UNIVERSIDADE FEDERAL DE SANTA CATARINA
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA

Vitor Lorizete de Lima Filho

Análise da Qualidade da Avaliação Automatizada de Criatividade de Aplicativos App Inventor com Modelos Grandes de Linguagens

Vitor Lorizete de Lima Filho

Florianópolis
2023

ANÁLISE DA QUALIDADE DA AVALIAÇÃO AUTOMATIZADA DE CRIATIVIDADE DE APLICATIVOS APP INVENTOR COM MODELOS GRANDES DE LINGUAGENS

Trabalho de conclusão de curso apresentado como parte dos requisitos para obtenção do grau de Bacharel em Ciências da Computação.
Orientador(a): Prof. Dr. rer. nat. Christiane A. Gresse von Wangenheim, PMP

Florianópolis
2023

AGRADECIMENTOS

Primeiramente aos meus pais, meus sinceros agradecimentos por tudo que fizeram por mim, não apenas durante o período da minha faculdade, mas durante minha vida inteira. Agradeço especialmente pelo apoio e amor incondicional em todas as fases da minha vida. E por terem me ajudado a nunca desistir mesmo em frente a todos os obstáculos enfrentados.

À minha orientadora, Christiane, sou muito grato por ter me aceitado como orientando e por não ter medido esforços em me ajudar em tudo o que eu precisava. Obrigado por todas as sugestões, ajuda e paciência comigo. Um agradecimento em especial por mesmo em momentos difíceis nunca deixar de comparecer para ajudar em qualquer coisa que eu precisava.

Agradeço aos membros da banca, Nathalia e Adriano, muito obrigado por terem aceitado o convite e por terem me orientado em tudo que precisei no decorrer do projeto inteiro.

A todos os amigos que fiz durante essa jornada, sem vocês a faculdade teria sido muito mais difícil e desgastante. Um agradecimento em especial a Beatriz, Manu e Mateus, que foram meus companheiros durante toda essa reta final.

Agradeço a Universidade Federal de Santa Catarina, por ter sido o palco de toda essa jornada e por tudo que contribui para a formação da pessoa que sou hoje.

RESUMO

O século XXI requer novas habilidades, como a criatividade, que são cada vez mais valorizadas para um desenvolvimento pessoal e no mercado de trabalho. Assim, a importância do incentivo da criatividade na educação básica é cada vez mais reconhecida. Uma das formas de desenvolver essa habilidade é o ensino de computação por meio de desenvolvimento de aplicativos móveis. Atualmente já existem diversas iniciativas voltadas a esse fim, porém observa-se em geral a falta de modelos para avaliar a criatividade de forma válida e eficiente, que possam ser facilmente aplicados na prática. Uma alternativa para automatizar a avaliação da aprendizagem de criatividade pode ser o uso de *Large Language Models* (LLMs), que, entre outras tarefas, também podem avaliar a criatividade de apps. Porém, atualmente ainda não existem estudos avaliando a qualidade das avaliações de criatividade de apps pelos LLMs. Assim, visa-se neste trabalho comparar a qualidade da avaliação automatizada de criatividade de aplicativos App Inventor com LLMs com avaliações feitas por juízes humanos e o modelo automatizado Creassessment. Assim espera-se obter conhecimento sobre a possibilidade em utilizar LLMs para esta tarefa apoiando a aprendizagem de criatividade como parte do ensino da computação em escolas no Brasil.

Palavras-chave: Criatividade, Ensino de computação, Avaliação, Educação Básica, *Large Language Model*

SUMÁRIO

1 Introdução

1.1 Contextualização

Criatividade é um substantivo que vem do latim creare, que indica a capacidade de criar, produzir ou inventar coisas novas. Na educação de jovens, o desenvolvimento da criatividade é fundamental (P21, 2021)(WEF, 2023). Nessa fase, é essencial estimulá-la uma vez que isso pode ajudá-los a se tornarem mais curiosos, imaginativos e a encontrar soluções inovadoras para problemas e desafios que enfrentam no dia a dia. Além disso, a criatividade também ajuda no desenvolvimento de habilidades essenciais para a vida, como pensamento crítico, resolução de problemas e adaptação às mudanças (P21, 2021)(WEF, 2023). Ao desenvolver a criatividade desde cedo, os estudantes estarão melhor preparados para enfrentar as demandas do mundo atual. Além disso, a criatividade é uma habilidade cada vez mais valorizada no mercado de trabalho atual, empresas que valorizam a criatividade tendem a buscar profissionais com essa habilidade, pois eles são capazes de contribuir de forma significativa para a inovação e o crescimento da organização. Portanto, investir no desenvolvimento da criatividade já nas escolas pode trazer benefícios não só para a vida pessoal dos estudantes, mas também para suas carreiras profissionais.

Existem várias formas de desenvolver e estimular a criatividade em jovens, incluindo como uma alternativa o ensino de computação (Saunders e Thagard, 2005) por meio de desenvolvimentos de aplicativo móveis com App Inventor (Harunani, Patton e Tissenbaum, 2019). A criação e desenvolvimento de aplicativos é uma forma de estimular a criatividade, pois a mesma envolve diferentes competências fazendo com que os estudantes pensem e construam soluções novas para problemas recorrentes (Mishra e Yadav, 2013). Ao criar um aplicativo, o aluno demonstra não apenas o conhecimento adquirido, mas também a capacidade de aplicá-lo de forma criativa (O’quin, 1998).

Já existem algumas iniciativas para estimular a criatividade do estudante por meio do ensino de desenvolvimento de aplicativos móveis (Alves, 2023)(MIT App Inventor). Porém ainda há uma lacuna em relação à avaliação da aprendizagem de criatividade neste contexto. Uma forma de avaliar o desenvolvimento da criatividade

dos alunos é por meio da avaliação do desempenho com base em artefatos desenvolvidos como resultado da aprendizagem, como uma forma eficiente e autêntica de medir o progresso e a habilidade do aluno (Hattie, 2007). Dessa forma, no contexto de ensino por meio do desenvolvimento de aplicativos, o desenvolvimento da criatividade pode ser avaliado por meio da avaliação da criatividade dos aplicativos criados pelos estudantes como resultado da aprendizagem.

Tal avaliação pode ser feita de diferentes maneiras. A forma mais convencional é manualmente por humanos simplesmente avaliando a criatividade como um fator, usando rubricas ou *checklists*. Um exemplo dessa abordagem pode ser visto na competição App Inventor *App of the Month*, em que especialistas avaliam a criatividade dos aplicativos criados com o App Inventor. Esses especialistas avaliam os aplicativos com base em uma rubrica com um critério voltado a criatividade, definindo os níveis de desempenho considerando a originalidade do conceito e a relevância para o mundo real (https://appinventor.mit.edu/explore/app-month-gallery). Embora a avaliação humana seja uma abordagem eficaz, ela pode ser demorada e cara, especialmente quando se trata de avaliar grandes quantidades de aplicativos.

Uma alternativa é a avaliação automatizada, como p.ex. o modelo de avaliação Creassessment (Alves, 2023), que com base em um modelo conceitual avalia aplicativos criados com *App Inventor* em termos de originalidade, fluência e flexibilidade em conformidade com a definição de criatividade. Com o objetivo de uma avaliação abrangente, essas dimensões são avaliadas em relação à originalidade dos componentes da interface do usuário, das funcionalidades e tópicos do aplicativo, bem como dos comandos da programação do app.

Além de modelos automáticos de avaliação, recentemente emergiu uma nova possibilidade de avaliação de criatividade de apps, por meio de *Large Language Models* (LLMs), que fornecendo informações sobre o aplicativo a LLM pode avaliar o mesmo. LLMs são modelos de linguagem computacionais que utilizam técnicas de aprendizado de máquina para entender e gerar texto em linguagem natural (Birhane et al., 2023). Esses modelos são treinados com grandes quantidades de dados de diversas bases para aprender as regras da linguagem e as relações entre as

palavras. Com base nesse conhecimento, eles podem gerar um novo texto, responder perguntas, traduzir idiomas e realizar outras tarefas linguísticas como avaliar a criatividade de um aplicativo. Atualmente as LLMS mais proeminentes são o ChatGPT (Brown et al., 2020) e suas variantes. Porém, atualmente ainda não existem estudos avaliando a qualidade de avaliações de criatividade de apps por LLMs.

Assim, a pergunta de pesquisa deste TCC é: Qual a qualidade da avaliação automatizada de criatividade de aplicativos App Inventor com LLMs comparado com avaliações realizadas por juízes humanos e/ou o modelo automatizado *Creassessment*?

1.2 Objetivos

Objetivo Geral

O objetivo geral deste trabalho é avaliar a qualidade da avaliação automatizada de criatividade de aplicativos App Inventor com LLMs no contexto do ensino de computação na educação básica. São incluídos modelos grandes de linguagem populares como ChatGPT e também modelos de código aberto. A qualidade da avaliação com estes modelos é comparada com avaliações realizadas por juízes humanos e/ou o modelo automatizado *Creassessment*.

Objetivos Específicos

Os objetivos específicos deste trabalho são:

O1. Sintetizar a fundamentação teórica relacionada à avaliação da criatividade no contexto do ensino de computação na educação básica, e LLMs.
O2. Levantar o estado da arte em relação a avaliação de LLMs para a avaliação da criatividade com LLMs no contexto de ensino da computação na educação básica.
O3. Realizar estudos empíricos para avaliar a qualidade de avaliação da criatividade com LLMs.

Delimitações

Como premissas desse projeto tem-se que a presente pesquisa é realizada ao final de cursos de computação do projeto Computação na Escola. Assim como o projeto será realizado de acordo com o regulamento vigente do Departamento de Informática e Estatística da UFSC em relação ao trabalho de conclusão de curso.

O escopo do presente trabalho é voltado à avaliação da criatividade, não englobando a avaliação a aprendizagem de outras competências. Também se limita a avaliação da criatividade no contexto do ensino de computação pelo desenvolvimento de apps com *App Inventor*.

1.3 Metodologia

Neste trabalho é aplicada a metodologia multi-método, dividida nas seguintes etapas:

Etapa 1. Síntese dos conceitos fundamentais e teóricos com base na literatura: Nesta etapa é analisado e definido o que é criatividade e LLMs com base na análise da literatura.

A1.1 Sintetizar conceitos sobre a avaliação da criatividade no ensino de computação na educação básica.

A1.2 Sintetizar conceitos sobre LLMs.

Etapa 2. Levantamento do estado da arte: Nesta etapa é realizado o mapeamento sistemático existente seguindo o procedimento proposto por Petersen et al. (2008) para identificar estudos de avaliação de criatividade existentes que sejam aplicados no ensino da computação na educação básica. Na primeira etapa é definido o protocolo de pesquisa. A segunda etapa, de execução da pesquisa, consiste em identificar os trabalhos existentes relacionados. Selecionado os trabalhos relevantes, são extraídas as informações relevantes do trabalho encontrado. Estas informações são analisadas em relação às perguntas de análise.

A2.1 Definir o protocolo da revisão da literatura.

A2.2 Executar busca.

A2.3 Extrair a informação dos trabalhos relevantes.

A2.4 Analisar e interpretar a informação.

Etapa 3. Realização de um estudo empírico para avaliação da qualidade das avaliações dos LLMs: Nesta etapa é analisada a qualidade das avaliações feitas pelos LLMs em comparação com avaliações feitas por humanos e modelos automatizados (*Creassessment* (Alves, 2023)). Este estudo é realizado por meio de um estudo de caso (Yin, 2009)(Wohlin et al., 2016). O estudo é definido em termos do *research design* e as perguntas de análise. É definido o escopo dos aplicativos sendo utilizados na avaliação, incluindo os LLMs. Os métodos de coleta de dados serão definidos e durante a execução da avaliação os dados serão coletados conforme definidos. Após a coleta, os dados são avaliados de forma estatística, comparando a qualidade das diferentes formas de avaliação.

A3.1 Definir o estudo.

A3.2 Definir o(s) prompts para os LLMs para coleta de dados.

A3.3 Executar a avaliação e coletar dados.

A3.4 Análise e interpretação dos dados coletados.

1.4 Estrutura do documento

Para garantir uma compreensão aprofundada e atingir plenamente os objetivos deste trabalho, foi adotada uma estrutura dividida em capítulos interligados de forma complementar. O Capítulo 2 apresenta uma fundamentação teórica, explorando conceitos essenciais relacionados ao objetivo do trabalho, como a criatividade, o *App Inventor* e os modelos grandes de linguagens (LLM). Em seguida, o Capítulo 3 apresenta o estado da arte em relação à qualidade da avaliação da criatividade de artefatos com LLMs. No Capítulo 4 é apresentado o estudo de casos do trabalho. No Capítulo 5 é apresentada a conclusão do trabalho.

2. Fundamentação teórica

2.1 Avaliação da criatividade no ensino de computação na educação básica

No contexto da avaliação da criatividade, é fundamental compreender a definição e a importância dessa habilidade. Em relação à importância, a criatividade vem sendo considerada uma das principais competências do século XX, sendo ela essencial para o sucesso tanto profissional, quanto pessoal do ser humano (Kaufman; Beghetto, 2009). Com isso torna-se evidente a necessidade de promover cada vez mais o ensino e evolução da criatividade desde a educação básica (Kaufman; Beghetto, 2013). Em relação a uma definição concreta da criatividade, não se é algo fácil de obter, por ser uma habilidade um tanto quanto subjetiva, existe uma dificuldade em estabelecer um significado simples e fixo a mesma (Walia, 2019). No entanto, existe uma definição amplamente aceita por especialistas que consideram que a criatividade engloba a capacidade de gerar ideias e soluções inovadoras para problemas, tarefas ou projetos. De forma geral, define-se a criatividade de um artefato considerando as dimensões de (Guilford, 1950): fluência, flexibilidade e originalidade. A fluência refere-se à quantidade de ideias e pensamentos geradas para um determinado conteúdo, evidenciando sua capacidade de pensar e gerar novas ideias, uma faceta facilmente quantificável e mensurável (Renzulli, 2018). A flexibilidade pode se resumir à adaptabilidade de uma pessoa, ou nesse caso do aluno, em suas abordagens ao enfrentar um desafio, sendo a capacidade de alternar entre diferentes métodos, perspectivas e/ou estratégias para resolver esse desafio, podendo assim ter mais de uma abordagem para chegar a uma solução (Renzulli, 2018). A originalidade destaca-se quando as soluções, ou respostas dos alunos, destacam-se por serem diferentes do convencional e pelo elemento surpresa que podem gerar, podendo uma ideia fomentada de originalidade ser caracterizada de várias formas, como incomum, rara, singular, inusitada e/ou simplesmente original (Renzulli, 2018). Porém, uma ideia original além de se destacar pela sua originalidade, precisa ter utilidade e ter eficácia em diferentes aspectos, para assim ser considerada criativa (Runco e Jaeger, 2012).

Segundo Rhodes (1961), criatividade pode ser observada por meio do

Produto, a qual refere-se aos artefatos tangíveis produzidos utilizando do processo criativo, por exemplo, livros, arte ou programas de computador, como os apps, entre outras formas como a criatividade de Pessoa, de Processo ou de Ambiente. Neste contexto, a criatividade de Produto é a que mais se assemelha ao foco do presente trabalho, a qual se concentra na capacidade de gerar ideias e artefatos originais e úteis, sendo assim a criação de algo novo que tenha valor e utilidade prática, podendo por meio desses artefatos solucionar problemas de formas criativas, sendo um desses possíveis artefatos os apps criados com o *App Inventor*.

2.1.1 Avaliação de criatividade de aplicativos no ensino de computação

Um dos métodos conhecidos para medir o nível de criatividade dos aplicativos é a avaliação por humanos. Conhecido como *Consensual Assessment Technique* (CAT, (Amabile, 1996; Alves, 2023)), método no qual um grupo de especialistas avalia os artefatos produzidos. A avaliação humana da criatividade pode ser feita manualmente por meio de uma simples pergunta (o *App* é criativo?) ou por meio de rubricas ou checklists. A avaliação por meio de rubricas consiste em uma série de critérios com níveis de desempenho bem definidos a qual cada artefato será julgado. Por exemplo, na competição App Inventor *App of the Month* (https://appinventor.mit.edu/explore/app-month-gallery), especialistas avaliam a criatividade dos aplicativos criados com o App Inventor a partir de uma série de categorias conforme apresentado na Tabela 1.

Tabela 1 - Categorias de avaliação *App of the Month*

Critério	**Explicação**
Best Design	O aplicativo esteticamente mais agradável.
Most Innovative	O aplicativo que usa a tecnologia *App Inventor* da maneira mais interessante e única.
Most Creative	o aplicativo que melhor utiliza elementos criativos, como arte, cor, som ou movimento.
App Inventor of the Month	Um *App Inventor* que demonstra grande potencial e criatividade

Fonte: https://appinventor.mit.edu/explore/app-month-gallery

Essa forma de avaliação de criatividade de artefatos pode exigir muito em relação a recursos, necessitando de uma série de avaliadores e tempo.

Outra forma de avaliação da criatividade de aplicativos no ensino da computação, é por meio de modelos automatizados. Desta forma, artefatos criados como resultado da aprendizagem são usados como entrada e então avaliados automaticamente, como *Creassessment* (Alves, 2023). *Creassessment* (Alves, 2023) é um modelo automatizado que busca oferecer uma base padronizada e objetiva para medir a criatividade em aplicativos criados com o *App Inventor*. O modelo usa como base de avaliação da criatividade as dimensões de originalidade, fluência e flexibilidade. A originalidade é avaliada com base nas funcionalidades, componentes de UI, tópicos e *tags* relevantes. A flexibilidade é avaliada com base na quantidade de diferentes componentes no aplicativo, na quantidade de diferentes blocos de programação e na quantidade de diferentes funções. A fluência é avaliada com base na quantidade de componentes e na quantidade de blocos de programação. As técnicas de extração e avaliação automatizada pelo modelo são apresentados na Tabela 2.

Tabela 2 - Técnicas de extração e pontuação do modelo Creassessment

Dimensão	**Item**	**Extração e identificação da técnica**	**Técnica de avaliação**
Originalidade	Funcionalidade	Extração baseada em regras (Alves et al., 2023) a partir de XML e JSON	Frequência única e combinada
	Componentes UI	Extração a partir de um JSON	Frequência única e combinada
	Tópicos	Processamento de linguagem natural usando o modelo de *Machine Learning* para classificação Naive Bayes (Rennie, Shih, et al., 2003) a partir de textos extraídos de XML e JSON	Frequência única
	Tags	Processamento de linguagem natural usando o extrator de palavras-chave (Campos, Mangaravite, et al., 2020)	Frequência única
Fluência	Componentes	Contagem do número de componentes extraídos de um JSON	Frequência única
	Programação	Contagem do número de blocos de programação extraídos de um XML	Frequência única
	Componentes	Contagem do número de	Contagem

Flexibilidade		componentes diferentes extraídos de um JSON	
	Programação	Contagem do número de blocos de programação diferentes extraídos de um XML	Contagem
	Funcionalidade	Contagem do número de diferentes funções extraídos de XML e JSON, usando regras já definidas	Contagem

Fonte: (Alves, 2023)

Em relação a confiabilidade e validade do modelo *Creassessment*, ele se mostrou satisfatório. A confiabilidade foi analisada a partir do ω (coeficiente ómega) (Hayes & Coutts, 2020). De acordo com a literatura ω entre 0,8 e 0,9 indica um bom nível de confiabilidade e acima de 0,9 um nível muito alto de confiabilidade, o modelo *Creassessment* apresentou um ω de 0,86. Para analisar a validade do modelo, foi utilizado o coeficiente de correlação de Pearson, o qual entende correlações entre 0,3 a 0,5 como fracas, 0,5 a 0,7 moderadas, 0,7 a 0,9 fortes e acima de 0,9 muito forte (Devellis, 2017). As oito variáveis do modelo, originalidade de funcionalidade, componentes UI e tags; fluência de componentes e programação; e flexibilidade de componentes, programação e funcionalidades foram correlacionados entre si. A correlação entre as variáveis é apresentada na Tabela 3.

Tabela 3 - Correlações de Pearson do modelo *Creassessment*.

		Originalidade			Fluência		Flexibilidade	
Dimensão	Item	Funcionalidade Componentes de interface gráfica *Tags*			Componentes Programação		Componentes Programação Funções	
Originalidade	Funcionalidade	1.00						
	Componentes de interface gráfica	**0.47**	1.00					
	Tags	0.00	0.18	1.00				
Fluência	Componentes	0.25	**0.40**	**0.30**	1.00			
	Programação	**0.33**	**0.33**	0.20	**0.72**	1.00		

Flexibilidade	Componentes	**0.69**	**0.40**	0.01	**0.40**	**0.46**	1.00		
	Programação	**0.60**	**0.45**	0.05	**0.48**	**0.73**	**0.78**	1.00	
	Funções	**0.71**	**0.56**	0.00	**0.39**	**0.45**	**0.81**	**0.73**	1.00

*Correlações acima de 0.29 foram marcadas em negrito.

Fonte: (Alves, 2023)

As variáveis de flexibilidade mostraram as correlações mais fortes com r > 0,7 para uma dimensão, os itens de fluência também mostraram alta correlação interna, com r > 0,7, diferente das variáveis de originalidade mostraram uma fraca correlação interna. Ainda assim, tendo uma média no mínimo moderada, mostrando uma confiabilidade e validade satisfatória.

Para analisar a validade do modelo em relação à nota de juízes humanos, foi usado um conjunto de *Apps* do *App Inventor* da *App of the Month*, o qual era dividido em dois grupos, *Winners* e *No-Winners*. Para analisar a correlação entre o modelo *Creassessment* e os juízes humanos foi analisada a média e a mediana dos dois grupos. Para ambos os testes a nota do grupo dos *Winners* era maior que a nota do grupo dos *No-Winners*, sendo essa uma primeira evidência que o *Creassessment* consegue diferenciar o grupo *Winners* e *No-Winners*. Além disso, foi analisada a convergência das notas por meio de um gráfico Boxplot, o qual em todos os níveis do gráfico as notas dos *Winners* tendiam a ser maiores. Mostrando assim resultados satisfatórios em relação aos juízes humanos.

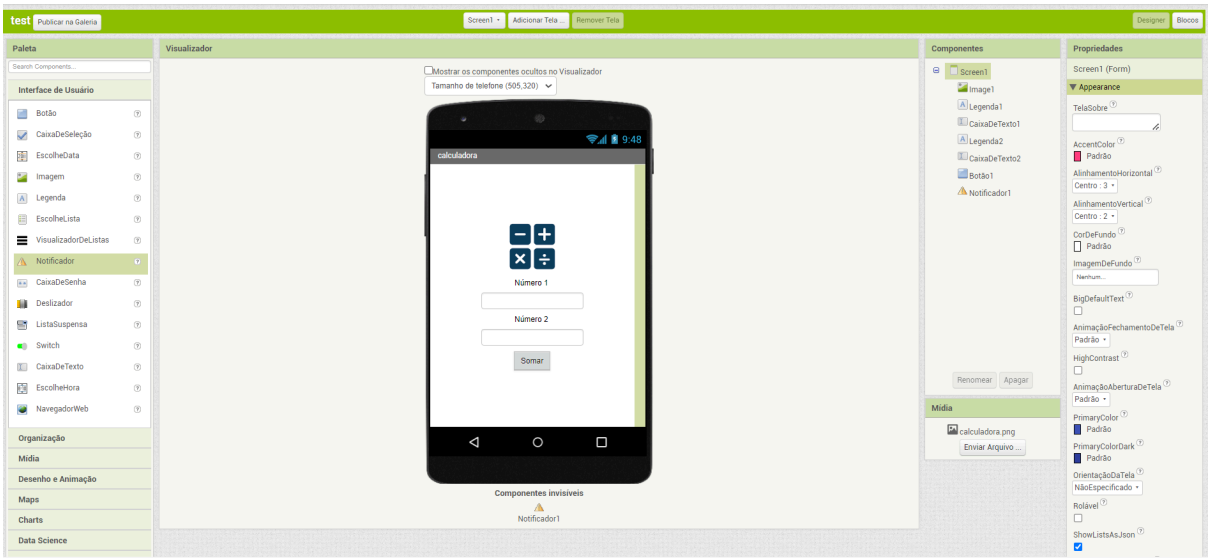
2.2 *App Inventor*

App Inventor (https://appinventor.mit.edu/) é uma aplicação de código aberto mantida pelo *Massachusetts Institute of Technolog*y (MIT). É uma plataforma de desenvolvimento de aplicativos móveis projetada com o intuito de simplificar e democratizar a programação e criação de aplicativos para Android em dispositivos móveis. Este ambiente de programação caracteriza-se por sua facilidade de uso, tornando-o assim acessível para iniciantes e facilitando o desenvolvimento (Harunani, Patton e Tissenbaum, 2019). Seu funcionamento baseia-se em uma interface gráfica de programação baseada em blocos para criar aplicativos móveis

onde seus usuários podem arrastar e soltar blocos para a criação da lógica do software.

O desenvolvimento de um aplicativo no *App Inventor* é separado em duas partes, a parte do Designer (Figura 1) e a parte dos Blocos (Figura 2). A parte do Designer refere-se a toda a logística envolvendo a parte de interface e de UI design do aplicativo, já a parte de blocos consiste na programação lógica do seu aplicativo, na ligação entre as estruturas da sua interface, na lógica dos componentes e eventos que vão ocorrer durante a execução do seu software.

Figura 1- Designer do *App Inventor*



Fonte: autor

Na esquerda da imagem pode-se ver os diversos componentes disponibilizados pelo Designer do *App Inventor* (Tabela 4).

Tabela 4 - Categorias, componentes e funções do Designer no *App Inventor*

Categoria	Componente	Função e breve descrição
	Botão	Cria um botão que pode ser clicado para executar ações
	CaixaDeSeleção	Permite selecionar opções em uma lista
	EscolheData	Permite colocar uma data
	Imagem	Permite adicionar imagens ao aplicativo
	Legenda	Adiciona textos ao aplicativo
	EscolheLista	Permite selecionar itens de uma lista
	VisualizadorDeListas	Exibe listas em um painel rolável
	Notificador	Envia mensagens e notificações ao usuário
	CaixaDeSenha	Permite inserir senhas

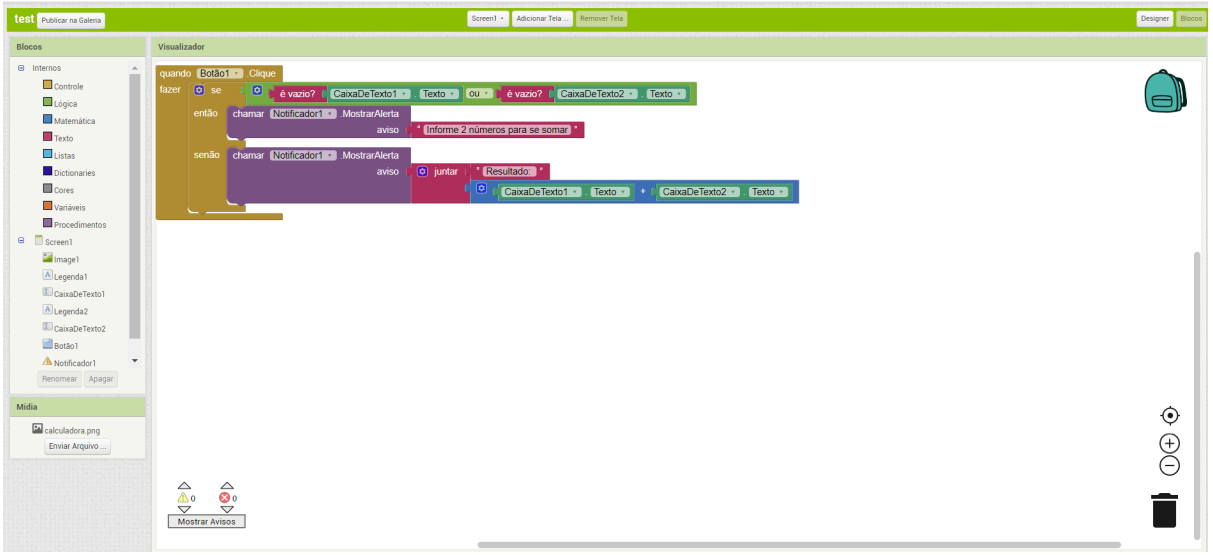
Interface de Usuário	Deslizador	Uma barra de progresso que permite ao usuário arrastar para ajustar a posição e ou o valor
	ListaSuspensa	Uma lista suspensa (pop-up) que permite ao usuário escolher itens
	Switch	Botão de alternância que permite o usuário executar eventos
	CaixaDeTexto	Permite ao usuário inserir texto
	EscolheHora	Permite ao usuário escolher um horário
	NavegadorWeb	Permite a exibição de páginas web dentro do aplicativo
Organização	OrganizaçãoHorizontal	Organiza os componentes horizontalmente na tela
	HorizontaScrollArrangement	Organiza os componentes de forma que sejam roláveis horizontalmente
	OrganizaçãoEmTabela	Organiza os componentes em forma de tabela
	OrganizaçãoVertical	Organiza os componentes verticalmente na tela
	VerticalScrollArrangement	Organiza os componentes de forma que sejam roláveis verticalmente
Mídia	CâmeraDeVídeo	Permite capturar e exibir vídeos diretamente no aplicativo
	Câmera	Permite capturar e exibir fotos diretamente no aplicativo
	FilePicker	Permite escolher imagens ou arquivos
	EscolheImagem	Permite ao usuário escolher imagens da galeria (Imagens já adicionadas ao projeto)
	Tocador	Reproduzir arquivos de áudio
	Som	Reproduz sons e ou arquivos de áudio
	Gravador	Permite gravar a áudio no aplicativo
	ReconhecedorDeVoz	Permite converter fala em texto
	TextoParaFalar	Permite converter texto em áudio
	Translator	Permite traduzir textos entre diferentes idiomas diretamente dentro do aplicativo
	ReprodutorDeVídeo	Permite a reprodução de vídeos diretamente dentro do aplicativo
Desenho e animação	Bola	Cria uma representação visual de uma bola, podendo ser usada para animações e jogos
	Pintura	Permite ao usuário desenhar diretamente no aplicativo
	SpriteImagem	Permite a exibição de imagens dentro do aplicativo
Maps	Circle	Permite desenhar círculos em um mapa
	FeatureCollection	Permite a representação de elementos geoespaciais em um mapa
	LineString	Permite desenhar linhas em um mapa
	Map	Permite a exibição de mapas diretamente no aplicativo
	Marker	Permite adicionar marcadores em um mapa
	Navigation	Permite a navegação por um mapa, como movimentação em relação a ele
	Polygon	Permite desenhar polígonos em um mapa
	Rectangle	Permite a criação de retângulos em um mapa
Charts	Chart	Permite a exibição de gráficos no aplicativo
	ChartData2D	Fornece dados para a criação de gráficos
Data Science	AnomalyDetection	Detecta anomalias em conjuntos de dados
	Regression	Permite realizar análises de regressão em conjuntos de dados
	SensorAcelerômetro	Detecta movimento do dispositivo
	CódigoDeBarras	Permite a leitura de códigos de barras usando a câmera do seu dispositivo móvel ou computador

Sensores	Barômetro	Permite medir a pressão atmosférica
	Temporizador	Permite a temporização e agendamento
	GyroscopeSensor	Permite a detecção da orientação e movimento do dispositivo
	Hygrometer	Permite medir a umidade do ambiente
	LightSensor	Permite a detecção de níveis de luminosidade no ambiente do aplicativo
	SensorDeLocalização	Oferece dados sobre a localização do dispositivo o qual está usando o aplicativo
	MagneticFieldSensor	Permite a detecção de campos magnéticos
	NearField	Permite ao aplicativo o uso de recursos NFC
	SensorDeOrientação	Permite determinar a orientação espacial do dispositivo
	Pedometer	Permite a contagem de passos através do aplicativo
	SensorDeProximidade	Permite medir a proximidade de objetos em relação a tela do dispositivo que está usando o aplicativo
	Thermometer	Permite medir a temperatura do ambiente
Social	EscolheContato	Um botão o qual ao clicar nele permite escolher um contato entre uma lista
	EscolheEmail	Permite adicionar um e-mail a um contato
	Ligação	Permite a ligação para um número especificado
	EscolheNúmeroDeTelefone	Semelhante ao EscolheContato, um botão que ao clicar nele permite escolher um número entre uma lista
Compartilhamento	CloudDB	Um componente o qual permite compartilhamento de dados em banco de dados
	DataFile	Permite a manipulação de dados
	Arquivo	Permite a leitura ou escrita de arquivos no dispositivo
	Spreadsheet	Permite a leitura ou escrita de dados através do Google Sheets
	TinyDB	Permite armazenar dados
	TinyWebDB	Permite a comunicação com um serviço Web para armazenar e recuperar dados
Conectividade	IniciadorDeAtividades	Permite a iniciação de atividades usando o método StartActivity
	ClienteBluetooth	Permite a conexão entre dispositivos usando Bluetooth
	ServidorBluetooth	Permite a transformação do dispositivo em um servidor que recebe conexões de outros aplicativos através de Bluetooth
	Serial	Componente para serial
	Web	Permite o uso de funções para solicitações HTTP GET, POST, PUT e DELETE
Experimental	ChatBot	Permite a comunicação com chatbots de inteligência artificial, o chatbot usado é o ChatGPT
	FirebaseDB	Permite a conexão com serviços web para armazenar e recuperar informação
	ImageBot	Componente que usa a tecnologia DALL-E 2 para criar e editar imagens

Fonte: autor

Já *LEGO MINDSTORMS* no *App Inventor* é um kit de robótica educacional que permite aos usuários criar e programar robôs, por meio de sensores e blocos de construção.

Figura 2 - Blocos do *App Inventor*



Fonte: Autor

Na parte dos blocos (Figura 2), à esquerda tem-se a estrutura que disponibiliza toda a parte lógica do aplicativo e a partir dela se constrói a base programacional do software. Existem diversos outros tipos de blocos possíveis de usar, como blocos específicos de matemática, blocos de texto, listas, procedimentos e entre outros, possibilitando assim a construção de um aplicativo complexo e da mesma forma fácil de implementar. Na Tabela 5 são apresentados esses blocos.

Tabela 5 - Blocos de controle, lógica, matemática e texto do *App Inventor*

Categoria	**Bloco**	**Função e breve descrição**
Controle	if & else if	Testa uma determinada condição
	for each number from to	Executa os blocos para cada valor número no intervalo entre *from* e *to*
	for each item in list	Executa os blocos para cada item da lista
	for each key with value in dictionary	Executa os blocos para cada entrada de valor-chave no dicionário
	while	Testa uma condição
	if then else	Testa uma determinada condição
	open another screen	Abre uma determinada tela
	close screen	Fecha a tela atual
	close application	Fecha o aplicativo
	break	Permite o escape de *loops*
Lógica	true	Representa a constante com valor *true*
	false	representa a constante com valor *false*
	not	Fornece a lógica de negação
	=	Testa se argumentos são iguais
	~=	Testa se argumentos são diferentes

	and	Fornece a lógica *and*
	or	Fornece a lógica *or*
Matemática	Biblioteca de matemática	Fornece toda uma biblioteca de matemática envolvendo operações básicas até operações complexas
Texto	Biblioteca de texto e *string*	Fornece toda uma biblioteca para o trabalho com texto e *strings*

Fonte: autor

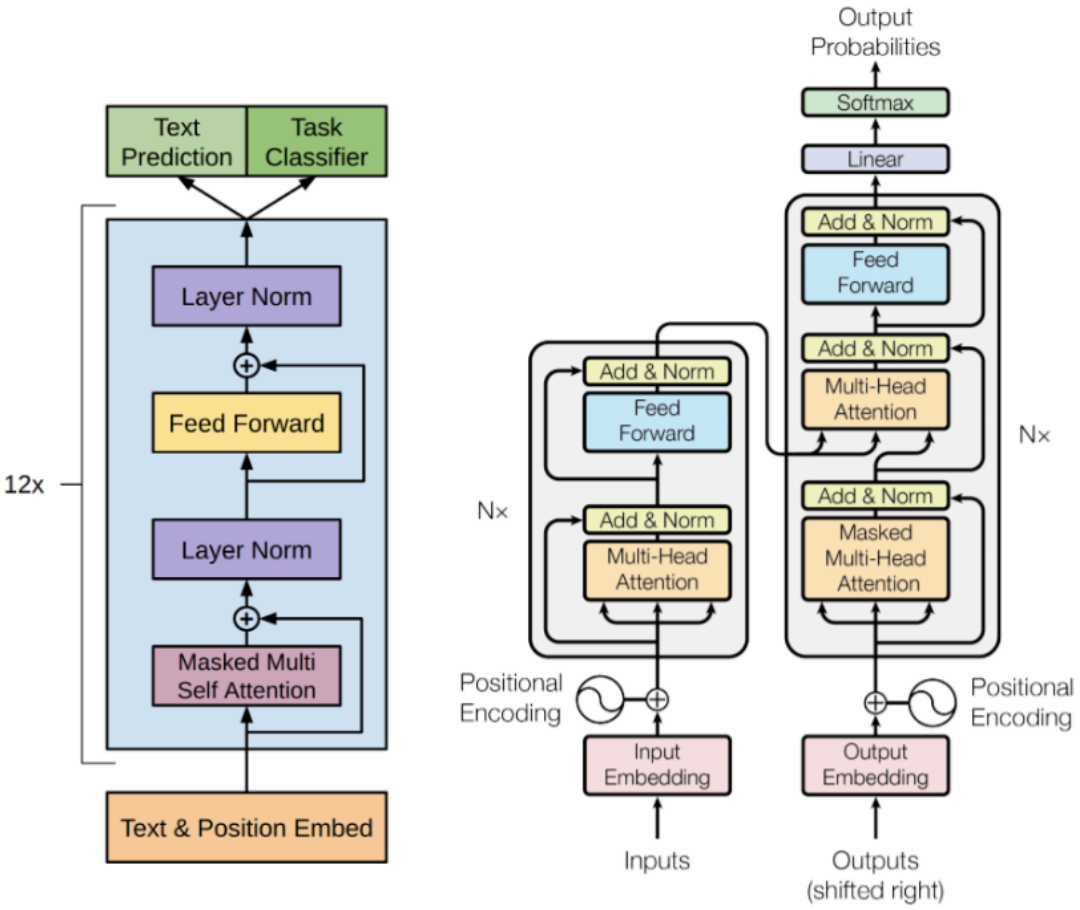
O projeto pode ser compilado em um arquivo .apk. Também pode ser exportado no formato .aia. O arquivo .aia contém todos os recursos e informações necessárias para o aplicativo, permitindo o compartilhamento e backup do projeto. Os recursos incluem arquivos internos relacionados ao projeto, arquivos de mídia usados pelo aplicativo, um arquivo .bky com o XML dos blocos lógicos do aplicativo e um arquivo .scm para cada tela do aplicativo com um JSON para os componentes de interface de usuário.

2.3 *Large Language Models*

Large Language Models (LLMs) são modelos de linguagem computacionais que utilizam técnicas de aprendizado de máquina para entender e gerar texto em linguagem natural (BIRHANE et al., 2023). Esses modelos são treinados com grandes quantidades de dados diversos para aprender as regras da linguagem e as relações entre as palavras.

As LLMs são construídas utilizando redes neurais profundas. Tem como base principal a arquitetura *Transformer* (Vaswaniet al., 2017). O *Transformer* é uma arquitetura com várias camadas de codificadores e decodificadores que são blocos de redes neurais responsáveis por aprender a representação de palavras e capturar a dependência entre elas. É muito usado para tarefas de processamento de linguagem devido a sua capacidade de modelar sequências eficientemente. Cada camada do *transformer* possui múltiplas cabeças de atenção, *multi-head self-attention mechanism*, que permitem que o modelo se concentre em diferentes partes da sequência de entrada em um mesmo momento.

Figura 3 - Arquitetura *Transformer*



Fonte: (Vaswani. et al., 2017)

O treinamento de uma LLM com a arquitetura *Transformer* é feito a partir de um aprendizado supervisionado. Primeiramente é feito um pré-processamento dos dados, separando um conjunto muito grande de dados e textos que serão usados para treinar a LLM. A partir desse texto é introduzido uma sequência de entradas e saídas para a LLM que tenta prever a próxima palavra da sequência, tarefa a qual é feita milhões de vezes até que a saída/resposta da LLM seja satisfatória. E por fim, após o treinamento a LLM é testada utilizando um conjunto de dados que não foi utilizado no treinamento para avaliar se o desempenho é satisfatório.

Existem LLMs de código aberto e/ou prioritário e alguns disponibilizam o acesso gratuitamente via web e/ou APIs. Na Tabela 6, será apresentado algumas das mais proeminentes atualmente, junto com suas opções de acesso.

Tabela 6 - Exemplos de LLMs

LLM	Link	Licença
GPT 3.5	openai.com	online: gratuito API: gratuito, com limites
Gemini Pro	gemini.google.com	online: gratuito API: gratuito, com limites
Llama3 70b	https://chat.lmsys.org	online: gratuito API: gratuito
Mistral 7b	https://chat.lmsys.org	online: gratuito API: gratuito
Vicuna 13b	https://chat.lmsys.org	online: gratuito API: gratuito

Fonte: (Signity, 2023)

As *Large Language Models* (LLMs) têm sido usadas cada vez mais em diferentes tipos de aplicações. Inicialmente, esses modelos eram designados para tarefas de processamento de linguagem, como tradução de idiomas, resumo e geração automática de textos, facilitando assim a comunicação e a disseminação de informações (Zhao et al, 2023). Esses modelos têm se mostrado valiosos em assistentes virtuais e *chatbots*, ou sendo usados como ferramentas poderosas na pesquisa científica, auxiliando na análise de grandes volumes de dados textuais (Wei et al. 2023). Além disso, esses modelos também podem ser empregados como avaliadores em diferentes contextos. Um exemplo é a avaliação da criatividade na criação de aplicativos, analisando a originalidade e inovação das ideias (Goes et al. 2022).

2.3.1 *Prompt engineering*

Apesar de as LLMs normalmente serem conhecidas por sua objetividade e facilidade de uso, existe o conceito de *prompt*, que são, essencialmente, instruções ou solicitações utilizadas para interagir com as LLMs (Giray, 2023). O objetivo dos prompts é obter respostas específicas às perguntas ou comandos fornecidos às LLMs. *Prompts* podem ser compostos por algumas partes (Dair.ai, 2023), as quais são apresentadas na Tabela 7.

Tabela 7 - Composição de um *Prompt*

Itens	Descrição
Instrução	Uma tarefa ou instrução específica que você deseja que o modelo execute.
Contexto	Informações externas ou contexto adicional que podem orientar o modelo para melhores respostas.
Dados de entrada	A entrada ou pergunta para a qual se tem o interesse em encontrar uma resposta
Indicador de saída	O tipo ou o formato da saída.

Fonte: (Dair.ai, 2023)

Prompt engineering é o termo que refere-se ao processo de criação de *prompts*. Existem várias técnicas de *prompt engineering* (Giray 2023). *Zero-shot prompting* (Dair.ai, 2023) é um método que envolve por meio de uma simples pergunta, a busca pela resposta da LLM sem a necessidade de um treinamento prévio específico. Ainda assim, em certas ocasiões, o modelo *zero-shot* não é o suficiente para extrair a informação necessária do modelo de linguagem. *Few-shot prompting* (Dair.ai, 2023), é um método que permite ao modelo de linguagem o aprendizado no contexto da pergunta, por meio de informações no *prompt* para orientar o modelo, se obtêm um melhor desempenho da LLM. *Chain-of-Thought Prompting* (CoT) (Dair.ai, 2023), é o método que envolve criar uma cadeia de pensamentos com demonstrações para a LLM, envolvendo a elaboração manual de exemplos para o modelo de linguagem aprender e entender como chegar a resposta final, ou induzindo a LLMs a pensar de passo em passo, para chegar a sua resposta. Na Tabela 8 são apresentados alguns exemplos de prompts para os modelos apresentados.

Tabela 8 - Exemplos de técnicas de *prompt engineering*, em negrito será demonstrado o uso da técnica

Técnica	Exemplo de prompt
Zero-shot	Classifique o artefato em neutro, negativo ou positivo. Artefato: Acho que as férias estão boas. **Sentimento:**
Few-shot	**Um "whatpu" é um animal pequeno e peludo nativo da Tanzânia. Um exemplo de frase que usa a palavra whatpu é:** **Estávamos viajando pela África e vimos esses whatpus muito fofos.**

	Fazer um "farduddle" significa pular para cima e para baixo muito rápido. Um exemplo de frase que usa palavra “farduddle” é:
Chain-of-Thought (CoT)	P: Roger tem 5 bolas de tênis. Ele compra mais 2 latas de bolas de tênis. Cada lata contém 3 bolas de tênis. Quantas bolas de tênis ele tem agora ? **R: Roger começou com 5 bolas. 2 latas de 3 bolas de tênis equivalem a 6 bolas de tênis cada. 5 + 6 = 11. A resposta é 11.** P: A cafeteria tinha 23 maçãs, se usaram 20 para fazer o almoço e compraram mais 6, quantas maçãs eles têm ?

Fonte: Autor, usando como referência os exemplos de (Dair.ai, 2023)

LLMs podem ser usadas para avaliação da criatividade em artefatos (Goes et al. 2023), sendo uma nova forma de avaliação. Junto com as técnicas de *prompt engineering* é possível extrair informações e avaliações do modelo de linguagem sobre o artefato em questão. Na Tabela 9 serão apresentados alguns exemplos de *prompts* para a avaliação de artefatos, relacionando com o objetivo deste trabalho.

Tabela 9 - Exemplos de prompts para a avaliação de artefatos

Técnica e explicação	**Prompt**	**Referência**
Prompt para classificação de piadas com base em *zero-shot prompting*	1 - Classifique a seguinte [Piada] como Engraçada ou $Oposto. Piada: $DescriçãoDaPiada Classificação: 2 - Classifique a seguinte [Piada] como Engraçada ou $Oposta. Piada: $DescriçãoDePiadaEngraçada Classificação: Engraçada Piada: $DescriçãoDePiadaNãoEngraçada Classificação: $Oposto. Piada: $DescriçãoDaPiada Classificação: 3 - Classifique a seguinte [Joke] como engraçada ou $Oposto como se fosse uma pessoa que gosta do seguinte tipo de humor $TipoDeHumor Piada: $DescriçãoDaPiada Classificação: 4 - Classifique a seguinte [Joke] como engraçada ou $Oposto como se fosse uma pessoa que gosta do seguinte tipo de humor $TipoDeHumor	(Goes et al. 2022).

	Piada: $DescriçãoDaPiada Classificação: $EngraçadaOuOposto Vamos pensar passo a passo o porque essa [Piada] é $EngraçadaOuOposto para uma pessoa que gosta do seguinte tipo de humor $TipoDeHumor	
Prompt para a classificação de sentenças	1 - autscore question: What is a surprising use for X response: Y. 2 - How original each use of X is: 3 - Below is a list of uses for a X. On a scale of 10-50, judge how original each use for a X is, where 10 is ‘not all creative’ and 50 is ‘very creative’ Uses: [Uses] Ratings: [Ratings for uses]	(Organisciak et al., 2023).

3. Estado da arte

Para levantar o estado da arte em relação a estudos que comparam a qualidade de avaliação de criatividade de apps (*App Inventor*) com LLMs com outros modelos e métodos, foi realizado um estudo de mapeamento sistemático seguindo o procedimento proposto por Petersen et al. (2008).

3.1 Definição do protocolo de revisão

Questão de Pesquisa: Existem estudos que analisam e comparam a qualidade de avaliação de criatividade de apps (*App Inventor*) com LLMs com outros modelos de avaliação?

Essa questão de pesquisa é refinada nas seguintes perguntas de análise:

PA1. Quais estudos existem e quais artefatos são avaliados e quais fatores de qualidade avaliam?

PA2. Quais LLMs foram utilizados para a avaliação da criatividade e com quais *prompts*?

PA3. Quais as características dos estudos de avaliação?

PA4. Quais as principais descobertas em relação aos fatores de qualidade analisados e/ou comparados?

Critério de inclusão/exclusão. Pela falta de estudos focados especificamente em apps como resultados de aprendizagem, foi estendido o escopo a qualquer artefato, tendo como prioridade artefatos computacionais. São considerados somente artigos em inglês publicados nos últimos 5 anos considerando o tema emergente de LLMs.

Fonte de dados. A busca foi realizada nas principais bases de dados da área da computação: ACM, Arxiv, IEEE, Scopus, Science Direct, Google Scholar. A pesquisa foi realizada no Google Scholar pelo fato de ser considerada aceitável como uma fonte adicional, visando minimizar os riscos de omissão de artigos, por meio de uma busca mais ampla de diversas áreas de conhecimento (Piasecki et al., 2018).

Definição da string de busca. A string de busca é composta de conceitos relacionados à questão de pesquisa considerando também sinônimos (Tabela 10).

Tabela 10: Termos do string de busca

Palavra chave	**Sinônimos**
creativity	
assess*	evaluat*, grading, scoring
“large language model”	chatGPT, llm

Fonte: autor

Considerando as palavras chaves e de seus sinônimos, definiu-se a seguinte string de busca genérico:

(creativity) AND (assess* OR evaluat* OR grading OR scoring) AND (“large language model” OR chatGPT OR llm)

Esse string de busca genérico foi adaptado conforme necessário para cada base de dados (Tabela 11).

Tabela 11: Strings de busca para cada base de dados

Fonte de Dados	**String de Busca**
ACM	[Abstract: creativity] AND [[Abstract: assess*] OR [Abstract: evaluat*] OR [Abstract: grading] OR [Abstract: scoring] OR [Abstract: rating]] AND [[Abstract: chatgpt] OR [Abstract: llm] OR [Abstract: "large language model"]] AND [E-Publication Date: (01/01/2018 TO 12/31/2023)]
Arxiv	-announced_date_first; size: 50; date_range: from 2018-01-01 to 2023-12-31; include_cross_list: True; terms: AND abstract=creativity; AND abstract=assess* OR evaluat* OR grading OR scoring OR rating; AND abstract=chatgpt OR llm OR "large language model"
IEEE	("Abstract":creativity) AND ("Abstract":assess* OR "Abstract":evaluat* OR "Abstract":grading OR "Abstract":scoring OR "Abstract":rating) AND ("Abstract":chatgpt OR "Abstract":llm OR "Abstract":"large language model") Filters Applied: 2018 - 2024
Scopus	TITLE-ABS-KEY ((assess* OR evaluat* OR grading OR scoring OR rating) AND (chatgpt OR llm OR "large language model") AND (creativity)) AND PUBYEAR > 2018 AND (LIMIT-TO (LANGUAGE, “English”))
Science Direct	(creativity) AND (assess OR evaluat OR grading OR scoring OR rating) AND (chatgpt OR llm OR "large language model") Year 2018-2023
Google Scholar	creativity, OR assess*, OR evaluate*, OR grading, OR rating, OR scoring, OR chatgpt, OR llm, OR "large language models"

Fonte: autor

3.2 Execução da busca

A pesquisa foi realizada em Setembro de 2023 pelo autor e revisada pela orientadora e co-orientadora.

Tabela 12 - Resultado da busca

Base de dados	**No. de artigos do resultado de busca**	**No. de artigos analisados**	**No. de artigos potencialmente relevantes**	**No. dos artigos relevantes**
Scopus	18	18	1	1
Arxiv	21	21	3	1
IEEE	2	2	0	0
Science Direct	81	81	1	1
ACM	4	4	0	0
Google Scholar	55	55	2	0
Total (Sem duplicatas)				**2**

Fonte: autor

Na primeira fase de análise, títulos e resumos foram analisados, resultando em 7 artigos potencialmente relevantes. No segundo estágio, os materiais foram lidos na íntegra, para assegurar sua relevância levando em conta os critérios de inclusão e exclusão. Documentos duplicados foram eliminados e aqueles que descrevem a mesma unidade instrucional foram unificados. Vários dos potencialmente relevantes não se referiram à criatividade, e por isso não foram incluídos. Ao final foram identificados 2 trabalhos relevantes.

3.3. Análise de dados

Como resultado foram encontrados 2 artigos relevantes (Tabela 13).

Tabela 13 - Artigos relevantes encontrados.

Referência Bibliográfica	**Artefato avaliado**
Organisciak, P., Acar, S., Dumas, D., & Berthiaume, K. (2023). Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. Thinking Skills and Creativity,	Sentenças sobre a criatividade no uso de

49, 101356.	objetos
Goes, F., Zhou, Z., Sawicki, P., Grzes, M., & Brown, D. G. (2022). Crowd Score: A Method for the Evaluation of Jokes using Large Language Model AI Voters as Judges. ArXiv Preprint, arXiv:2212.11214 [cs.AI].	Piadas

Fonte: autor

Organisciak et al. (2023) avaliam sentenças sobre o uso de forma criativa de objetos, como o uso criativo de facas, tijolos e etc. Goes et al. (2022) avaliam o nível de criatividade de piadas.

3.3.1 Quais LLMs foram utilizados para a avaliação da criatividade, com quais *prompts*?

Tabela 14 - Visão geral sobre as LLMs utilizadas, prompts e escala de nota/feedback.

Referências	**LLM(s)**	***Prompts* utilizados**	**Qual escala de nota/*feedback* foi utilizada**
(Organisciak et al., 2023).	GPT-3 (Brown et al., 2020) e T5 (Raffel et al., 2020)	1 - autscore question: What is a surprising use for X response: Y. 2 - How original each use of X is: 3 - Below is a list of uses for a X. On a scale of 10-50, judge how original each use for a X is, where 10 is 'not all creative' and 50 is 'very creative' Uses: [Uses] Ratings: [Ratings for uses]	Em relação a nota avaliada pelos humanos, foi feito uma média com base em um grande escala de respostas, já na avaliação das LLMs, através de *prompt engineering* foi solicitado uma nota entre 1 e 10 e dividido por 2, sendo 10 a mais criativa e 1 a menos criativa.
(Goes et al., 2022).	GPT-3 (Brown et al., 2020)	1 - Classifique a seguinte [Piada] como Engraçada ou $Oposto. Piada: $DescriçãoDaPiada Classificação: 2 - Classifique a seguinte [Piada] como Engraçada ou $Oposta. Piada: $DescriçãoDePiadaEngraçada Classificação: Engraçada Piada: $DescriçãoDePiadaNãoEngraçada Classificação: $Oposto. Piada: $DescriçãoDaPiada Classificação: 3 - Classifique a seguinte [Joke]	Primeiramente as piadas foram avaliadas de forma binária entre 1 e 0, sendo 1 engraçadas e 0 não engraçadas. Para fins de comparação com o *dataset* já obtido com avaliações humanas foi feito uma avaliação entre 1 e 4, sendo 4 a mais criativa e engraçada e 1 menos criativa e engraçada.

		como engraçada ou $Oposto como se fosse uma pessoa que gosta do seguinte tipo de humor $TipoDeHumor Piada: $DescriçãoDaPiada Classificação: 4 - Classifique a seguinte [Joke] como engraçada ou $Oposto como se fosse uma pessoa que gosta do seguinte tipo de humor $TipoDeHumor Piada: $DescriçãoDaPiada Classificação: $EngraçadaOuOposto Vamos pensar passo a passo o porque essa [Piada] é $EngraçadaOuOposto para uma pessoa que gosta do seguinte tipo de humor $TipoDeHumor	

Organisciak et al. (2023) apresentam um modelo chamado *Ocsai*, que propõe o treinamento de LLMs para avaliar a criatividade. Tem como objetivo melhorar a avaliação automatizada de testes de pensamento divergente a partir de ajustes e treinamento nos modelos de linguagem com base em avaliações humanas. Como artefato, foram utilizadas sentenças e frases sobre quais usos têm um determinado objeto, como faca, colher, entre outros. Foram utilizadas duas LLMs para a avaliação: GPT-3 (Brown et al., 2020) e T5 (Raffel et al., 2020). Foram utilizados prompts baseados em *zero-shot*, sem nenhum exemplo adicional para a LLM e prompts baseados em *few-shot*, treinando a LLM por meio de exemplos. Foi solicitado aos LLMs gerar uma resposta entre um intervalo de 1 e 10, sendo 1 a menos criativa e 10 a mais criativa, e foi feita a divisão da nota por 2. Utilizando o conjunto de dados de avaliações por juízes humanos, foi calculada uma média para obter a nota de criatividade das sentenças e comparado a nota da LLM.

Goes et al. (2022) apresentaram um modelo chamado *Crowd Score* para avaliar a criatividade de artefatos usando inteligência artificial. O modelo baseia-se em induzir diferentes personalidades às LLMs, para ver os diferentes resultados. O modelo usou piadas como primeiro artefato a ser testado, sendo testadas 52 piadas, que foram avaliadas por IAs com diferentes estilos de humor. Foi utilizada a LLM GPT-3 (Brown et al., 2020) para a avaliação das piadas. Foram usados diferentes

tipos de *prompts*, incluindo: *Prompts* simples baseados em *zero-shot*, *Prompts* baseados em *few-shot*, apresentando exemplos de piadas engraçadas a LLM, além de induzir personalidade ao modelo de linguagem. Foram utilizados *prompts* baseados em *Chain-of-Thought* (CoT), utilizando a frase “vamos pensar passo a passo”. Para uma primeira avaliação, foi solicitado para as LLMs por meio de prompts apenas uma avaliação binária: engraçada ou não engraçada. Foram feitas também avaliações num intervalo de 1 e 4 (sendo 4 a mais criativa e engraçada e 1 a menos criativa e engraçada). Cada teste foi realizado 3 vezes e foi calculada uma média para chegar ao valor final.

3.3.2 **Quais as características dos estudos de avaliação, quais fatores de qualidade foram avaliados e quais descobertas foram encontradas?**

Tabela 15 - Visão geral sobre as características dos estudos, fatores de qualidade avaliados e descobertas encontradas.

Referência	**Quais fatores de qualidade foram avaliados?**	**Design de estudo**	**Quantidade de artefatos considerados na avaliação**	**Métodos estatísticos utilizados para análise de dados**	**Descobertas**
(Organisciak et al., 2023).	• Correlação entre as respostas das LLMs e a avaliação dos juízes humanos. • Diferença entre as duas avaliações por meio do *Root Mean Square Error* (RMSE).	Foi feita a comparação entre um conjunto de dados de avaliações humanas anteriormente obtidas com a avaliação das LLMs.	20.202 artefatos avaliadas	Correlação de Pearson, *Root Mean Square Erro*r (RMSE).	Os resultados mostraram uma grande capacidade das LLMs de imitar a avaliação de juízes humanos, sendo assim, satisfatório.
(Goes et al., 2022).	• Verificação se a LLM produziu uma avaliação baseada em raciocínio lógico utilizando de CoT (*Chain-of-Thought*) • Na acurácia das avaliações, foram utilizadas duas métricas: *f-score* e *balanced accuracy*.	Foi feita a comparação entre um conjunto de dados de avaliações humanas anteriormente obtidas com a avaliação das LLMs.	52 piadas avaliadas	Foram usadas as métricas *F-score* e *balanced accuracy*. Não foi mencionado nenhum método estatístico. Foi feita uma comparação com a avaliação dos juízes humanos por meio de gráficos.	Os resultados apontam que as LLMs podem ser usadas como juízes de humor no processo de avaliação de piadas.

Organisciak et al. (2023) avaliaram a correlação das respostas da LLM com as respostas de avaliadores humanos utilizando a correlação de Pearson. Foram usados diferentes *c*onjuntos de dados já avaliados por humanos, que cada um deles tinha seus objetos para fins da pergunta. Os dados utilizados foram compilados por meio de vários estudos anteriores com avaliações humanas disponíveis. Foram avaliados um total de 20.202 artefatos, divididos entre diferentes conjuntos de dados. Como eram diferentes conjuntos de dados, cada um teve sua correlação avaliada separadamente. Entre as previsões do modelo e as avaliações humanas obteve-se um intervalo de **r** entre **0.12** a **0.81**. Para as avaliações na mesma escala das avaliações humanas, foi avaliado também o *Root Mean Square* (RMSE). Tiveram poucas avaliações de RMSE, tendo um range entre **0.518** a **0.595,** indicando um desempenho muito bom. Os resultados de comparação se mostraram satisfatórios, com desempenhos altos de correlação como **r = 0.76** e **r = 0.81**. Foi evidenciado também que o treinamento das LLMs por meio de *prompts*, para induzir e treinar como chegar a resposta, foi muito eficaz.

Goes et al. (2022) realizaram testes para validar a confiabilidade das respostas das LLMs. Primeiro foram coletadas respostas das LLMs com base em *prompts* do tipo *Chain-of-Thought* (CoT), que forçam a LLM a dar uma explicação do porquê a piada foi avaliada daquela forma. Após os dados coletados por meio desse primeiro *prompt*, foi feita uma segunda pergunta a LLM, perguntando se a explicação dada pela própria LLM era condizente com a classificação dada por ela mesma. Com isso, caso houvesse contradição na resposta, a avaliação era dada como inválida. Cada teste foi realizado três vezes e o resultado final foi obtido por meio do cálculo da média. Foram usadas as métricas *f-score* e *balanced accuracy* para analisar o desempenho da avaliação das LLMs. Foram feitos vários testes classificando as piadas numa escala ordinal: *Funny* ou *Opposite*. Foram testadas várias palavras como *Opposite*, para a partir das métricas verificar qual trazia uma melhor acurácia. Na Tabela 16 são apresentados os testes de acurácia entre as palavras usadas como *Opposite* com este modelo.

Tabela 16 - Testes de acurácia do método *Crowd Score* para as palavras *Opposite*

	F-Score	***Balanced Accuracy***

	Zero-Shot	***Few-Shot***	***Zero-Shot***	***Few-Shot***
Funny / Boring	0.89	1	0.88	1
Funny / Dull	0.89	1	0.88	1
Funny / Serious	0.8	0.8	0.75	0.75
Funny / Sad	0.8	0.8	0.75	0.75
Funny / Not Amusing	0.75	0.86	0.75	0.88
Funny / Unfunny	0.67	0.86	0.62	0.88
Funny / Dumb	0.67	0.86	0.62	0.88
Funny / Not Funny	0.67	0.86	0.67	0.88

Fonte: (Goes et al., 2022)

A palavra com a maior acurácia no lugar de *Opposite* foi a palavra *Boring* e a de pior acurácia foi a palavra *Not Funny*, mostrando uma possível redundância. Os *prompts* utilizando a técnica *few-shot* mostraram melhor desempenho em relação a todos os testes. Para o experimento final, foram utilizados um total de 52 piadas, as quais foram geradas por um comediante profissional. Ao final, foi mencionado que o método *Crowd Score* seguiu as mesmas tendências de avaliações que os juízes humanos, atribuindo pontuação mais alta às piadas que também são consideradas mais engraçadas pelos juízes humanos. Com isso, as descobertas apontam que as LLMs podem ser usadas como juízes de humor no processo de avaliação de piadas. A partir disso, podendo reduzir o gargalo da necessidade de juízes humanos para a avaliação. Também permitindo que os próprios comediantes tenham um sistema de *feedback* instantâneo para suas piadas.

3.4. Discussão

De forma geral, observa-se uma dificuldade em encontrar artigos relevantes sobre o tema devido à recente evolução dos LLMs. Não foi encontrada nenhuma pesquisa deste tipo aplicado a artefatos computacionais e assim os artigos encontrados tem como artefato avaliado um diferente do objetivo deste trabalho: sentenças e piadas. Isso demonstra que ainda não existem muitas pesquisas com este objetivo.

Ao final foram encontrados dois estudos que avaliam a criatividade de artefatos, um que apresenta o modelo *Ocsai* (Organisciak et al., 2023) e outro que

apresenta o modelo *The Crowd Score* (Goes et al., 2022).

A principal LLM utilizada para as avaliações foi a GPT-3 (Brown et al., 2020), usada em ambas as pesquisas. Diversos *prompts* foram utilizados, nos dois trabalhos encontrados foram usados mais de um tipo de *prompt* para chegar a avaliação final, sendo eles: *prompts* de modelo *zero-shot*, *few-shot* e *Chain-of-Thought* (CoT). *Prompts* do tipo *few-shot* se mostraram os mais eficazes, pelo fato de introduzirem a LLM exemplos de como fazer e como responder a avaliação, e a partir disso treinando a LLM para o fim que se deseja.

A qualidade dos resultados gerados pelos LLMs foi avaliada a partir de fatores de qualidade como correlação, utilizando o coeficiente de correlação de Pearson, a diferença entre a avaliação humana e das LLMs por meio da medida do *Root Mean Square Error* (RMSE), além da acurácia das avaliações utilizando as métricas *f-score* e *balance accuracy*.

Tanto no modelo *Ocsai* (Organisciak et al., 2023), quanto no modelo *The Crowd Score* (Goes et al., 2022), as características do estudo se resumiram em comparar as avaliações das LLMs com avaliações humanas. No modelo *Ocsai* foi utilizado um grande conjunto de dados de respostas já avaliadas com a recém avaliação feita pelas LLMs. O modelo *The Crowd Score* comparou um total de 52 piadas já avaliadas por um profissional da área com as avaliações das LLMs.

Tanto no modelo *Ocsai* (Organisciak et al., 2023) quanto no modelo *The Crowd Score* (Goes et al., 2022), as descobertas foram satisfatórias. Estes resultados apontam que as LLMs podem ser usadas como alternativa para a avaliação da criatividade de artefatos.

Ameaças à validade. Como em qualquer mapeamento sistemático de pesquisa, existem possíveis ameaças à validade dos resultados. Mapeamentos sistemáticos podem ter resultados tendenciosos, visto que resultados negativos podem acabar não sendo publicados.

Devido à pouca influência que a tendência destes artigos tem sobre este mapeamento, uma vez que se refere a uma busca pelo estado da arte e foram encontrados poucos artigos sobre a pesquisa em questão, essas ameaças não representam uma alta ameaça.

Para tentar mitigar as ameaças e a omissão de estudos relevantes, foi definida uma string de busca utilizando termos relevantes e incluindo sinônimos.

Junto com isso foram feitas buscas em diferentes bases de dados científicas, reduzindo assim as possibilidades de omissão de estudos relevantes.

A seleção dos artigos e trabalhos relevantes foi feita seguindo os critérios de inclusão e exclusão explicitamente definidos e revisados pela orientadora até que um consenso geral foi atingido.

4. Estudos de avaliação

4.1 Definição dos estudos

A série de estudos de caso nesta pesquisa tem como objetivo comparar as avaliações da criatividade de apps criados com *App Inventor* por meio de LLMs em comparação com a avaliação humana e avaliação automatizada usando o modelo *Creassessment*.

Perguntas de pesquisa:

PP1: Qual o grau de concordância/confiabilidade da avaliação da criatividade entre LLMs?

PP2: Qual a correlação da avaliação da criatividade comparando a avaliação por LLMs vs. avaliação por juízes humanos vs. avaliação usando o modelo automatizado *Creassessment?*

Para responder essas perguntas de pesquisa é realizada uma série de estudos de casos seguindo o procedimento proposto por Yin (2009).

Tabela 17. Visão geral dos estudos a serem realizados.

Estudo	Objetivo	Tamanho da amostra	LLMs	Prompt automático	Especialistas	Creassessment
1A	Analisar interrater agreement/reliability entre LLMs	1.078 apps	ChatGPT 3.5 e Gemini Pro	X	--	--
1B		35 apps	ChatGPT 3.5, Gemini Pro, Llama3 70b, Mistral 7b e Vicuna 13b	X	--	--
2A	Analisar a correlação entre LLMs vs. especialistas vs. modelo automatizado (Creassessment)	1.078 apps	ChatGPT 3.5 e Gemini Pro	X	MIT App of the month (winner/no-winner)	X
2B		35 apps	ChatGPT 3.5, Gemini Pro, Llama3 70b, Mistral 7b e Vicuna 13b	X	MIT App of the season (Creativity)	X

Amostra de apps. Para a análise dos resultados dos estudos, foram usados dois conjuntos diferentes de *Apps* como artefato a serem avaliados. Conjuntos de dados os quais foram escolhidos por já termos a avaliação dos juízes humanos do MIT para os aplicativos:

- 1.078 *apps* do *App Inventor* aleatoriamente selecionados da competição MIT App of the month (de 2016 até 2022), os quais são divididos em dois grupos *Winners* e *No-Winners*, sendo 832 *No-Winners* e 246 *Winners*.
- 35 *apps* do *App Inventor* selecionados do *appathon* do *App of the Season* 2023, os quais tiveram sua criatividade avaliada por um grupo de juízes da competição App of the Month do MIT/EUA.

LLMs utilizados. Para a avaliação dos conjuntos de *Apps* foram selecionadas diferentes LLMs como avaliadores e foi solicitado resposta via interface web e via API.

- ChatGPT 3.5 - proprietário - https://chat.openai.com/
- Gemini Pro - proprietário - https://gemini.google.com/app
- Llama3 - código aberto - https://chat.lmsys.org/
- Mistral 7b - código aberto - https://chat.lmsys.org/
- Vicuna 13b - código aberto - https://chat.lmsys.org/

Prompts utilizados. Para a solicitação da avaliação das LLMs foram utilizados prompts criados automaticamente com base nos dados automaticamente extraídos do código dos *apps* App Inventor (.aia) utilizando o modelo *Creassessment* (Anexo A). Foram extraídas informações como funcionalidades, tags, componentes UI, conteúdo textual e blocos de código automaticamente para todos os aplicativos, e a partir disso foi montado automaticamente o *prompt* para solicitação das respostas às LLMs usando como base a técnica zero-shot de *prompt engineering*.

Research design. Foi adotado um *Mixed Factorial Design*, analisando a qualidade das respostas dos LLMs baseado em uma amostra de aplicativos *App Invento*r e em um grupo fixo de avaliadores, sempre todos os avaliadores avaliando todos os apps.

Coleta de dados. Foram coletados os seguintes dados:

LLMs	Escala de [0 a 10] com 0 nada criativo e 10 totalmente criativo com arredondamento para o meio ponto (arredonda os valores para o múltiplo mais próximo de 0,5 (ou meio ponto).
Creassessment	Escala de [0 a 10] com 0 nada criativo e 10 totalmente criativo com arredondamento para o meio ponto. (arredonda os valores para o múltiplo mais próximo de 0,5 (ou meio ponto).
Juízes humanos da competição MIT *App of the Month*	Escala nominal [*Winner*, *No-Winner]*
Juízes humanos da competição MIT Appathon 2023	Escala de [0 a 4] a qual foi convertida para [0 a 10] por meio de uma transformação linear simples para comparação com os outros modelos de avaliação.

Métodos de análise de dados. Para a análise dos resultados dos estudos, foram usados vários tipos de análises estatísticas.

Com o objetivo de analisar a confiabilidade entre os avaliadores foram utilizados:

- Coeficiente de correlação de Pearson. O coeficiente de correlação Pearson será usado para todas as análises que envolvem variáveis quantitativas. Ele indica a confiabilidade entre as avaliações dos diferentes avaliadores. O valor da correlação de Pearson é interpretado da seguinte forma (Devellis, 2017): correlações acima de 0.9 indicam uma correlação muito forte, entre 0.7 a 0.9 correlação forte, 0.5 a 0.7 correlação moderada, 0.3 a 0.5 correlação fraca, 0 a 0.3 correlação desprezível.
- *Root Mean Square Error* (RMSE). O *Root Mean Square Error* (RMSE) indica a confiabilidade entre as avaliações dos diferentes avaliadores. O RMSE é interpretado de forma distinta ao coeficiente de correlação de Pearson, não seguindo uma escala específica como a interpretação de Pearson. Ele é usado para avaliar a qualidade do modelo em fazer previsões precisas em relação aos dados reais. Nesse caso, quanto menor o RMSE e mais próximo de 0, mais semelhante um conjunto de dados é em relação ao outro.

Voltado à análise de concordância entre os avaliadores foram utilizados:

- Coeficiente de correlação intraclasse (ICC). O ICC serve para avaliar a concordância entre os diferentes avaliadores para um mesmo conjunto de dados. Existem diferentes tipos de ICC, o escolhido para a análise nesse

projeto foi o ICC3, o qual entende que todos os dados foram avaliados por todo o conjunto de avaliadores. O coeficiente de correlação intraclasse pode ser interpretado da seguinte forma (Cichetti, 1994): correlação entre 0.75 e 1.0 indicam uma correlação excelente, 0.6 a 0.75 correlação boa, correlação 0.4 a 0.6 correlação razoável e menor que 0.4 correlação fraca. Seu intervalo de confiança (CI), indica dentro de qual faixa de valores o valor do ICC provavelmente está contido, por exemplo, um intervalo de confiança de 95% indica que há uma probabilidade de 95% de que o ICC daquela análise esteja contida dentro da faixa de valores mínimo e máximo apresentada.

- Dispersão por meio de um gráfico *Boxplot*. A partir do gráfico *Boxplot* pode se observar a dispersão das notas dos dois conjuntos de dados os quais serão comparados, e por meio do coeficiente pode-se ver se as notas tendem a diminuir ou aumentar nos diferentes níveis do gráfico. Um coeficiente negativo mostra que as notas tendem a diminuir naquele mesmo nível de um conjunto para o outro. Um coeficiente igual a 0 mostra que as notam tendem a permanecer na mesma média nos dois conjuntos. Um coeficiente positivo mostra que as notas tendem a aumentar naquele mesmo nível de um conjunto de dados para o outro.
- Gráfico de Bland-Altman. O gráfico Bland-Altman (Bland; Altman, 1986) avalia a concordância entre duas medidas quantitativas, examinando a diferença média entre elas e estabelecendo limites de concordância. Esses limites permitem identificar se há um viés sistemático nas diferenças médias e estimar um intervalo de concordância dentro do qual 95% das diferenças entre os dois métodos estão incluídas. Este método, é uma ferramenta para determinar a intercambialidade entre métodos e para detectar discrepâncias consistentes entre eles.

Além destas métricas, com o objetivo de analisar alguns dados, foram utilizados também o cálculo da média, mediana e diferença máxima.

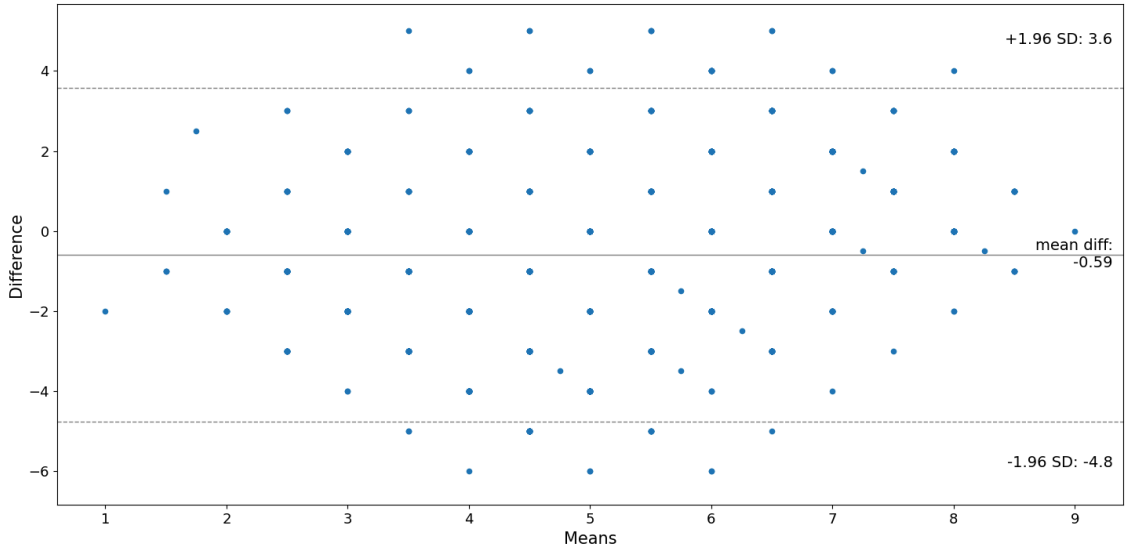
4.2 Estudo de caso 1A (PP1) - *interrater agreement/reliability* entre LLMs

<table>
<tr><td>Artefatos avaliados</td><td colspan="3">1.078 apps do App Inventor da App of the month.</td></tr>
<tr><td>LLMs utilizados</td><td colspan="3">Avaliação via API das LLMs: ChatGPT 3.5 e Gemini Pro</td></tr>
<tr><td>Prompt utilizado</td><td colspan="3">Avalie o grau de criatividade do seguinte App em uma escala de [0 a 10].
Responda de forma simples e objetiva com base na descrição abaixo.
Funcionalidades do aplicativo: [Funcionalidades extraídas do aplicativo usando creassessment]
Palavras-chaves: [tags extraídas do aplicativo usando creassessment]
Interface: [componentes UI extraídas do aplicativo usando creassessment]
Conteúdo textual: [Conteúdo textual extraído do aplicativo usando creassessment]
Blocos de código: [Blocos extraídos do aplicativo usando creassessment]</td></tr>
<tr><td>Parâmetros de respostas das LLMs</td><td colspan="3">Foi solicitada uma resposta simples e objetiva com base na descrição dos aplicativos, usando os parâmetros de default das LLMS e não foi realizado ajuste fino (fine-tuning).</td></tr>
<tr><td>Escala de resposta</td><td colspan="3">Escala de [0 a 10] com 0 nada criativo e 10 totalmente criativo com arredondamento para o meio ponto.</td></tr>
<tr><td rowspan="5">Análise estatística dos dados</td><td colspan="3">A correlação entre as avaliações das LLMs é avaliada pelo coeficiente de Pearson, pelo Root Mean Square Error (RMSE) e análise do gráfico Bland-Altman. É analisado também o coeficiente de correlação intraclasse (ICC), usando o modelo ICC3 (Mixed Factorial Design) em que todos apps foram avaliados por todos os avaliadores. E um gráfico Bland-Altman para analisar o intervalo de confiança entre as notas.</td></tr>
<tr><td colspan="3">Resultados</td></tr>
<tr><td>Correlação de Pearson</td><td>r = 0.4</td><td>Correlação fraca</td></tr>
<tr><td>RMSE</td><td>RMSE = 2.2</td><td>RMSE alto, ou seja, muita divergência em relação às notas dadas pelo ChatGPT 3.5 e Gemini Pro</td></tr>
<tr><td>ICC</td><td>ICC = 0.374
CI = [0.3, 0.44]</td><td>ICC pobre, mostrando pouca convergência nas notas. E com um intervalo de confiança entre 0.3 a 0.44.</td></tr>
</table>

A partir da análise, a correlação de *Pearson* e o RMSE indicam uma falta de confiabilidade entre as duas LLMs (ChatGPT 3.5 e Gemini Pro), pelo baixo coeficiente de *Pearson* **r = 0.4** e pelo alto RMSE (**RMSE = 2.2)**.

A análise do coeficiente de correlação intraclasse (ICC modelo 3), também indica um valor fraco **ICC = 0.374,** indicando a falta de uma concordância entre os LLMs avaliadores.

Figura 3 - Gráfico Bland-Altman para a avaliação ChatGPT 3.5 vs Gemini Pro



A partir da análise do gráfico Bland-Altman, percebe-se que o intervalo de confiança foi de -4.8 até 3.6, demonstrando que 95% das diferenças entre as medidas dos dois LLMs estão contidas em um intervalo de 8.4 unidades. Isso sugere que as diferenças nas avaliações do ChatGPT 3.5 e do Gemini Pro estão dentro de um intervalo muito grande, tendo em vista a escala usada, indicando novamente uma baixa concordância entre elas.

Desta forma, com os resultados de todas as análises (*Pearson*, RMSE, ICC e Bland-Altman) indicam uma baixa confiabilidade e concordância entre os LLMs avaliadores ChatGPT 3.5 vs Gemini Pro.

4.3 Estudo de caso 1B (PP1) - *interrater agreement/reliability* entre LLMs

Artefatos avaliados	35 apps *App Inventor appathon* do *App of the Season* 2023.
LLMs utilizados	Avaliação via interface web, usando ChatGPT 3.5, Gemini Pro, Llama3 70b, Mistral 7b e Vicuna13b
Prompt utilizado	Avalie o grau de criatividade do seguinte App em uma escala de [0 a 10]. Responda de forma simples e objetiva com base na descrição abaixo. Funcionalidades do aplicativo: [Funcionalidades extraídas do aplicativo usando *creassessment*] Palavras-chaves: [tags extraídas do aplicativo usando *creassessment*] Interface: [componentes UI extraídas do aplicativo usando *creassessment*] Conteúdo textual: [Conteúdo textual extraído do aplicativo usando *creassessment*] Blocos de código: [Blocos extraídos do aplicativo usando *creassessment*]
Parâmetros de LLMs	Foi solicitada uma resposta simples e objetiva com base na descrição dos aplicativos, usando os parâmetros de *default* das LLMS e não realizando ajuste fino (*fine-tuning*).
Escala de resposta	Escala de [0 a 10] com 0 nada criativo e 10 totalmente criativo com arredondamento para o meio ponto.
Tratamento de dados	Para cada aplicativo foram solicitadas 3 avaliações para cada LLM em momentos diferentes, e a partir das 3 avaliações foi calculada a média para calcular a nota final da avaliação do app.
Análise estatística dos dados	A correlação entre as avaliações das LLMs foi avaliada pelo coeficiente de Pearson, pelo *Root Mean Square Error* (RMSE) e pelo coeficiente de correlação intraclasse (ICC), usando modelo ICC3 (Mixed Factorial Design), em que todos *os* apps foram avaliados por todos os avaliadores.

Tratamento dos dados coletados. Para a avaliação dos 35 apps foi solicitado 3 avaliações para cada LLM utilizando o mesmo *prompt*, e a partir das 3 avaliações foi calculada a média para calcular a nota final da avaliação do app. Com o objetivo de analisar o grau com que a própria LLM concorda com ela mesmo, se para um mesmo *prompt* gera resultados semelhantes, foi calculada a média da diferença máxima das três notas atribuídas para cada app (Tabela 18).

Tabela 18 - Análise da média da diferença máxima entre as três notas das avaliações para cada LLM

Média da diferença máxima	**ChatGPT 3.5**	**Gemini Pro**	**Llama3 70b**	**Mistral 7b**	**Vicuna 13b**
	2	1.87	**0.94**	1.31	1.34

Através da média da diferença máxima das notas das LLMs, conseguimos analisar a variabilidade entre as notas de uma LLM com ela mesma, mostrando a

amplitude média de diferentes avaliações de uma LLM para um mesmo aplicativo. Por meio da escala [0..10] que está sendo usada o maior intervalo possível entre uma nota e outra é de 10 pontos. A menor média da diferença máxima foi da LLM Llama3, com uma média próxima do 1 ponto, mostrando-se á com a maior concordância entre suas próprias notas. A com a maior média de diferença máxima foi a do ChatGPT 3.5, com 2 pontos de média entre suas próprias avaliações.

Para a análise da confiabilidade das notas entre as LLMs foi analisado o coeficiente de correlação de Pearson (Tabela 19).

Tabela 19 - Análise do coeficiente de correlação de *Pearson*

Pearson	**ChatGPT 3.5**	**Gemini Pro**	**Llama3 70b**	**Mistral 7b**	**Vicuna 13b**
Chat GPT	-	0.13	0.27	**0.61**	0.28
Gemini Pro	0.13	-	0.38	0.1	-0.02
Llama3 70b	0.27	0.38	-	0.26	-0.15
Mistral 7b	**0.61**	0.1	0.26	-	0.18
Vicuna 13b	0.28	-0.02	-0.15	0.18	-

A maior correlação para Pearson entre as LLMs foi Mistral 7b vs ChatGPT 3.5, com uma correlação moderada de **r = 0.61**. As de mais todas apresentam correlações fracas ou desprezíveis, inclusive tendo uma correlação negativa entre as notas alocadas pelo Llama3 70b vs Vicuna 13b e Vicuna 13b vs Gemini Pro. Estes resultados mostram uma baixa confiabilidade na avaliação da criatividade entre as LLMs.

Analisando a confiabilidade entre as LLMs também pelo *Root Mean Square Error*, os resultados são apresentados na Tabela 20.

Tabela 20 - Análise usando *Root Mean Square Error*

RMSE	**ChatGPT 3.5**	**Gemini Pro**	**Llama3 70b**	**Mistral 7b**	**Vicuna 13b**
ChatGPT 3.5	-	1.7	2.17	1.57	1.52

Gemini Pro	1.7	-	1.76	1.48	**1.24**
Llama3 70b	2.17	1.76	-	1.48	2.19
Mistral 7b	1.57	1.48	1.48	-	1.53
Vicuna 13b	1.52	**1.24**	2.19	1.53	-

Analisando os resultados do *Root Mean Square Error* entre a avaliação das LLMs o menor resultado é **RMSE = 1.24** entre Vicuna vs Gemini.

O resultado do coeficiente de correlação intraclasse (modelo 3) **ICC = 0.166**, é um resultado fraco, com um intervalo de confiança (CI) entre 0.05 a 0.33, mostrando assim também uma baixa concordância entre as LLMs.

A partir desses resultados observa-se de forma geral que há baixa confiabilidade e concordância entre os LLMs, com melhores convergências entre Mistral 7b e ChatGPT 3.5.

4.4 Estudo de caso 2A (PP2) - LLMs vs. humanos vs. *Creassessment*

<table>
<tr><th></th><th colspan="2">Avaliação com LLMs</th><th colspan="2">Avaliação humana</th><th>Avaliação com modelo automatizado</th></tr>
<tr><td>Artefatos avaliados</td><td colspan="5">1.078 apps do App Inventor da App of the month.</td></tr>
<tr><td rowspan="3">Técnica de Avaliação</td><td>LLMs utilizados</td><td>Avaliação via API das LLMs: ChatGPT 3.5 e Gemini Pro.</td><td>Avaliadores</td><td>Foram comparados 1078 apps disponibilizados pelo App Inventor team, dos quais 246 (23%) foram identificados como winners e 832 (73%) como non-winners da competição App of the Month, sendo julgados pelos membros do MIT App Inventor</td><td rowspan="3">*Creassessment* (Alves, 2023)</td></tr>
<tr><td>Prompt utilizado</td><td>Avalie o grau de criatividade do seguinte App em uma escala de [0 a 10]. Responda de forma simples e objetiva com base na descrição abaixo.
Funcionalidades do aplicativo: [Funcionalidades extraídas do aplicativo usando creassessment]
Palavras-chaves: [tags extraídas do aplicativo usando creassessment]
Interface: [componentes UI extraídas do aplicativo usando creassessment]
Conteúdo textual: [Conteúdo textual extraído do aplicativo usando creassessment]
Blocos de código: [Blocos extraídos do aplicativo usando creassessment]</td><td>Rubricas e critérios de avaliação</td><td>Os apps foram revisados pelos membros do MIT App Inventor considerando:

Design - The app is the most aesthetically pleasing.

Innovation - The app uses App Inventor technology in the most interesting/unique way.

Creativity - The app that best uses creative elements such as art, color, sound, or movement</td></tr>
<tr><td>Parâmetro</td><td>Foi solicitada uma resposta</td><td>Quantidade</td><td>Não informado.</td></tr>
</table>

	s de LLMs	simples e objetiva com base na descrição dos aplicativos, serão usados os parâmetros de *default* das LLMS e não será realizado ajuste fino (*fine-tuning*).	**de avaliadores humanos**		
Escala de resposta	**LLMs**: Escala de [0 a 10] com 0 nada criativo e 10 totalmente criativo com arredondamento para o meio ponto. **Creassessment**: Escala de [0 a 10] com 0 nada criativo e 10 totalmente criativo com arredondamento para o meio ponto. **Juízes appathon**: Escala nominal [*Winner* ou *No-Winner]*.				
Análise estatística dos dados	A análise foi feita média/mediana e dispersão por meio de um gráfico *Boxplot* para avaliar LLMs vs juízes humanos e foi calculado o coeficiente de correlação Pearson, RMSE, ICC usando modelo ICC3 (Mixed Factorial Design) e Bland-Altman para analisar LLMs vs *Creassessment*.				

Com o objetivo de avaliar a correlação entre LLMs e juízes humanos, foram feitos testes a partir da média, mediana e dispersão por meio de um gráfico *Boxplot* para analisar se as LLMs conseguem diferenciar os grupos de *Winners* e *No-Winners* classificados por juízes humanos.

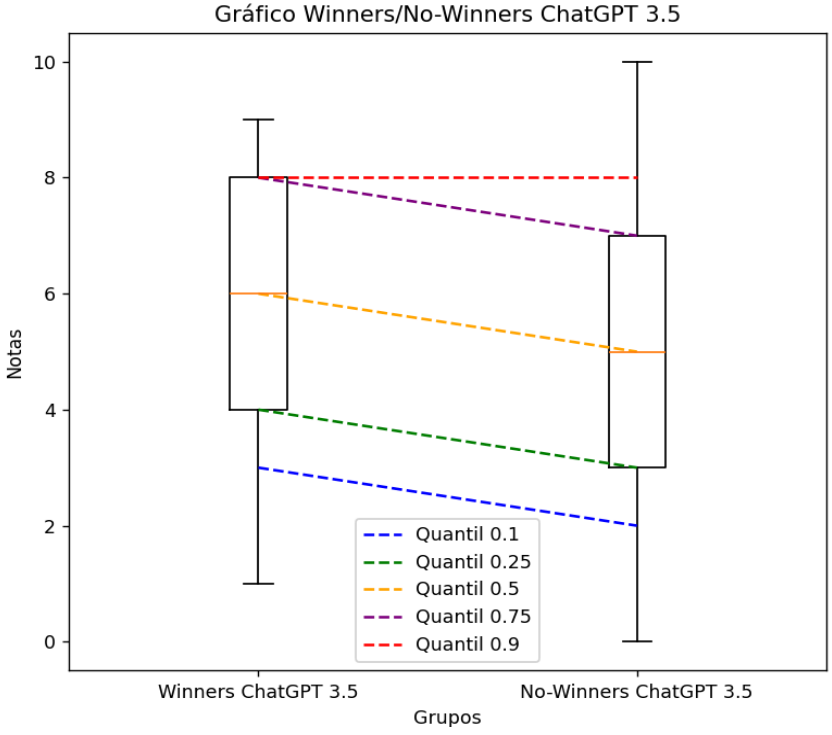
Tabela 21 - Análise da média e mediana entre grupos *Winners* e *No-Winners*

	Média		**Mediana**	
	Winner	***No-Winner***	***Winner***	***No-Winner***
ChatGPT 3.5	5.63	4.92	6	5
Gemini Pro	6	5.57	6	6

Como resultado observa-se que a média dos *Winner* para as duas LLMs é maior do que a média para os *No-Winner*, porém

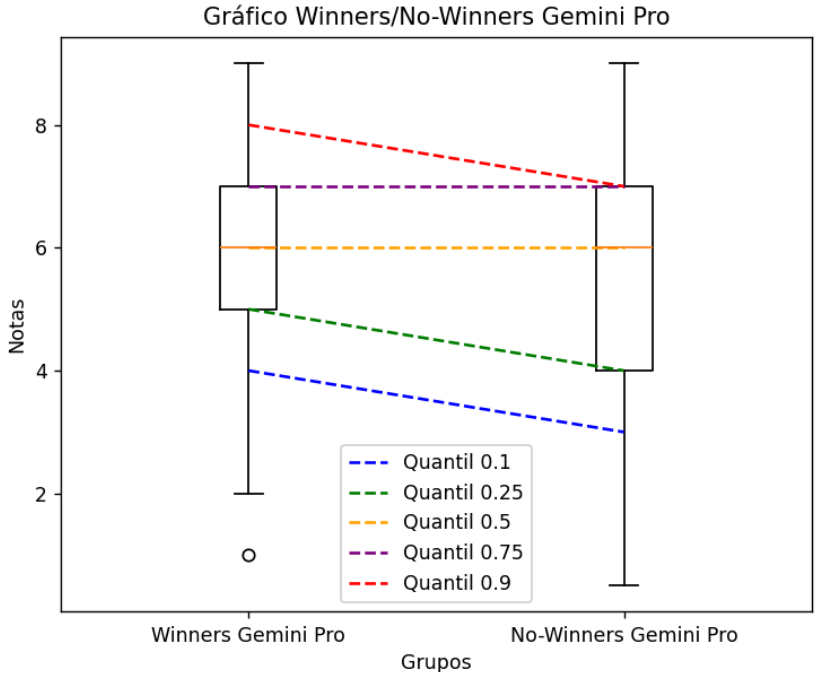
com pouca diferença. A mediana segue uma tendência parecida, com o grupo de *Winner* tendo uma mediana maior ou no mínimo igual ao grupo de Winner. Assim, observa-se que as LLMs tendem a dar notas um pouco mais altas para o grupo de *Winner* do que ao grupo de *No-Winner*. Isso mostra uma possível diferenciação nas notas de apps de grupos diferentes com graus de criatividade diferentes, porém com pouca diferença.

Figura 4 - Gráfico *Boxplot* para dispersão das notas por grupo do ChatGPT 3.5



Quantis	Coeficiente
0.1	-1
0.25	-1
0.5	-1
0.75	-1
0.9	0

Figura 5 - Gráfico Boxplot para dispersão das notas por grupo do Gemini



Quantis	Coeficiente
0.1	-1
0.25	-1
0.5	0
0.75	0
0.9	-1

Por meio dos dois gráficos Boxplot considerando as medianas, percebe-se que em todos os quantis temos coeficientes nulos ou negativos passando do grupo de *Winner* para o grupo *No-Winner*, evidenciando que em todos os quantias ou as notas baixam ou as notas permanecem iguais em relação a *Winner* para *No-Winner*. Percebe-se também que o ChatGPT 3.5 consegue diferenciar um pouco melhor os apps nos grupos do que o Gemini Pro. Pois tanto na média quanto na mediana teve uma diferença mais significativa, quanto no gráfico Boxplot apenas no quantis 0.9 teve um coeficiente nulo, de resto, todos foram coeficientes negativos.

Comparando também as avaliações das LLMs com as notas alocadas pelo modelo de avaliação automático *Creassessment* (Alves, 2023) considerado um *ground truth*, foi analisado a confiabilidade e concordância entre as notas dos LLMs e *Creassessment* (Tabela 22).

Tabela 22 - Análise do Coeficiente *Pearson*, RMSE e ICC (modelo ICC3) e seu intervalo de confiança (CI) entre *Creassessment* vs ChatGPT 3.5 e Gemini Pro

LLMs	***Creassessment***		
	Pearson	**RMSE**	**ICC (modelo ICC3)**
ChatGPT 3.5	**0.51**	2.0	ICC = 0.37, CI = 0.28 a 0.46
Gemini Pro	0.46	**1.48**	**ICC = 0.42, CI = 0.37 a 0.47**

Observa-se coeficientes de correlação Pearson semelhantes entre ChatGPT 3.5 vs *Creassessment* e Gemini Pro vs *Creassessment*, sendo ambos próximos do limiar entre um coeficiente baixo e moderado. Sendo ChatGPT 3.5 vs *Creassessment* apresentando uma maior correlação (**r = 0.51**). Estes resultados mostram uma baixa confiabilidade nas avaliações das LLMs em relação ao modelo *Creassessment*.

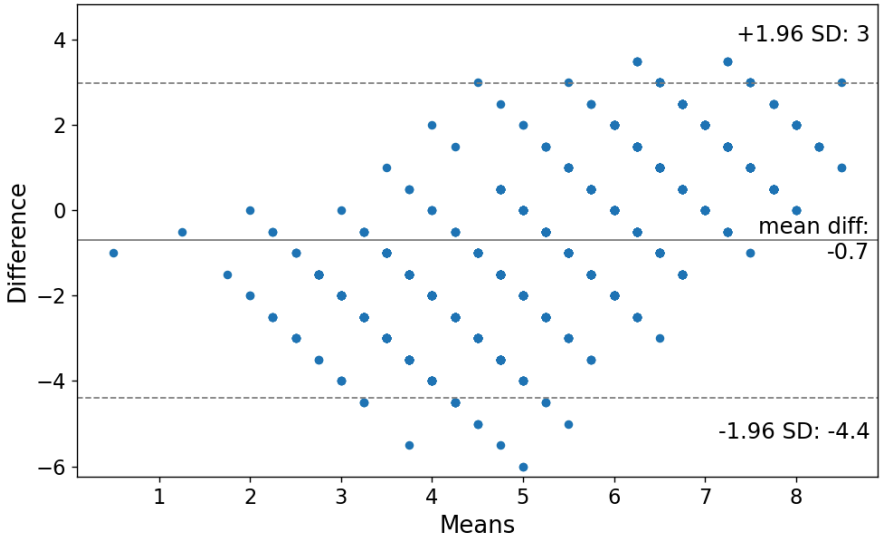
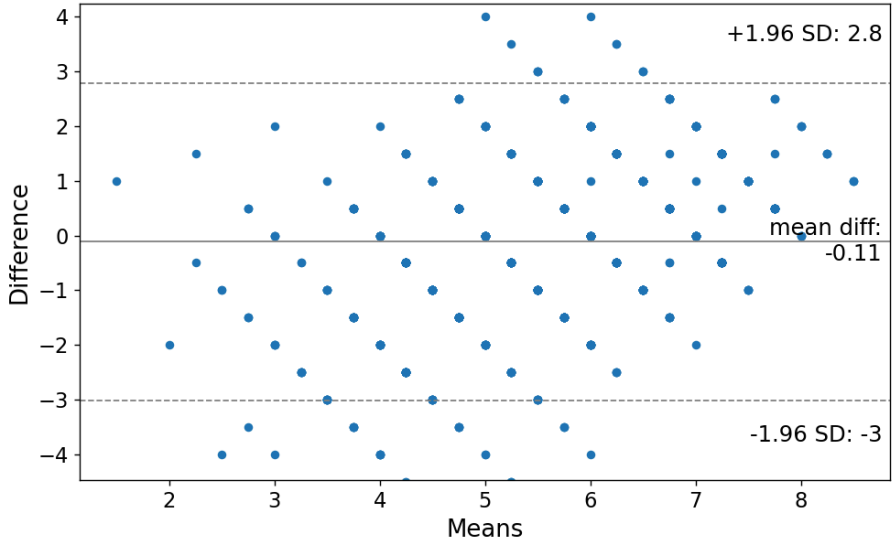
Os resultados da análise do RMSE também confirmam esta falta de confiabilidade nas avaliações dos LLMs em

comparação com as notas do modelo *Creassessment*, como ambas as LLMs, ChatGPT 3.5 e Gemini Pro, mostraram altos resultados de RMSE em relação às avaliações alocadas pelo modelo *Creassessment*.

Os resultados do ICC (modelo ICC3) para ChatGPT 3.5 vs *Creassessment* e Gemini Pro vs *Creassessment* mostraram baixa correlação. Ambos os resultados estão próximos do limiar entre um ICC moderado e um ICC baixo. Estes resultados mostram novamente uma baixa concordância entre as notas das LLMs e do modelo *Creassessment*.

Tabela 23 apresenta os gráficos Bland-Altman entre ChatGPT 3.5 vs *Creassessment* e Geminio Pro vs *Creassessment.*

Tabela 23 - Gráficos Bland-Altman para as avaliações *Creassessment* vs ChatGPT 3.5 e Gemini Pro

Bland-Altman ChatGPT 3.5 vs. *Creassessment*	Bland-Altman Gemini Pro vs. *Creassessment*
	

A partir da análise dos gráficos Bland-Altman, percebe-se que o intervalo de confiança foi de -4.4 até 3 para ChatGPT 3.5

vs *Creassessment* e -3 até 2.8 para Gemini Pro vs *Creassessment,* demonstrando que 95% das diferenças entre as medidas das LLMs e do modelo *Creassessment* estão contidas em um intervalo de 7.4 e 5.8 respectivamente. Sendo o menor intervalo entre Gemini Pro vs *Creassessment*, porém ainda assim os dois resultados mostram um intervalo muito grande, tendo em vista a escala de avaliação usada, indicando novamente uma baixa concordância entre elas.

4.5 Estudo de caso 2B (PP2) - LLMs vs. humanos vs. *Creassessment*

<table>
<tr><th></th><th colspan="2">Avaliação com LLMs</th><th colspan="2">Avaliação humana</th><th>Avaliação com modelo automatizado</th></tr>
<tr><td>Artefatos avaliados</td><td colspan="5">35 apps App Inventor appathon do App of the Season 2023.</td></tr>
<tr><td rowspan="3">Técnica de Avaliação</td><td>LLMs utilizados</td><td>Avaliação via interface web, usando as LLMs: ChatGPT 3.5, Gemini Pro, Llama3 70b, Mistral 7b e Vicuna 13b.</td><td>Avaliadores</td><td>Avaliadores do appathon do App of the Season 2023.</td><td rowspan="3">Creassessment (Alves, 2023)</td></tr>
<tr><td>Prompt utilizado</td><td>Avalie o grau de criatividade do seguinte App em uma escala de [0 a 10]. Responda de forma simples e objetiva com base na descrição abaixo.
Funcionalidades do aplicativo: [Funcionalidades extraídas do aplicativo usando creassessment]
Palavras-chaves: [tags extraídas do aplicativo usando creassessment]
Interface: [componentes UI extraídas do aplicativo usando creassessment]
Conteúdo textual: [Conteúdo textual extraído do aplicativo usando creassessment]
Blocos de código: [Blocos extraídos do aplicativo usando creassessment]</td><td>Rúbricas e critérios de avaliação</td><td>4 - App idea is very unique or a novel take on other ideas - App demonstrates significant potential impact related to the theme and can effectively help the target audience.
3 - App has some novel aspects to it - App demonstrates potential impact related to the theme and has potential to help the target audience.
2 - App idea is somewhat novel, or extends an existing idea slightly - App demonstrates some potential impact and/or relation to the theme.
1 - Very little in terms of unique or new ideas - App demonstrates little potential impact and/or relation to the theme.
0 - Remake of an existing app - No impact or relation to the theme.</td></tr>
<tr><td>Parâmetros de LLMs</td><td>Foi solicitada uma resposta simples e objetiva com base na descrição dos aplicativos, foram usados os parâmetros de default das LLMS e não foi realizado</td><td>Quantidade de avaliadores humanos</td><td>Não informado.</td></tr>
</table>

		ajuste fino (*fine-tuning*).			
Escala de resposta	**LLMs**: Escala de [0 a 10] com 0 nada criativo e 10 totalmente criativo com arredondamento para o meio ponto. **Creassessment**: Escala de [0 a 10] com arredondamento para o meio ponto **Juízes appathon**: Escala de [0 a 4] a qual foi normalizada para [0 a 10] com arredondamento para o meio ponto para comparação com os outros modelos.				
Tratamento de dados	Para cada aplicativo foram solicitadas 3 avaliações para cada LLM, e a partir das 3 avaliações foi calculada a média usando essa média na análise.				
Análise estatística dos dados	A correlação entre as avaliações das LLMs, do *Creassessment* e das notas de juízes humanos foi avaliada pelo coeficiente de *Pearson*, *Root Mean Square Error* (RMSE) e coeficiente de correlação intraclasse usando o modelo ICC3 em que todos *apps* foram avaliados por todos os avaliadores.				

Tratamento dos dados coletados. Os dados das notas das LLMs deste estudo são os mesmos do estudo 1B, o cálculo das médias das 3 notas obtidas por cada aplicativo e a diferença máxima entre as notas são as mesmas como no estudo 1B.

Considerando as avaliações feitas pelos juízes especialistas no *appathon* e as notas alocadas pelo modelo *Creassessment* como *ground truth*, sao comparadas as notas dadas pelos LLMs por meio da análise da correlação entre LLMs vs juízes humanos e LLMs vs *Creassessment* (Tabela 24).

Tabela 24 - Análise do Coeficiente *Pearson*

Pearson	**Chat GPT 3.5**	**Gemini Pro**	**Llama3 70b**	**Mistral 7b**	**Vicuna 13b**
Juízes humanos	0.18	0.04	0.3	**0.4**	-0.09
Creassessment	0.28	-0.02	0.0002	**0.56**	0.13

Observa-se de forma geral coeficientes de correlação Pearson baixas comparando as notas das LLMs com as notas alocadas pelos juízes humanos. A maior correlação foi da LLM Mistral 7b, com um **r = 0.4** com os juízes humanos. Observa-se novamente até uma correlação negativa no caso do Vicuna 13b.

Comparando as notas do LLMs com as notas alocadas pelo modelo Creassessment, observam-se em geral coeficientes de correlação Pearson mais baixas ainda, também com uma correlação negativa do Gemini. A única exceção novamente é o Mistral 7b com um coeficiente de correlação moderada (**r = 0.56)**.

Tabela 25 - Análise do *Root Mean Square Error*

RMSE	ChatGPT 3.5	Gemini Pro	Llama3 70b	Mistral 7b	Vicuna 13b
Juízes humanos	1.89	1.65	2.03	**1.61**	1.63
Creassessment	1.75	1.41	1.61	**0.73**	1.41

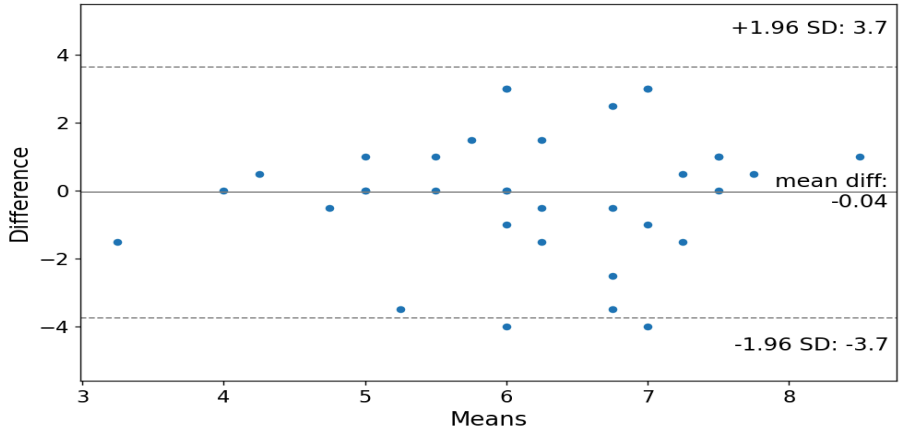
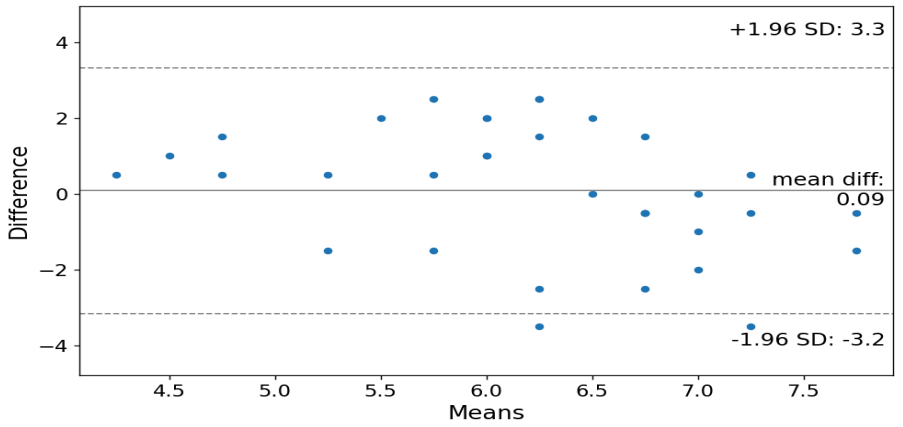
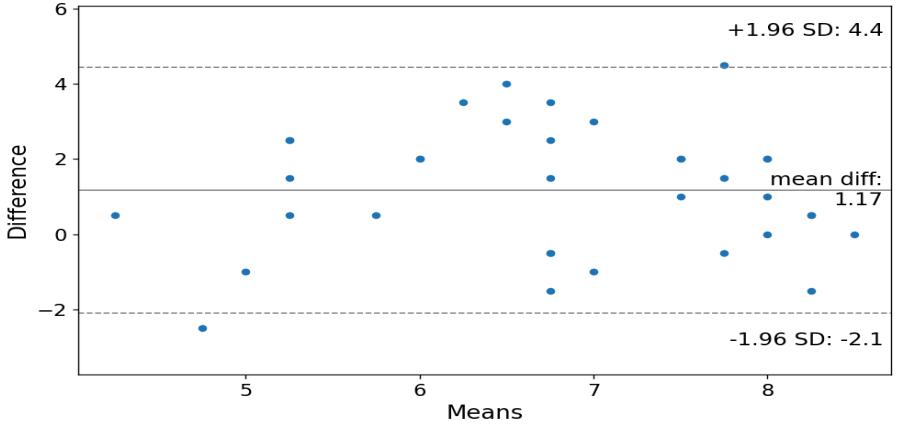
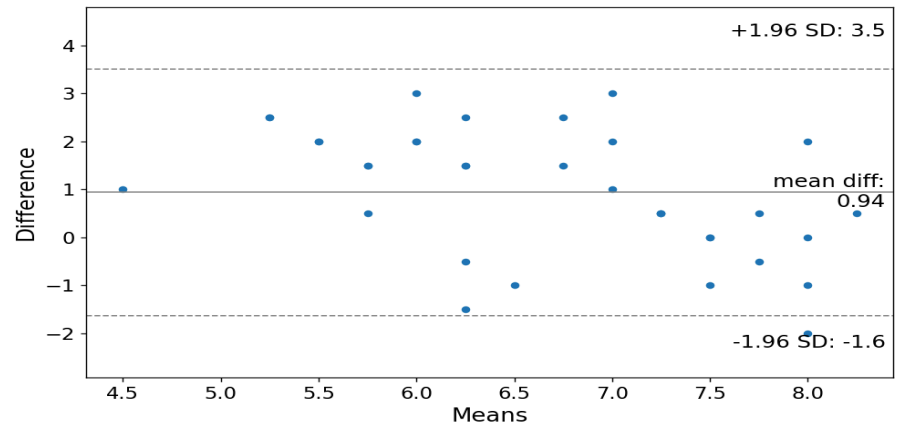
Os resultados da análise do RMSE também confirmam os resultados de *Pearson*. Novamente a LLM que mostrou o melhor desempenho no resultado do RMSE com os juízes humanos e o *Creassessment* foi a Mistral, resultando em um **RMSE = 1.61** em comparação aos juízes humanos e **RMSE = 0.73** em comparação ao *Creassessment*.

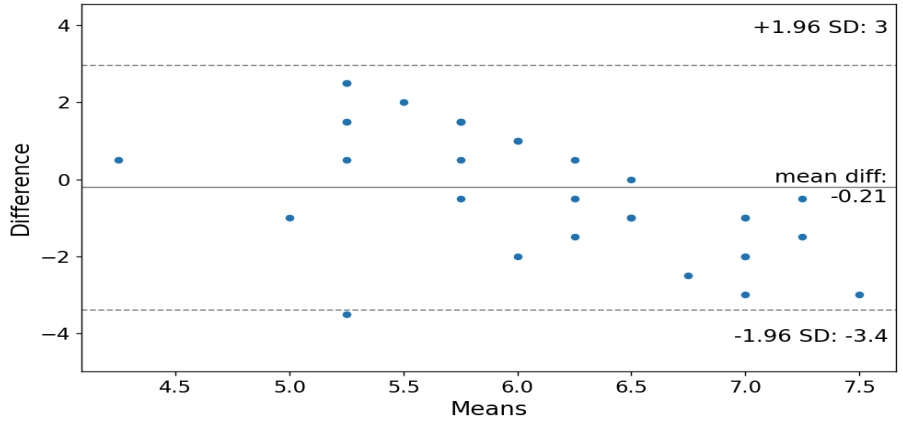
Tabela 26 - Análise do ICC (modelo ICC3) e do seu intervalo de confiança (CI)

ICC	ChatGPT 3.5	Gemini Pro	Llama3 70b	Mistral 7b	Vicuna 13b
Juízes humanos	ICC = 0.18, CI = -0.16 a 0.49	ICC = 0.04, CI = -0.3 a 0.37	ICC = 0.23, CI = -0.06 a 0.5	**ICC = 0.27, CI = -0.04 a 0.54**	ICC = -0.07, CI = -0.4 a 0.26
Creassessment	ICC = 0.16, CI = -0.1 a 0.44	ICC = -0.017, CI = -0.23 a 0.24	ICC = 0, CI = -0.32 a 0.33	**ICC = 0.55, CI = 0.27 a 0.75**	ICC = 0.06, CI = -0.09 a 0.27

Assim como para o coeficiente de correlação de Pearson e o RMSE, a LLM com o melhor resultado em relação ao ICC (modelo 3) foi a Mistral 7b, com um ICC pobre (**ICC = 0.27)** e um CI entre -0.04 a 0.54 em comparação com os juízes humanos e um ICC razoável próximo ao bom (**ICC = 0.55**) e um CI entre 0.25 a 0.75 em comparação com o *Creassessment*. Porém de forma geral observa-se novamente que os resultados apontam uma concordância muito fraca dos LLMs com as notas alocadas pelos juízes humanas e/ou alocadas pelo modelo *Creassessment*, inclusive com valores negativos no caso da Vicuna 13b em comparação aos juízes humanos e do Gemini Pro em comparação ao Creassessment.

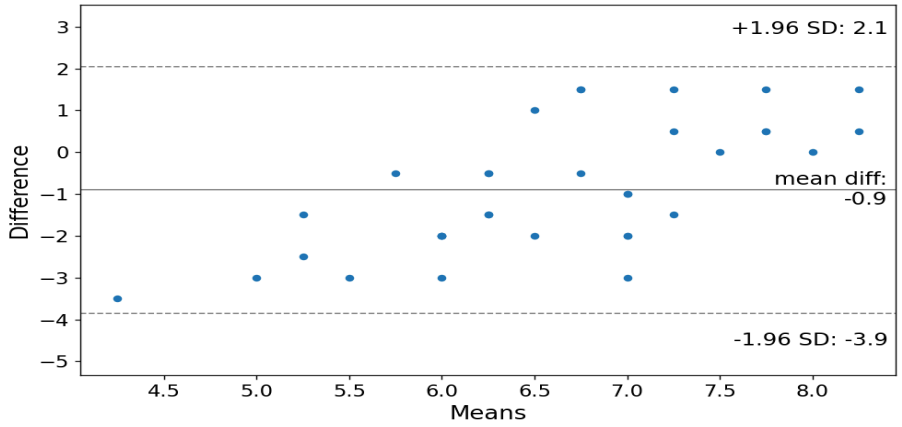
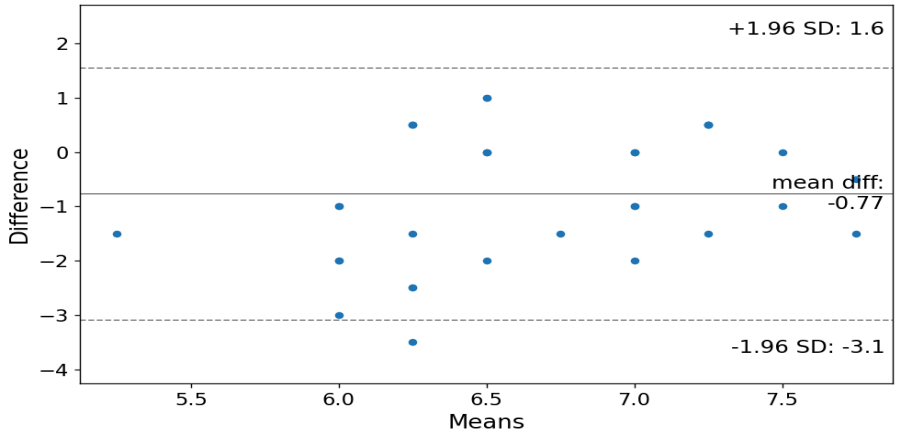
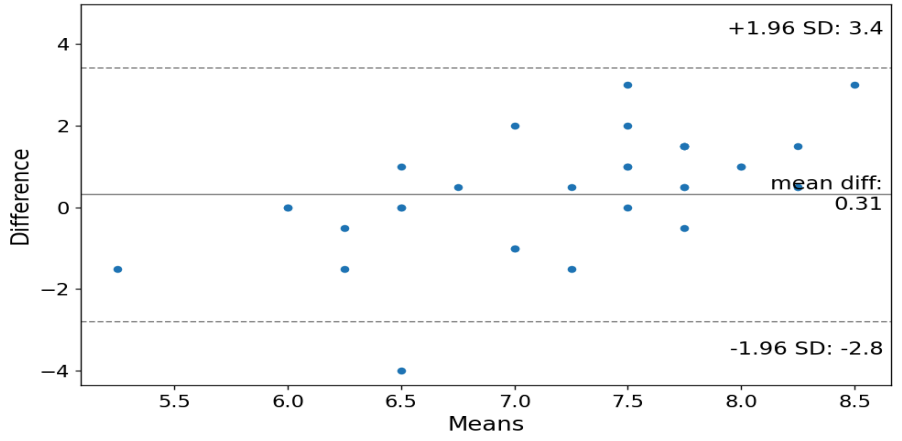
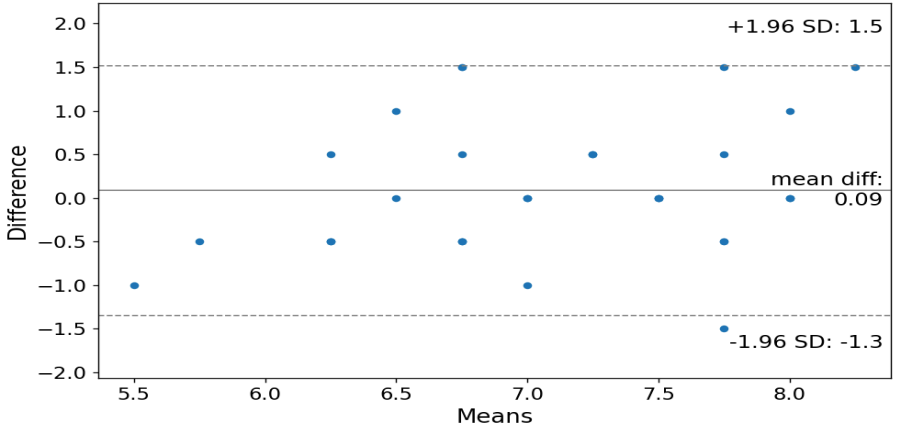
Tabela 27 - Gráficos Bland-Altman: juízes humanos vs. LLMs

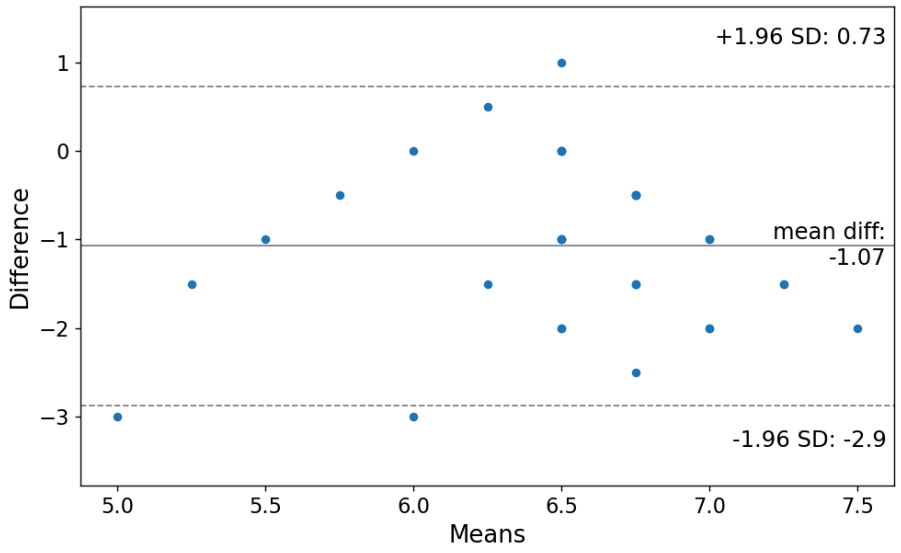
Juízes humanos vs. ChatGPT 3.5	Juízes humanos vs. Gemini Pro
	
Juízes humanos vs. Llama3 70b	**Juízes humanos vs. Mistral 7b**
	

Juízes humanos vs. Vicuna 13b	
	

A partir da análise dos gráficos Bland-Altman, percebe-se que o menor intervalo de confiança foi o da LLM Mistral 7b, com um intervalo de -1.6 até 3.5, demonstrando que 95% das diferenças entre as medidas das LLMs e das avaliações feitas por juízes humanos estão contidas em um intervalo de 5.1 pontos. Mesmo sendo o menor intervalo ainda é um intervalo muito grande, tendo em vista a escala da avaliação usada sendo de 0 a 10. Desta forma indicando novamente uma baixa concordância entre juízes humanos e LLMs.

Tabela 28 - Gráficos Bland-Altman: *Creassessment* vs LLMs

Creassessment vs. ChatGPT 3.5	*Creassessment* vs. Gemini Pro
	
Creassessment vs. Llama3 70b	*Creassessment* vs. Mistral 7b
	

Creassessment vs. Vicuna 13b	
	

A partir da análise dos gráficos Bland-Altman, percebe-se que o menor intervalo de confiança foi novamente da LLM Mistral 7b, com um intervalo de -1.3 até 1.5, demonstrando que 95% das diferenças entre as medidas das LLMs e das avaliações feitas pelo modelo *Creassessment* estão contidas em um intervalo de 2.8 pontos. Sendo um intervalo menor que o apresentado pela LLM Mistral 7b vs juízes humanos, porém ainda sendo um intervalo de quase 3 pontos, indicando uma baixa concordância entre LLMs vs *Creassessment* novamente.

Desta forma percebe-se que a LLM Mistral 7b se mostrou a melhor em resultados em todos os estudos , tanto em comparação às avaliações feitas pelos juízes humanos quanto em comparação às notas alocadas pelo o *Creassessment*. Mostrando resultados moderados, porém melhores em comparação com as outras LLMs. Já para as outras LLMs ChatGPT 3.5,

Gemini Pro, Llama3 70b e Vicuna 13b apresentaram resultados insatisfatórios, com resultados mostrando uma baixa correlação com os juízes humanos e o modelo automatizado *Creassessment*.

4.6 Discussão

Analisando os dados coletados, pode-se concluir que o nível de concordância e confiabilidade de avaliação de criatividade entre as LLMs não é satisfatório. Para ambos os estudos, 1A e 1B, os testes de concordância e confiabilidade entre as LLMs demonstraram resultados fracos. Resultados os quais podem vir pelo fato das LLMs serem treinadas por diferentes conjuntos de dados, tanto em forma quantitativa quanto em relação ao conteúdo desses dados. Podendo ter assim cada LLM uma noção diferente do que é um app criativo, algo aceitável de se esperar, visto que criatividade pode ser algo mais subjetiva de se interpretar.

Porém, em relação à concordância entre suas próprias notas para um mesmo aplicativo, os resultados não foram ruins, tendo em vista os testes da média da diferença máxima entre suas próprias notas (Tabela 18), com a maior diferença foi de 2 unidades, da LLM ChatGPT 3.5, e a menor diferença foi de aproximadamente 1 unidade, da LLM Llama3 70b, tendo em vista uma escala de 0 a 10, pode-se dizer que foram diferenças máximas baixas, entendendo assim que existe uma concordância entre as notas da própria LLM para um mesmo aplicativo.

Em relação a correlação da avaliação da criatividade comparando a avaliação por LLMs com avaliação por juízes humanos e avaliação usando o modelo automatizado *Creassessment*, a partir dos dados coletados pode-se concluir que no geral houve uma baixa correlação entre as avaliações dos juízes humanos com as LLMs e do modelo *Creassessment* com as LLMs. A maioria das correlações foram fracas e moderadas. A LLM que se mostrou com a melhor correlação foi a LLM Mistral 7b, que obteve os melhores resultados em praticamente todas avaliações de correlação e concordância com os outros dois modelos de avaliação. Tendo correlações *Pearson* moderadas com o *Creassessment* no estudo 2B e perto de moderada com os juízes humanos, além de um ICC razoável e próximo de bom com o *Creassessment* também no estudo 2B. Porém um ponto positivo é o resultado do estudo 2A indicando que as LLMs conseguem distinguir os grupos de apps entre *Winners* e *No-Winners,* ainda assim com pouca diferença. A partir disso, com todos os testes apresentados, podemos concluir que a melhor LLM em relação à concordância com os outros modelos de avaliação para avaliação de criatividade de aplicativos criados com *App Inventor* foi a LLM Mistral 70b, porém no geral tivemos baixas concordâncias entre as LLMs e os outros modelos de avaliação.

Limitações. Para minimizar ameaças a validade de conclusão foram realizados estudos adicionais com tamanho de amostra maior (1.078 apps), além dos estudos com somente 35 do *App of the Month App of the Season*. Foram também escolhidos métodos estatísticos apropriados tanto para a pergunta de pesquisa quanto em relação ao tipo de dados e quantidade de dados coletados.

Em relação à validade interna se criou protocolos de coleta de dados explícitos. A seleção dos apps avaliados foi feita tanto da galeria do *App Inventor* quanto do conjunto de apps do *App of the Month*. Os prompts foram gerados de forma automatizada pelos dados extraídos automaticamente dos apps utilizando o modelo *Creassessment*. Porém, apesar do modelo *Creassessment* ser um modelo já validado, o uso das informações automaticamente extraídas pelo modelo para geração do *prompt* automatizado pode ter influenciado nos resultados.

Para minimizar ameaças em termos de validade externa foram utilizadas várias LLMs diferentes. Em termos de avaliação humana foram utilizadas as classificações dos apps em *Winners* e *No-Winners* e também a criatividade avaliada por um grupo de juízes da competição *App of the Month* do MIT/EUA.

5. Conclusão

Como resultado do presente trabalho foi feita uma fundamentação teórica relacionada à avaliação da criatividade no contexto do ensino de computação na educação básica, além dos diferentes meios de avaliação da criatividade e sobre as *Large Language Models* (LLMs) e sua capacidade de entender e gerar textos (OE1). Foi levantado o estado da arte em relação a avaliação de LLMs para a avaliação da criatividade no contexto de ensino da computação na educação básica, indicando que ainda existem muito poucos estudos deste tipo e nenhum relacionado a avaliação de criatividade de aplicativos (OE2). Foram realizados diversos estudos empíricos para avaliar a qualidade de avaliação da criatividade com LLMs em relação a outros modelos de avaliação já validados, como juízes humanos e o modelo automatizado *Creassessment* (Alves, 2023), além de estudos empíricos

para avaliar o grau de confiabilidade e concordância entre as avaliações de diferentes LLMs (OE3).

A contribuição deste trabalho está na comparação da qualidade da avaliação da criatividade de aplicativos *App Invento*r com LLMs, comparadas com modelos de avaliação já validados, como juízes humanos e o modelo automatizado *Creassessment*, comparação a qual mostrou resultados insatisfatórios, com testes de correlação e concordância demonstrando resultados fracos e/ou moderados. Além disso, este trabalho também contribuiu para a análise da confiabilidade e concordância entre as respostas de diferentes LLMs, análises as quais demonstraram também baixa concordância e confiabilidade. Portanto, mostrando que a princípio e com a forma com que foram feitos os testes, as LLMs não seguem a mesma tendência de avaliação dos outros modelos de avaliação já validados, além de não terem entre si a mesma concordância para avaliação da criatividade de apps *App Inventor*.

Como sugestão para trabalhos futuros, sugere-se realizar novas aplicações e análises de testes utilizando *prompts* diferentes, podendo ser criados manualmente e/ou com mais dados e características dos aplicativos, ou com outras técnicas de *prompt engineering*. Além disso, também sugere-se utilizar diferentes LLMs das utilizadas neste trabalho, para verificar se existe alguma com um desempenho melhor ao ser comparada com outros modelos de avaliação já validados.

Referências

Amabile, T. M. The context of creativity. Boulder, CO: Westview, 1996.

Alves, N. C. Assessing mobile apps creativity in computing education. Tese (Doutorado em em Ciência da Computação) – PPGCC/Universidade Federal de Santa Catarina. 2023.

Birhane, A., Kasirzadeh, A., Leslie, D., Wachter S.. Science in the Age of Large Language Models. Nature Reviews Physics 5, 277–280, 2023.

Bland, J. Martin; Altman, Douglas G. Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. en. The Lancet, 327(8476), p. 307–310, 1986.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A.. Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901, 2020.

Campos, R. et al. YAKE! Keyword Extraction from Single Documents using Multiple Local Features. Information Sciences Journal, 509, p. 257-289, 2020.

Cichetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment, 6(4), 284-290.

Dair.ai. Elements of a prompt. 2023. https://www.promptingguide.ai/introduction/elements

Devellis, R. F. Scale development: theory and applications. 4th. ed. SAGE, 2017.

Goes, F., Zhou, Z., Sawicki, P., Grzes, M., & Brown, D. G. . Crowd Score: A Method for the Evaluation of Jokes using Large Language Model AI Voters as Judges. ArXiv Preprint, arXiv:2212.11214 [cs.AI], 2022.

Giray, L. Prompt Engineering with ChatGPT: A Guide for Academic Writers. Annals of Biomedical Engineering, v. 51, n. 12, 2023.

Guilford, J. P. Creativity. American Psychologist, v. 5, 1950.

Harunani, F., Patton, E. W., Tissenbaum, M.. MIT App Inventor: Objectives, Design, and Development. Computational Thinking Education, p. 31-49, 2019.

Hattie, J., Timperley, H.. The power of feedback. Review of Educational Research, 7(1), 81--112, 2007.

Hayes, A. F.; Coutts, J. J. Use omega rather than Cronbach's alpha for estimating reliability. But…. Communication Methods and Measures, 14(1), p. 1-24, 2020.

Kaufman, J. C.; Beghetto, R. A. Beyond Big and Little: The Four C Model of Creativity. Review of General Psychology, 13(1), p. 1-12, 2009.

Kaufman, J. C.; Beghetto, R. A. In praise of Clark Kent: Creative metacognition and the importance of teaching kids when (not) to be creative. Roeper Review, 35(3), p. 155-165, 2013.

Mishra, P.; Yadav, A. Of art and algorithms: Rethinking technology and creativity in the 21st century. TechTrends , 57(3), p. 10–14, 2013.

MIT App Inventor. Disponível em: https://appinventor.mit.edu/. Acesso em: 18 de novembro de 2023.

Organisciak, P., Acar, S., Dumas, D., & Berthiaume, K. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. Thinking Skills and Creativity, 49, 101356, 2023.

Piasecki, J., Waligora, M., Dranseika, V. Google Search as an Additional Source in Systematic Reviews. Science and Engineering Ethics, 24, p. 809- 810, 2018.

P21, Partnership for 21st century learning, P21 Framework Definitions, 2021. Disponivel em: http://www.p21.org/

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J.. Exploring the limits of transfer learning with a unified text-totext transformer. Journal of Machine Learning Research, 21(140), 1–67, 2020.

Renzulli, J. The Malleability of Creativity: A Career in Helping Students Discover and Nurture Their Creativity. In: Sernberg R.; Kaufman, J. The Nature of Human Creativity. Cambridge: Cambridge University Press, p. 209-223, 2018.

Rennie, J. D. et al. Tackling the poor assumptions of naive bayes text classifiers. Proceedings of the 20th International Conference on Machine Learning (ICML-03), Washington, D.C:ICML, p. 616-623, 2003.

Rhodes, M. An Analysis of Creativity. The Phi Delta Kappan, v.42, n. 7, p. 305-310, 1961.

Runco, M. A., Jaeger, G. J. The Standard Definition of Creativity. Creativity Research Journal, 24(1), p. 92-96, 2012.

Saunders, D., Thagard, P. Creativity in Computer Science. In Creativity across domains: Faces of the muse. J. C. Kaufman and J. Baer. Mahwah (eds.), NJ, Lawrence Erlbaum Associates, 2005.

Signity Solutions. Top Large Language Models. Disponível em: https://www.signitysolutions.com/blog/top-large-language-models. Acesso em: 15 de novembro de 2023.

Torrance, E. P.; Torrance Tests of Creative Thinking. Princeton, N.J.: Personnel Press, 1966.

Vaswani, A. et al. Attention Is All You Need. arxiv:1706.03762 [cs.CL], 15p., 2017.

WEF, World Economic Forum, Defining Education 4.0: A Taxonomy for the Future of Learning, 2023. Disponível em: https://www.weforum.org/publications/defining-education-4-0-a-taxonomy-for-the-future-of-learning/

Walia, C. A Dynamic Definition of Creativity. Creativity Research Journal, v. 31, n. 3, p. 237-247, 2019.36

Wei, J., Kim, S., Jung, H., Kim, Y.-H. Leveraging Large Language Models to Power Chatbots for Collecting User Self-Reported Data. arXiv:2301.05843 [cs.HC], 22 p., 2023.

Wohlin, C., et al. Experimentation in Software Engineering, Heidelberg: Springer, p. 236, 2012.

Yin, K., Case study research: design and methods, 4. ed. Beverly Hills: Sage Publications, p. 312, 2009.

Zhao, W. X. et al. A Survey of Large Language Models. arXiv:2303.18223 [cs.CL], 97 p., 2023.

ANEXO A - Exemplos de *prompts* automaticamente criados

Prompt para a avaliação da criatividade do Xô Dengue

O app Xô Dengue, criado pela Computação na Escola/INCoD/INE/UFSC, é um dos 35 aplicativos aleatoriamente escolhidos usados nos testes apresentados nos estudos 1B e 2B.

Prompt:

Avalie o grau de criatividade do seguinte App criado com App Inventor em uma escala de [0..10], com 0 indicando nada criativo e 10 muito criativo. Responda de forma simples e objetiva com base na descrição abaixo.

Nome do aplicativo: Xô Denque

Funcionalidades do aplicativo: DisplayInformation, SaveDataLocally

Palavras-chaves: font, larva, dengue, fechar, prevenção, color, classificação, uso, computação, qualquer

Interface: 57 Label, 13 Button, 4 Spinner, 8 Image, 1 WebViewer, 1 CheckBox, 1 Notifier

Blocos de codigo: Control, Logic, Math, Text, Lists, Dictionaries, Variables, Procedures, User_Interface, Layout, Storage, Screen, Helpers, Extensions

Conteudo texual: None Fechar menu Início Classificação de larva Prevenção de dengue Sobre o app Fechar app Sobre o app About Fechar app Início Prevenção de dengue About Classificação de larva None Fechar menu Início Classificação de larva Prevenção de dengue Sobre o aplicativo Fechar o aplicativo Sobre o aplicativo Fechar app Início Prevenção de dengue Classificação de larva Sobre o aplicativo Fechar o aplicativo Início Prevenção de dengue Classificação de larva None None None Fechar menu Início Classificação de larva Prevenção de dengue Sobre o aplicativo Fechar o aplicativo Nonmosquito Isto não parece ser uma larva de mosquito Não foi possível identificar a espécie dessa larva. Tente novamente tirando outra foto Isto provavelmente é uma larva de
 (% de confiança) Isto muito provavelmente é uma larva de
 (% de confiança) Aedes aegypti Screen2 <FONT COLOR="#700A05">Aceitação.</FONT> Este “Termo de Uso de Aplicativo” rege o uso do aplicativo Android Xô Dengue, gratuitamente disponibilizado. Se você não concordar com estes termos não use este aplicativo. Você reconhece que analisou e aceitou as condições de uso. Leia-as atentamente, pois o uso deste aplicativo significa que você aceitou todos os termos e concorda em cumpri-los. Se você, usuário, for menor de idade ou declarado incapaz em quaisquer aspectos, precisará da permissão de seus pais ou responsáveis que também deverão concordar com estes mesmos termos e condições.
<FONT COLOR="#700A05">Licença Limitada.</FONT> Você recebeu uma licença limitada, não transferível, não exclusiva, livre de royalties e revogável para baixar, instalar, executar e utilizar este aplicativo em seu dispositivo e desta forma não lhe transfere os direitos sobre o produto. O aplicativo deverá ser utilizado por você, usuário. A venda, transferência, modificação, engenharia reversa ou distribuição bem como a cópia de textos, imagens ou quaisquer partes nele contidas é expressamente proibida.
<FONT COLOR="#700A05">Propriedade Intelectual.</FONT> O conteúdo desenvolvido pela iniciativa Computação na Escola está protegido pelas leis e tratados de direitos autorais nacionais e internacionais, assim como outras leis e tratados de propriedade intelectual. O conteúdo é licenciado, não vendido. Você não pode fazer cópias não autorizadas ou usar qualquer conteúdo exceto como especificado aqui e de acordo com as leis aplicáveis. Você concorda em cumprir com todas as leis de proteção dos direitos autorais relacionadas ao uso dos Serviços e do Conteúdo. Não é permitido aos Usuários tentar reconfigurar, desmontar ou fazer engenharia

reversa no aplicativo. Você não está autorizado a utilizar nenhuma marca e/ou sinais distintivos encontrados no aplicativo. Você não pode copiar, exibir ou usar nenhuma das marcas sem consentimento prévio por escrito do seu proprietário. Qualquer uso não autorizado poderá violar as leis de propriedade industrial, leis de privacidade, de propriedade intelectual e ainda estatutos civis e/ou criminais.
<FONT COLOR="#700A05">Alterações, Modificações e Rescisão.</FONT> A qualquer tempo, estes termos podem ser modificados, seja incluindo, removendo ou alterando quaisquer de suas cláusulas com efeito imediato como a distribuição deste aplicativo. Após publicadas tais alterações, ao continuar com o uso do aplicativo você terá aceitado e concordado em cumprir os termos modificados. A iniciativa Computação na Escola não fornece nenhum serviço de suporte ou atualização para este aplicativo e não se responsabiliza por quaisquer modificações, suspensões ou descontinuidade do aplicativo.
<FONT COLOR="#700A05">Isenção de Garantias e Limitações de Responsabilidade.</FONT> Você usa este aplicativo a seu próprio e exclusivo risco. A iniciativa Computação na Escola se isenta de garantias e condições de qualquer tipo, sejam expressas, implícitas ou instituídas, incluindo, mas não se limitando a, garantias relacionadas à segurança, confiabilidade, conveniência e performance do aplicativo. Você entende e concorda que ao fazer download ou obter qualquer outra forma de acesso ao aplicativo você o faz ao seu próprio critério e risco e que você será o único responsável por quaisquer danos ao seu dispositivo ou perda de dados que resulte do download deste aplicativo. Você concorda que, em nenhuma circunstância, a iniciativa Computação na Escola poderá ser responsabilizada por qualquer dano direto, indireto, incidental, especial, consequencial ou punitivo, incluindo, mas não se limitando a, perdas e danos, lucros cessantes, perda de uma chance, outras perdas e danos intangíveis, relacionado ao uso do aplicativo, nem com relação à incapacidade e/ou impossibilidade de usá-los (incluindo hipóteses de negligência). Não se garante, declara e assegura que o uso deste aplicativo será ininterrupto ou livre de erros e você concorda que o aplicativo poderá ser removido por períodos indefinidos ou ser cancelado a qualquer momento sem que você seja avisado. Não se garante, declara nem assegura que este aplicativo esteja livre de perda, interrupção, ataque, vírus, interferência, pirataria ou outra invasão de segurança e isenta-se de qualquer responsabilidade em relação a essas questões. Você é responsável pelo backup do seu próprio dispositivo.
<FONT COLOR="#700A05">Indenização.</FONT> Você concorda em defender, indenizar e proteger a iniciativa Computação na Escola contra toda e qualquer reclamação, perda, dano, responsabilização, julgamento, avaliação, multa, custo e outras despesas, incluindo honorários advocatícios, resultantes ou relacionados ao uso feito por você dos Serviços ou de qualquer violação dos Termos de Uso, uso da Plataforma, Suas Informações, violação de qualquer lei, estatuto, ordem ou regulamento ou direitos de terceiros. aceitou 1 Para poder utilizar o aplicativo, você precisa confirmar que leu e aceita o termo de uso aceitou 0 1 Sobre o aplicativo Prevention Fechar o aplicativo Início Prevenção de dengue Prevention Classificação de larva None Fechar menu Início Classificação de larva Prevenção de dengue Sobre o aplicativo Fechar o aplicativo Sobre o app Prevention Instruções Screen2 UsageTerm COMO PREVENIR DENGUE A forma mais eficaz de prevenção é o combate ao mosquito aedes aegypti Seguem algumas ações que você pode fazer:
 • Verificar se a caixa d’água está bem tampada
• Deixar as lixeiras bem tampadas
• Colocar areia nos pratos de plantas
• Recolher e acondicionar o lixo do quintal
• Limpar as calhas
• Cobrir piscinas
• Tapar os ralos
 • Baixar as tampas dos vasos sanitários
• Limpar a bandeja externa da geladeira
• Limpar e guardar as vasilhas dos pets

• Limpar a coletora de água do ar-condicionado
• Cobrir bem a cisterna
• Cobrir bem todos os reservatórios de água
 INSTRUÇÕES Tirar foto da larva Com uma colher, coloque a larva em um recipiente branco como um prato, tampa de pote, etc. Pode ser com um pouco de água. Use o zoom máximo da sua câmera de celular. Mantenha a larva em foco. TERMO DE USO Text for Label1 CLASSIFICAÇÃO Tire uma foto da larva Result Este resultado é apenas uma indicação. Em caso de dúvida entre em contato com a vigilância epidemiológica SOBRE O APLICATIVO <FONT COLOR="#700A05"> Este aplicativo classifica larvas de mosquitos das espécies aedes aegypti, aedes albopictus e culex automaticamente com uma acurácia de 93% utilizando Inteligência Artificial. O modelo de Deep Learning (Mobilenet) foi treinado com 1997 imagens rotuladas por biólogos. </FONT> O app possibilita a participação da população para combater a dengue na inspeção das suas residências em busca de larvas para eliminar os criadouros dos mosquitos transmissores. O app foi desenvolvido pela iniciativa Computação na Escola/INCoD/INE/UFSC em cooperação com o LTH/UFSC com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Mais informações: computacaonaescola.ufsc.br em cooperação com LTH/UFSC App em fase de TESTE! Achou uma larva de mosquito? Este app ajuda você a verificar se é um mosquito <FONT COLOR="#700A05">aedes aegypti </FONT>que pode transmitir dengue Este app consegue fazer esta classificação com 95% de confiança em cooperação com LTH/UFSC Como tirar a foto? Como prevenir dengue? Eu li e aceito o termo de uso About About Prevention Prevention Screen1 Screen1 Screen2 Screen2 ScreenRoteiro ScreenRoteiro UsageTerm UsageTerm

Resultados da avaliação da criatividade do aplicativo Xô Dengue

LLM	Média das notas das avaliações das LLMs
ChatGPT 3.5	8
Gemini Pro	6.5
Llama3 70b	6
Mistral 7b	6
Vicuna 13b	6.5

APÊNDICE A

Neste apêndice será apresentado o artigo no formato SBC, referente ao presente projeto.

Análise da Qualidade da Avaliação Automatizada de Criatividade de Aplicativos App Inventor com Modelos Grandes de Linguagens

Vitor Lorizete de Lima Filho, Christiane Gresse von Wangenheim

Departamento de Informática e Estatística, Universidade Federal de Santa Catarina, Florianópolis, SC, Brasil

lvlfilhor@gmail.com,c.wangenheim@ufsc.br

Abstract. *The 21st century demands new skills, such as creativity, which are increasingly valued for personal development and in the job market. Therefore, the importance of fostering creativity in basic education is increasingly recognized. One way to develop this skill is through teaching computing via mobile app development. Currently, there are several initiatives aimed at this goal, but there is generally a lack of models to validly and efficiently assess creativity that can be easily applied in practice. An alternative to automate the assessment of creativity learning could be the use of Large Language Models (LLMs), which, among other tasks, can also evaluate the creativity of apps. However, there are currently no studies evaluating the quality of creativity assessments of apps by LLMs. Thus, this work aims to compare the quality of automated creativity assessment of App Inventor applications using LLMs with assessments made by human judges and the automated Creassessment model. In this way, we expect to gain insight into the possibility of using LLMs for this task to support creativity learning as part of computing education in schools in Brazil.*

Resumo. *O século XXI requer novas habilidades, como a criatividade, que são cada vez mais valorizadas para um desenvolvimento pessoal e no mercado de trabalho. Assim, a importância do incentivo da criatividade na educação básica é cada vez mais reconhecida. Uma das formas de desenvolver essa habilidade é o ensino de computação por meio de desenvolvimento de aplicativos móveis. Atualmente já existem diversas iniciativas voltadas a esse fim, porém observa-se em geral a falta de modelos para avaliar a criatividade de forma válida e eficiente, que possam ser facilmente aplicados na prática. Uma alternativa para automatizar a avaliação da aprendizagem de criatividade pode ser o uso de Large Language Models (LLMs), que, entre outras tarefas, também podem avaliar a criatividade de apps. Porém, atualmente ainda não existem estudos avaliando a qualidade das avaliações de criatividade de apps pelos LLMs. Assim, visa-se neste trabalho comparar a qualidade da avaliação automatizada de criatividade de aplicativos App Inventor com LLMs com avaliações feitas por juízes humanos e o modelo automatizado Creassessment. Assim espera-se obter conhecimento sobre a*

possibilidade em utilizar LLMs para esta tarefa apoiando a aprendizagem de criatividade como parte do ensino da computação em escolas no Brasil.

Introdução

Criatividade é um substantivo que vem do latim creare, que indica a capacidade de criar, produzir ou inventar coisas novas. Na educação de jovens, o desenvolvimento da criatividade é fundamental (P21, 2021)(WEF, 2023). Nessa fase, é essencial estimulá-la uma vez que isso pode ajudá-los a se tornarem mais curiosos, imaginativos e a encontrar soluções inovadoras para problemas e desafios que enfrentam no dia a dia. Além disso, a criatividade também ajuda no desenvolvimento de habilidades essenciais para a vida, como pensamento crítico, resolução de problemas e adaptação às mudanças (P21, 2021)(WEF, 2023). Ao desenvolver a criatividade desde cedo, os estudantes estarão melhor preparados para enfrentar as demandas do mundo atual. Além disso, a criatividade é uma habilidade cada vez mais valorizada no mercado de trabalho atual, empresas que valorizam a criatividade tendem a buscar profissionais com essa habilidade, pois eles são capazes de contribuir de forma significativa para a inovação e o crescimento da organização. Portanto, investir no desenvolvimento da criatividade já nas escolas pode trazer benefícios não só para a vida pessoal dos estudantes, mas também para suas carreiras profissionais.

Existem várias formas de desenvolver e estimular a criatividade em jovens, incluindo como uma alternativa o ensino de computação (Saunders e Thagard, 2005) por meio de desenvolvimentos de aplicativo móveis com App Inventor (Harunani, Patton e Tissenbaum, 2019). A criação e desenvolvimento de aplicativos é uma forma de estimular a criatividade, pois a mesma envolve diferentes competências fazendo com que os estudantes pensem e construam soluções novas para problemas recorrentes (Mishra e Yadav, 2013). Ao criar um aplicativo, o aluno demonstra não apenas o conhecimento adquirido, mas também a capacidade de aplicá-lo de forma criativa (O'quin, 1998).

Já existem algumas iniciativas para estimular a criatividade do estudante por meio do ensino de desenvolvimento de aplicativos móveis (Alves, 2023)(MIT App Inventor). Porém ainda há uma lacuna em relação à avaliação da aprendizagem de criatividade neste contexto. Uma forma de avaliar o desenvolvimento da criatividade dos alunos é por meio da avaliação do desempenho com base em artefatos desenvolvidos como resultado da aprendizagem, como uma forma eficiente e autêntica de medir o progresso e a habilidade do aluno (Hattie, 2007). Dessa forma, no contexto de ensino por meio do desenvolvimento de aplicativos, o desenvolvimento da criatividade pode ser avaliado por meio da avaliação da criatividade dos aplicativos criados pelos estudantes como resultado da aprendizagem.

Tal avaliação pode ser feita de diferentes maneiras. A forma mais convencional é manualmente por humanos simplesmente avaliando a criatividade como um fator, usando rubricas ou *checklists*. Um exemplo dessa abordagem pode ser visto na competição App Inventor *App of the Month*, em que especialistas avaliam a criatividade dos aplicativos criados com o App Inventor. Esses especialistas avaliam os aplicativos com base em uma rubrica com um critério voltado a criatividade, definindo os níveis de desempenho

considerando a originalidade do conceito e a relevância para o mundo real (https://appinventor.mit.edu/explore/app-month-gallery). Embora a avaliação humana seja uma abordagem eficaz, ela pode ser demorada e cara, especialmente quando se trata de avaliar grandes quantidades de aplicativos.

Uma alternativa é a avaliação automatizada, como, p.ex., o modelo de avaliação Creassessment (Alves, 2023), que com base em um modelo conceitual avalia aplicativos criados com *App Inventor* em termos de originalidade, fluência e flexibilidade em conformidade com a definição de criatividade. Com o objetivo de uma avaliação abrangente, essas dimensões são avaliadas em relação à originalidade dos componentes da interface do usuário, das funcionalidades e tópicos do aplicativo, bem como dos comandos da programação do app.

Além de modelos automáticos de avaliação, recentemente emergiu uma nova possibilidade de avaliação de criatividade de apps, por meio de *Large Language Models* (LLMs), que fornecendo informações sobre o aplicativo a LLM pode avaliar o mesmo. LLMs são modelos de linguagem computacionais que utilizam técnicas de aprendizado de máquina para entender e gerar texto em linguagem natural (Birhane et al., 2023). Esses modelos são treinados com grandes quantidades de dados de diversas bases para aprender as regras da linguagem e as relações entre as palavras. Com base nesse conhecimento, eles podem gerar um novo texto, responder perguntas, traduzir idiomas e realizar outras tarefas linguísticas como avaliar a criatividade de um aplicativo. Atualmente as LLMS mais proeminentes são o ChatGPT (Brown et al., 2020) e suas variantes. Porém, atualmente ainda não existem estudos avaliando a qualidade de avaliações de criatividade de apps por LLMs.

Assim, a pergunta de pesquisa deste trabalho é: Qual a qualidade da avaliação automatizada de criatividade de aplicativos App Inventor com LLMs comparado com avaliações realizadas por juízes humanos e/ou o modelo automatizado *Creassessment*?

Definição dos estudos

A série de estudos de caso nesta pesquisa tem como objetivo comparar as avaliações da criatividade de apps criados com *App Inventor* por meio de LLMs em comparação com a avaliação humana e avaliação automatizada usando o modelo *Creassessment*.

Perguntas de pesquisa

PP1: Qual o grau de concordância/confiabilidade da avaliação da criatividade entre LLMs?
PP2: Qual a correlação da avaliação da criatividade comparando a avaliação por LLMs vs. avaliação por juízes humanos vs. avaliação usando o modelo automatizado *Creassessment?*

Para responder essas perguntas de pesquisa é realizada uma série de estudos de casos seguindo o procedimento proposto por Yin (2009).

Tabela 1. Visão geral dos estudos a serem realizados.

Estudo	Objetivo	Tamanho da amostra	LLMs	Prompt automático	Especialistas	Creassessment
1A	Analisar interrater agreement/reliability entre LLMs	1.078 apps	ChatGPT 3.5 e Gemini Pro	X	--	--
1B		35 apps	ChatGPT 3.5, Gemini Pro, Llama3 70b, Mistral 7b e Vicuna 13b	X	--	--
2A	Analisar a correlação entre LLMs vs. especialistas vs. modelo automatizado (Creassessment)	1.078 apps	ChatGPT 3.5 e Gemini Pro	X	MIT App of the month (winner/no-winner)	X
2B		35 apps	ChatGPT 3.5, Gemini Pro, Llama3 70b, Mistral 7b e Vicuna 13b	X	MIT App of the season (Creativity)	X

Amostra de apps

Para a análise dos resultados dos estudos, foram usados dois conjuntos diferentes de *Apps* como artefato a serem avaliados. Conjuntos de dados os quais foram escolhidos por já termos a avaliação dos juízes humanos do MIT para os aplicativos:

- 1.078 *apps* do *App Inventor* aleatoriamente selecionados da competição MIT App of the month (de 2016 até 2022), os quais são divididos em dois grupos *Winners* e *No-Winners*, sendo 832 *No-Winners* e 246 *Winners*.
- 35 *apps* do *App Inventor* selecionados do *appathon* do *App of the Season* 2023, os quais tiveram sua criatividade avaliada por um grupo de juízes da competição App of the Month do MIT/EUA.

LLMs utilizados

Para a avaliação dos conjuntos de *Apps* foram selecionadas diferentes LLMs como avaliadores e foi solicitado resposta via interface web e via API.

- ChatGPT 3.5 - proprietário - https://chat.openai.com/
- Gemini Pro - proprietário - https://gemini.google.com/app

- Llama3 - código aberto - https://chat.lmsys.org/
- Mistral 7b - código aberto - https://chat.lmsys.org/
- Vicuna 13b - código aberto - https://chat.lmsys.org/

Prompts utilizados

Para a solicitação da avaliação das LLMs foram utilizados prompts criados automaticamente com base nos dados automaticamente extraídos do código dos *apps* App Inventor (.aia) utilizando o modelo *Creassessment*. Foram extraídas informações como funcionalidades, tags, componentes UI, conteúdo textual e blocos de código automaticamente para todos os aplicativos, e a partir disso foi montado automaticamente o *prompt* para solicitação das respostas às LLMs usando como base a técnica zero-shot de *prompt engineering*.

Research design

Foi adotado um *Mixed Factorial Design*, analisando a qualidade das respostas dos LLMs baseado em uma amostra de aplicativos *App Invento*r e em um grupo fixo de avaliadores, sempre todos os avaliadores avaliando todos os apps.

Coleta de dados

Foram coletados os seguintes dados:

LLMs	Escala de [0 a 10] com 0 nada criativo e 10 totalmente criativo com arredondamento para o meio ponto (arredonda os valores para o múltiplo mais próximo de 0,5 (ou meio ponto).
Creassessment	Escala de [0 a 10] com 0 nada criativo e 10 totalmente criativo com arredondamento para o meio ponto. (arredonda os valores para o múltiplo mais próximo de 0,5 (ou meio ponto).
Juízes humanos da competição MIT *App of the Month*	Escala nominal [*Winner*, *No-Winner]*
Juízes humanos da competição MIT Appathon 2023	Escala de [0 a 4] a qual foi convertida para [0 a 10] por meio de uma transformação linear simples para comparação com os outros modelos de avaliação.

Métodos de análise de dados

Para a análise dos resultados dos estudos, foram usados vários tipos de análises estatísticas.

Com o objetivo de analisar a confiabilidade entre os avaliadores foram utilizados:

- Coeficiente de correlação de Pearson. O coeficiente de correlação Pearson será usado para todas as análises que envolvem variáveis quantitativas. Ele indica a confiabilidade entre as avaliações dos diferentes avaliadores. O valor da correlação de Pearson é interpretado da seguinte forma (Devellis, 2017): correlações acima de 0.9 indicam uma correlação muito forte, entre 0.7 a 0.9 correlação forte, 0.5 a 0.7 correlação moderada, 0.3 a 0.5 correlação fraca, 0 a 0.3 correlação desprezível
- *Root Mean Square Error* (RMSE). O *Root Mean Square Error* (RMSE) indica a confiabilidade entre as avaliações dos diferentes avaliadores. O RMSE é interpretado de forma distinta ao coeficiente de correlação de Pearson, não seguindo uma escala específica como a interpretação de Pearson. Ele é usado para avaliar a qualidade do modelo em fazer previsões precisas em relação aos dados reais. Nesse caso, quanto menor o RMSE e mais próximo de 0, mais semelhante um conjunto de dados é em relação ao outro.

Voltado à análise de concordância entre os avaliadores foram utilizados:

- Coeficiente de correlação intraclasse (ICC). O ICC serve para avaliar a concordância entre os diferentes avaliadores para um mesmo conjunto de dados. Existem diferentes tipos de ICC, o escolhido para a análise nesse projeto foi o ICC3, o qual entende que todos os dados foram avaliados por todo o conjunto de avaliadores. O coeficiente de correlação intraclasse pode ser interpretado da seguinte forma (Cichetti, 1994): correlação entre 0.75 e 1.0 indicam uma correlação excelente, 0.6 a 0.75 correlação boa, correlação 0.4 a 0.6 correlação razoável e menor que 0.4 correlação fraca. Seu intervalo de confiança (CI), indica dentro de qual faixa de valores o valor do ICC provavelmente está contido, por exemplo, um intervalo de confiança de 95% indica que há uma probabilidade de 95% de que o ICC daquela análise esteja contida dentro da faixa de valores mínimo e máximo apresentada.
- Dispersão por meio de um gráfico *Boxplot*. A partir do gráfico *Boxplot* pode se observar a dispersão das notas dos dois conjuntos de dados os quais serão comparados, e por meio do coeficiente pode-se ver se as notas tendem a diminuir ou aumentar nos diferentes níveis do gráfico. Um coeficiente negativo mostra que as notas tendem a diminuir naquele mesmo nível de um conjunto para o outro. Um coeficiente igual a 0 mostra que as notam tendem a permanecer na mesma média nos dois conjuntos. Um coeficiente positivo mostra que as notas tendem a aumentar naquele mesmo nível de um conjunto de dados para o outro.

- Gráfico de Bland-Altman. O gráfico Bland-Altman (Bland; Altman, 1986) avalia a concordância entre duas medidas quantitativas, examinando a diferença média entre elas e estabelecendo limites de concordância. Esses limites permitem identificar se há um viés sistemático nas diferenças médias e estimar um intervalo de concordância dentro do qual 95% das diferenças entre os dois métodos estão incluídas. Este método, é uma ferramenta para determinar a intercambialidade entre métodos e para detectar discrepâncias consistentes entre eles.

Além destas métricas, com o objetivo de analisar alguns dados, foram utilizados também o cálculo da média, mediana e diferença máxima.

Estudo de caso 1A (PP1) - *interrater agreement/reliability* entre LLMs

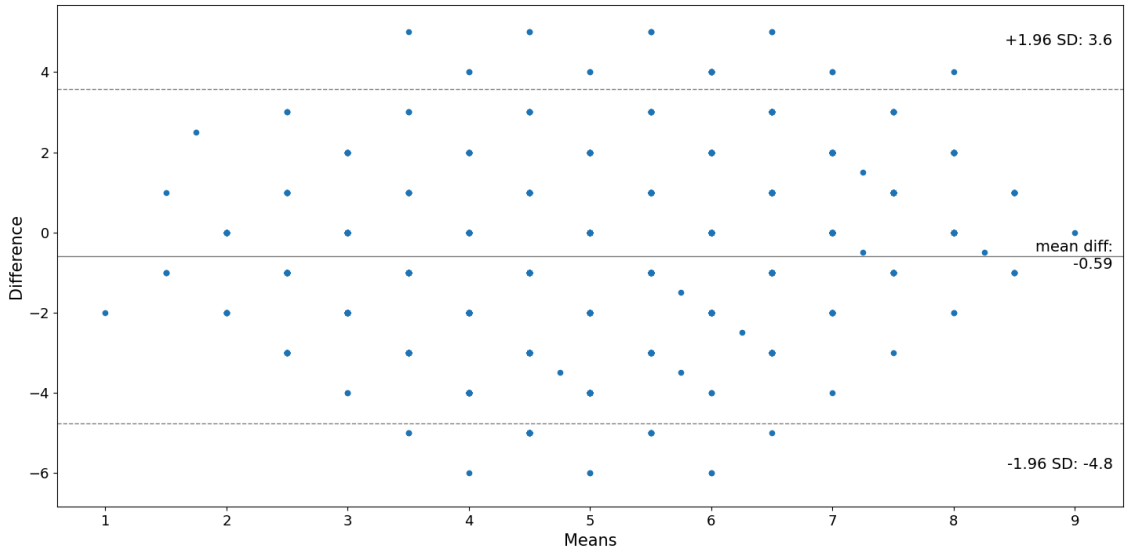
<table>
<tr><td>Artefatos avaliados</td><td colspan="3">1.078 apps do App Inventor da App of the month.</td></tr>
<tr><td>LLMs utilizados</td><td colspan="3">Avaliação via API das LLMs: ChatGPT 3.5 e Gemini Pro</td></tr>
<tr><td>Prompt utilizado</td><td colspan="3">Avalie o grau de criatividade do seguinte App em uma escala de [0 a 10]. Responda de forma simples e objetiva com base na descrição abaixo.
Funcionalidades do aplicativo: [Funcionalidades extraídas do aplicativo usando creassessment]
Palavras-chaves: [tags extraídas do aplicativo usando creassessment]
Interface: [componentes UI extraídas do aplicativo usando creassessment]
Conteúdo textual: [Conteúdo textual extraído do aplicativo usando creassessment]
Blocos de código: [Blocos extraídos do aplicativo usando creassessment]</td></tr>
<tr><td>Parâmetros de respostas das LLMs</td><td colspan="3">Foi solicitada uma resposta simples e objetiva com base na descrição dos aplicativos, usando os parâmetros de default das LLMS e não foi realizado ajuste fino (fine-tuning).</td></tr>
<tr><td>Escala de resposta</td><td colspan="3">Escala de [0 a 10] com 0 nada criativo e 10 totalmente criativo com arredondamento para o meio ponto.</td></tr>
<tr><td rowspan="4">Análise estatística dos dados</td><td colspan="3">A correlação entre as avaliações das LLMs é avaliada pelo coeficiente de Pearson, pelo Root Mean Square Error (RMSE) e análise do gráfico Bland-Altman. É analisado também o coeficiente de correlação intraclasse (ICC), usando o modelo ICC3 (Mixed Factorial Design) em que todos apps foram avaliados por todos os avaliadores. E um gráfico Bland-Altman para analisar o intervalo de confiança entre as notas.</td></tr>
<tr><td colspan="3">Resultados</td></tr>
<tr><td>Correlação de Pearson</td><td>r = 0.4</td><td>Correlação fraca</td></tr>
<tr><td>RMSE</td><td>RMSE = 2.2</td><td>RMSE alto, ou seja, muita divergência em relação às notas dadas pelo ChatGPT 3.5 e Gemini Pro</td></tr>
</table>

	ICC	**ICC = 0.374** **CI = [0.3, 0.44]**	ICC pobre, mostrando pouca convergência nas notas. E com um intervalo de confiança entre 0.3 a 0.44.

A partir da análise, a correlação de *Pearson* e o RMSE indicam uma falta de confiabilidade entre as duas LLMs (ChatGPT 3.5 e Gemini Pro), pelo baixo coeficiente de *Pearson* **r = 0.4** e pelo alto RMSE (**RMSE = 2.2)**.

A análise do coeficiente de correlação intraclasse (ICC modelo 3), também indica um valor fraco **ICC = 0.374,** indicando a falta de uma concordância entre os LLMs avaliadores.

Figura 1 - Gráfico Bland-Altman para a avaliação ChatGPT 3.5 vs Gemini Pro



A partir da análise do gráfico Bland-Altman, percebe-se que o intervalo de confiança foi de -4.8 até 3.6, demonstrando que 95% das diferenças entre as medidas dos dois LLMs estão contidas em um intervalo de 8.4 unidades. Isso sugere que as diferenças nas avaliações do ChatGPT 3.5 e do Gemini Pro estão dentro de um intervalo muito grande, tendo em vista a escala usada, indicando novamente uma baixa concordância entre elas.

Desta forma, com os resultados de todas as análises (*Pearson*, RMSE, ICC e Bland-Altman) indicam uma baixa confiabilidade e concordância entre os LLMs avaliadores ChatGPT 3.5 vs Gemini Pro.

Estudo de caso 1B (PP1) - *interrater agreement/reliability* entre LLMs

Artefatos avaliados	35 apps *App Inventor appathon* do *App of the Season* 2023.

LLMs utilizados	Avaliação via interface web, usando ChatGPT 3.5, Gemini Pro, Llama3 70b, Mistral 7b e Vicuna13b
Prompt utilizado	Avalie o grau de criatividade do seguinte App em uma escala de [0 a 10]. Responda de forma simples e objetiva com base na descrição abaixo. Funcionalidades do aplicativo: [Funcionalidades extraídas do aplicativo usando *creassessment*] Palavras-chaves: [tags extraídas do aplicativo usando *creassessment*] Interface: [componentes UI extraídas do aplicativo usando *creassessment*] Conteúdo textual: [Conteúdo textual extraído do aplicativo usando *creassessment*] Blocos de código: [Blocos extraídos do aplicativo usando *creassessment*]
Parâmetros de respostas das LLMs	Foi solicitada uma resposta simples e objetiva com base na descrição dos aplicativos, usando os parâmetros de *default* das LLMS e não foi realizado ajuste fino (*fine-tuning*).
Escala de resposta	Escala de [0 a 10] com 0 nada criativo e 10 totalmente criativo com arredondamento para o meio ponto.
Tratamento dos dados	Para cada aplicativo foram solicitadas 3 avaliações para cada LLM em momentos diferentes, e a partir das 3 avaliações foi calculada a média para calcular a nota final da avaliação do app.
Análise estatística dos dados	A correlação entre as avaliações das LLMs foi avaliada pelo coeficiente de Pearson, pelo *Root Mean Square Error* (RMSE) e pelo coeficiente de correlação intraclasse (ICC), usando modelo ICC3 (Mixed Factorial Design), em que todos *os* apps foram avaliados por todos os avaliadores.

Para a avaliação dos 35 apps foi solicitado 3 avaliações para cada LLM utilizando o mesmo *prompt*, e a partir das 3 avaliações foi calculada a média para calcular a nota final da avaliação do app. Com o objetivo de analisar o grau com que a própria LLM concorda com ela mesmo, se para um mesmo *prompt* gera resultados semelhantes, foi calculada a média da diferença máxima das três notas atribuídas para cada app (Tabela 2).

Tabela 2. Análise da média da diferença máxima entre as três notas das avaliações para cada LLM

Média da diferença máxima	**ChatGPT 3.5**	**Gemini Pro**	**Llama3 70b**	**Mistral 7b**	**Vicuna 13b**
	2	1.87	**0.94**	1.31	1.34

Através da média da diferença máxima das notas das LLMs, conseguimos analisar a variabilidade entre as notas de uma LLM com ela mesma, mostrando a amplitude média de diferentes avaliações de uma LLM para um mesmo aplicativo. Por meio da escala [0..10] que está sendo usada o maior intervalo possível entre uma nota e outra é de 10 pontos. A menor média da diferença máxima foi da LLM Llama3, com uma média próxima do 1 ponto, mostrando-se á com a maior concordância entre suas próprias notas. A com a maior média de diferença máxima foi a do ChatGPT 3.5, com 2 pontos de média entre suas próprias avaliações.

Para a análise da confiabilidade das notas entre as LLMs foi analisado o coeficiente de correlação de Pearson (Tabela 3).

Tabela 3. Análise do coeficiente de correlação de *Pearson*

Pearson	ChatGPT 3.5	Gemini Pro	Llama3 70b	Mistral 7b	Vicuna 13b
Chat GPT	-	0.13	0.27	**0.61**	0.28
Gemini Pro	0.13	-	0.38	0.1	-0.02
Llama3 70b	0.27	0.38	-	0.26	-0.15
Mistral 7b	**0.61**	0.1	0.26	-	0.18
Vicuna 13b	0.28	-0.02	-0.15	0.18	-

A maior correlação para Pearson entre as LLMs foi Mistral 7b vs ChatGPT 3.5, com uma correlação moderada de **r = 0.61**. As de mais todas apresentam correlações fracas ou desprezíveis, inclusive tendo uma correlação negativa entre as notas alocadas pelo Llama3 70b vs Vicuna 13b e Vicuna 13b vs Gemini Pro. Estes resultados mostram uma baixa confiabilidade na avaliação da criatividade entre as LLMs.

Analisando a confiabilidade entre as LLMs também pelo *Root Mean Square Error*, os resultados são apresentados na Tabela 4.

Tabela 4. Análise usando *Root Mean Square Error*

RMSE	ChatGPT 3.5	Gemini Pro	Llama3 70b	Mistral 7b	Vicuna 13b
Chat GPT	-	1.7	2.17	1.57	1.52
Gemini Pro	1.7	-	1.76	1.48	**1.24**
Llama3 70b	2.17	1.76	-	1.48	2.19
Mistral 7b	1.57	1.48	1.48	-	1.53
Vicuna 13b	1.52	**1.24**	2.19	1.53	-

Analisando os resultados do *Root Mean Square Error* entre a avaliação das LLMs o menor resultado é **RMSE = 1.24** entre Vicuna vs Gemini.

O resultado do coeficiente de correlação intraclasse (modelo 3) **ICC = 0.166**, é um resultado fraco, com um intervalo de confiança (CI) entre 0.05 a 0.33, mostrando assim também uma baixa concordância entre as LLMs.

A partir desses resultados observa-se de forma geral que há baixa confiabilidade e concordância entre os LLMs, com melhores convergências entre Mistral 7b e ChatGPT 3.5.

Estudo de caso 2A (PP2) - LLMs vs. humanos vs. *Creassessment*

<table>
<tr><th></th><th colspan="2">Avaliação com LLMs</th><th colspan="2">Avaliação humana</th><th>Avaliação com modelo automatizado</th></tr>
<tr><td>Artefatos avaliados</td><td colspan="5">1.078 apps do App Inventor da App of the Month.</td></tr>
<tr><td rowspan="3">Técnica de Avaliação</td><td>LLMs utilizados</td><td>Avaliação via API das LLMs: ChatGPT 3.5 e Gemini Pro.</td><td>Avaliadores</td><td>Foram comparados 1078 apps disponibilizados pelo App Inventor team, dos quais 246 (23%) foram identificados como winners e 832 (73%) como non-winners da competição App of the Month, sendo julgados pelos membros do MIT App Inventor</td><td rowspan="3">Creassessment (Alves, 2023)</td></tr>
<tr><td>Prompt utilizado</td><td>Avalie o grau de criatividade do seguinte App em uma escala de [0 a 10]. Responda de forma simples e objetiva com base na descrição abaixo.
Funcionalidades do aplicativo: [Funcionalidades extraídas do aplicativo usando creassessment]
Palavras-chaves: [tags extraídas do aplicativo usando creassessment]
Interface: [componentes UI extraídas do aplicativo usando creassessment]
Conteúdo textual: [Conteúdo textual extraído do aplicativo usando creassessment]
Blocos de código: [Blocos extraídos do aplicativo usando creassessment]</td><td>Rubricas e critérios de avaliação</td><td>Os apps foram revisados pelos membros do MIT App Inventor considerando:

Design - The app is the most aesthetically pleasing.

Innovation - The app uses App Inventor technology in the most interesting/unique way.

Creativity - The app that best uses creative elements such as art, color, sound, or movement</td></tr>
<tr><td>Parâmetros de LLM</td><td>Foi solicitada uma resposta simples e objetiva com base na descrição dos aplicativos, serão usados os parâmetros de default das LLMS e não será realizado ajuste fino (fine-tuning).</td><td>Quantidade de avaliadores humanos</td><td>Não informado.</td></tr>
<tr><td>Escala de resposta</td><td colspan="5">LLMs: Escala de [0 a 10] com 0 nada criativo e 10 totalmente criativo com arredondamento para o meio ponto.

Creassessment: Escala de [0 a 10] com 0 nada criativo e 10 totalmente criativo com arredondamento para o meio ponto.

Juízes appathon: Escala nominal [Winner ou No-Winner].</td></tr>
<tr><td>Análise</td><td colspan="5">A análise foi feita média/mediana e dispersão por meio de um gráfico Boxplot para avaliar LLMs vs juízes humanos e</td></tr>
</table>

estatística dos dados	foi calculado o coeficiente de correlação Pearson, RMSE, ICC usando modelo ICC3 (Mixed Factorial Design) e Bland-Altman para analisar LLMs vs *Creassessment*.

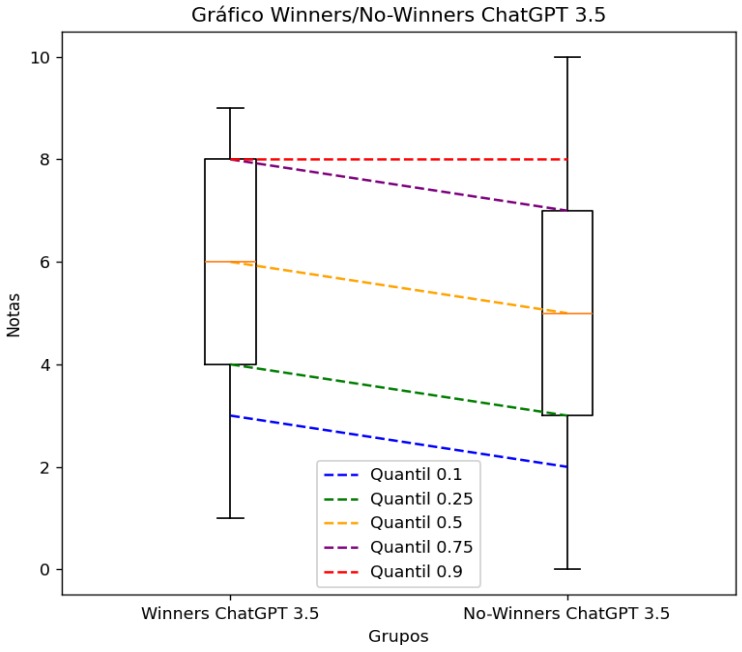
Com o objetivo de avaliar a correlação entre LLMs e juízes humanos, foram feitos testes a partir da média, mediana e dispersão por meio de um gráfico *Boxplot* para analisar se as LLMs conseguem diferenciar os grupos de *Winners* e *No-Winners* classificados por juízes humanos.

Tabela 5. Análise da média e mediana entre grupos *Winners* e *No-Winners*

	Média		**Mediana**	
	Winner	*No-Winner*	***Winner***	***No-Winner***
ChatGPT 3.5	5.63	4.92	6	6
Gemini Pro	6	5.57	6	6

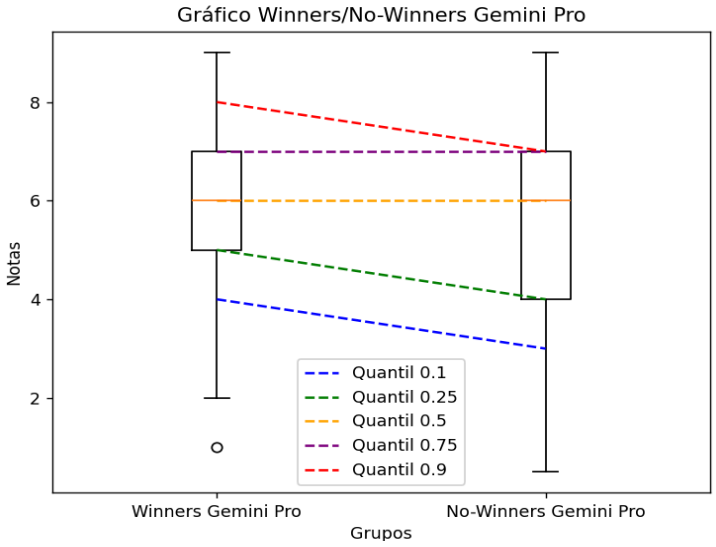
Como resultado observa-se que a média dos *Winner* para as duas LLMs é maior do que a média para os *No-Winner*, porém com pouca diferença. A mediana segue uma tendência parecida, com o grupo de *Winner* tendo uma mediana maior ou no mínimo igual ao grupo de Winner. Assim, observa-se que as LLMs tendem a dar notas um pouco mais altas para o grupo de *Winner* do que ao grupo de *No-Winner*. Isso mostra uma possível diferenciação nas notas de apps de grupos diferentes com graus de criatividade diferentes, porém com pouca diferença.

Figura 2 - Gráfico *Boxplot* para dispersão das notas por grupo do ChatGPT 3.5



Quantis	Coeficiente
0.1	-1
0.25	-1
0.5	-1
0.75	-1
0.9	0

Figura 3 - Gráfico *Boxplot* para dispersão das notas por grupo do Gemini



Quantis	Coeficiente
0.1	-1
0.25	-1
0.5	0
0.75	0
0.9	-1

Por meio dos dois gráficos Boxplot considerando as medianas, percebe-se que em todos os quantis temos coeficientes nulos ou negativos passando do grupo de *Winner* para o grupo *No-Winner*, evidenciando que em todos os quantias ou as notas baixam ou as notas permanecem iguais em relação a *Winner* para *No-Winner*. Percebe-se também que o ChatGPT 3.5 consegue diferenciar um pouco melhor os apps nos grupos do que o Gemini Pro. Pois tanto na média quanto na mediana teve uma diferença mais significativa, quanto no gráfico Boxplot apenas no quantis 0.9 teve um coeficiente nulo, de resto, todos foram coeficientes negativos.

Comparando também as avaliações das LLMs com as notas alocadas pelo modelo de avaliação automático *Creassessment* (Alves, 2023) considerado um *ground truth*, foi analisado a confiabilidade e concordância entre as notas dos LLMs e *Creassessment* (Tabela 6).

Tabela 6. Análise do Coeficiente *Pearson*, RMSE e ICC (modelo ICC3) e seu intervalo de confiança (CI) entre *Creassessment* vs ChatGPT 3.5 e Gemini pro

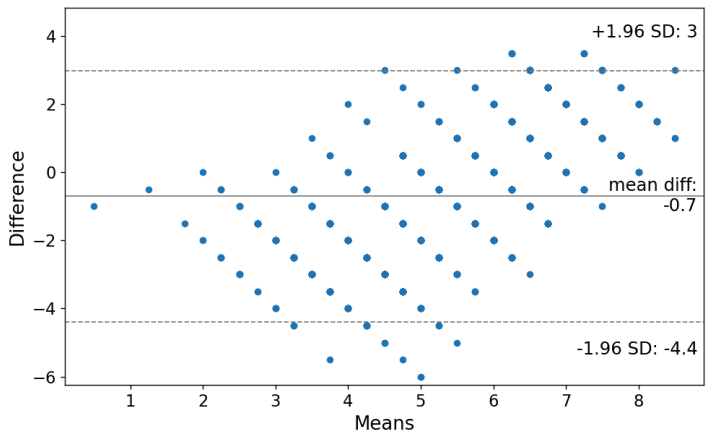
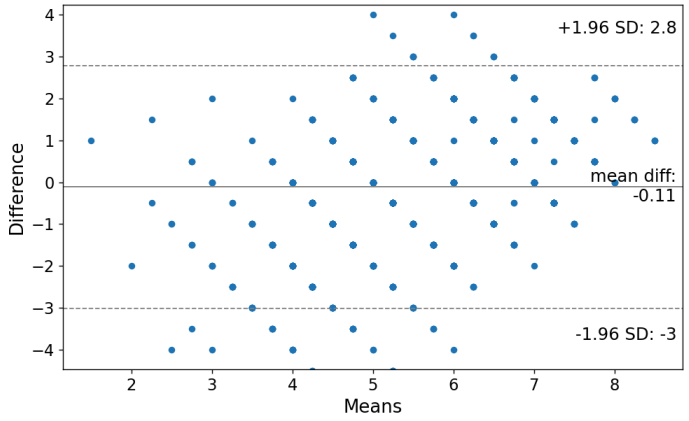
LLMS	***Creassessment***		
	Pearson	**RMSE**	**ICC (modelo ICC3)**
ChatGPT 3.5	**0.51**	2.0	ICC = 0.37, CI = 0.28 a 0.46
Gemini Pro	0.46	**1.48**	**ICC = 0.42, CI = 0.37 a 0.47**

Observa-se coeficientes de correlação Pearson semelhantes entre ChatGPT 3.5 vs *Creassessment* e Gemini Pro vs *Creassessment*, sendo ambos próximos do limiar entre um coeficiente baixo e moderado. Sendo ChatGPT 3.5 vs *Creassessment* apresentando uma maior correlação (**r = 0.51**). Estes resultados mostram uma baixa confiabilidade nas avaliações das LLMs em relação ao modelo *Creassessment*.

Os resultados da análise do RMSE também confirmam esta falta de confiabilidade nas avaliações dos LLMs em comparação com as notas do modelo *Creassessment*, como ambas as LLMs, ChatGPT 3.5 e Gemini Pro, mostraram altos resultados de RMSE em relação às avaliações alocadas pelo modelo *Creassessment*.

Os resultados do ICC (modelo ICC3) para ChatGPT 3.5 vs *Creassessment* e Gemini Pro vs *Creassessment* mostraram baixa correlação. Ambos os resultados estão próximos do limiar entre um ICC moderado e um ICC baixo. Estes resultados mostram novamente uma baixa concordância entre as notas das LLMs e do modelo *Creassessment*.

Tabela 7. Gráficos Bland-Altman para as avaliações *Creassessment* vs. ChatGPT e Gemini

Bland-Altman ChatGPT 3.5 vs. *Creassessment*	Bland-Altman Gemini vs. *Creassessment*
	

A partir da análise dos gráficos Bland-Altman, percebe-se que o intervalo de confiança foi de -4.4 até 3 para ChatGPT 3.5 vs *Creassessment* e -3 até 2.8 para Gemini Pro vs *Creassessment,* demonstrando que 95% das diferenças entre as medidas das LLMs e do modelo *Creassessment* estão contidas em um intervalo de 7.4 e 5.8 respectivamente. Sendo o menor intervalo entre Gemini Pro vs *Creassessment*, porém ainda assim os dois resultados mostram um intervalo muito grande, tendo em vista a escala de avaliação usada, indicando novamente uma baixa concordância entre elas.

Estudo de caso 2B (PP2) - LLMs vs. humanos vs. *Creassessment*

<table>
<tr><th></th><th colspan="3">Avaliação com LLMs</th><th colspan="2">Avaliação humana</th><th>Avaliação com modelo automatizado</th></tr>
<tr><td>Artefatos avaliados</td><td colspan="6">35 apps App Inventor appathon do App of the Season 2023.</td></tr>
<tr><td rowspan="3">Técnica de Avaliação</td><td>LLMs utilizados</td><td colspan="2">Avaliação via interface web, usando as LLMs: ChatGPT 3.5, Gemini Pro, Llama3 70b, Mistral 7b e Vicuna 13b.</td><td>Avaliadores</td><td>Avaliadores do appathon do App of the Season 2023.</td><td rowspan="3">Creassessment (Alves, 2023)</td></tr>
<tr><td>Prompt utilizado</td><td colspan="2">Avalie o grau de criatividade do seguinte App em uma escala de [0 a 10]. Responda de forma simples e objetiva com base na descrição abaixo.
Funcionalidades do aplicativo: [Funcionalidades extraídas do aplicativo usando creassessment]
Palavras-chaves: [tags extraídas do aplicativo usando creassessment]
Interface: [componentes UI extraídas do aplicativo usando creassessment]
Conteúdo textual: [Conteúdo textual extraído do aplicativo usando creassessment]
Blocos de código: [Blocos extraídos do aplicativo usando creassessment]</td><td>Rubricas e critérios de avaliação</td><td>4 - App idea is very unique or a novel take on other ideas - App demonstrates significant potential impact related to the theme and can effectively help the target audience.
3 - App has some novel aspects to it - App demonstrates potential impact related to the theme and has potential to help the target audience.
2 - App idea is somewhat novel, or extends an existing idea slightly - App demonstrates some potential impact and/or relation to the theme.
1 - Very little in terms of unique or new ideas - App demonstrates little potential impact and/or relation to the theme.
0 - Remake of an existing app - No impact or relation to the theme.</td></tr>
<tr><td>Parâmetros de LLM</td><td colspan="2">Foi solicitada uma resposta simples e objetiva com base na descrição dos aplicativos, serão usados os parâmetros de default das LLMS e não será realizado ajuste fino (fine-tuning).</td><td>Quantidade de avaliadores humanos</td><td>Não informado.</td></tr>
<tr><td>Escala de resposta</td><td colspan="6">LLMs: Escala de [0 a 10] com 0 nada criativo e 10 totalmente criativo com arredondamento para o meio ponto.

Creassessment: Escala de [0 a 10] com arredondamento para o meio ponto

Juízes appathon: Escala de [0 a 4] a qual foi normalizada para [0 a 10] com arredondamento para o meio ponto para comparação com os outros modelos.</td></tr>
<tr><td>Tratamento dos dados</td><td colspan="6">Para cada aplicativo foram solicitadas 3 avaliações para cada LLM, e a partir das 3 avaliações foi calculada a média usando essa média na análise.</td></tr>
<tr><td>Análise estatística</td><td colspan="6">A correlação entre as avaliações das LLMs, do Creassessment e das notas de juízes humanos foi avaliada pelo coeficiente de Pearson, Root Mean Square Error (RMSE) e coeficiente de correlação intraclasse usando o modelo</td></tr>
</table>

dos dados	ICC3 em que todos *apps* foram avaliados por todos os avaliadores.

Os dados das notas das LLMs deste estudo são os mesmos do estudo 1B, o cálculo das médias das 3 notas obtidas por cada aplicativo e a diferença máxima entre as notas são as mesmas como no estudo 1B.

Considerando as avaliações feitas pelos juízes especialistas no *appathon* e as notas alocadas pelo modelo *Creassessment* como *ground truth*, sao comparadas as notas dadas pelos LLMs por meio da análise da correlação entre LLMs vs juízes humanos e LLMs vs *Creassessment* (Tabela 8).

Tabela 8. Análise do Coeficiente *Pearson*

Pearson	**ChatGPT 3.5**	**Gemini Pro**	**Llama3 70b**	**Mistral 7b**	**Vicuna 13b**
Juízes humanos	0.18	0.04	0.3	**0.4**	-0.09
Creassessment	0.28	-0.02	0.0002	**0.56**	0.13

Observa-se de forma geral coeficientes de correlação Pearson baixas comparando as notas das LLMs com as notas alocadas pelos juízes humanos. A maior correlação foi da LLM Mistral 7b, com um **r = 0.4** com os juízes humanos. Observa-se novamente até uma correlação negativa no caso do Vicuna 13b.

Comparando as notas do LLMs com as notas alocadas pelo modelo Creassessment, observam-se em geral coeficientes de correlação Pearson mais baixas ainda, também com uma correlação negativa do Gemini. A única exceção novamente é o Mistral 7b com um coeficiente de correlação moderada (**r = 0.56)**.

Tabela 9. Análise do *Root Mean Square Error*

RMSE	**ChatGPT 3.5**	**Gemini Pro**	**Llama3 70b**	**Mistral 7b**	**Vicuna 13b**
Juízes humanos	1.89	1.65	2.03	**1.61**	1.63
Creassessment	1.75	1.41	1.61	**0.73**	1.41

Os resultados da análise do RMSE também confirmam os resultados de *Pearson*. Novamente a LLM que mostrou o melhor desempenho no resultado do RMSE com os juízes humanos e o *Creassessment* foi a Mistral, resultando em um **RMSE = 1.61** em comparação aos juízes humanos e **RMSE = 0.73** em comparação ao *Creassessment*.

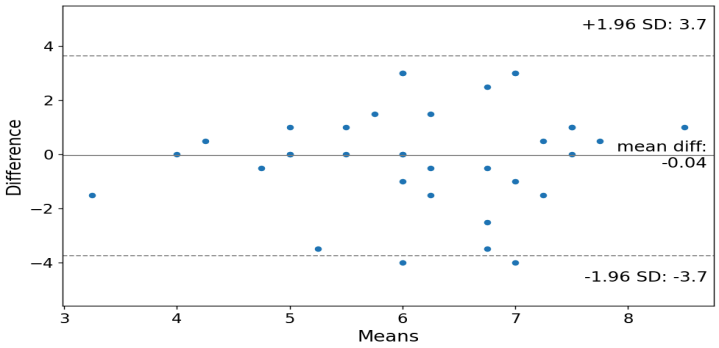
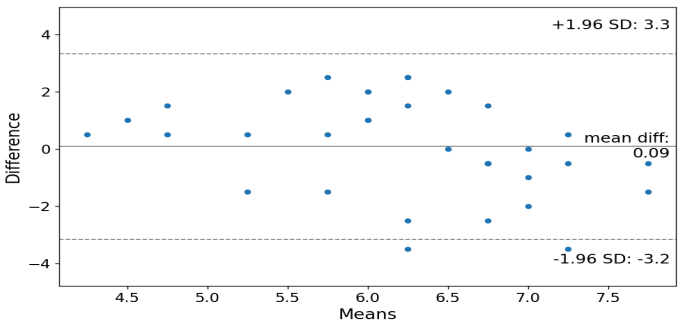
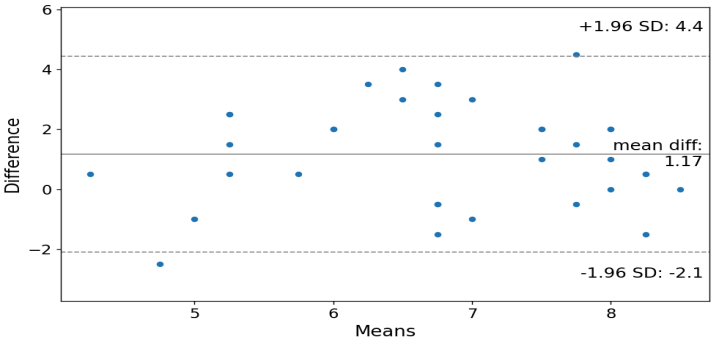
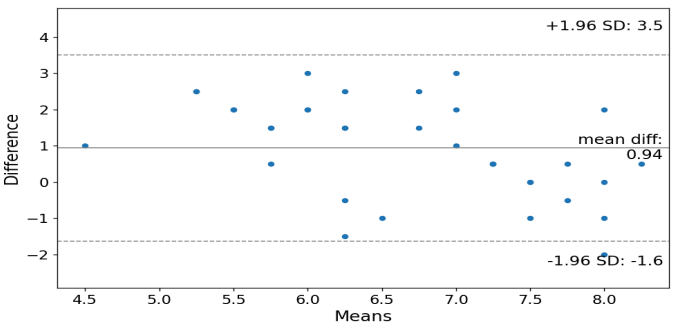
Tabela 9. Análise do ICC (modelo ICC3) e do seu intervalo de confiança (CI)

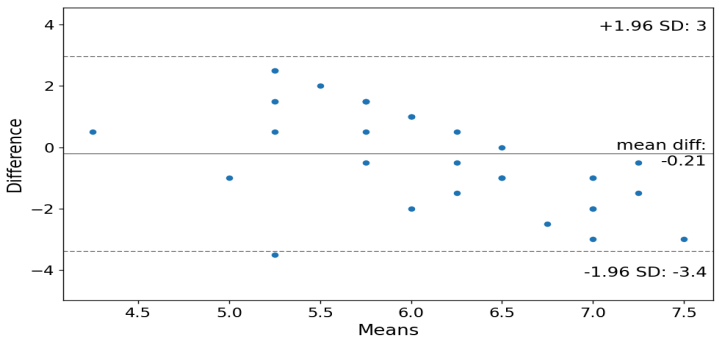
ICC	**ChatGPT 3.5**	**Gemini Pro**	**Llama3 70b**	**Mistral 7b**	**Vicuna 13b**
Juízes humanos	ICC = 0.18, CI = -0.16 a	ICC = 0.04, CI = -0.3 a	ICC = 0.23, CI = -0.06 a 0.5	**ICC = 0.27, CI = -0.04 a**	ICC = -0.07, CI = -0.4 a

	0.49	0.37		**0.54**	0.26
Creassessment	ICC = 0.16, CI = -0.1 a 0.44	ICC = -0.017, CI = -0.23 a 0.24	ICC = 0, CI = -0.32 a 0.33	**ICC = 0.55, CI = 0.27 a 0.75**	ICC = 0.06, CI = -0.09 a 0.27

Assim como para o coeficiente de correlação de Pearson e o RMSE, a LLM com o melhor resultado em relação ao ICC (modelo 3) foi a Mistral 7b, com um ICC pobre (**ICC = 0.27)** e um CI entre -0.04 a 0.54 em comparação com os juízes humanos e um ICC razoável próximo ao bom (**ICC = 0.55**) e um CI entre 0.25 a 0.75 em comparação com o *Creassessment*. Porém de forma geral observa-se novamente que os resultados apontam uma concordância muito fraca dos LLMs com as notas alocadas pelos juízes humanas e/ou alocadas pelo modelo *Creassessment*, inclusive com valores negativos no caso da Vicuna 13b em comparação aos juízes humanos e do Gemini Pro em comparação ao Creassessment.

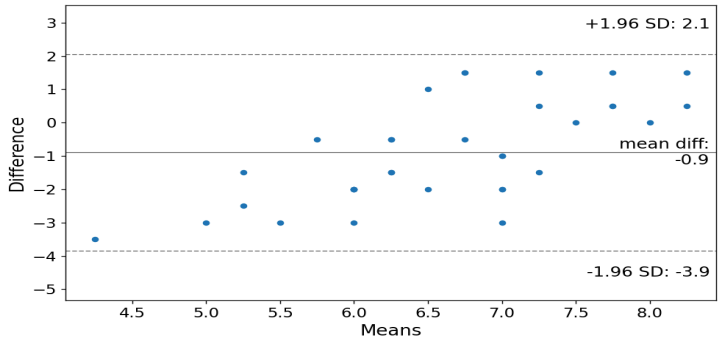
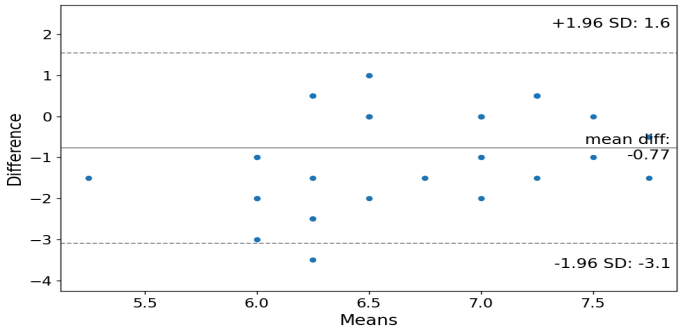
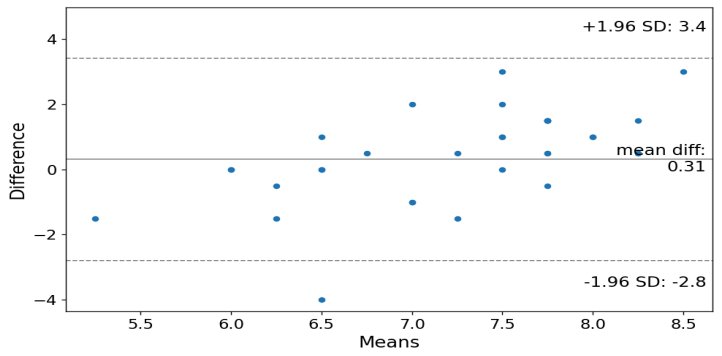
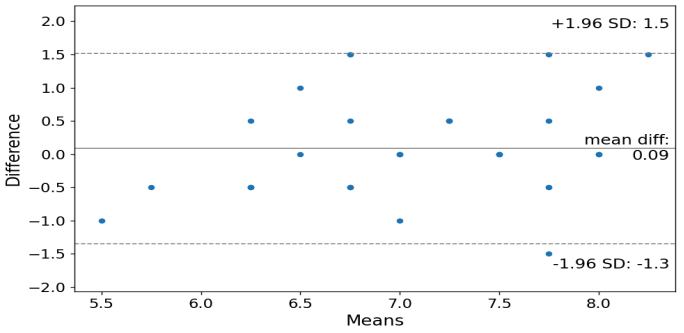
Tabela 10. Gráficos Bland-Altman: juízes humanos vs. LLMs

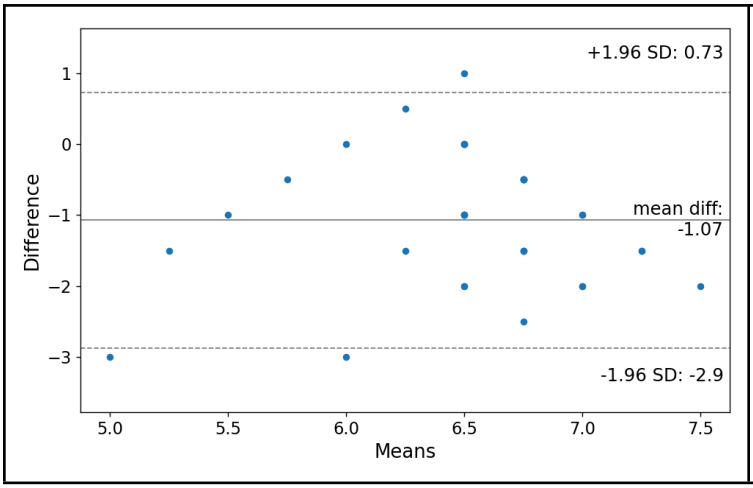
<table>
<tr><td>Juízes humanos vs. ChatGPT 3.5</td><td>Juízes humanos vs. Gemini Pro</td></tr>
<tr><td> 
</td><td> 
</td></tr>
<tr><td>Juízes humanos vs. Llama3 70b</td><td>Juízes humanos vs. Mistral 7b</td></tr>
<tr><td> 
</td><td> 
</td></tr>
<tr><td>Juízes humanos vs. Vicuna 13b</td><td></td></tr>
</table>



A partir da análise dos gráficos Bland-Altman, percebe-se que o menor intervalo de confiança foi o da LLM Mistral 7b, com um intervalo de -1.6 até 3.5, demonstrando que 95% das diferenças entre as medidas das LLMs e das avaliações feitas por juízes humanos estão contidas em um intervalo de 5.1 pontos. Mesmo sendo o menor intervalo ainda é um intervalo muito grande, tendo em vista a escala da avaliação usada sendo de 0 a 10. Desta forma indicando novamente uma baixa concordância entre juízes humanos e LLMs.

Tabela 10. Gráficos Bland-Altman: *Creassessment* vs. LLMs

Creassessment vs. ChatGPT 3.5	*Creassessment* vs. Gemini Pro
	
Creassessment vs. Llama3 70b	*Creassessment* vs. Mistral 7b
	
Creassessment vs. Vicuna 13b	



A partir da análise dos gráficos Bland-Altman, percebe-se que o menor intervalo de confiança foi novamente da LLM Mistral 7b, com um intervalo de -1.3 até 1.5, demonstrando que 95% das diferenças entre as medidas das LLMs e das avaliações feitas pelo modelo *Creassessment* estão contidas em um intervalo de 2.8 pontos. Sendo um intervalo menor que o apresentado pela LLM Mistral 7b vs juízes humanos, porém ainda sendo um intervalo de quase 3 pontos, indicando uma baixa concordância entre LLMs vs *Creassessment* novamente.

Desta forma percebe-se que a LLM Mistral 7b se mostrou a melhor em resultados em todos os estudos , tanto em comparação às avaliações feitas pelos juízes humanos quanto em comparação às notas alocadas pelo o *Creassessment*. Mostrando resultados moderados, porém melhores em comparação com as outras LLMs. Já para as outras LLMs ChatGPT 3.5, Gemini Pro, Llama3 70b e Vicuna 13b apresentaram resultados insatisfatórios, com resultados mostrando uma baixa correlação com os juízes humanos e o modelo automatizado *Creassessment*.

Discussão

Analisando os dados coletados, pode-se concluir que o nível de concordância e confiabilidade de avaliação de criatividade entre as LLMs não é satisfatório. Para ambos os estudos, 1A e 1B, os testes de concordância e confiabilidade entre as LLMs demonstraram resultados fracos. Resultados os quais podem vir pelo fato das LLMs serem treinadas por diferentes conjuntos de dados, tanto em forma quantitativa quanto em relação ao conteúdo desses dados. Podendo ter assim cada LLM uma noção diferente do que é um app criativo, algo aceitável de se esperar, visto que criatividade pode ser algo mais subjetiva de se interpretar.

Porém, em relação à concordância entre suas próprias notas para um mesmo aplicativo, os resultados não foram ruins, tendo em vista os testes da média da diferença máxima entre suas próprias notas (Tabela 2), com a maior diferença foi de 2 unidades, da LLM ChatGPT 3.5, e a menor diferença foi de aproximadamente 1 unidade, da LLM Llama3 70b, tendo em vista uma escala de 0 a 10, pode-se dizer que foram diferenças

máximas baixas, entendendo assim que existe uma concordância entre as notas da própria LLM para um mesmo aplicativo.

Em relação a correlação da avaliação da criatividade comparando a avaliação por LLMs com avaliação por juízes humanos e avaliação usando o modelo automatizado *Creassessment*, a partir dos dados coletados pode-se concluir que no geral houve uma baixa correlação entre as avaliações dos juízes humanos com as LLMs e do modelo *Creassessment* com as LLMs. A maioria das correlações foram fracas e moderadas. A LLM que se mostrou com a melhor correlação foi a LLM Mistral 7b, que obteve os melhores resultados em praticamente todas avaliações de correlação e concordância com os outros dois modelos de avaliação. Tendo correlações *Pearson* moderadas com o *Creassessment* no estudo 2B e perto de moderada com os juízes humanos, além de um ICC razoável e próximo de bom com o *Creassessment* também no estudo 2B. Porém um ponto positivo é o resultado do estudo 2A indicando que as LLMs conseguem distinguir os grupos de apps entre *Winners* e *No-Winners,* ainda assim com pouca diferença. A partir disso, com todos os testes apresentados, podemos concluir que a melhor LLM em relação à concordância com os outros modelos de avaliação para avaliação de criatividade de aplicativos criados com *App Inventor* foi a LLM Mistral 70b, porém no geral tivemos baixas concordâncias entre as LLMs e os outros modelos de avaliação.

Limitações

Para minimizar ameaças a validade de conclusão foram realizados estudos adicionais com tamanho de amostra maior (1.078 apps), além dos estudos com somente 35 do *App of the Month App of the Season*. Foram também escolhidos métodos estatísticos apropriados tanto para a pergunta de pesquisa quanto em relação ao tipo de dados e quantidade de dados coletados.

Em relação à validade interna se criou protocolos de coleta de dados explícitos. A seleção dos apps avaliados foi feita tanto da galeria do *App Inventor* quanto do conjunto de apps do *App of the Month*. Os prompts foram gerados de forma automatizada pelos dados extraídos automaticamente dos apps utilizando o modelo *Creassessment*. Porém, apesar do modelo *Creassessment* ser um modelo já validado, o uso das informações automaticamente extraídas pelo modelo para geração do *prompt* automatizado pode ter influenciado nos resultados.

Para minimizar ameaças em termos de validade externa foram utilizadas várias LLMs diferentes. Em termos de avaliação humana foram utilizadas as classificações dos apps em *Winners* e *No-Winners* e também a criatividade avaliada por um grupo de juízes da competição *App of the Month* do MIT/EUA.

Conclusão

Como resultado do presente trabalho foi feita uma fundamentação teórica relacionada à avaliação da criatividade no contexto do ensino de computação na educação básica, além dos diferentes meios de avaliação da criatividade e sobre as *Large Language Models* (LLMs) e sua capacidade de entender e gerar textos. Foi levantado o estado da arte em

relação a avaliação de LLMs para a avaliação da criatividade no contexto de ensino da computação na educação básica, indicando que ainda existem muito poucos estudos deste tipo e nenhum relacionado a avaliação de criatividade de aplicativos. Foram realizados diversos estudos empíricos para avaliar a qualidade de avaliação da criatividade com LLMs em relação a outros modelos de avaliação já validados, como juízes humanos e o modelo automatizado *Creassessment* (Alves, 2023), além de estudos empíricos para avaliar o grau de confiabilidade e concordância entre as avaliações de diferentes LLMs.

A contribuição deste trabalho está na comparação da qualidade da avaliação da criatividade de aplicativos *App Invento*r com LLMs, comparadas com modelos de avaliação já validados, como juízes humanos e o modelo automatizado *Creassessment*, comparação a qual mostrou resultados insatisfatórios, com testes de correlação e concordância demonstrando resultados fracos e/ou moderados. Além disso, este trabalho também contribuiu para a análise da confiabilidade e concordância entre as respostas de diferentes LLMs, análises as quais demonstraram também baixa concordância e confiabilidade. Portanto, mostrando que a princípio e com a forma com que foram feitos os testes, as LLMs não seguem a mesma tendência de avaliação dos outros modelos de avaliação já validados, além de não terem entre si a mesma concordância para avaliação da criatividade de apps *App Inventor*.

Como sugestão para trabalhos futuros, sugere-se realizar novas aplicações e análises de testes utilizando *prompts* diferentes, podendo ser criados manualmente e/ou com mais dados e características dos aplicativos, ou com outras técnicas de *prompt engineering*. Além disso, também sugere-se utilizar diferentes LLMs das utilizadas neste trabalho, para verificar se existe alguma com um desempenho melhor ao ser comparada com outros modelos de avaliação já validados.

Referências

Alves, N. C. Assessing mobile apps creativity in computing education. Tese (Doutorado em em Ciência da Computação) – PPGCC/Universidade Federal de Santa Catarina. 2023.

Birhane, A., Kasirzadeh, A., Leslie, D., Wachter S.. Science in the Age of Large Language Models. Nature Reviews Physics 5, 277–280, 2023.

Bland, J. Martin; Altman, Douglas G. Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. en. The Lancet, 327(8476), p. 307–310, 1986.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A.. Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901, 2020.

Cichetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment, 6(4), 284-290.

Devellis, R. F. Scale development: theory and applications. 4th. ed. SAGE, 2017.

Harunani, F., Patton, E. W., Tissenbaum, M.. MIT App Inventor: Objectives, Design, and Development. Computational Thinking Education, p. 31-49, 2019.

Hattie, J., Timperley, H.. The power of feedback. Review of Educational Research, 7(1), 81--112, 2007.

Mishra, P.; Yadav, A. Of art and algorithms: Rethinking technology and creativity in the 21st century. TechTrends , 57(3), p. 10–14, 2013.

MIT App Inventor. Disponível em: https://appinventor.mit.edu/. Acesso em: 18 de novembro de 2023.

P21, Partnership for 21st century learning, P21 Framework Definitions, 2021. Disponivel em: http://www.p21.org/

Saunders, D., Thagard, P. Creativity in Computer Science. In Creativity across domains: Faces of the muse. J. C. Kaufman and J. Baer. Mahwah (eds.), NJ, Lawrence Erlbaum Associates, 2005.

WEF, World Economic Forum, Defining Education 4.0: A Taxonomy for the Future of Learning, 2023. Disponível em:

https://www.weforum.org/publications/defining-education-4-0-a-taxonomy-for-the-future-of-learning/

Yin, K., Case study research: design and methods, 4. ed. Beverly Hills: Sage Publications, p. 312, 2009.