Generative AI perspectives: Design e IHC

Heloisa Candello
Manager of Human-centered and responsible tech group
hcandello@br.ibm.com

IBM Research

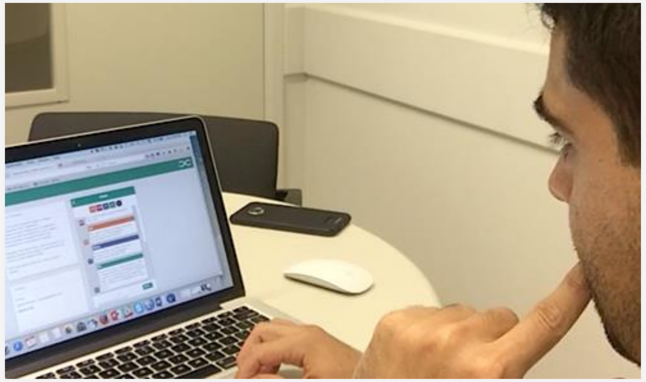
Responsible and Inclusive Tech

Leading, publishing and conducting **research to understand peoples' contexts and motivations to use conversational technologies.**

Manager of Human-centered and responsible tech group

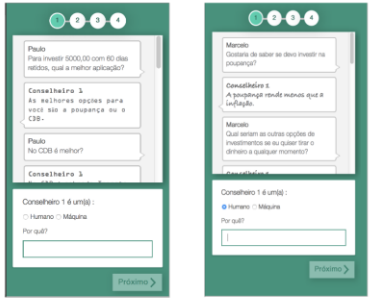
PhD in Computer Science, Brighton University, UK, 2012.
Master in Multimedia, UNICAMP, 2006.
Design, UFSC, 2003.

Human-Computer Interaction
Publications in several conferences
CSCW, CHI, CUI, DUXU, IHC, AAAI, IDC.



IBM **Research**

Conversational Systems



1 2 3 4

Paulo
Para investir 5000,00 com 60 dias retidos, qual a melhor aplicação?

Conselheiro 1
As melhores opções para você são a poupança ou o CDB.

Paulo
No CDB é melhor?

Conselheiro 1

Conselheiro 1 é um(a) :
○ Humano ○ Máquina
Por quê?

Próximo

1 2 3 4

Marcelo
Gostaria de saber se devo investir na poupança?

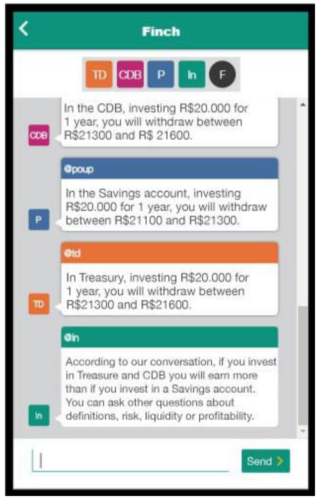
Conselheiro 1
A poupança rende menos que a inflação.

Marcelo
Qual seriam as outras opções de investimentos se eu quiser tirar o dinheiro a qualquer momento?

Conselheiro 1 é um(a) :
● Humano ○ Máquina
Por quê?

Próximo

How do typefaces affect the perception of humanness in chatbots?
[CHI2017]



Finch

TD CDB P In F

CDB: In the CDB, investing R$20.000 for 1 year, you will withdraw between R$21300 and R$ 21600.

@poup
In the Savings account, investing R$20.000 for 1 year, you will withdraw between R$21100 and R$21300.

@td
In Treasury, investing R$20.000 for 1 year, you will withdraw between R$21300 and R$21600.

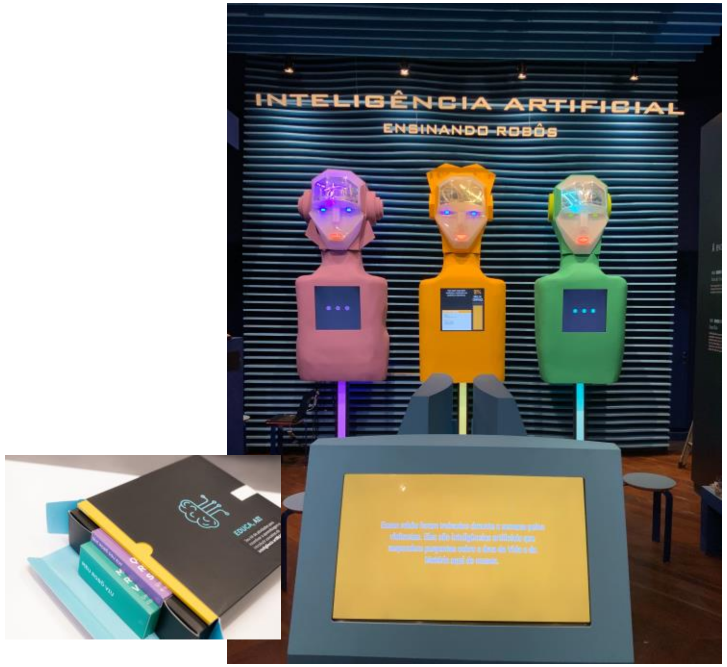
@in
According to our conversation, if you invest in Treasure and CDB you will earn more than if you invest in a Savings account. You can ask other questions about definitions, risk, liquidity or profitability.

Send

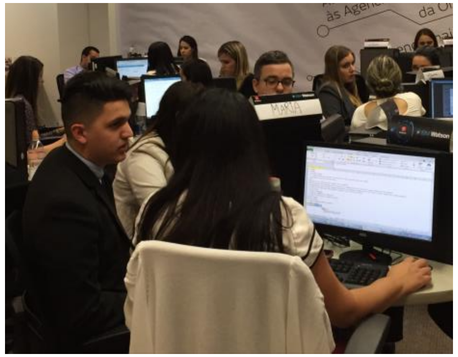
Should we design multi or single bots?
[*DeepDial'18,*]

What are the audience effects?
[CUI2021, CHI2019, CHI2020]

What are the popular questions in Art galleries?
[CUI 2021}



How can we teach ML for children in 30 min?
[IDC 2020, CSCW 2021]



How do curators articulate their work using strategies that are not previously supported by the conversational platforms?
[CHI2021]. [CSCW2022]

Responsible and Inclusive Tech

Frameworks, Methods, & Tools

Mitigation of Harmful Tech Impacts

Responsible Quantum Computing

Tech for Justice

Tech for Justice - Mission

To advance and leverage our research and development capabilities in partnership with organizations, institutions and communities whose focus is to address racial and social injustices with solutions, tools, and technical resources.

We aim to support positive social impact across disciplinary domains and lived experiences.

IBM **Research**

How can Artificial Intelligence leverage alternative criteria for creditworthiness and help to measure social impact?

Financial local practices & AI

Why is this work needed?

Vulnerable and low-income communities

Brazil has a large low-income population that rely on cash, if they have opportunity for microcredit **this can enhance** their economic life, business & social outcomes.

Limited credit access

Lack of formal credit score, creditworthiness proof, financial education, documentation, awareness and knowledge about their own business

High interest rates

The interest rate of banks and financial agencies are high and it can increase poverty . Credit for GOV beneficiaries is irresponsible and lead to over-indebtedness

Alternative data

19 million adults as credit unscorable and 26 million as credit invisible. Combined, these 45 million represent almost 20 percent of the adult population, with African Americans and Latinx more likely to be **credit unscorable** or invisible than white people and Asian Americans. (CFPB)

How can AI leverage alternative criteria and suggest a better way to measure social and economic impact of microcredit actions?

Our research

1. Support the decision-making processes on the microcredit framework
2. **Calculate social impact of microcredit actions;**
3. **Microcredit access for low-income entrepreneurs by enhancing non-traditional financial practices with AI;**
4. AI conversational technologies to promote **entrepreneurs' business health** “awareness”

Use cases:

1. Social impact in low-income communities
2. Machine learning pipeline for microcredit

Supporting Technologies:

1. Conversational models
2. Data model/analysis
3. Machine learning devops

Heloisa Candello
Challenge lead and HCI

Emílio Brazi
Machine Learning

Rogerio De Paula
Manager – Social Science

Cassia Sanctos
RSE - ML

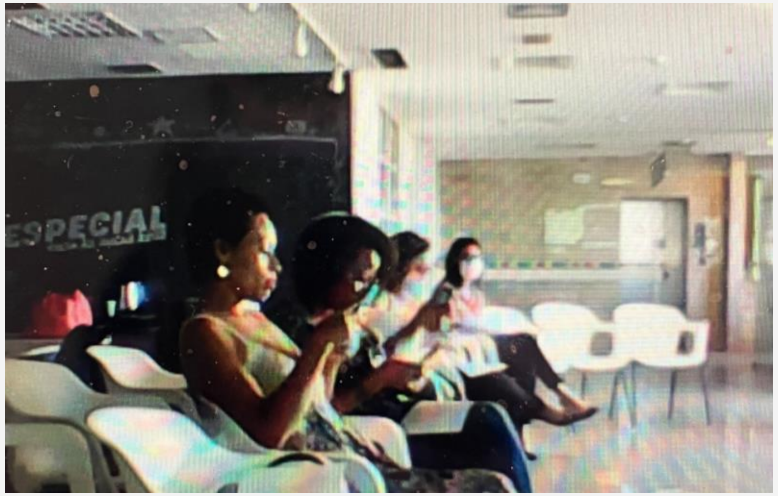
Melina Guerra
RSE

Marcelo Grave
RSE – Conversational tech

Gabriel Soella
Research Intern

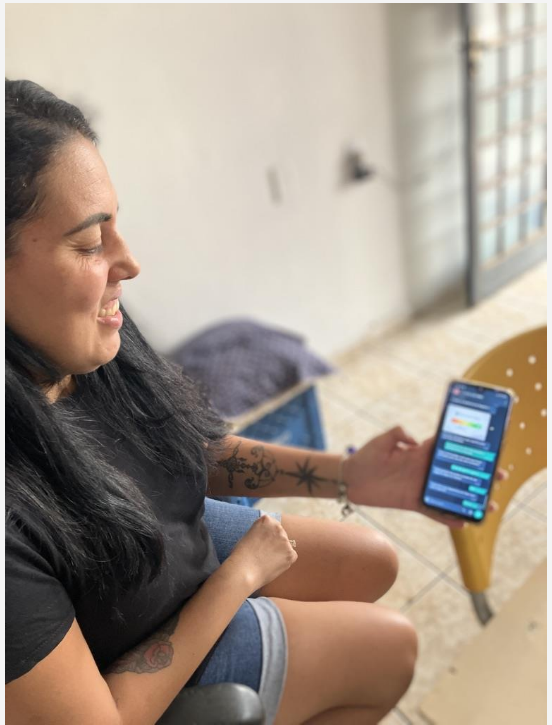
Adinan Brito
Undergrad Intern

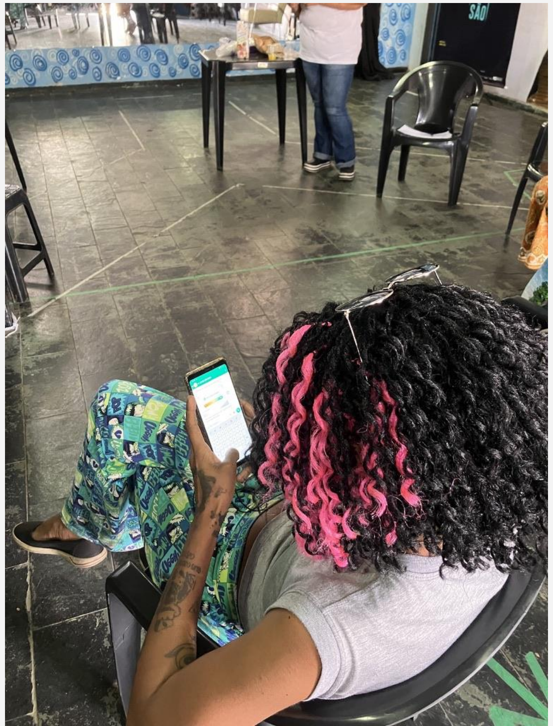
What questions does AI need to answer to be useful, effective and Trustworthy?



Situation: Small business owners are not always aware of their business status.

It can affect their assessment of credit need and how to apply the loan.





Liao, Q. V., Gruen, D., & Miller, S. (2020, April). Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-15).

IBM **Research**

Business health index

Por que deu zero?

(deu 0 a 500) “Que eu tenho muito o que melhorar” e Pq? “Porque eu não tinha nada estruturado”

Está considerando mais a renda e organização

“Ele olhou a renda da casa “ e quantas pessoas?” E os seus gastos. É que o aluguel não sou que pago, mulhé”

Visão de futuro achei muito interessante essa pergunta aqui. Eu queria continuar

Acho muito importante educação, entrar em uma loja e alguém te trata bem

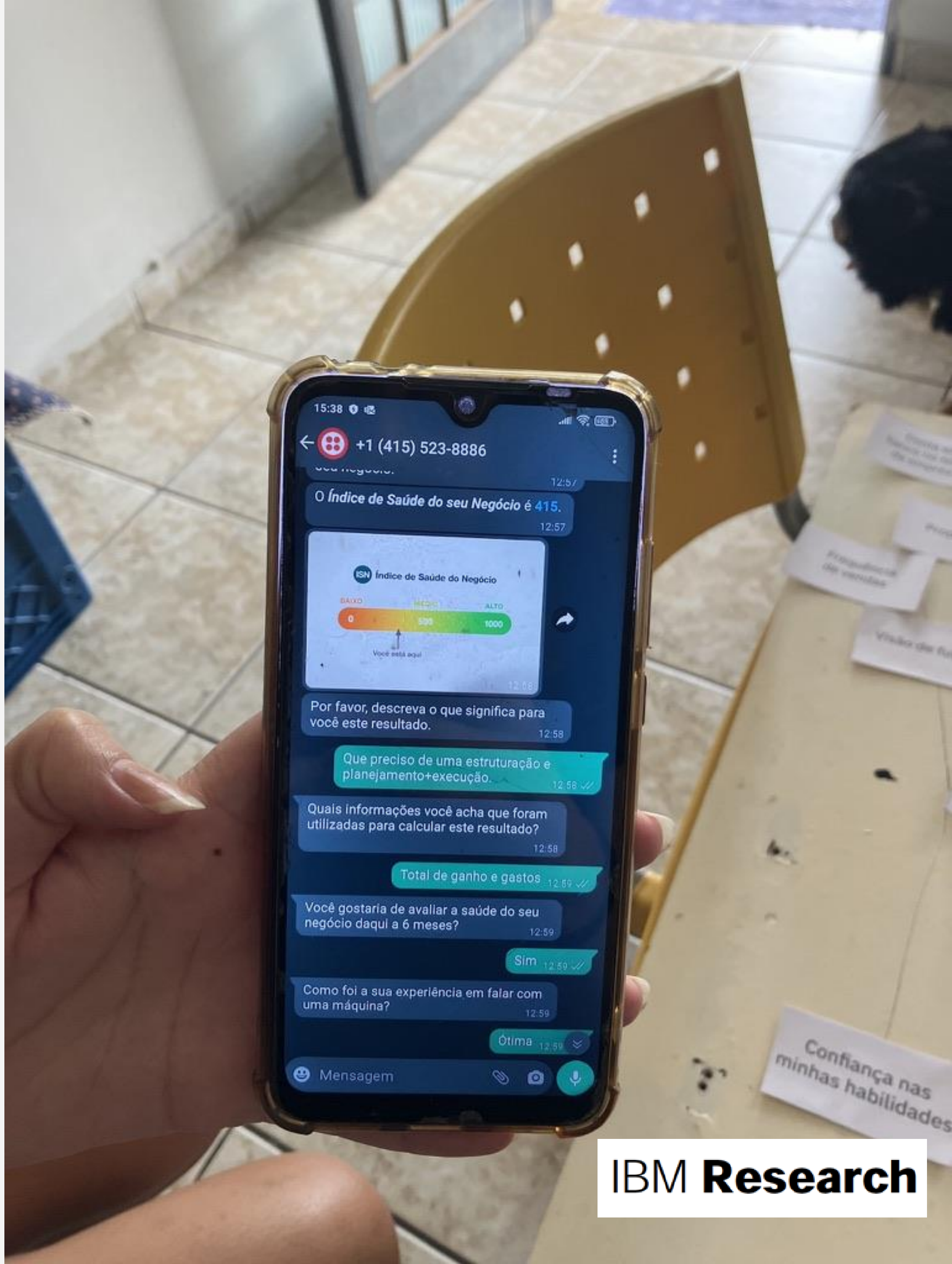
Esse que tipo de identificação que o seu negócio possui e o nome?

Eu não gostaria de responder sobre religião. Eu não sigo nenhuma religião, e isso e pessoal e não interfere no negócio

Um treinamento para saber se eu estou no caminho certo. igual tinha no todas elas uma consultoria, mais específica para mim, tem a boleira tem eu que vendo roupa.

Se eu estou indo no caminho certo, se eu preciso melhorar?

Como está a minha propaganda? Se eu estou errando em alguma coisa



O *Índice de Saúde do seu Negócio* é 415. 12:57

Por favor, descreva o que significa para você este resultado. 12:58

Que preciso de uma estruturação e planejamento+execução. 12:58

Quais informações você acha que foram utilizadas para calcular este resultado? 12:58

Total de ganho e gastos 12:59

Você gostaria de avaliar a saúde do seu negócio daqui a 6 meses? 12:59

Sim 12:59

Como foi a sua experiência em falar com uma máquina? 12:59

Ótima 12:59

IBM **Research**

Generative AI – What are the challenges?

Unprecedented levels of *performance*

Three main capabilities are driving their uptake

Scale

- Ability to ingest/process huge amounts of data

Homogenization

- Built once and adapted to multiple tasks; multi-modality; multi-disciplinarity

Emergence

- Scale created unprecedented and unexpected capabilities (e.g. unparallel fluency, multi-step reasoning, etc.)

Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." arXiv preprint arXiv:2108.07258 (2021).

Risks associated with language models.

Hate speech and exclusion

The LM accurately reflects unjust, toxic, and oppressive speech present in the training data.

Malicious uses

Humans intentionally use the LM to cause harm.

Human-computer interaction harms

Humans are deceived or made vulnerable via direct interaction with a powerful conversational agent.

Information hazards

The LM leaks or correctly infers sensitive information.

Misinformation harms

The LM provides false, misleading, nonsensical or poor-quality information.

Discrimination and socioeconomic harms

LMs are used to underpin widely used downstream applications that disproportionately benefit and harm different groups.

Source: L. Weidinger, et al. “Taxonomy of Risks Posed by Language Models.” *ACM Conference on Fairness, Accountability and Transparency*, pp. 214–229, Jun. 2022.

Specific Issues concerning FMs

Generative Nature

– Hallucination, false and harmful language generation due to lack of adequate model control & safeguards

Misalignment of Expectations

– Generate contents that are not aligned with expected social, cultural values and norms

Lack of transparency

– Hard to inspect & audit and may obscure potential societal and other harms

Specifically in our Work

Our Mission is to

“Enable customers to safely harness the power of foundation models to do enterprise NLP better, faster, and cheaper, while enabling opportunities for new capabilities.”

David Cox
VP of FMs Technologies
IBM Research

Safe-First

Safety is the minimization of the probability of expected harms and the possibility of unexpected harms.

+

Responsible

To a critical, reflective, and community-centered perspective on the impact of LLMs on society

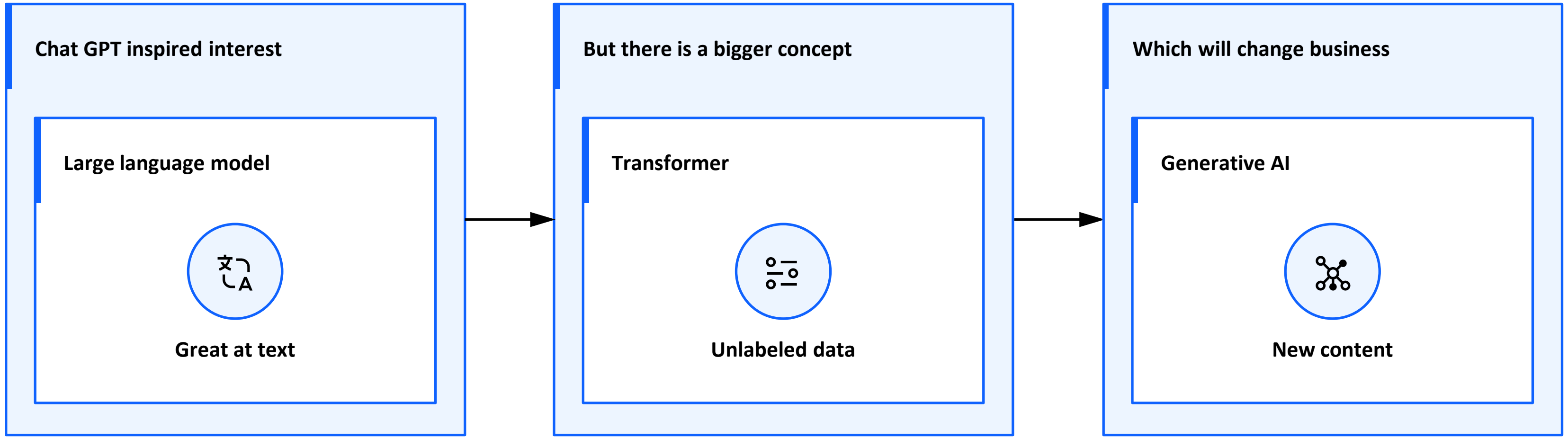
+

Enterprise-Oriented

In contrast to existing LLMs openly available, which can be thought of as “open domain systems,” IBM focuses on “closed domains”

Source: K. R. Varshney and H. Alemzadeh. “On the Safety of Machine Learning: Cyber-Physical Systems, Decision Sciences, and Data Products.” *Big Data*, vol. 5, no. 3, pp. 246 – 255, Sep. 2017.

Foundational models are unlocking new opportunities



Chat GPT inspired interest

Large language model

Great at text

But there is a bigger concept

Transformer

Unlabeled data

Which will change business

Generative AI

New content

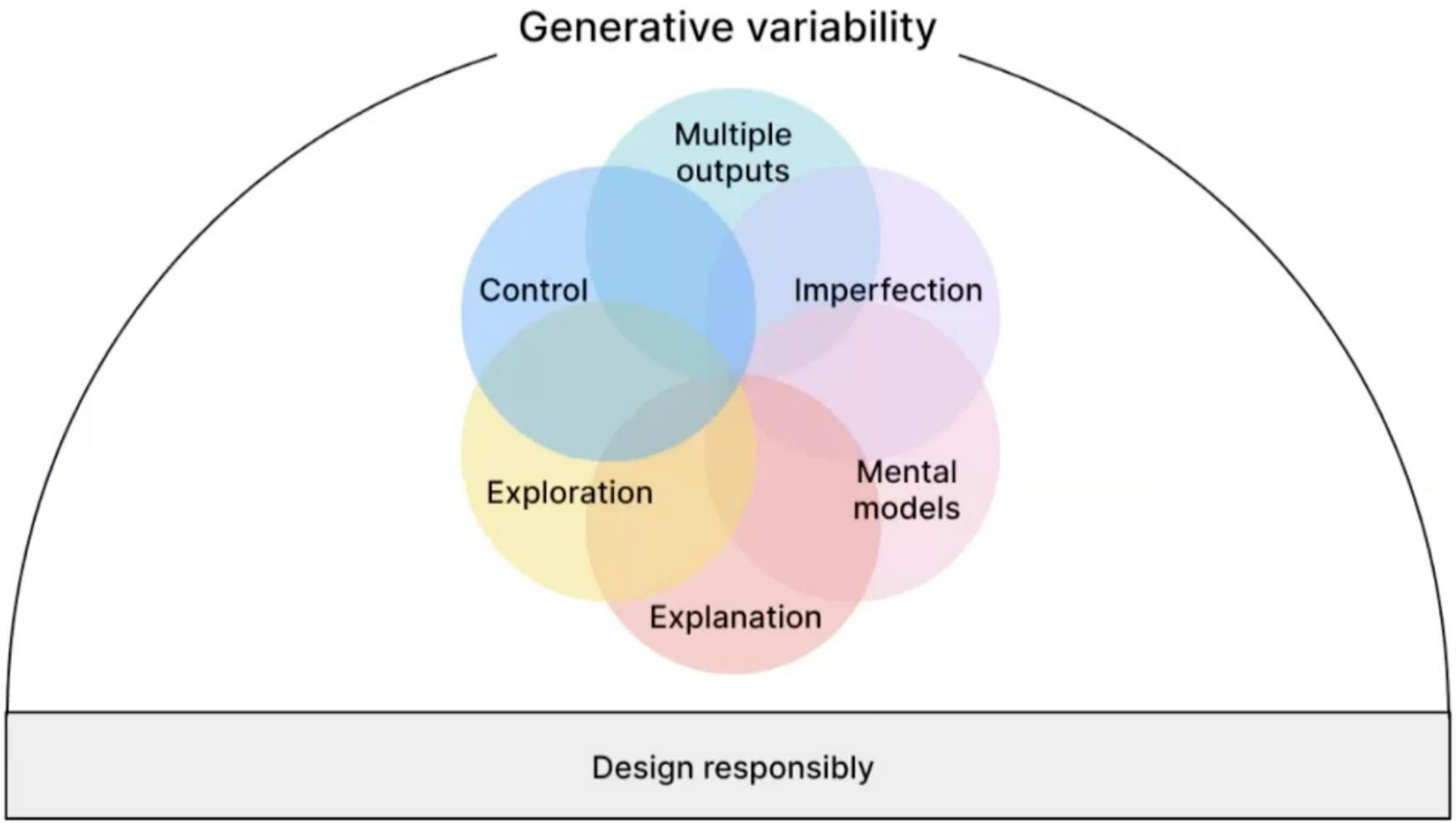
Foundation models for business

- Large language model
- IT automation models
- Digital labor models
- Cybersecurity models

Design principles

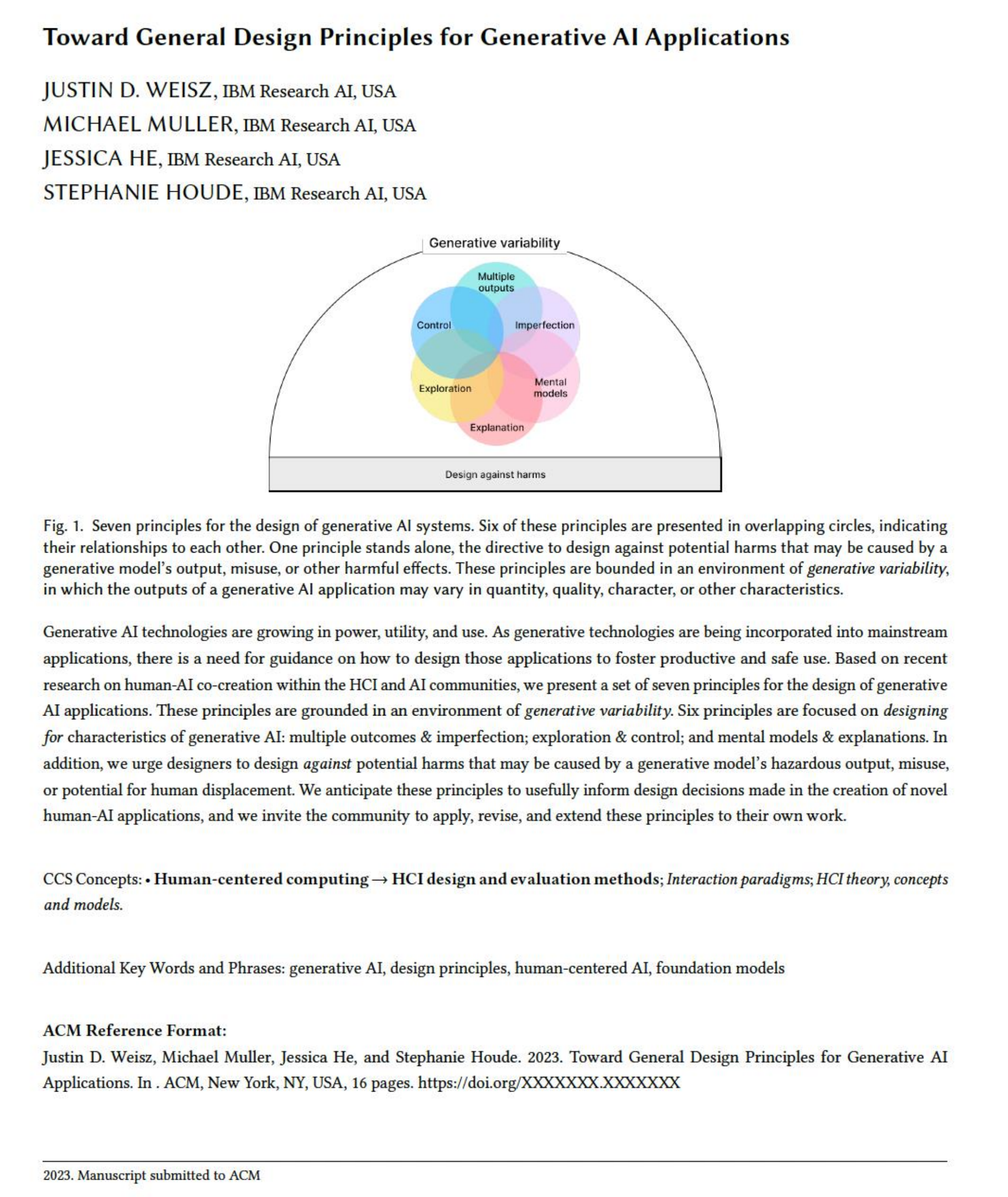
What concepts do designers of AI systems need to understand about generative AI?

Seven design principles grounded in an environment of generative variability



Weisz, J. D., Muller, M., He, J., & Houde, S. (2023). Toward General Design Principles for Generative AI Applications. In Joint Proceedings of the IUI 2023 Workshops: HAI-GEN, ITAH, MILC, SHAI, SketchRec, SOCIALIZE, co-located with the ACM International Conference on Intelligent User Interfaces (IUI 2023). Virtual Event, Sydney, Australia, March 27-31, 2023, CEUR-WS.org/Vol-3359

IBM **Research**



Toward General Design Principles for Generative AI Applications

JUSTIN D. WEISZ, IBM Research AI, USA
MICHAEL MULLER, IBM Research AI, USA
JESSICA HE, IBM Research AI, USA
STEPHANIE HOUDE, IBM Research AI, USA

Fig. 1. Seven principles for the design of generative AI systems. Six of these principles are presented in overlapping circles, indicating their relationships to each other. One principle stands alone, the directive to design against potential harms that may be caused by a generative model's output, misuse, or other harmful effects. These principles are bounded in an environment of *generative variability*, in which the outputs of a generative AI application may vary in quantity, quality, character, or other characteristics.

Generative AI technologies are growing in power, utility, and use. As generative technologies are being incorporated into mainstream applications, there is a need for guidance on how to design those applications to foster productive and safe use. Based on recent research on human-AI co-creation within the HCI and AI communities, we present a set of seven principles for the design of generative AI applications. These principles are grounded in an environment of *generative variability*. Six principles are focused on *designing for* characteristics of generative AI: multiple outcomes & imperfection; exploration & control; and mental models & explanations. In addition, we urge designers to design *against* potential harms that may be caused by a generative model's hazardous output, misuse, or potential for human displacement. We anticipate these principles to usefully inform design decisions made in the creation of novel human-AI applications, and we invite the community to apply, revise, and extend these principles to their own work.

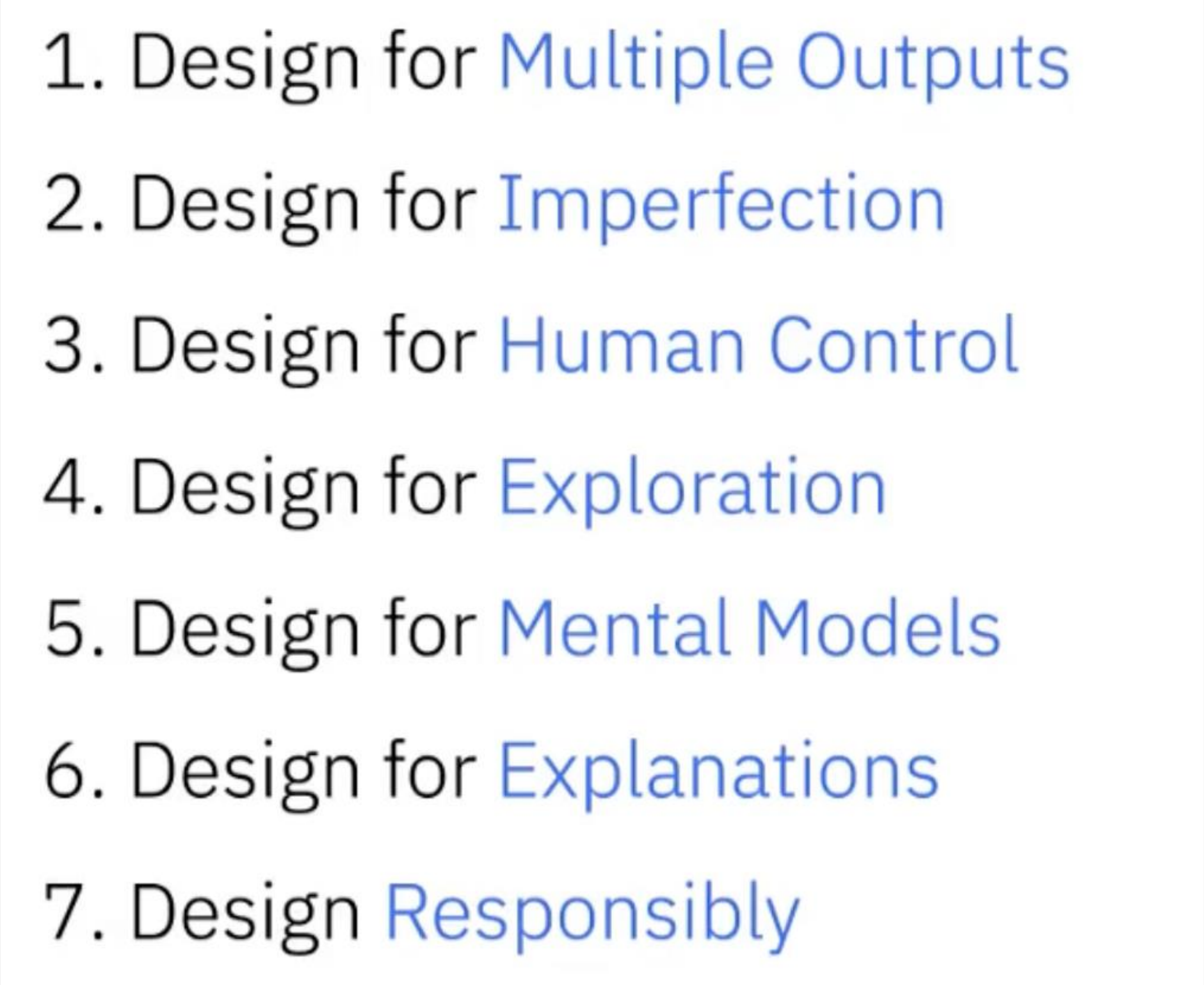
CCS Concepts: • **Human-centered computing → HCI design and evaluation methods**; *Interaction paradigms*; *HCI theory, concepts and models*.

Additional Key Words and Phrases: generative AI, design principles, human-centered AI, foundation models

ACM Reference Format:

Justin D. Weisz, Michael Muller, Jessica He, and Stephanie Houde. 2023. Toward General Design Principles for Generative AI Applications. In . ACM, New York, NY, USA, 16 pages. https://doi.org/XXXXXXX.XXXXXXX

2023. Manuscript submitted to ACM



1. Design for Multiple Outputs
2. Design for Imperfection
3. Design for Human Control
4. Design for Exploration
5. Design for Mental Models
6. Design for Explanations
7. Design Responsibly

Design Principles for Generative AI Applications

Principle 1: Design for Multiple Outputs

Versioning

Help users keep track of the work they produce

Visualizing Differences

Help users understand how multiple outputs differ from each other

Principle 2: Design for Imperfection

Multiple Outputs

Sometimes the model needs to try more than once to produce the right solution

Co-Creation

Enable users to collaboratively edit with the AI

Evaluation & Identification

Use domain-specific metrics to evaluate artifacts or identify promising candidates

Principle 3: Design for Human Control

Generic Controls

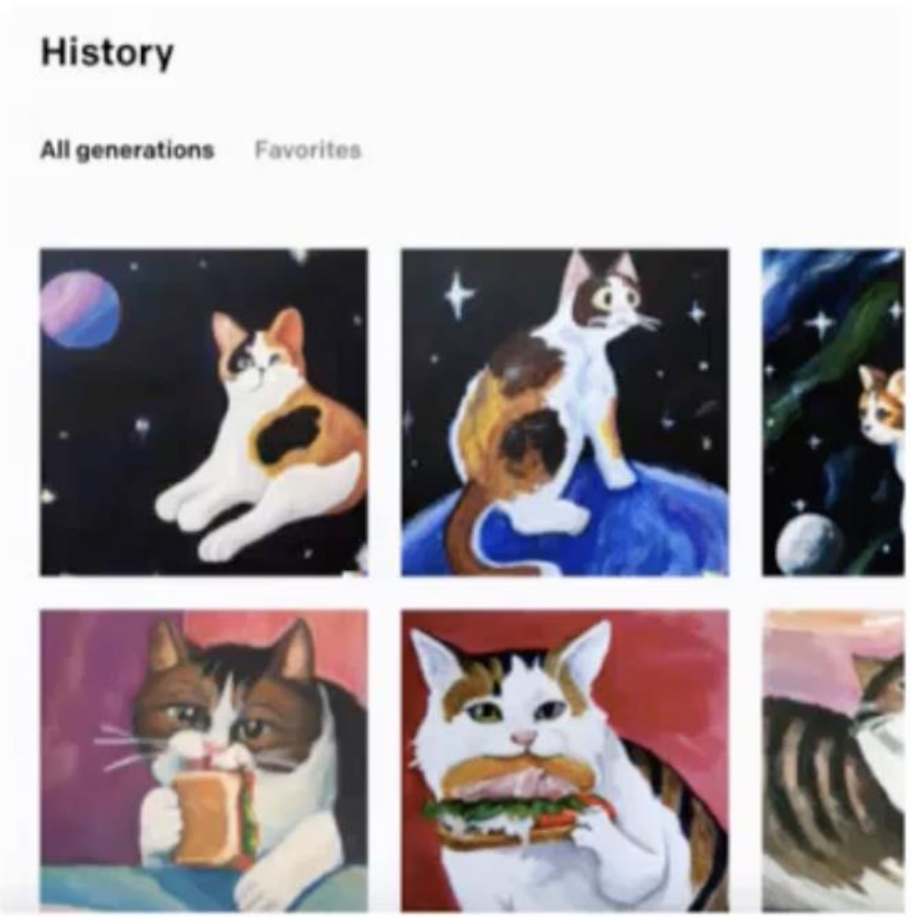
Number of outputs, variability (e.g. temperature), random seed

Technology-specific Controls

Enc/Dec: Semantic sliders (e.g. Liu & Chilton 2021)
Transformer: Prompt engineering (e.g. CoT)

Domain-specific Controls

Molecules: Water solubility, molecular weight
Code: Run-time, memory



DALL·E

Weisz, J. D., Muller, M., He, J., & Houde, S. (2023). Toward General Design Principles for Generative AI Applications. In Joint Proceedings of the IUI 2023 Workshops: HAI-GEN, ITAH, MILC, SHAI, SketchRec, SOCIALIZE, co-located with the ACM International Conference on Intelligent User Interfaces (IUI 2023). Virtual Event, Sydney, Australia, March 27-31, 2023, CEUR-WS.org/Vol-3359

Design Principles for Generative AI Applications

Principle 4: Design for Exploration

Multiple Outputs

Show users different possibilities or options

Control

Yes, have them 😌

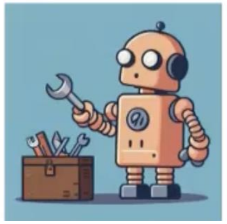
Sandbox / Playground

Let users explore before committing to a particular artifact

Visualization

Show the space of possibilities and what the user has explored (Kreminski et al. 2022; Rost & Andreasson 2023)

Rezwana & Maher (2023)



https://medium.com/human-centered-ai/on-ai-anthropomorphism-abff4cecc5ae

Shneiderman & Muller (2023)

Principle 5: Design for Mental Models

Orient Users to Generative Variability

Users should understand that the system may produce multiple outputs, those outputs may be imperfect, and their effort may be required to produce the desired result

Consider the Role of the AI

Is it a tool or partner? Does it act proactively or just respond to the user? Does it make changes to an artifact directly or simply make recommendations to the user?

Weisz, J. D., Muller, M., He, J., & Houde, S. (2023). Toward General Design Principles for Generative AI Applications. In Joint Proceedings of the IUI 2023 Workshops: HAI-GEN, ITAH, MILC, SHAI, SketchRec, SOCIALIZE, co-located with the ACM International Conference on Intelligent User Interfaces (IUI 2023). Virtual Event, Sydney, Australia, March 27-31, 2023, CEUR-WS.org/Vol-3359

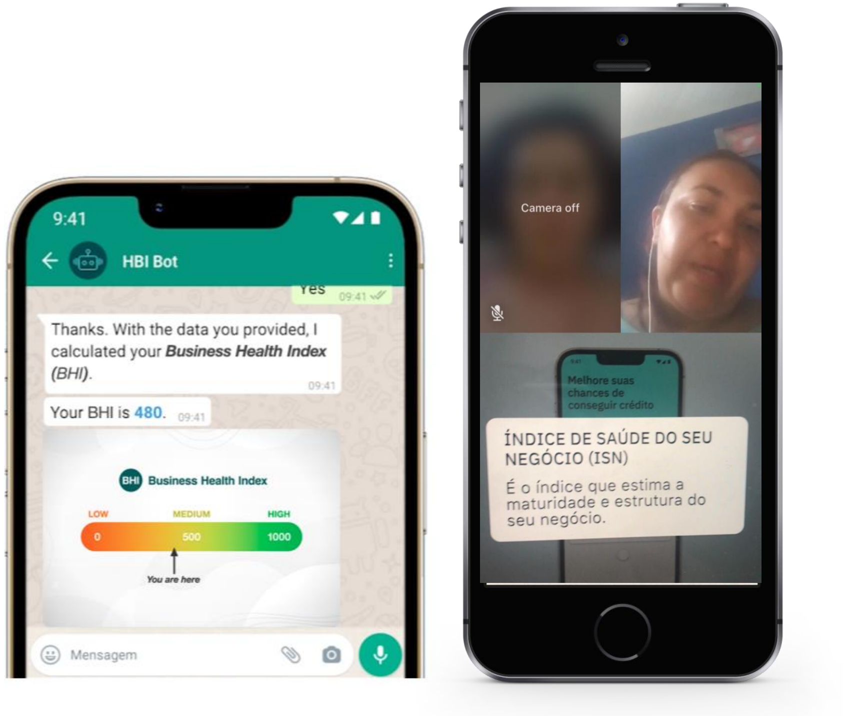
Principle 6: Design for Explanations

Communicate capabilities and limitations

Help users calibrate their trust by understanding what the system can and cannot do

Create useful mental models

Use explanations to help users understand how the system works and how to work with it effectively (Weisz et al. 2021)



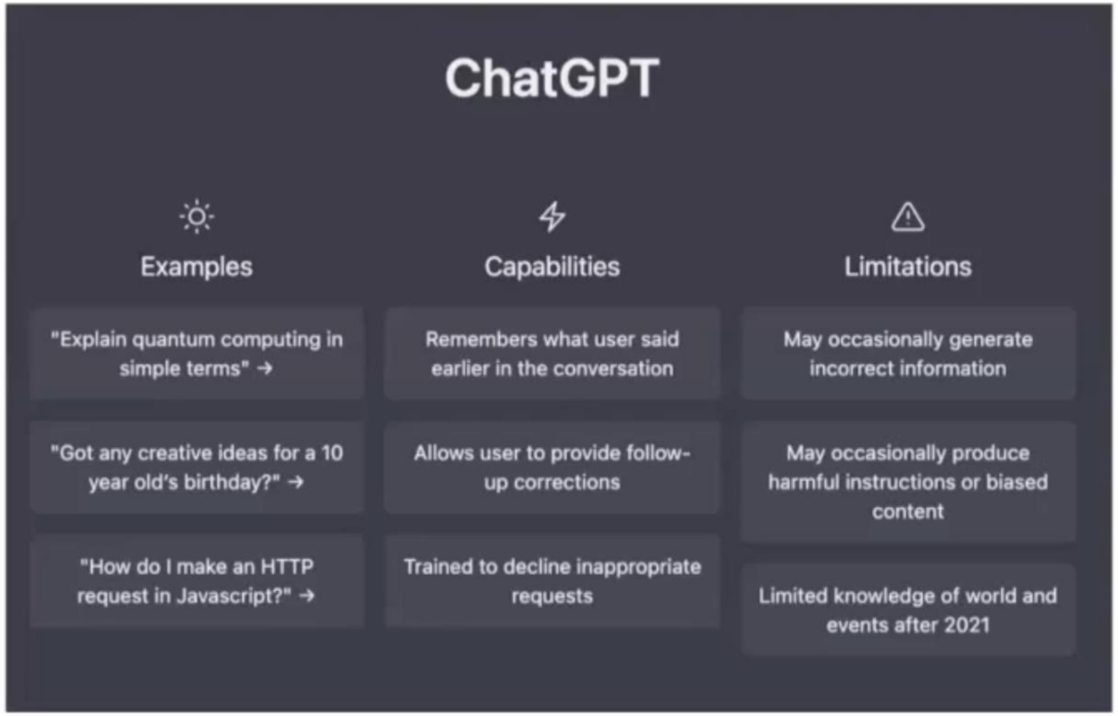
Thanks. With the data you provided, I calculated your ***Business Health Index*** *(BHI)*. 09:41

Your BHI is 480. 09:41

ÍNDICE DE SAÚDE DO SEU NEGÓCIO (ISN)

É o índice que estima a maturidade e estrutura do seu negócio.

How can AI leverage alternative criteria and suggest a better way to measure credit worthiness and economic growth ?
[CHI 22 (panel), FAccT22 (panel) CUI 22 (demo) ;
HCI@NeurIps21, AAAI workshop]



ChatGPT

Examples	Capabilities	Limitations
"Explain quantum computing in simple terms" →	Remembers what user said earlier in the conversation	May occasionally generate incorrect information
"Got any creative ideas for a 10 year old's birthday?" →	Allows user to provide follow-up corrections	May occasionally produce harmful instructions or biased content
"How do I make an HTTP request in Javascript?" →	Trained to decline inappropriate requests	Limited knowledge of world and events after 2021

Weisz, J. D., Muller, M., He, J., & Houde, S. (2023). Toward General Design Principles for Generative AI Applications. In Joint Proceedings of the IUI 2023 Workshops: HAI-GEN, ITAH, MILC, SHAI, SketchRec, SOCIALIZE, co-located with the ACM International Conference on Intelligent User Interfaces (IUI 2023). Virtual Event, Sydney, Australia, March 27-31, 2023, CEUR-WS.org/Vol-3359

Thank you.

Heloisa Candello
Manager of Human-centered and responsible tech group
hcandello@br.ibm.com

watsonx.ai

A proven, trusted enterprise studio that brings together Machine Learning and Generative AI for builders

Combine the power of predictive, prescriptive and generate AI in a single integrated studio to optimize the AI lifecycle.

Empower all users including coders and non-coders to use open source and visual modeling tools on a unified studio.

Leverage existing cloud and data investments and avoid lock-in with flexible deployment.

Connect to and analyze data across the business, no matter where it resides

Operationalize AI with an integrated environment across any cloud

Accelerate and Scale your business using next-gen foundation models

Leverage foundation models and generative AI

Build AI applications in a fraction of the time with a fraction of the data.

Foundation model Libraries: Easy access to IBM-proprietary and open-source Foundation Models

Prompt Lab: Experiment with zero/few-shot learning for enterprise tasks

Tuning Studio: Tailor pre-trained Foundation Models for complex downstream tasks on enterprise data

+ AI Builder API / SDK Toolkit
Workbench tooling and models can be used via GUI or APIs that integrate directly into enterprise applications

Use cases:

– Generate customized marketing emails

– Summarize Webex meeting transcripts

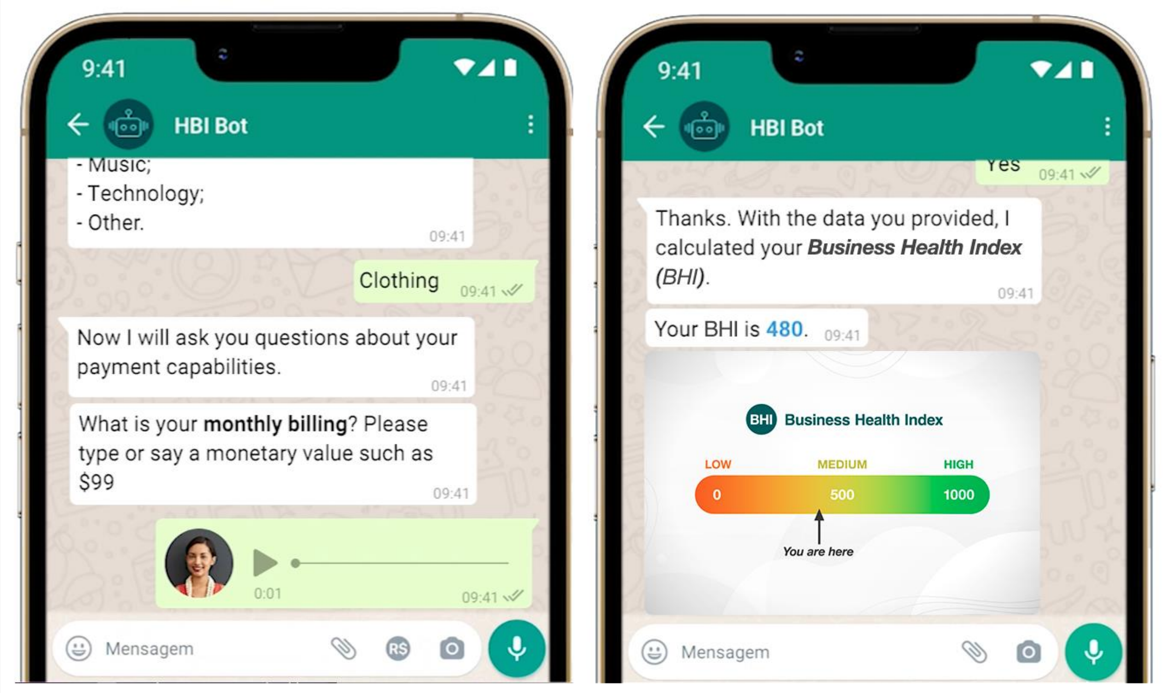
– Classify customer complaints without labeled data

– Translate code from markdown to html

– Extract key facts from unstructured financial documents

Empowering minority communities by enhancing local financial practices with Artificial Intelligence

-bots



- Music;
- Technology;
- Other. 09:41

Clothing 09:41

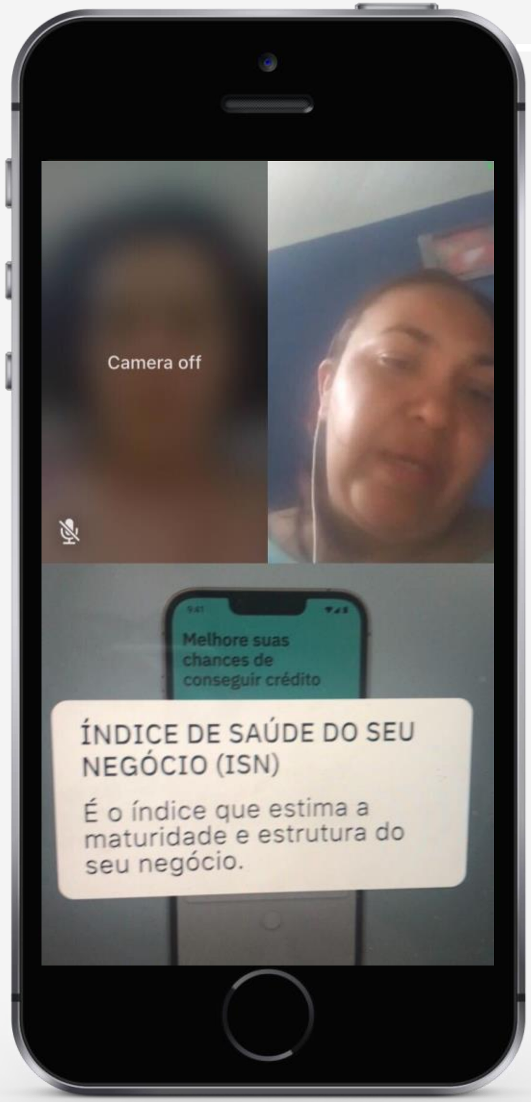
Now I will ask you questions about your payment capabilities. 09:41

What is your **monthly billing**? Please type or say a monetary value such as $99 09:41

Yes 09:41

Thanks. With the data you provided, I calculated your ***Business Health Index (BHI)***. 09:41

Your BHI is 480. 09:41



Melhore suas chances de conseguir crédito

ÍNDICE DE SAÚDE DO SEU NEGÓCIO (ISN)

É o índice que estima a maturidade e estrutura do seu negócio.

How can AI leverage alternative criteria and suggest a better way to measure credit worthiness and economic growth ?
[CHI 22 (panel), FAccT22 (panel) CUI 22 (demo) ;
HCI@NeurIps21, AAAI workshop]

Future expected impacts

Individuals and Communities

1

FINANCIAL INCLUSION AND ECONOMIC GROWTH OPPORTUNITIES

2

A SOCIAL IMPACT ASSESSMENT OF FINANCIAL ACTIONS IN VULNERABLE COMMUNITIES

3

ENABLING INFORMED DECISIONS TO MITIGATE POSSIBLE SOCIAL AND ECONOMICAL PROBLEMS