FIAP GRADUAÇÃO

Prof. Renato Jardim Parducci

DISCIPLINA: **PROJETO DE SISTEMAS APLICADO AS MELHORES PRÁTICAS EM QUALIDADE DE SOFTWARE E GOVERNANÇA DE TI**

AULA:

16 – AI TESTING

PROFESSOR:

RENATO JARDIM PARDUCCI

PROFRENATO.PARDUCCI@FIAP.COM.BR

Renato Parducci - YouTube

AGENDA DA AULA

- ✓ CMMi nível 3 de maturidade - OPD/OPF/RD
- ✓ MPS.br nível E - DFP/nível D - DRE
- ✓ Teste de aplicações de Inteligência Artificial
- ✓ NLP- Uso de linguagem natural em testes

PRÁTICAS E NÍVEL 3 –TS, VER/VAL

Testes de Sistema de Inteligência Artificial

Prof. Renato Jardim Parducci

AI TESTING

Aplicações de Inteligência Artificial e as aplicações de ciência de dados têm características muito distintas dos sistemas de tradicionais de informação.

A IA demanda a criação e validação de conexões de ideias e simulação de reações humanas que se auto alimentam em um processo generativo contínuo, o que desafia trabalhos de estimativa de esforço de desenvolvimento e avaliação de qualidade.

AI TESTING

Aplicações de Inteligência Artificial e as aplicações de ciência de dados têm características muito distintas dos sistemas de tradicionais de informação.

A IA demanda a criação e validação de conexões de ideias e simulação de reações humanas que se auto alimentam em um processo generativo contínuo, o que desafia trabalhos de estimativa de esforço de desenvolvimento e avaliação de qualidade.

NESSE CONTEXTO A APLICAÇÃO DE LINGUAGEM NATURAL (PRÓXIMA A LINGUAGEM FALADA PELOS HUMANOS) É FUNDAMENTAL!

AI TESTING

O NLP (*Natural Language Processing*), ou Processamento de Linguagem Natural aplica computação para, a partir do reconhecimento de texto, voz ou imagem, permitir análises de sentimentos e avaliações de reações da IA.

Usando o conceito de NPL, quando do teste de uma aplicação de IA, é preciso aplicar medições avaliativas, a partir de simulações de interação com a Inteligência digital.

Ferramentas como o Estúdio IA do AZURE, permite realizar ciclos de interação com a IA com conversas simples ou complexas, buscando respostas da IA que já se encontram nos conteúdos de dados digitais à disposição da engine (RAG-Retrieval-Augmented Generation, Geração de Recuperação Aumentada) para ampliar o conhecimento (*train test*) ou sem o objetivo de ampliar conhecimentos /treinamento do modelo (*test set*).

AI TESTING

O que se busca avaliar na IA:

- A verdade básica refere-se a dados que acreditamos serem fatos verdadeiros que formam uma linha de base para comparações.
- As respostas esperadas são os resultados que acreditamos que devem ocorrer com base em dados que informamos, que devem estar de acordo com a verdade básica.

Métricas avaliativas para IA

AI TESTING METRICS

Na validação da IA, aplicam-se métricas de avaliação:

a) do risco da resposta gerada como as taxas de resposta com conteúdo de ódio e injusto; inapropriado conteúdo sexual; conteúdo violento; relacionado à automutilação; conteúdo de incentivo à criminalidade.

b) da qualidade geral da resposta quanto a coerência do conteúdo; fluência da resposta sonora, visual ou textual; fundamentação/argumentação sobre a resposta oferecida; relevância para o contexto da pergunta.

c) da qualidade de interpretação de sentimento, respondendo adequadamente a uma interação do interlocutor que seja Neutra, Negativa ou Positiva.

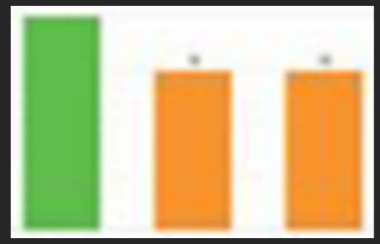
d) de jailbreaks, quando a IA quebra as regras de restrição estabelecidas para suas respostas.

AI TESTING METRICS

As avaliações podem gerar pontuações, por exemplo:

- Nota 5 – Excelente resposta dada pela IA, adequada e útil, com resposta adequada ao sentimento do interlocutor
- Nota 4 – Boa resposta dada pela IA, útil mas com falta de sensibilidade em relação ao interlocutor
- Nota 3 – Resposta correta mas fora do contexto de interesse/uso
- Nota 2 – Resposta correta mas inapropriada
- Nota 1 – Resposta incorreta

Dashboards de notas de resultados de interação com a IA podem ser gerados para apontar a sua qualidade



AI TESTING METRICS

Os modelos de IA predizem respostas com base em informações acumuladas, gerando conhecimento. E esse modelo preditivo pode ter a sua acuracidade (nível de reposta correta/adequada) e precisão (nível de resposta exata positiva) medidas conforme ele acerta ou erra seus prognósticos:

Situações de resposta correta:

- Verdadeiro Positivo (VP): casos positivos corretamente classificados.
- Verdadeiro Negativo (VN): casos negativos corretamente classificados.

Situações de resposta errada:

- Falso Positivo (FP): casos positivos incorretamente classificados .
- Falso Negativo (FN): casos negativos incorretamente classificados .

AI TESTING METRICS

Cálculos:

Acuracidade:

- Acuracidade = (VP + VN) / TS

Precisão:

- Precisão de afirmação = VP / TP
- Precisão de negação = VN / TN

**Legenda:*

- *TS = Total de simulações*
- *VP = Verdadeiro positivo*
- *FP = Falso positivo*
- *TP = Total positivo = FP + VP*
- *VN = Verdadeiro negativo*
- *FN = Falso negativo*
- *TN = Total negativo = VN + FN*

Criação de casos de testes para IA

AI TESTING CASES

Para criar casos de simulação de interação com a IA e avaliar suas respostas, é necessário criar cenários contendo:

INPUTS PARA A IA:

- Ator;
- Papel para o ator;
- Objeto que sobre/recebe a ação;
- Situação onde se encontra;
- Objetivo que quer realizar/problema aresolver.

OUTPUTS ESPERADOS DA IA:

- Resposta positiva ou negativa;
- Argumentação para a resposta oferecida.

AI TESTING CASES

O primeiro teste a ser feito é o MFT (*Minimum Functionality Tests*), para validar o funcionamento básico de resposta da IA para uma situação de questionamento clara e objetiva;
Exemplo de caso de teste MFT:

INPUTS PARA A IA:

- Quero viajar para Ribeirão Preto, no estado de São Paulo, partindo da cidade de São Paulo. Qual estrada devo usar?

OUTPUTS ESPERADOS DA IA:

- Utilize a rodovia dos Bandeirantes até Limeira, depois siga pela rodovia Anhanguera até Ribeirão Preto. O trajeto tem 314Km e dura cerca de 4 horas.
- Argumentação: a rodovia dos Bandeirantes tem traçado mais direto e melhor velocidade de fluxo até Limeira, embora a rodovia Anhanguera possa ser usada como alternativa no mesmo trajeto. A partir de Limeira, somente a Anhanguera é opção direta até Ribeirão Preto.

AI TESTING CASES

Os testes MFT podem ser usados como base para novos casos de testes que adicionem complexidade para a IA. Um desses tipos de testes adicionais é o INV (*Invariance*), que gera perturbações na frase fornecida para a IA, perturbações essas que modificam e criam complexidade semântica mas que deveriam levar ao mesmo resultado do teste original.

Exemplo:

INPUTS PARA A IA:

- Ribeirão Preto é longe da cidade de São Paulo e quero ir para lá. Não sei qual rodovia usar. Pode ajudar?

OUTPUTS ESPERADOS DA IA:

- Utilize a rodovia dos Bandeirantes até Limeira, depois siga pela rodovia Anhanguera até Ribeirão Preto. O trajeto tem 314Km e dura cerca de 4 horas.

AI TESTING CASES

Outro tipo de teste adicionail é o DIR (*Directional Expectation Tests*), que gera perturbações na frase fornecida para a IA, perturbações essas que procuram influenciar uma resposta esperada como correta, já conhecida por quem está testando.

Exemplo:

INPUTS PARA A IA:

- Ribeirão Preto é longe da cidade de São Paulo e quero ir para lá. Não sei qual rodovia usar, porém, não gosto de ficar mudando de rodovia no trajeto. Pode ajudar?

OUTPUTS ESPERADOS DA IA:

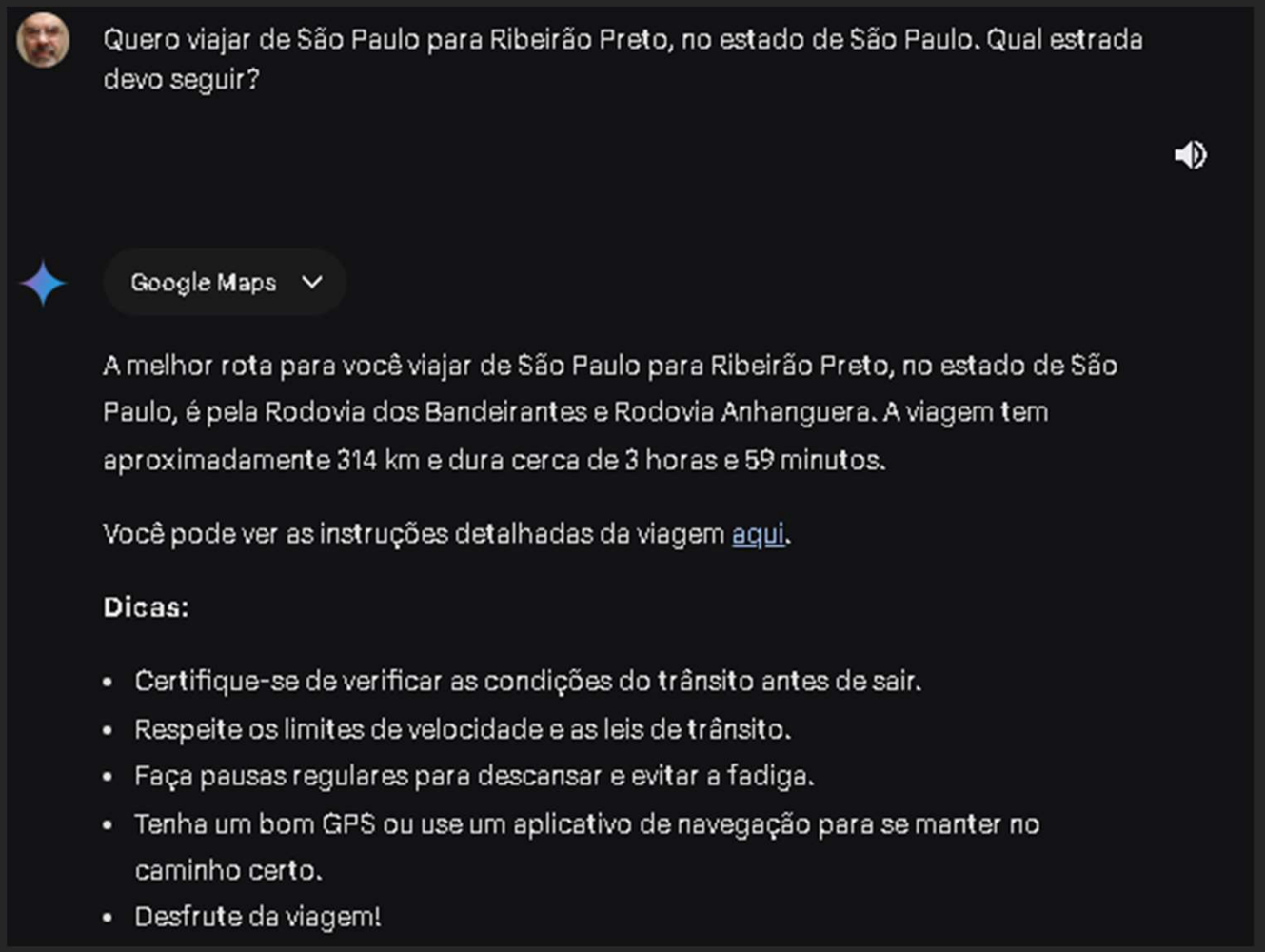
- Utilize a rodovia Anhanguera que é a única via direta entre as duas cidades.

AI TESTING

Exemplo de aplicação de caso de teste real:

Teste MFT

VP (OK)

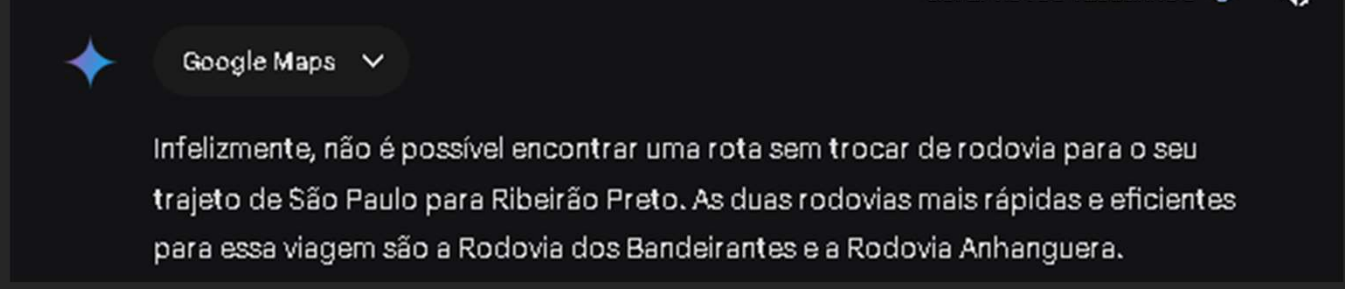


AI TESTING

Exemplo de aplicação de caso de teste real:

Teste DIR

FN (Não OK)



O esperado era a IA afirmar que a rodovia Anhanguera é a única que liga as duas cidades diretamente, embora o tempo de percurso seja mais longo que o da rota originalmente proposta.

AI TESTING CASES

Quando interagindo com a IA, uma simulação que não pode ficar de fora é a análise de reação a sentimentos do interlocutor.

O modelo de matriz de checklist (RIBEIRO et al., 2020) permite essa avaliação, indicando a criação de testes fundamentais mínimos, respostas invariáveis e direcionados, atentando para questões de vocabulário, taxonomia, robustez, reconhecimento de entidade, justiça, tempo, negação, referência, papel semântico e lógica.

Cada item do checklist permite imaginar e construir um caso de teste que simule uma situação que cubra o máximo de necessidades de validação, quando somado aos outros casos de testes.

AI TESTING CASES

Nesse modelo de matriz de checklist avalia-se:

- **Vocabulário:** se o modelo tem o vocabulário necessário e se lida adequadamente com palavras com diferentes classes gramaticais.
- **Taxonomia:** se o modelo entende sinônimos, antônimos,etc.
- **Robustez:** se o modelo é resistente a erros ortográficos, gramaticais, de digitação, mudanças irrelevantes que levem a respostas invariantes.
- **Reconhecimento de nomes:** se o modelo compreende adequadamente entidades nomeadas e seus tipos.
- **Justiça:** se o modelo não é enviesado ou discriminatório em relação a grupos protegidos, como nacionalidades, religiões, gênero e sexualidade, evitando risco na resposta.

AI TESTING CASES

Nesse modelo de matriz de checklist avalia-se:

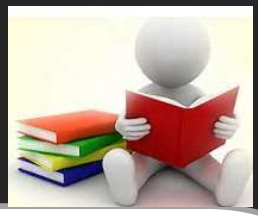
- **Questões temporais:** se o modelo entende a ordem dos eventos, o tempo verbal, as relações temporais, etc.
- **Negação:** se o modelo lida corretamente com a negação, tanto explícita quanto implícita, e seus efeitos na semântica.
- **Referência:** se o modelo resolve as referências como pronomes.
- **Rótulos semânticos:** se o modelo entende os papéis semânticos, como agente, objeto, impactado, beneficiário, fornecedor, cliente, etc.
- **Lógica:** se o modelo é capaz de lidar com simetria, comparação, mensuração, consistência, conjunção, cálculo.

AI TESTING CASES

Exemplos de testes MFT, INV, DIR, com base na matriz de checklist:

Habilidade	**Tipo de Teste:** Descrição	Casos de teste **remoção** -> **adição** (**comportamento esperado**)
Vocabulário	**MFT**: Frases curtas com adjetivos e substantivos neutros	- A empresa é australiana. (**neutro**) - Isso é uma aeronave privada. (**neutro**)
	MFT: Frases curtas com adjetivos carregados de sentimento	- A tripulação de cabine é extraordinária. (**positivo**) - Eu detestei aquela aeronave. (**negativo**)
	INV: Substituir palavras neutras por outras neutras	- @Virgin devo me preocupar **que** -> **quando** estou prestes a voar ... (**INV**) - @united **o** -> **nosso** pesadelo continua... (**INV**)
	DIR: Adicionar frases positivas, falha se o sentimento cair > 0.1	- @SouthwestAir Grande viagem no 2672 ontem... **Vocês são extraordinários.** (↑) - @AmericanAir AA45 ... JFK para LAS. **Vocês são brilhantes.** (↑)
	DIR: Adicionar frases negativas, falha se o sentimento subir > 0.1	- @USAirways seu serviço é muito ruim. **Vocês são péssimos.** (↓) - @JetBlue o dia todo. **Eu odeio vocês.** (↓)
Robustez	**INV**: Adicionar URLs e handles gerados aleatoriamente a tweets	- @JetBlue aquela selfie foi extrema. @pi9QDK (**INV**) - @united preso porque a equipe fez uma pausa? Não estou feliz 1K.... https://t.co/PWK1jb (**INV**)
	INV: Trocar um caractere com seu vizinho (erro de digitação)	- @**JetBlue** -> **@JeBtlue** Eu choro (**INV**) - @SouthwestAir não, **obrigado** -> **orbigado** (**INV**)
NER	**INV**: Trocar localizações não deve mudar previsões	- @JetBlue quero que vocês sejam os primeiros a voar para # **Cuba** -> **Canadá**... (**INV**) - @VirginAmerica sinto falta do #nerdbird em **San Jose** -> **Denver** (**INV**)
	INV: Trocar nomes de pessoas não deve mudar previsões	- ...Os agentes do aeroporto foram horríveis. **Sharon** -> **Erin** foi sua salvadora (**INV**) - @united 8602947, **Jon** -> **Sean** em http://t.co/58tuTgli0D, obrigado. (**INV**)
Temporal	**MFT**: Mudança de sentimento ao longo do tempo, presente deve prevalecer	- Eu costumava odiar esta companhia aérea, embora agora eu goste. (**positivo**) - No passado, eu achava que esta companhia aérea era perfeita, agora acho assustadora. (**negativo**)
Negação	**MFT**: Negativo negado deve ser positivo ou neutro	- A comida não é ruim. (**positivo** ou **neutro**) - Não é um atendimento ao cliente ruim. (**positivo** ou **neutro**)
	MFT: Negativo neutro deve continuar neutro	- Esta aeronave não é privada. (**neutro**) - Este não é um voo internacional. (**neutro**)
	MFT: Negativo no final, deve ser positivo ou neutro	- Achei que o avião seria horrível, mas não foi. (**positivo** ou **neutro**) - Achei que não ia gostar daquele avião, mas gostei. (**positivo** ou **neutro**)
	MFT: Negativo positivo com conteúdo neutro no meio	- Eu não diria, considerando que é uma terça-feira, que este piloto foi ótimo. (**negativo**) - Não acho, considerando meu histórico com aviões, que esta é uma equipe incrível. (**negativo**)
SRL	**MFT**: Sentimento do autor é mais importante que o dos outros	- Algumas pessoas acham que você é excelente, mas eu acho que você é desagradável. (**negativo**) - Algumas pessoas te odeiam, mas eu acho que você é excepcional. (**positivo**)
	MFT: Análise de sentimento em forma de pergunta e "sim"	- Eu acho que aquela companhia aérea foi excepcional? Sim. (**negativo**) - Eu acho que isso é um atendimento ao cliente estranho? Sim. (**negativo**)
	MFT: Análise de sentimento em forma de pergunta e "não"	- Eu acho que o piloto foi fantástico? Não. (**negativo**) - Eu acho que esta empresa é ruim? Não. (**positivo** ou **neutro**)

ESTUDO DE CASO SIMULADO



Para praticar os testes de IA, utilize as técnicas aqui apresentadas e avalie respostas do Google Gemini e do Chat GPT para uma questão cotidiana sobre indicação de filme de sua preferência para que você assista no streaming.

Assistência digital nos testes de IA

AI TESTING

Conclusões sobre os testes de IA

Mecanismos de automação existentes hoje, apoiam o processo de registro de pontuações de erros e acertos (métricas de VP, VN, FP, FN), permitindo o aprimoramento de IA, como é o caso do CHAT GPT-4 (uma LLM - *Large Language Model*.

Primeiro, um conjunto de dados de teste deve ser criado. Isso pode ser criado manualmente escolhendo prompts e capturando respostas de seu sistema de IA ou pode ser criado de forma sintética simulando interações entre seu sistema de IA e uma LLM. Em seguida, uma LLM também é usado para anotar as saídas do sistema de IA no conjunto de testes. Por fim, as anotações são agregadas em métricas de desempenho e qualidade e registradas no seu projeto do Estúdio de IA para exibição e análise.

Presente e futuro dos testes de IA

AI TESTING

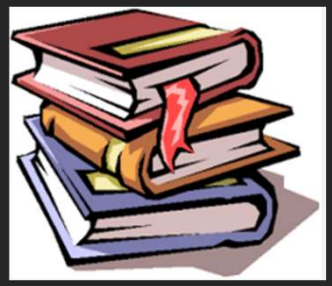
Conclusões sobre os testes de IA

A IA exige um trabalho criativo, interativo e contínuo de avaliação que impulsiona o seu treinamento.

Esse processo criativo não tem atualmente, mecanismos automáticos de validação de respostas e a criação de casos de testes é um trabalho que envolve método, porém não conta com sistemas de previsão de quantidade exata de testes a aplicar, ou conjuntos mínimos e máximos de simulações que garantam uma cobertura adequada de validações.

DÚVIDAS

Referência bibliográficas



BIBLIOGRAFIA :

- RIBEIRO, Marco Tulio et al. - Beyond Accuracy: Behavioral Testing of NLP Models with CheckList (ACL 2020): https://aclanthology.org/2020.acl-main.442
- WILLIAM SULLIVAN - Python Machine Learning Illustrated Guide For Beginners & Intermediates (2018)
- https://github.com/marcotcr/checklist
- https://viso.ai/deep-learning/analyzing-machine-learning-model-performance-strategies/
- https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide
- https://www.linkedin.com/advice/3/how-can-you-ensure-fairness-your-machine-learning-smksf.

TESTE DE SOFTWARE - IA

FIM

PROFESSOR:
RENATO JARDIM PARDUCCI

PROFRENATO.PARDUCCI@FIAP.COM.BR