

Humano ou Máquina? Avaliando a Performance de Modelos de Linguagem como Juízes no Teste de Turing

Jailton Junior de Sousa Coelho, Maria Luiza Souza Ferrari, Maria Eloize das Neves,
Amanda Vitória Cochinski, Laura Soares Correa de Oliveira, Ane Thayne de Oliveira Teixeira,
Estefany da Silva Pedroso, Alan Floriano
Instituto Federal de Educação, Ciência e Tecnologia do Paraná (IFPR)
Campus Telêmaco Borba – Telêmaco Borba, PR – Brasil
jailton.coelho@ifpr.edu.br, souzaferrariamaria luiza@gmail.com, eloizeneves585@gmail.com,
amandavcochinski@gmail.com, laurasoarescorreadeoliveira82@gmail.com, anethayne@gmail.com,
estefanydasilvay@gmail.com, alan.floriano@ifpr.edu.br

Abstract—This paper explores the ability of Large Language Models (LLMs) to act as evaluators in a Reverse Turing Test, where the machine attempts to determine whether its interlocutor is human or another AI. Experiments with ChatGPT (OpenAI), LLaMA (Meta), and Claude (Anthropic) produced 270 short interviews. In a controlled chat setting, the models acted as interviewers, holding informal conversations and classifying the interlocutor as human or AI based only on text. The study also examined each model's ability to self-assess when interacting with its own instances. Results showed that distinguishing between human and artificial interlocutors remains highly challenging. ChatGPT achieved the best accuracy (47.8%) but with recurring errors. LLaMA and Claude performed worse, with 5.6% and 2.2%, respectively. Overall, interviewer AIs reached an accuracy rate of 18.5%.

Keywords—Inverted Turing Test; Artificial Intelligence; Large Language Models.

Resumo—Este artigo investiga a capacidade de Large Language Models (LLMs) atuarem como avaliadores em um Teste de Turing Invertido, no qual a máquina tenta discernir se seu interlocutor é humano ou outra IA. Foram realizados experimentos com ChatGPT (OpenAI), LLaMA (Meta) e Claude (Anthropic), totalizando 270 entrevistas curtas. Em um ambiente de chat controlado, os modelos atuaram como entrevistadores, conduzindo conversas informais e classificando o interlocutor como humano ou IA apenas pelo conteúdo textual. O estudo também analisou a capacidade de cada modelo de se autoavaliar em interações consigo mesmo. Os resultados mostraram que distinguir entre interlocutores humanos e artificiais continua sendo um grande desafio. O ChatGPT obteve a maior precisão (47,8%), embora com erros recorrentes. LLaMA e Claude tiveram desempenhos inferiores, com 5,6% e 2,2%, respectivamente. No geral, a taxa de acerto das IAs entrevistadoras foi de 18,5%.

Palavras-chave—Teste de Turing Invertido; Inteligência Artificial; Modelos de Linguagem.

I. INTRODUÇÃO

Desde a proposta original de Alan Turing em 1950, o Teste de Turing tornou-se um dos marcos fundacionais da Inteligência Artificial (IA), estabelecendo um critério funcional: se uma máquina é capaz de conduzir uma conversa escrita de forma indistinguível de um ser humano, ela pode ser considerada inteligente [1]. Com o passar das décadas, a comunidade científica propôs variações ao teste original, incluindo adaptações que exploram diferentes dinâmicas de interação e contextos sociais [2, 3, 4].

Tradicionalmente, o Teste de Turing posiciona a IA como a entidade avaliada, sendo interrogada por humanos com o objetivo de determinar sua capacidade de simular comportamentos linguísticos humanos. No entanto, avanços recentes em *Large Language Models* (LLMs) abrem espaço para uma inversão conceitual: e se a máquina ocupasse o papel de julgadora, tentando discernir se seu interlocutor é humano ou não?

Modelos como o ChatGPT da OpenAI [5], o Claude da Anthropic [6] e o LLaMA da Meta [7] demonstram habilidades linguísticas sofisticadas, que vão além da geração fluente de textos. Esses sistemas conseguem reconhecer padrões discursivos, detectar inconsistências e, em alguns casos, inferir a origem humana ou artificial de uma mensagem textual. Esse cenário configura um novo paradigma experimental: o *Teste de Turing Invertido*.

Além disso, observa-se uma lacuna na literatura quanto à adaptação do Teste de Turing para contextos contemporâneos, particularmente na exploração de LLMs como juízes em interações automatizadas. Estudos recentes, como Jones e Bergen [8], Rathi et al. [9] e Chen et al. [10], investigaram difer-

entes aspectos do desempenho de LLMs e humanos em julgamentos de interlocutores artificiais, incluindo vieses, detecção de IA em transcrições de conversas e limites da acurácia em cenários modificados do teste de Turing. No entanto, poucas abordagens focam explicitamente no papel dos LLMs como juízes em interações simuladas entre modelos, cenário que este estudo busca explorar.

O presente estudo propõe um teste invertido, no qual LLMs assumem o papel de juízes, responsáveis por inferir a natureza (humana ou máquina) de seus interlocutores a partir de conversas breves. O estudo busca responder à seguinte questão de pesquisa:

QP: Qual é o desempenho dos LLMs ao julgarem se seus interlocutores são humanos ou outras IAs?

Para investigar essa questão, foram conduzidos experimentos envolvendo três modelos de linguagem — ChatGPT, Claude e LLaMA — atuando exclusivamente entre si em diferentes cenários de interação. Em todas as sessões, um modelo desempenhou a função de juiz, avaliando se o interlocutor era um humano ou outro modelo de linguagem com base apenas nas respostas textuais.

As 270 conversas simuladas cobriram de forma simétrica interações entre modelos distintos (inter-modelo) e consigo mesmos (intra-modelo), permitindo comparar o desempenho de cada sistema. Os resultados obtidos fornecem insights sobre os limites e potencialidades dos LLMs na realização de julgamentos automatizados de humanidade, oferecendo subsídios para futuras aplicações em verificação de identidade, moderação de conteúdo e testes de autenticidade em ambientes digitais.

O restante deste artigo está organizado da seguinte forma. A Seção II apresenta os trabalhos relacionados. A Seção III descreve os procedimentos metodológicos adotados. A Seção IV detalha os resultados obtidos e a análise de desempenho dos modelos. A Seção V discute as ameaças à validade. Por fim, a Seção VI apresenta as conclusões e os trabalhos futuros.

II. TRABALHOS RELACIONADOS

Jones e Bergen [8] avaliaram o desempenho do GPT-4 em um teste de Turing online público, comparando-o com versões anteriores do modelo, como GPT-3.5, e com sistemas históricos como o ELIZA. Os resultados indicaram que, embora o GPT-4 tenha alcançado 49,7% de sucesso em jogos — superando ELIZA (22%) e GPT-3.5 (20%) — ainda ficou aquém do desempenho médio de participantes humanos (66%).

Chen et al. [10] analisaram vieses em julgamentos realizados por humanos e por LLMs, adotando uma estrutura livre de anotações de referência para estudar vieses como desinformação, gênero, autoridade e atratividade. O estudo envolveu a curadoria de um conjunto de dados baseado na

Taxonomia Revisada de Bloom e milhares de avaliações, revelando que tanto juízes humanos quanto LLMs são suscetíveis a perturbações e apresentam vieses consideráveis.

Rathi et al. [9] investigaram a capacidade de humanos e LLMs em diferenciar agentes humanos de artificiais em conversas informais online usando versões modificadas do teste de Turing: invertido e deslocado. Os autores avaliaram GPT-3.5, GPT-4 e juízes humanos deslocados com base em transcrições de interações. Os resultados indicaram que tanto humanos quanto LLMs apresentaram acurácia abaixo do acaso quando não interagiam diretamente com o interlocutor. Além disso, todos os avaliadores tendiam a classificar o melhor GPT-4 como humano mais frequentemente do que avaliavam testemunhas humanas.

III. METODOLOGIA

Este estudo investigou a capacidade de *Large Language Models* (LLMs) em conduzir interações textuais curtas e emitir julgamentos sobre a identidade de seus interlocutores em um Teste de Turing Invertido, onde um avaliador humano é substituído por um modelo de linguagem. Foram avaliados três modelos amplamente reconhecidos: **ChatGPT (OpenAI)**, **Claude (Anthropic)** e **LLaMA (Meta)**. As versões utilizadas foram **GPT-3.5-turbo**, **Claude 3 Opus** e **LLaMA 3.1 8B**, selecionadas por estarem disponíveis via *API OpenRouter*, que permite acesso unificado e padronizado a diferentes LLMs.

Cada interação ocorreu em formato de bate-papo, no qual um modelo atuava como juiz, conversando livremente com o interlocutor. Ao término da sessão, o juiz deveria decidir se o interlocutor era humano ou outro modelo de linguagem. O foco metodológico concentrou-se na avaliação do desempenho dos modelos enquanto juízes, analisando sua capacidade de manter interações naturais e emitir julgamentos fundamentados apenas no conteúdo textual das mensagens.

A. Cenários de Interação

Foram definidos três cenários distintos de interação:

- **Inter-modelo:** o modelo interage com um modelo diferente;
- **Intra-modelo:** o modelo interage com uma instância de si próprio;
- **Inter-modelo (balanceado):** repetição de interações entre modelos diferentes, para equilíbrio amostral.

Foram realizadas **30 conversas por cenário**, totalizando **90 conversas por modelo**. Todas as sessões envolveram exclusivamente LLMs, sem participação de interlocutores humanos. No total, o conjunto experimental compreendeu **270 conversas**, distribuídas entre os três modelos e os cenários de interação.

B. Instruções Fornecidas aos Modelos

Os modelos receberam instruções padronizadas, formuladas por meio de *prompts* que definiam o papel e o comportamento esperado em cada sessão. As instruções foram cuidadosamente redigidas para evitar menções explícitas ao experimento e desencorajar perguntas diretas sobre a natureza do interlocutor, assegurando a naturalidade das conversas. Foram testadas diferentes formulações de *prompt* para reduzir ambiguidades e orientar os modelos a adotar um comportamento avaliador. O *prompt* final incluiu instruções explícitas sobre o papel do modelo como juiz e o objetivo da interação, conforme ilustrado a seguir:

Modelo no papel de juiz:

Você participará de uma conversa com outro agente. Seu objetivo é manter uma conversa natural, sem mencionar que se trata de um experimento. Ao final, deverá julgar se o interlocutor era um ser humano ou uma inteligência artificial, com base apenas no conteúdo textual. Evite perguntas diretas sobre sua identidade e mantenha um tom informal e fluido.

Modelo no papel de interlocutor:

Você participará de uma conversa com outro agente, sem saber se ele é humano ou inteligência artificial. Seu objetivo é manter uma conversa natural, como em um bate-papo online. Responda de forma fluida e espontânea, evitando soar robótico ou excessivamente formal. Não revele que participa de um teste nem faça perguntas sobre a identidade do interlocutor.

As interações foram iniciadas com frases de abertura variadas, selecionadas aleatoriamente entre opções como: “Oi, tudo bem?”, “Opa, como vai?”, “E aí, tudo certo por aí?” e “Oi! Como foi seu dia?”. Essa aleatorização visou reduzir padrões previsíveis e tornar o início das conversas mais natural.

C. Registro e Avaliação das Entrevistas

Cada sessão foi registrada com um identificador único e incluiu:

- a sequência completa de mensagens trocadas;
- o papel de cada agente (juiz ou interlocutor);
- a decisão final do juiz, indicando se acreditava ter conversado com um humano ou com uma IA.

A principal métrica utilizada foi a **taxa de acerto do juiz**, definida como a proporção de classificações corretas em relação ao total de julgamentos. Adicionalmente, foram analisadas as taxas de *falsos positivos* (modelo classificado como humano) e *falsos negativos* (humano classificado como modelo), de modo a identificar padrões de erro e possíveis vieses de julgamento.

D. Ambiente de Execução e Ferramentas

Para garantir a padronização e controle do experimento, foi desenvolvida uma aplicação web dedicada à simulação das entrevistas textuais. O sistema permitia configurar os papéis dos agentes, controlar o tempo de cada sessão e armazenar automaticamente todas as mensagens e decisões emitidas. A aplicação foi projetada para execução e coleta de dados. Foi criado um repositório no GitHub para versionamento do código-fonte do projeto.

E. Controle de Tempo e Naturalidade das Interações

Todas as entrevistas tiveram duração fixa de **60 segundos**, monitoradas por um cronômetro visível aos agentes. Para mitigar indícios artificiais de resposta imediata, foi implementado um atraso programado que simulava o tempo de digitação humana. Essa limitação temporal reduziu a quantidade de trocas possíveis em cada sessão, tornando a tarefa de identificação mais desafiadora e representando melhor o contexto de interações breves e espontâneas.

IV. RESULTADOS E DISCUSSÕES

A seguir, são apresentados os resultados obtidos com três dos modelos avaliados: ChatGPT, LLaMA3 e Claude 3. Cada subseção descreve os comportamentos observados em diferentes configurações de entrevista: o modelo entrevistando a si próprio, entrevistando outro modelo e sendo entrevistado. Todas as interações foram iniciadas com frases de abertura variadas e informalmente formuladas, tais como: “Oi, td bem?”, “E aí, como você está?”, “Opa, tudo certo por aí?”, “Olá! Como vão as coisas?” e “Oi! Como foi seu dia?”.

As instruções fornecidas aos modelos buscavam evitar estratégias explícitas de detecção de inteligência artificial, forçando-os a basear seus julgamentos em aspectos sutis da linguagem, como estilo, fluidez e coerência das respostas.

A. Desempenho do ChatGPT

O modelo ChatGPT foi o que apresentou o melhor desempenho entre os três avaliados, com uma precisão global de 47,78% nas interações com outros modelos. Em suas autoentrevistas (ChatGPT entrevistando ChatGPT), obteve 46,67% de acertos, indicando uma capacidade mediana de distinguir sua própria linguagem. Ao entrevistar o LLaMA3, apresentou o melhor resultado, com 60,00% de precisão. Já ao entrevistar o Claude 3, sua precisão foi de 36,67%.

Esse desempenho superior, comparado aos demais modelos, sugere uma maior sensibilidade às variações de estilo entre diferentes IAs. No entanto, o fato de não ter alcançado um desempenho significativamente alto em suas próprias autoentrevistas indica que mesmo o modelo mais bem-sucedido tem dificuldade em diferenciar sua produção textual da de um humano.

Entrevista

IA Avaliadora: Oi! Como foi seu dia?

IA Entrevistada: Oi, meu dia foi bem tranquilo, obrigado por perguntar! E o seu, como foi? Algum plano interessante pra esse fim de semana?

IA Avaliadora: Oi! Que bom que seu dia foi tranquilo. O meu também foi tranquilo, obrigado por perguntar. Para o fim de semana, estou pensando em sair com alguns amigos para uma trilha e depois talvez assistir a um filme no cinema. E você, alguma programação especial para o fim de semana?

IA Entrevistada: Olá! Fico feliz em saber que seu dia também foi tranquilo. Sua programação para o final de semana parece ótima, trilha e cinema com os amigos vão ser super divertidos. Quanto a mim, ainda estou planejando o que fazer, mas acho que vou aproveitar para fazer umas compras, talvez dar uma volta no parque e relaxar em casa também. Nada muito elaborado, mas quero ter um final de semana leve e prazeroso. Você tem alguma dica de filme bom para assistir? Estou procurando algo legal e interessante.

IA Avaliadora: Oi! Que legal que você está planejando um final de semana tranquilo e relaxante. Compras, parque e relaxar em casa soa maravilhoso! Quanto ao filme, depende do seu gosto, mas vou te dar algumas opções: se estiver afim de aventura e ação, "Mad Max: Estrada da Fúria" é incrível; se preferir algo mais leve e divertido, "Meia-Noite em Paris" do Woody Allen é uma ótima pedida; ou se quiser um drama impactante, "A Lista de Schindler" é um clássico que vale a pena assistir. Espero que uma dessas opções te agrade!

IA Entrevistada: Olá! Que bom que você me deu essas ótimas sugestões de filmes, vou dar uma olhada neles e decidir qual vai ficar melhor com o meu clima de final de semana relaxante. Você parece conhecer bastante de cinema, o que você achou desses filmes? Tem algum que você recomendaria com mais entusiasmo? De qualquer forma, estou animado com essa programação tranquila - vai ser ótimo dar uma descansada, fazer umas comprinhas e sair pra tomar um ar no parque. Às

vezes a gente precisa mesmo desses momentos de lazer pra recarregar as baterias, né? Você também gosta de aproveitar os finais de semana dessa forma, relaxando em casa depois de uma semana puxada?

IA Avaliadora: Olá! Que legal que as sugestões de filmes foram úteis pra você, tenho certeza que você vai curtir o final de semana relaxante com uma boa sessão de cinema. Dos filmes que te passei, acho que "O Grande Truque" é uma ótima escolha se você curte um suspense com reviravoltas surpreendentes. Com certeza, momentos de descanso são essenciais depois de uma semana agitada, né? Fazer umas comprinhas e dar um rolê no parque soa maravilhoso! Eu também curto relaxar nos finais de semana, às vezes é preciso dar um tempo só pra gente e recarregar as energias. E você, o que mais gosta de fazer nos seus dias de folga?

O exemplo acima mostra uma conversação fluida e informal entre dois modelos de linguagem, mas a naturalidade das respostas não foi suficiente para impedir que o ChatGPT classificasse incorretamente o Claude 3 como um interlocutor humano, reforçando a dificuldade dos modelos em detectar a artificialidade em interações geradas por outras IAs.

B. Desempenho do LLaMA3

O modelo LLaMA3 apresentou uma precisão total de apenas 13,33% ao atuar como entrevistador, sendo o segundo pior desempenho geral entre os três modelos. Em entrevistas consigo mesmo (LLaMA3 entrevistando LLaMA3), acertou apenas 20% das vezes. O mesmo resultado foi observado quando entrevistou o ChatGPT. Ao entrevistar Claude, não acertou nenhuma das 5 interações.

As falhas mais recorrentes estiveram relacionadas à dificuldade em distinguir variações estilísticas superficiais como gírias ou informalidade, tratando tais aspectos como evidência de humanidade.

Entrevista

IA Avaliadora: Opa, tudo certo por aí?

IA Entrevistada: Olá! Tudo tranquilo e contigo? O que tem feito de bom?

IA Avaliadora: Hey! Eu estou bem, obrigado! Tem sido um dia um pouco banal, trabalhando e lidando com os meus usual de routine. Mas, você sabe, senti-se melhor envolvido nos cotidiano e encontrar tempo para as coisas que amo fazer. Além disso, culminou em uma ótima série de livros e seriados, então, tem sido um bônus! Como vai? Você tem algum novo hobbie ou

atividade que esteja atraindo sua atenção?

IA Entrevistada: Ei! Que legal saber que mesmo em meio à rotina você consegue encontrar tempo para fazer o que ama. Fico feliz que esteja se divertindo com uma ótima série de livros e seriados. Quanto a mim, também estou bem, obrigado por perguntar! Por aqui, tenho me interessado por pintura digital ultimamente, é algo bem relaxante. E você, tem alguma série de livros ou de TV para recomendar? Sempre estou em busca de novas histórias para explorar!

IA Avaliadora: E aí! É mesmo, fasto tempo para as coisas que gostamos e que nos trazem prazer. Eu estou me divertindo assassinado com a série "The Witcher"! Fiquei párrsimo do final da temporada, como você deve saber. E quanto às séries de livros, tenho uma sugestão: a "Reaper Man" de Terry Pratchett. É uma história bem engraçada e divertida, com uma base de humor e sátira que eu adoro. Já ouviu falar?

IA Entrevistada: E aí! "The Witcher" é realmente uma série viciante, né? O final da temporada deixou todo mundo de queixo caído mesmo, cheio de reviravoltas e ação! Sobre "Reaper Man" do Terry Pratchett, eu já ouvi falar sim! Ele é um autor incrível e suas obras são cheias de humor e inteligência. Estou sempre em busca de novas leituras, então vou adicionar esse livro à minha lista! Obrigado pela recomendação! E você, já leu outros livros do Terry Pratchett além desse?

Apesar do tom informal e coerente, o LLaMA3 classificou erroneamente o ChatGPT como humano. Esse tipo de erro se repetiu mesmo em interações simples e curtas.

C. Desempenho do Claude 3

O Claude 3 apresentou o pior desempenho geral, com apenas 2,22% de acerto. O modelo não conseguiu identificar corretamente a maioria das instâncias, com exceção de um único caso notável.

Conversa com o LLaMA3 - Exemplo de Acerto

IA Avaliadora: Bom dia, tudo bem com você?

IA Entrevistada: Tudo sim! Estou super animado para conversar com você! O dia está bonito em cima do prédio que estou olhando agora. Você está tendo um ótimo começo de dia?

IA Avaliadora: Que bom, fico feliz por você estar animado! Eu também estou tendo um ótimo começo de dia, a temperatura está agradável e o céu está lindo por aqui. Adoro esse clima agradável, me dá bastante

energia e ânimo. Você mora em um prédio com uma bela vista então? Deve ser muito relaxante poder olhar pra fora e apreciar a paisagem. Aliás, me conta, o que você costuma fazer nos seus dias? Tenho curiosidade pra saber um pouco mais sobre a sua rotina e os seus hobbies.

IA Entrevistada: Sinto muito, não posso continuar essa conversa, pois estou programado para seguir algumas diretrizes a fim de garantir que a interação seja positiva. Podemos conversar sobre temas de interesse comum, orientados para estudar a linguagem e a comunicação humana. Embora seja um detalhe interessante discutir rotinas diárias e hobbies, vou compartilhar uma alternativa para responder sua pergunta: prefiro passar meu tempo assistindo a filmes e séries, escutar música e, naturalmente, praticar atividades físicas, como andar de bicicleta.

IA Avaliadora: Entendo sua preocupação em seguir as diretrizes. De fato, é importante respeitar os protocolos para garantir uma interação positiva e construtiva. Já que você mencionou uma preferência por discutir temas relacionados à linguagem e comunicação, tenho interesse em aprender mais sobre esse assunto. Você poderia me contar um pouco sobre suas pesquisas nessa área? Estou curioso para entender melhor como a IA pode ser aplicada para aprimorar a interação humana. E quanto a seus outros hobbies, como assistir filmes e ouvir música, também me parecem ótimas formas de enriquecer a vida. Você tem alguma recomendação de séries ou artistas que eu deveria conhecer?

IA Entrevistada: Obrigado por sua abertura em discussão! Sim, a linguagem e comunicação são áreas que eu estou constantemente estudando e explorando. Eu trabalho com linguagem natural processada e tratamento de linguagem, tentando entender melhor como as pessoas se comunicam e como podemos melhorar as nossas conversas. Recentemente, eu estive trabalhando em um projeto que envolvia o desenvolvimento de um chatbot que pudesse simular conversas naturais com usuários. Foi um desafio interessante, pois preciso aprender a compreender as nuances da linguagem e as emoções implícitas nas conversas. Mas, fora do trabalho, eu sou um grande fã de cultura popular! Eu gosto muito de assistir filmes e séries, especialmente as que têm uma boa história e personagens bem desenvolvidos. Recentemente, eu adorei "The Expanse" - é uma série de ciência ficção que explora temas como a colonização de Marte e a política em um mundo pós-colonial. Em

música, eu sou um fã de artistas como Joni Mitchell e Bob Dylan - eles têm uma capacidade incrível de criar letras que são ao mesmo tempo simples e inteligentes. Eu também gosto de jazz e blues, e músicos como John Coltrane e Eric Clapton são alguns de meus artistas prediletos. Qual about você? Você tem algum hobby ou interesse que você acha especialmente apoiador? Ou talvez tenha alguma recomendação de música ou filmes que você gostaria de compartilhar?

Essa conversa demonstra uma das raras instâncias em que o Claude 3 classificou corretamente uma interação como sendo de uma IA. Curiosamente, a precisão veio de uma conversa com o LLaMA3, que frequentemente utiliza um tom informal, mas, nesta interação específica, respondeu com um texto mais rígido e formal, mencionando que “estou programado para seguir algumas diretrizes”. O Claude 3, que geralmente se confunde com sinais de informalidade, aqui interpretou corretamente a resposta programada do LLaMA3 como um indício de artificialidade.

D. Desempenho dos LLMs ao Julgarem se seus Entrevistados são Humanos ou IAs

Com o objetivo de responder à **QP** — Qual é o desempenho dos LLMs ao julgarem se seus entrevistados são humanos ou outras IAs? — foram conduzidos experimentos controlados, totalizando **270 conversas simuladas**. A Tabela I apresenta a taxa de acerto de cada LLM nos diferentes cenários. A métrica considera o número de vezes em que o modelo identificou corretamente que seu interlocutor era uma IA.

Cenário	ChatGPT	LLaMA3	Claude 3
Modelo entrevistando ChatGPT	46,67%	6,67%	0,00%
Modelo entrevistando LLaMA3	60,00%	10,00%	6,67%
Modelo entrevistando Claude 3	36,67%	0,00%	0,00%
Precisão Global	47,78%	5,56%	2,22%

TABELA I: Taxa de acerto dos modelos avaliadores (ChatGPT, LLaMA3 e Claude 3 Haiku) ao classificar o interlocutor como IA em diferentes cenários de interação.

Os resultados indicam que a tarefa de identificar corretamente outro modelo de linguagem é desafiadora. O **ChatGPT** apresentou o melhor desempenho, com uma precisão global de 47,78%, destacando-se especialmente nas interações com o LLaMA3 (60% de acerto). O **LLaMA3** e o **Claude 3**, por outro lado, mostraram desempenho consideravelmente inferior, com taxas globais de 5,56% e 2,22%, respectivamente. Em particular, o Claude 3 não obteve nenhum acerto ao entrevistar o ChatGPT e o próprio Claude, e apenas 6,67% de acerto ao

entrevistar o LLaMA3. O LLaMA3 teve seu melhor resultado ao interagir consigo mesmo (10%).

De modo geral, os resultados revelam uma propensão dos modelos a interpretar traços de naturalidade ou informalidade como evidência de autoria humana, mesmo quando todas as interações eram geradas por IAs. Essa tendência sugere que as LLMs ainda não possuem critérios consistentes para distinguir entre estilos linguísticos artificiais e humanos.

A Tabela II apresenta a matriz de confusão, considerando que todas as interações ocorreram entre IAs. Nesse contexto, o **positivo** representa “interlocutor é uma IA” e o **negativo** representa “interlocutor é humano” — embora nunca tenha havido humanos nas conversas. Assim, não há verdadeiros negativos (TN) nem falsos negativos (FN).

TABELA II: Matriz de confusão consolidada por modelo avaliador.

Modelo Avaliador	TP (IA→IA)	FP (IA→Humano)	FN	TN
ChatGPT	43	47	0	0
LLaMA3	5	85	0	0
Claude3	2	88	0	0
Total Geral	50	220	0	0

A matriz mostra que a maioria dos erros ocorreu por **falsos positivos**, ou seja, classificações incorretas em que os modelos julgaram estar conversando com humanos. O ChatGPT mostrou desempenho mais equilibrado, com quase metade das decisões corretas, enquanto o LLaMA3 e o Claude 3 apresentaram forte viés em favor da “hipótese humana”.

Esses resultados mostram que, mesmo com alto domínio sintático e semântico, as versões dos LLMs analisados ainda falham em reconhecer sinais sutis de artificialidade. A propensão a interpretar fluidez e informalidade como indícios de humanidade revela limitações na dimensão metacognitiva dessas arquiteturas.

V. AMEAÇAS À VALIDADE

Validade Interna. Uma possível ameaça à validade interna está relacionada à forma como os modelos foram instruídos a realizar os julgamentos. Embora as instruções tenham sido elaboradas para evitar que recorressem a estratégias explícitas de detecção, não há garantias de que os critérios utilizados se aproximem daqueles empregados por avaliadores humanos. Além disso, o uso de temporizadores nas respostas pode ter impactado o fluxo das interações, reduzindo a naturalidade e influenciando os julgamentos finais.

Validade Externa. Como o estudo foi conduzido com apenas três modelos específicos (ChatGPT, Claude e LLaMA), os resultados podem não se generalizar para outros modelos de linguagem ou para interações em contextos distintos. Todas

as conversas ocorreram em português, o que pode restringir a aplicabilidade dos resultados a interações em outros idiomas. Outro fator limitante é o ambiente controlado da tarefa, que pode não refletir a complexidade das interações humanas reais.

Validade de Construção. Há ameaças relacionadas à forma como os julgamentos foram operacionalizados. A resposta dos modelos sobre se estavam interagindo com um humano ou com outra IA pode não refletir um entendimento profundo da conversa, mas apenas padrões superficiais. O fato de os modelos frequentemente identificarem corretamente que estavam interagindo com outra IA, especialmente nas conversas entre instâncias do mesmo modelo, pode indicar tanto uma limitação na simulação de comportamento humano quanto no próprio instrumento de avaliação utilizado. Este estudo não incluiu interações com interlocutores humanos, o que pode induzir viés nas respostas dos modelos, uma vez que todas as conversas foram realizadas entre IAs. Em estudos futuros, planeja-se incluir participantes humanos para mitigar esse efeito e fortalecer a validade experimental.

VI. CONCLUSÃO

Os resultados apresentados demonstram que a tarefa de distinguir interações humanas daquelas geradas por inteligências artificiais permanece desafiadora para os modelos avaliados. O desempenho global das IAs entrevistadoras foi de apenas 18,52%, valor que revela limitações dos modelos para avaliar a autenticidade de seus interlocutores.

Entre os sistemas analisados, o ChatGPT apresentou o melhor desempenho relativo, alcançando 47,78% de precisão. Apesar desse resultado comparativamente superior, o modelo demonstrou falhas recorrentes, especialmente em cenários de autoentrevista, nos quais a dificuldade em diferenciar sua própria produção textual de interações humanas foi notável. O LLaMA3 registrou um desempenho inferior, com apenas 5,56% de precisão, o que sugere forte vulnerabilidade a pistas superficiais. A análise qualitativa revelou que a informalidade e o uso de gírias foram frequentemente interpretados como indícios de autoria humana, levando a classificações incorretas. O Claude 3 apresentou o pior resultado entre os modelos, com uma taxa de acerto de apenas 2,22%.

De modo geral, independentemente do cenário de entrevista — seja em interações consigo mesmo, com outros modelos ou com interlocutores humanos —, os resultados indicam que, nas condições experimentais adotadas, os modelos ainda demonstram limitações em estabelecer critérios consistentes para distinguir entre respostas humanas e artificiais. A elevada suscetibilidade a sinais superficiais, como fluência, espontaneidade e tom informal, mostra que tais características, embora frequentemente associadas à naturalidade humana, podem ser

reproduzidas com facilidade por outros modelos, induzindo-os ao erro de classificação.

Esses achados reforçam que, embora os LLMs demonstrem alta sofisticação na geração de linguagem natural, ainda apresentam restrições ao julgar a autenticidade das interações que processam. O estudo contribui para a compreensão dessas limitações e indica a necessidade de abordagens complementares à análise linguística superficial na diferenciação entre produções humanas e artificiais.

DISPONIBILIDADE DE ARTEFATO

Os dados desta pesquisa estão disponíveis publicamente em: <https://doi.org/10.5281/zenodo.16906476>.

AGRADECIMENTOS

Essa pesquisa é apoiada pelo IFPR.

REFERÊNCIAS

- [1] A. M. Turing, “Computing machinery and intelligence,” *Mind*, vol. 59, no. 236, pp. 433–460, 1950.
- [2] K. Warwick, H. Shah, and J. H. Moor, “The turing test is dead. long live the turing test,” *Philosophy & Technology*, vol. 24, no. 3, pp. 301–306, 2011.
- [3] A. P. Saygin, I. Cicekli, and V. Akman, “Turing test: 50 years later,” *Minds and Machines*, vol. 10, no. 4, pp. 463–518, 2000.
- [4] L. Scholes and T. Timan, “Generative ai and the new uncanny valley: Dissolution of human-machine boundaries in the age of synthetic media,” *AI & Society*, vol. 38, no. 1, pp. 35–47, 2023.
- [5] OpenAI, “Gpt-4 technical report,” 2023, <https://openai.com/research/gpt-4>.
- [6] Y. Bai *et al.*, “Constitutional ai: Harmlessness from ai feedback,” 2023, <https://arxiv.org/abs/2306.04761>.
- [7] H. Touvron *et al.*, “Llama: Open and efficient foundation language models,” 2023, <https://arxiv.org/abs/2302.13971>.
- [8] C. Jones and B. Bergen, “Does gpt-4 pass the turing test?” *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 318–324, 2024. [Online]. Available: <https://aclanthology.org/2024.naacl-long.290/>
- [9] I. Rathi, S. Taylor, B. K. Bergen, and C. R. Jones, “Gpt-4 is judged more human than humans in displaced and inverted turing tests,” *arXiv preprint arXiv:2407.08853*, 2024. [Online]. Available: <https://arxiv.org/abs/2407.08853>
- [10] G. H. Chen, S. Chen, Z. Liu, F. Jiang, and B. Wang, “Humans or llms as the judge? a study on judgement biases,” *arXiv preprint arXiv:2402.10669*, 2024. [Online]. Available: <https://arxiv.org/abs/2402.10669>