AI-02 · Bài 1

Trợ lý AI đang làm gì khi nó trả lời bạn

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

- nói được bằng lời của mình cách một trợ lý AI dựng ra câu trả lời, và vì sao cách dựng đó làm cho một câu sai trông giống hệt một câu đúng.
- Chia được một câu trả lời thành ba loại chất liệu: thứ bạn đưa vào, thứ máy đã học sẵn, và thứ máy dựng thêm tại chỗ.
- Nói được vì sao giọng chắc chắn trong câu trả lời không phải là căn cứ để tin.

1 · Trợ lý AI dựng câu trả lời bằng cách đoán chữ đi tiếp, không bằng cách tra sổ

- Mô hình ngôn ngữ lớn
- Khi bạn gõ một câu hỏi, thứ cái máy nhận được không phải là một câu hỏi theo nghĩa bạn hiểu.
- Hãy so với một thứ bạn quen hơn: tra một bảng tra.
- Trợ lý AI không có chỗ trả về trống.
- Có một chỗ cần nói rõ, vì nó hay gây nhầm.
- Hai cách lấy một câu trả lời: tra bảng thì có thể không có dòng nào, còn đoán chữ thì bao giờ cũng ra một đoạn chữ. -->

2 · Trôi chảy và đúng là hai việc khác nhau, và máy chỉ được rèn một trong hai

- Cái máy được rèn để đoạn chữ nó viết ra giống văn người viết: đúng ngữ pháp, mạch lạc, đi thẳng vào câu hỏi, có thứ lớp.
- Nó không được rèn để biết lúc nào mình đang không biết.
- Khi một người không chắc, giọng họ đổi.
- Với trợ lý AI, mối liên hệ ấy đứt.
- Giới nghề gọi phần nội dung máy dựng ra nghe như thật mà không có thật là nội dung bịa (hallucination).
- Hệ quả thực tế rất đáng nhớ: bạn không bắt được loại lỗi này bằng cách đọc kỹ hơn.

3 · Trong một câu trả lời có ba loại chất liệu, và chúng không đáng tin như nhau

- Một câu trả lời trông như một khối liền.
- Loại thứ nhất là thứ bạn đưa vào
- Loại thứ hai là thứ nằm sẵn trong phần máy đã học
- Loại thứ ba là thứ máy dựng thêm tại chỗ
- Ba loại này không được dán nhãn trong câu trả lời.
- Một câu trả lời bổ ra làm ba lớp chất liệu, lớp càng xa thứ bạn đưa vào thì càng cần đối chiếu. -->

4 · Nó có chữ người ta đã viết về những việc giống việc của bạn, chứ không có việc của bạn

- Chất liệu của trợ lý AI là văn bản.
- Điều này sinh ra một tính chất bạn nên nhớ: câu trả lời của nó nghiêng về chỗ trung bình.
- Có ba vùng mà chất liệu luôn mỏng, và bạn nên dè chừng sẵn.
- Vùng thứ nhất là thứ chưa ai viết.
- Vùng thứ hai là thứ viết bằng tiếng ít người dùng, hoặc viết ở nơi ít được thu thập.
- Vùng thứ ba là thứ người ta không viết ra vì nó hiển nhiên với người trong nghề.

Bảng tổng hợp

Bảng tổng hợp ba loại chất liệu trong một câu trả lời

Loại chất liệu	Nó từ đâu ra	Dấu hiệu nhận ra	Tin tới mức nào	Việc bạn phải làm
Thứ bạn đưa vào	câu hỏi bạn viết, tài liệu bạn dán vào, tệp bạn đưa lên	câu trả lời nhắc lại đúng những thứ có trong tài liệu của bạn	khá cao, vì bản gốc đang trong tay bạn	đối chiếu với bản gốc, soi chỗ máy tóm tắt lệch hoặc bỏ sót
Thứ máy đã học sẵn	khối văn bản dùng để luyện, dừng ở một mốc thời gian	nội dung chung của nghề, định nghĩa, nguyên lý, cách làm phổ biến	vừa phải ở mức chung, thấp dần khi đi vào chỗ riêng và chỗ mới	hỏi lại xem thứ này có đổi sau mốc thời gian không, và có đúng với bối cảnh Việt Nam không

Bảng tổng hợp

Bảng tổng hợp ba loại chất liệu trong một câu trả lời tiếp

Loại chất liệu	Nó từ đâu ra	Dấu hiệu nhận ra	Tin tới mức nào	Việc bạn phải làm
Thứ máy dựng thêm tại chỗ	phần máy ghép thêm cho câu trôi chảy khi thiếu chất liệu	số hiệu, con số, tên riêng, trích dẫn trông rất sắc nét mà bạn không đưa vào	thấp, và trông càng cụ thể thì càng nên dè chừng	mở nguồn gốc ra kiểm, hoặc bỏ hẳn phần đó khỏi bản của bạn

Phần giá trị nhất của bài

Chỗ người ta hay làm sai

- **Đọc giọng chắc chắn thành dấu hiệu rằng máy đã kiểm**

  Cách tránh là đặt một luật cho mình, không phụ thuộc cảm giác: mọi con số, mọi số hiệu, mọi trích dẫn đều phải mở nguồn gốc ra xem, bất kể câu văn nghe chắc tới đâu.

- **Hỏi lại chính cái máy để kiểm cái máy**

  Cách tránh là kiểm bằng một nguồn ở ngoài: văn bản gốc, trang của cơ quan ban hành, một người biết việc.

- **Tưởng rằng bản mới hơn thì hết lỗi**

  Cách tránh là buộc việc soát vào loại nội dung, không buộc vào bản máy: con số và nguồn thì luôn kiểm, dù dùng bản nào.

Ví dụ để khái niệm sáng ra

Ví dụ minh hoạ

- Năm 1966, Joseph Weizenbaum, nhà khoa học máy tính ở Viện Công nghệ Massachusetts, công bố một chương trình tên là ELIZA trên tạp chí Communications of the ACM...
- Weizenbaum viết rõ ngay trong bài báo rằng chương trình không hiểu gì cả.

Điều rút ra cho bài này: người đọc gán sự hiểu biết cho một đoạn chữ dựa vào chỗ đoạn chữ ấy nghe có trôi chảy hay không. Phản xạ đó đã có từ sáu mươi năm trước...

Ví dụ dưới đây là ca thật trong lịch sử máy tính, có bài báo khoa học công khai, nên bạn tra lại được.

Bài tập tự chấm

Thử nhìn lại việc của mình

- [ ] Lấy một câu hỏi thật trong công việc tuần này của bạn, loại câu hỏi mà bình thường bạn sẽ đi hỏi đồng nghiệp hoặc đi tra tài liệu.
- [ ] Đặt câu hỏi ấy cho một trợ lý AI, đặt trơn, không đưa kèm tài liệu nào.
- [ ] In hoặc chép câu trả lời ra, rồi đánh dấu từng đoạn theo ba loại chất liệu trong bảng tổng hợp: thứ bạn đưa vào, thứ máy đã học sẵn, thứ máy dựng thêm.
- [ ] Sau đó chọn ba chỗ cụ thể nhất trong câu trả lời, là ba chỗ có số hoặc có tên riêng, và đi mở nguồn gốc của từng chỗ.
- [ ] Ghi lại kết quả: có thật, sai, hay không tìm thấy.

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-01.html

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