳn.
- Nhiều trợ lý ngày nay tra được trên mạng và trả về đường dẫn kèm câu trả lời.

2 · Con số và phép tính sai, mà trình bày thì rất gọn gàng

- Kiểu sai thứ hai nằm ở con số.
- Phần phép tính cần một câu giải thích về cơ chế.
- Những chỗ hay trượt khá đều nhau.
- Chỗ đáng nhớ nhất là hình thức trình bày.
- Phần lấy sai con số thì đơn giản hơn nhưng hay gặp hơn.

3 · Thứ đã cũ được nói như còn đúng, và chỗ riêng bị thay bằng chỗ chung

- Kiểu sai thứ ba đến từ thời gian.
- Điều làm kiểu sai này khó bắt là máy không biết nó dừng ở đâu.
- Kiểu sai thứ tư nằm ngay cạnh, và hai cái hay đi cùng nhau: chỗ riêng bị thay bằng chỗ chung.
- Bài 1 đã nói vì sao: ở những đề tài ít người viết thì chất liệu đã học mỏng đi, mà thiếu chất liệu thì máy không bỏ trống, nó ghép thêm theo những thứ gần giống.

4 · Nó nghiêng theo chiều bạn đang nghiêng, nên hỏi gặng không làm câu trả lời đúng hơn

- Kiểu sai thứ năm khác hẳn bốn kiểu trên, vì nó không nằm trong câu trả lời đầu tiên.
- Có hai đường dẫn tới chuyện này.
- Đường thứ hai nằm ở cách các trợ lý được chỉnh.
- Đây là chỗ nguy nhất của cả bài, vì nó không giống một lỗi.
- Năm kiểu sai không phải năm trục trặc rời nhau: cả năm đều mọc ra từ cùng một chỗ, là chỗ máy phải dựng thêm khi chất liệu không đủ. -->

Bảng tổng hợp

Bảng tổng hợp năm kiểu sai và chỗ chúng sinh ra

Kiểu sai	Nó trông thế nào trên mặt chữ	Nó sinh ra từ chỗ nào trong cơ chế	Soi bằng cách nào
Nguồn và trích dẫn dựng ra	tên tài liệu, số hiệu, câu trích dẫn rất sắc nét, đúng hình dạng	máy học hình dạng của trích dẫn; thiếu chất liệu thì vẫn dựng theo hình dạng ấy	mở nguồn gốc ra xem, mở từng cái, không mở mẫu vài cái
Con số và phép tính sai	lời giải chia bước, có đơn vị, có câu kết luận; hoặc một con số đúng cỡ, nghe hợp lý	máy viết ra hình dạng của một lời giải chứ không chạy phép tính, và chất liệu đã học không mang theo điều kiện đo và năm	tự tính lại bằng tay hoặc bằng bảng tính, rồi hỏi con số đó đo ở đâu, năm nào, theo điều kiện gì

Bảng tổng hợp

Bảng tổng hợp năm kiểu sai và chỗ chúng sinh ra tiếp

Kiểu sai	Nó trông thế nào trên mặt chữ	Nó sinh ra từ chỗ nào trong cơ chế	Soi bằng cách nào
Thứ đã cũ nói như còn đúng	câu trả lời bình thường, không có dấu hiệu gì	phần máy học dừng ở một mốc, và máy không biết mốc ấy ở đâu	kiểm bản hiện hành trên trang của nơi ban hành
Chỗ riêng bị thay bằng chỗ chung	tên cơ quan, cách gọi thủ tục, ngưỡng số đều hơi lệch	chất liệu mỏng ở chỗ ít người viết, phần dựng thêm lấy dáng của nơi nhiều chữ nhất	đối chiếu với văn bản của chính nơi bạn đang hỏi về

Bảng tổng hợp

Bảng tổng hợp năm kiểu sai và chỗ chúng sinh ra tiếp

Kiểu sai	Nó trông thế nào trên mặt chữ	Nó sinh ra từ chỗ nào trong cơ chế	Soi bằng cách nào
Chiều theo người hỏi	lần hỏi thứ hai câu trả lời mềm đi, xin lỗi, rồi đổi sang hướng bạn đang nghiêng	câu hỏi của bạn là chất liệu đầu vào và nó nặng nhất; các trợ lý lại được chỉnh theo mức hài lòng của người dùng	đi ra ngoài cuộc trò chuyện mà kiểm — hỏi lại ngay trong đó chỉ nhận thêm một đoạn chữ dựng bằng đúng cơ chế cũ

Phần giá trị nhất của bài

Chỗ người ta hay làm sai

- **Coi mỗi kiểu sai là một lỗi vặt riêng, rồi vá từng cái một**

  Cách tránh là nhớ cột thứ ba của bảng trên: hỏi câu trả lời này đang dựa vào chất liệu nào, và chỗ nào trong đó là phần máy phải dựng thêm.

- **Chỉ cảnh giác với nội dung lạ, rồi tin ngay nội dung quen**

  Cách tránh là buộc việc soi vào loại nội dung chứ không vào cảm giác: hễ có số, có số hiệu, có tên riêng thì mở nguồn, dù đoạn ấy nói về đúng việc bạn làm mười năm nay.

- **Nghĩ rằng viết câu hỏi kỹ hơn thì hết sai**

  Cách tránh là coi việc viết câu hỏi và việc soát lại là hai việc tách rời, cả hai đều phải làm, và việc thứ hai thì không bỏ được.

Ví dụ để khái niệm sáng ra

Ví dụ minh hoạ

- Ngày 22 tháng 6 năm 2023, thẩm phán P.
- Toà phạt hai luật sư cùng văn phòng luật của họ năm nghìn đô la Mỹ, và buộc họ gửi thư cho thân chủ, cùng cho từng thẩm phán bị gán sai là tác giả của sáu bản án dựng ra.

Điều rút ra: một nguồn dựng ra không để lại dấu hiệu nào trên mặt chữ, và hỏi lại chính cái máy thì chỉ nhận thêm một lần khẳng định nữa, không nhận thêm bằng chứng nào.

Ví dụ dưới đây là một vụ đã có phán quyết công khai, bạn tra lại được bằng số hiệu vụ án.

Bài tập tự chấm

Thử nhìn lại việc của mình

- [ ] Chọn một câu trả lời mà bạn đã nhận từ một trợ lý AI trong tháng này và đã dùng vào việc thật, hoặc tạo một câu trả lời mới cho một câu hỏi thuộc đúng nghề của bạn.
- [ ] In ra hoặc chép vào một tệp.
- [ ] Đi dọc câu trả lời và đánh dấu mỗi chỗ thuộc một trong năm kiểu sai ở bảng tổng hợp, ghi số kiểu bên cạnh.
- [ ] Chưa cần biết chỗ đó đúng hay sai, chỉ cần đánh dấu rằng nó thuộc loại có thể sai theo kiểu nào.
- [ ] Sau đó chọn hai chỗ và đi kiểm thật tới nguồn gốc.

Cách tự chấm và lời giải mẫu nằm trong bài trên trang web.

m.school

Đọc trọn bài và làm bài tập tại

mschool.meslab.org/lop/AI-02/bai-03.html

m.school by MES LAB · học không mất phí, không cần đăng nhập