Nome: _____ Número: _____

1 2 9 0

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D COIMBRA

Inteligência Artificial Centrada no Humano

2022/23 – 1º Semestre

Exame de Recurso
23/janeiro/2023 – 90 min

Mestrado em Engenharia e Ciência de Dados
Departamento de Engenharia Informática

Question:	1	2	3	4	Total
Points:	25	25	25	25	100
Score:					

Ler com atenção:

- **Importante**: A fraude denota uma grave falta de ética e constitui um comportamento não admissível num estudante do ensino superior e futuro profissional mestre. Qualquer tentativa de fraude leva à reprovação na disciplina, tanto do facilitador como do prevaricador.
- Este teste é individual e com consulta. Não pode usar dispositivos electrónicos, incluindo o telemóvel. Só são admitidos apontamento em papel. Não pode trocar apontamentos com colegas.
- Responda às perguntas nos espaços indicados para o efeito. Se precisar de papel adicional, contacte o docente vigilante da sua sala.
- Nas **perguntas Verdadeiro/Falso com pedido de justificação**, respostas sem justificação são cotadas a zero.
- Nas **perguntas Verdadeiro/Falso sem pedido de justificação**, respostas incorretas implicam uma penalização de 30% da cotação; se não responder, tem 0 na pergunta.

Pergunta 1 (25 %)

ChatGPT 3.5 (`https://chat.openai.com`) é um sistema de IA capaz de responder a questões em linguagem natural ou de produzir código numa linguagem de programação. Pelas muitas aplicações práticas que se vislumbram, poderá ter um forte impacto nas nossas vidas. Considere os seguintes excertos de descrições da metodologia usada no seu desenvolvimento, bem como das suas limitações, a maior parte das quais extraídas do website oficial (`https://openai.com/blog/chatgpt/`) da organização que desenvolveu o sistema, a OpenAI:

> Reinforcement learning from Human Feedback (also referenced as RL from human preferences - RLHF) is a challenging concept because it involves a multiple-model training process and different stages of deployment. RLHF's most recent success was its use in ChatGPT 3.5:
>
> We trained this model using RLHF, using the same methods as InstructGPT [previous model of OpenAI], but with slight differences in the data collection setup. We trained an initial model using supervised fine-tuning: human AI trainers provided conversations in which they played both sides—the user and an AI assistant.
>
> Limitations:
> ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers. Fixing this issue is challenging, as: (1) during RL training, there's currently no source of truth; (2) training the model to be more cautious causes it to decline questions that it can answer correctly; and (3) supervised training misleads the model because the ideal answer depends on what the model knows, rather than what the human demonstrator knows.
> Ideally, the model would ask clarifying questions when the user provided an ambiguous query. Instead, our current models usually guess what the user intended.
> When deploying a system using RLHF, gathering the human preference data is quite expensive due to the mandatory and thoughtful human component. RLHF performance is only as good as the quality of its human annotations, which takes on two varieties: human-generated text, such as fine-tuning the initial language model in InstructGPT, and labels of human preferences between model outputs.
> The second challenge of data for RLHF is that human annotators can often disagree, adding a substantial potential variance to the training data without ground truth.

Neste excerto da descrição do ChatGPT 3.5, constata-se explicitamente, e em alguns casos depreende-se implicitamente, o uso de conceitos/técnicas/tecnologias de IA que foram lecionadas na unidade curricular de IACH.

Pretende-se mesta questão que relacione os conceitos que foram lecionados na disciplina com o ChatGPT 3.5.

Dos seguintes conceitos/temas/tecnologias lecionadas na unidade curricular de IACH, identifique (circundando) e justifique sucintamente, eventualmente recorrendo a excertos do texto, os que poderão estar envolvidos no desenvolvimento e na utilização do ChatGPT 3.5.

Nome: _____ Número: _____

(a) Human in the Loop:

.....
.....
.....
.....

(b) Human out of the Loop:

.....
.....
.....
.....

(c) Machine in the Loop:

.....
.....
.....
.....

(d) Supervised Learning:

.....
.....
.....
.....

(e) Apprenticeship Learning/Learning by Demonstration/Imitation Learning:

.....
.....
.....
.....

Pergunta 2 (25 %)

Classifique as seguintes frases como Verdadeiro ou Falso (nota importante, para cada frase: se der a resposta correcta, recebe 100%; se não der uma resposta, recebe 0%; se der uma resposta incorrecta, recebe uma penalização de 30% do valor da pergunta):

(a) A semântica local das Redes Bayesianas define cada nodo como condicionalmente independente dos seus não-descendentes, tendo em conta os seus pais.

(a) _____

(b) As Redes Bayesianas são gráficos cíclicos com um conjunto de nodos, um por variável, associados a uma tabela de probabilidade condicional.

(b) _____

(c) No contexto de Sistemas de Recomendação, o processo de identificar utilizadores semelhantes e recomendar o que os utilizadores semelhantes gostam é chamado de filtragem colaborativa (Collaborative Filtering).

(c) _____

(d) Na maioria das aplicações, o sistema de recomendação não oferece aos utilizadores uma classificação de todos os itens, mas sugere antes os que o utilizador poderá valorizar mais.

(d) _____

(e) Para a interpretação de uma característica binária ou categórica num modelo linear representado na Equação 1, podemos dizer que um aumento da característica x_k por uma unidade aumenta a previsão para y em β_k unidades quando todos os outros valores da característica permanecem fixos.

$$g(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \tag{1}$$

(e) _____

(f) Um modelo de Aprendizagem Computacional é considerado interpretável se puder: (i) ser inspeccionado; (ii) ser entendido que uma determinada resposta (output) é obtida para uma dada entrada (input); (iii) ser entendido como a resposta mudaria se a entrada mudasse.

(f) _____

(g) Um modelo de Aprendizagem Computacional não pode ser explicável se não for interpretável.

(g) _____

(h) A interpretabilidade de um modelo de Aprendizagem Computacional envolve apenas este modelo e não requer um modelo extra, enquanto que a explicabilidade pode ser fornecida por um modelo separado.

(h) _____

Nome: _____ Número: _____

Pergunta 3 (25 %)

Pretende criar um modelo vetorial para representar documentos, que tenha por base os oito documentos na tabela 1, considerando como features **unigramas** e **bigramas**, mas apenas aqueles que ocorrem em pelo menos dois documentos.

uma experiência única
quero repetir todos os dias
experiência muito muito boa
repetir ou não repetir
experiência indiferente
péssima experiência
experiência muito má
não quero repetir

Tabela 1: Coleção de documentos.

(a) Enumere o vocabulário a considerar para este modelo:

(b) Com base no vocabulário e na ordem apresentada na resposta anterior, indique o vetor resultante da representação do seguinte documento:

não posso não repetir

1. Considerando apenas contagens:

2. Após a aplicação de pesos TF-IDF. Ignore o logaritmo no cálculo do DF e apresente os cálculos necessários:

(c) Relativamente às representações anteriores, indique uma vantagem de usar *embeddings* codificados por redes neuronais.

Nome: _____ Número: _____

(d) Indique um tipo de viés (bias) comum em grandes modelos de linguagem neuronais, ilustre-o com um exemplo.

(e) Imagine que, de forma a treinar um modelo para Análise de Sentimentos, dados como os da tabela 1 seriam anotados com recurso a microtrabalho (*microwork*). Explique em que consiste e como poderia ser aplicado a esta tarefa.

Pergunta 4 (25 %)

O sistema de síntese de voz *e-Parrot* gera um discurso semelhante ao humano a partir de um texto. Pode ser utilizado numa variedade de aplicações, tais como sistemas de assistência técnica, assistentes virtuais, e locução para publicidade e meios de comunicação. Pode também ser utilizado para criar vozes artificiais para animação ou para dispositivos de fala, tais como altifalantes "inteligentes", para proporcionar acessibilidade a pessoas com deficiências ou idosos que têm dificuldade em ler ou falar, e também para ajudar pessoas com barreiras linguísticas a comunicar de forma mais eficaz.

O e-Parrot pode ainda imitar a voz de uma pessoa específica através de um processo de *clonagem de voz*: dá-se ao sistema uma gravação de uma amostra da fala da pessoa, ele cria um modelo das caraterísticas identificadoras da voz, e usa depois o modelo no processo de síntese de voz, gerando assim um discurso na voz da pessoa em causa.

Suponha que a empresa que desenvolveu e vende o e-Parrot pretende que ele siga as Orientações Éticas para uma IA de Confiança definidas pelo Grupo Independente de Peritos de Alto Nível sobre a Inteligência Artificial criado por iniciativa da Comissão Europeia.

(a) Analise com atenção as seguintes frases que exprimem preocupações relativamente ao funcionamento de sistemas como o e-Parrot. Assinale, para cada uma delas, qual das três componentes da IA de Confiança estará a ser comprometida se a situação se verificar.
Para cada frase, assinale uma opção; se achar que a situação pode envolver mais do que uma componente, assinale a que acha que se enquadra melhor.
Justifique cada resposta.

i. *O e-Parrot pode ser usado para criar* deepfakes *de fala que possam ajudar alguém a fazer-se passar por uma pessoa específica a dizer coisas que ela não disse, para assim espalhar desinformação ou propaganda.*

Preocupação relativa a (assinale uma opção):

◯ Legalidade ◯ Ética ◯ Robustez

Justifique:

ii. *Se os recursos computacionais do e-Parrot não forem suficientes para fazer face a uma procura intensa, ele pode não ser capaz de processar todos os pedidos, pondo em causa o fornecimento do serviço, podendo com isso causar prejuízos aos utilizadores.*

Preocupação relativa a (assinale uma opção):

◯ Legalidade ◯ Ética ◯ Robustez

Justifique:

iii. *O uso de sistemas como o e-Parrot pode levantar preocupações com a recolha, armazenamento e uso de dados pessoais, especialmente se os sistemas forem usados para gravar e clonar as vozes dos utilizadores sem a expressa autorização destes.*

Preocupação relativa a (assinale uma opção):

◯ Legalidade ◯ Ética ◯ Robustez

Justifique:

iv. *O e-Parrot pode perpetuar e amplificar vieses e atitudes discriminatórias presentes nos dados de voz usados para treino, resultando em fala enviezada (p.ex., usando apenas vozes que identifiquem um grupo maioritário, podendo dessa forma promover a discriminação de grupos minoritários.*

Preocupação relativa a (assinale uma opção; se achar que a situação pode envolver mais do que uma componente, assinale a que acha que se enquadra melhor e justifique em conformidade):

◯ Legalidade ◯ Ética ◯ Robustez

Justifique:

(b) Classifique as seguintes afirmações como Verdadeiras ou Falsas (considerando as Orientações Éticas atrás referidas), justificando a sua resposta:

i. O elevado desempenho do e-Parrot pode retirar mercado a outras empresas concorrentes, o que viola a obrigação de respeito pela dignidade humana.

◯ Verdadeira ◯ Falsa

Justifique:

ii. A Prevenção de Danos é um imperativo ético para um sistema como o e-Parrot.

◯ Verdadeira ◯ Falsa

Justifique:

iii. Pode ocorrer tensão entre o direito indivisível da Liberdade Individual e o direito fundamental do Respeito Pela Dignidade Humana na operação do e-Parrot.

◯ Verdadeira ◯ Falsa

Justifique:

Nome: _____ Número: _____

Página extra:

Anexo - Tradução para português do excerto da descrição do ChatGPT3.5:

A Aprendizagem por Reforço a partir de Feedback Humano (também referido como RL a partir de preferências humanas - RLHF) é um conceito desafiante porque envolve um processo de treino de múltiplos modelos e diferentes fases de implantação. O sucesso mais recente da RLHF foi a sua utilização no ChatGPT 3.5.

Treinámos este modelo utilizando RLHF, utilizando os mesmos métodos que o InstructGPT [modelo anterior do OpenAI], mas com ligeiras diferenças na configuração da recolha de dados. Treinámos um modelo inicial usando aperfeiçoamento supervisionado: os treinadores humanos de IA forneceram conversas em que se posicionaram de ambos os lados - o utilizador e um assistente de IA.

Limitations:
ChatGPT por vezes escreve respostas plausíveis, mas incorrectas ou sem sentido. Resolver esta questão é um desafio, pois: (1) durante o treino RL, não há atualmente nenhuma fonte de verdade; (2) treinar o modelo para ser mais cauteloso faz com que ele recuse perguntas às quais possa responder corretamente; e (3) o treino supervisionado deturpa o modelo porque a resposta ideal depende do que o modelo sabe, e não do que o demonstrador humano sabe.
O ideal seria que o modelo fizesse perguntas esclarecedoras quando o utilizador apresentasse uma pergunta ambígua. Em vez disso, os nossos modelos atuais normalmente adivinham o que o utilizador pretendia.
Ao implementar um sistema utilizando RLHF, a recolha dos dados de preferências dos humanos é bastante dispendiosa devido à obrigatoriedade e necessidade de atenção por parte dos humanos. O desempenho do RLHF é apenas tão bom como a qualidade das suas anotações humanas, que assume duas variedades: texto gerado por humanos, tal como o aperfeiçoamento do modelo linguístico inicial em InstructGPT, e anotações de preferências humanas entre os resultados dos modelos.
O segundo desafio dos dados para RLHF é que os anotadores humanos podem frequentemente discordar, acrescentando uma possível variação substancial aos dados de treino sem verdade de base.