OFFICE OF ETHICAL AND HUMANE USE

Building Trust in AI with a Human at the Helm

We believe that humans and technology work best together. To ensure that AI technology supports human flourishing and fosters trust, we need to center humans and put them in the lead in human-AI interactions and processes. Salesforce user research has found that the outputs and impacts of generative AI use are better and more trustworthy when humans play key roles in the process.

Read on for research highlights and concrete ways to optimize AI use in your business by keeping a human at the helm.

salesforce

What is ‘Human in the Loop’?

Originally used in the military, nuclear energy, and aviation sectors, the phrase “human in the loop”[1] (HITL) referred to the ability for humans to intervene in automated systems to prevent disasters. Today, HITL has been widely repurposed to talk about how people both generate and approve content using AI and give feedback about which outputs are best in order to improve the system.

Examples of Human in the Loop:

Customer Support and Chatbots

Customer support systems rely on human agents to ensure accurate and contextually appropriate responses and to help resolve complex or sensitive customer queries. Human agents monitor and assist AI-powered chatbots (i.e., by editing or providing additional information).

Medical Imaging and Diagnosis

In radiology, for example, AI algorithms can assist in the analysis of medical images, such as X-rays, CT scans, or MRIs. However, human radiologists are essential to review the findings, interpret complex cases, and make final diagnoses, considering the broader clinical context and patient history.

Fraud Detection and Prevention

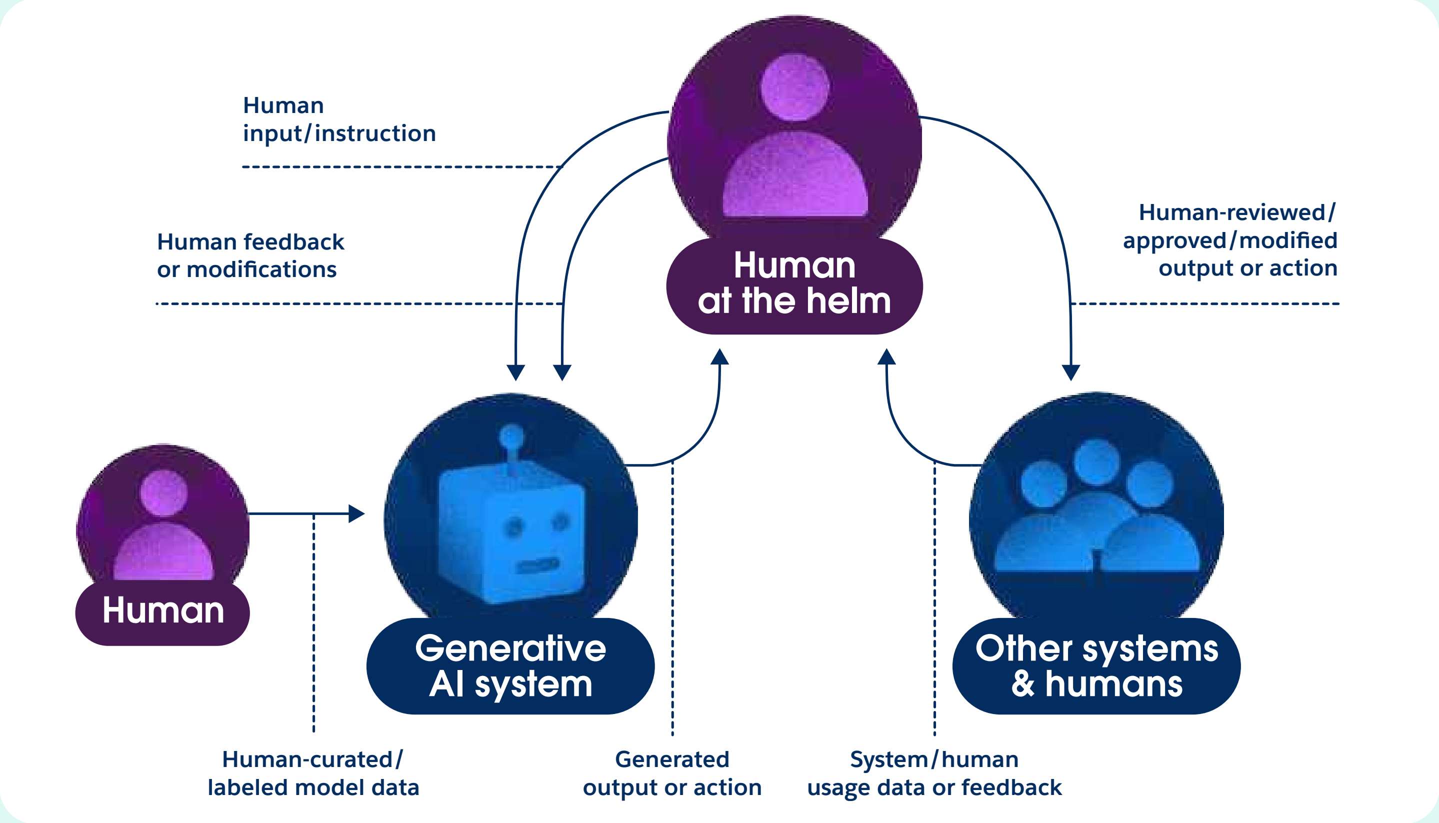
HITL is used in financial institutions and e-commerce to detect and prevent fraud. AI systems flag suspicious transactions and human analysts review, assess risk, investigate anomalies, and provide feedback to improve accuracy and effectiveness of fraud detection systems.

Our research has led Salesforce to believe that generative AI outcomes are better when humans are *more than* “in the loop.” To cultivate trust over time and build confidence in this still-emergent AI technology, generative AI systems should meaningfully integrate humans into decision-making, output refinement, or system feedback processes, so humans can take the lead in their partnership with AI.

[1] Marc Anderson, Karën Fort. Human Where? A New Scale Defining Human Involvement in Technology Communities from an Ethical Standpoint. International Review of Information Ethics, 2022, 31 (1), ff10.29173/irie477ff. ffhal-03762035f

Enter: Human at the Helm

With a "human at the helm" (HATH) approach, businesses and their employees can thoughtfully steer, review, and act upon AI-generated content to ensure safe, secure, and trustworthy AI.


Human input/instruction
Human feedback or modifications
Human at the helm
Human-reviewed/ approved/modified output or action
Human
Generative AI system
Other systems & humans
Human-curated/ labeled model data
Generated output or action
System/human usage data or feedback

Successful HATH requires:

1. Balancing the degree of human touch, depending on the context. Over-reliance on humans can be time-consuming and inefficient, while under-reliance may lead to suboptimal or harmful outputs.
2. Thoughtful design that ensures the human's role is intuitive, effective, fulfilling, and adds value without creating unnecessary friction.
3. **Training and guidance** to help humans understand the AI system's capabilities, limitations, and how to effectively collaborate with the AI.

Successful HATH enables:

1. **Trust and adoption** of AI systems by providing transparency, accountability, and a sense of control over the technology's outcomes.
2. **Accuracy and quality** of outputs by having humans (with domain expertise and context) review, validate, and edit the results.
3. **Ethical risk mitigation** by supporting humans in spotting bias and ensuring that outputs/decisions align with ethical, legal, and societal norms.
4. **Innovation and continuous improvement** of AI systems by harnessing human feedback and insights to iterate and enhance the algorithms, models, and processes.

The big takeaway

Humans should remain at the helm of generative AI experiences until we've cultivated trust and enabled users to learn the technology. Salesforce Research & Insights conducted a study that found that AI is more trusted and desirable when humans work in close partnership with AI. Participants in the study believed in the potential of AI to increase efficiency, bring inspiration and fulfillment, and improve how they organize their thoughts. But they also reported they aren't ready to fully trust generative AI without an empowered human taking the lead.

What drives users' distrust of generative AI?

1. Fear (and anticipation) of confabulation and errors:

Users, familiar with consumer AI tools, have encountered imperfections such as plagiarism and confabulation. These "oh no" moments have motivated them to more vigilantly review and refine generated outputs.

2. Lack of grounding* and/or confidence in the data source:

Until they see better grounding and better purpose-built solutions, users feel the need to be involved to fact-check and infuse context and expertise.

*Grounding involves using information from trusted knowledge sources to ensure generated responses are are accurate/verifiable and contextualized.

3. Ongoing learning and experimentation:

Users are still learning about the technology and how to best prompt the systems to offer them quality outputs. They do suspect that as they become more experienced in their prompt writing, the need for human review of outputs will decrease.

To move toward autonomous AI, trust must be cultivated by:

1. Consistent track record of accuracy

2. Availability of purpose-built models & grounding

3. Users becoming effective prompt writers

Generative AI that offers thoughtful HATH experiences drives trust today and forms the foundation for autonomous solutions down the line.

> "Generative AI makes humans work better. A human should be in the loop to foster better human value/output. This is what will increase trust in generative AI."
>
> **– End user, high-tech company**

Research approach

Salesforce Research & Insights set out to bring the voice of the user into our conversations on generative AI. We sought to answer: How might Salesforce's generative AI combine the best of human and automated intelligence to create more productive businesses, more empowered people, and more trustworthy AI?

Research Objectives:

1. Understand how users think about the value of a human in the loop.
2. Understand when and why generative AI should involve more or less human touch.
3. Explore how to best–and most inclusively–design for the human in the loop.

1 **Research Wall at Dreamforce**

Dreamforce attendees participated in an interactive experience populating 1K+ use cases for Salesforce generative AI and exploring the role of humans in those use cases.

2 **Deep-dive one-on-one interviews**

Salesforce researchers interviewed Salesforce buyers, enablers, and end users about their experiences with and perspectives toward generative AI.

3 **2-3 week generative AI experiments/homework**

Participants completed activities that involved experimenting with generative AI and reflecting on their professional use cases for AI and the role that a human should or should not play in those use cases.

4 **Focus group sessions to explore use cases and UI patterns for HITL**

Participants plotted their use cases on a continuum from "no human touch" to "high human touch" and gave feedback on text-based HITL design concepts.

Participants:

- Salesforce end users
- Salesforce enablers (admins, developers, architects)
- Salesforce buyers/decision makers

Participants had no experience with Salesforce's generative AI, but were all likely to use, administer, or buy Salesforce generative AI technology in the future.

The research intentionally included diverse participants, representing a mix of industries, company sizes, clouds used, skills and abilities (i.e., neurodivergent, english-as-a-second-language), and perspectives toward and experience with generative AI. The goal was to ensure that we seize the opportunity to design this new interaction paradigm ethically and inclusively from the beginning by accounting for the broad array of experiences and abilities of our customers and users.

Users value human involvement, potentially even long term

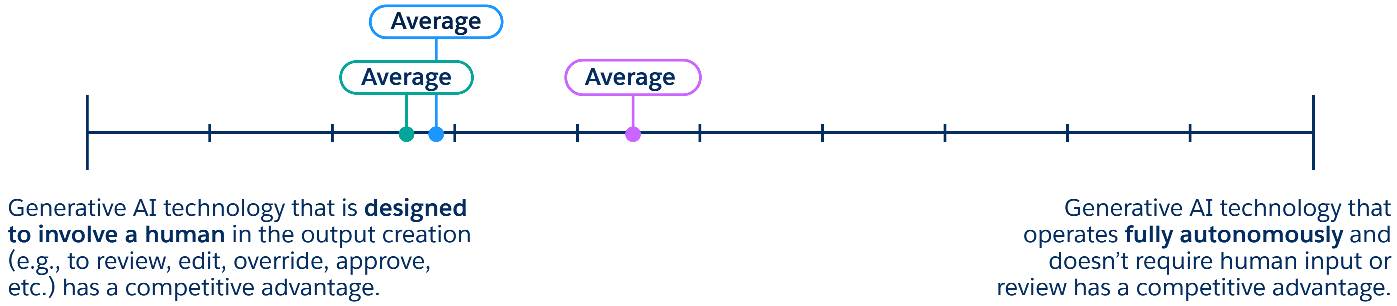
Participants plotted their perspectives on a series of continuums at the study's start and end.

We learned that participants considered thoughtful human involvement an advantage, and trusted outputs produced via human-AI collaboration over autonomously-produced outputs. As participants gained experience with generative AI, their desire for humans to be involved grew slightly, and their trust in generative AI's ability to autonomously produce accurate, quality outputs weakened slightly.

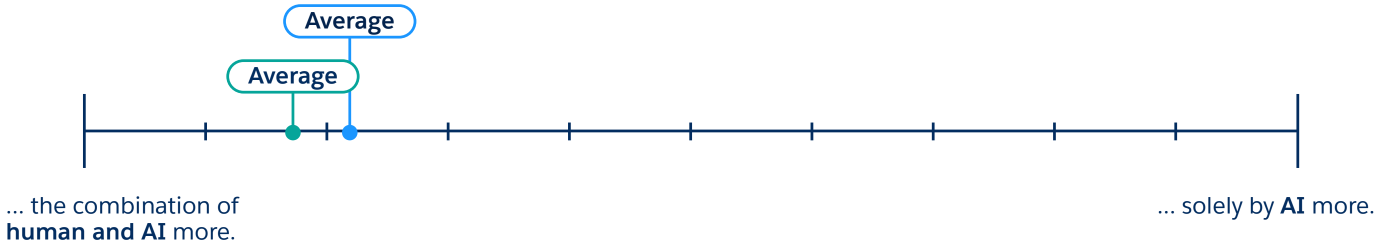
Participants especially valued human involvement in the nearterm but were optimistic about moving toward autonomous AI in the future (though they expected the need for a human to persist in some capacity).

● Beginning of first research session
● After spending three weeks using generative AI* and thinking about its applications for Salesforce-related work
● When question was reframed "in the long run"

*Most were using consumer-grade tools; A few were using purpose-built enterprise tools.

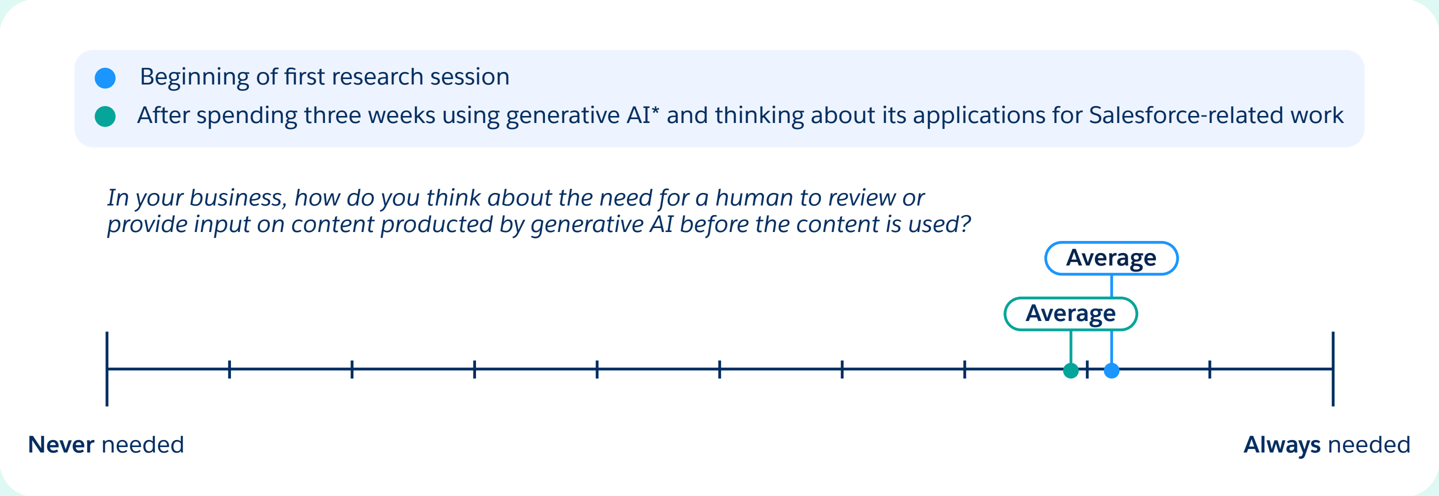


"In my line of work, if there were two pieces of content generated by AI, one that was producted by a human working with AI to create content, and one that was produced by autonomous AI (no human involvement) I would trust the output produced by..."

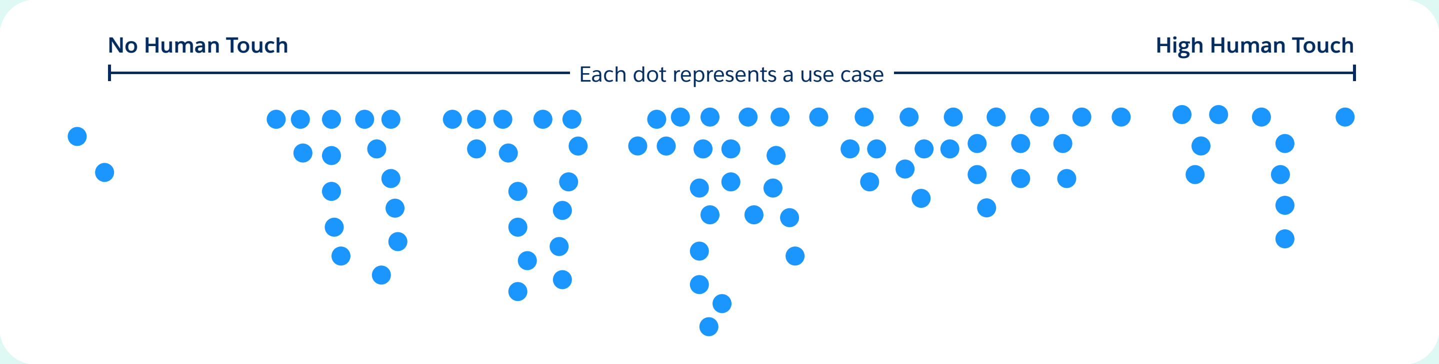


Users believe human touch is necessary for most use cases

As participants gained experience with generative AI, they felt a human in the loop was more often needed. Participants believed that human participation was almost always necessary.



In focus group sessions, participants mapped 90 generative AI use cases to a continuum from "no human touch needed" to "high human touch needed." Participants believed that all but two use cases needed some degree of human touch.



There were no significant variations by industry; all industries had use cases across the spectrum.

Situations that call for more human touch

When the stakes are high

(e.g., involving financial, ethical, legal risk)

For compound tasks

(e.g.., "Do X based on W" or "Do X then do Y")

For macro-level tasks

(e.g., "Define a sales forecasting strategy")

When the data source is unknown or has known gaps

(e.g., trained on data pre-2021)

When the user is still learning how to write effective prompts

(since poor prompts lead to more iterations on outputs)

When the generated output will be used externally/will be customer-facing

When the generated output will be used as the final deliverable

(as opposed to a starting point)

When the generated output is meant to impersonate

(e.g., a sales email)

When the output needs to be bespoke, persuasive, or emotional

(e.g., a marketing campaign)

When the inputs needed are qualitative, nuanced, or implicit

(e.g., a conversation not codified into a data set)

When complete automation contradicts the company's or user's brand

(e.g., a brand built on relationships)

When the work is appropriate for humans and humans enjoy the work

(as opposed to situations where technology is filling a resource or skill gap, or automating rote aspects of the work)

Better Together

If well designed, users believe that generative AI has the potential* to fulfill a lot of promises …

- **Efficiency**
  The ability to do more in less time
- **Fulfillment**
  The ability to focus on the pieces of the work that they enjoy
- **Inspiration**
  The ability to get started, to get unstuck, to see variations or builds
- **Consistency**
  The ability to get the same types of results every time
- **Organization**
  Support in structuring thoughts/ideas

But still, users believe that HUMANS bring a lot of important things like**...

- **Contextual understanding**
- **Subject-matter expertise**
- **Creativity/originality**
- **Human tone/personality**
- **Empathy/emotional intelligence**
- **Ethical/moral judgment**
- **Critical thinking in complex/ ambiguous situations**
- **Real-time or implicit knowledge** (e.g., the latest thinking from yesterday's all-hands call that is not codified into the system)
- **Strategic thinking** (e.g., what does this mean for my business?)

*It is important to note that these are user perceptions and it's unlikely that generative AI will deliver on all of these promises – particularly, the promise of consistency. We need to help users understand what kind of consistency they can and can not expect.

**While participants weren't actively thinking about their role in overall system governance, we (Salesforce) know that AI also needs a human for system governance/maintenance/guidance purposes, to prevent drift and improve results over time.

> “What is the human flavor? I can't explain it, but everyone knows it.... You can tell if someone (during a sales pitch) is interested in you or not. I don't believe AI can detect... intangible things (and) interpersonal dynamics.”
>
> **– IT buyer, high-tech company**

According to participants, today, generative AI can do 60-80% of the work (depending on the task); Humans still carry 20-40% of the load.

Best practices for building generative AI that works better for people

Our user conversations illuminate five best practices for designing generative AI that people trust and want to use.

1. **Trust is a marathon, not a sprint: Cultivate trust over time**
   Trust is not a one-time conversation. Incorporate continuous trust-building experiences to help humans build confidence in your AI technology.

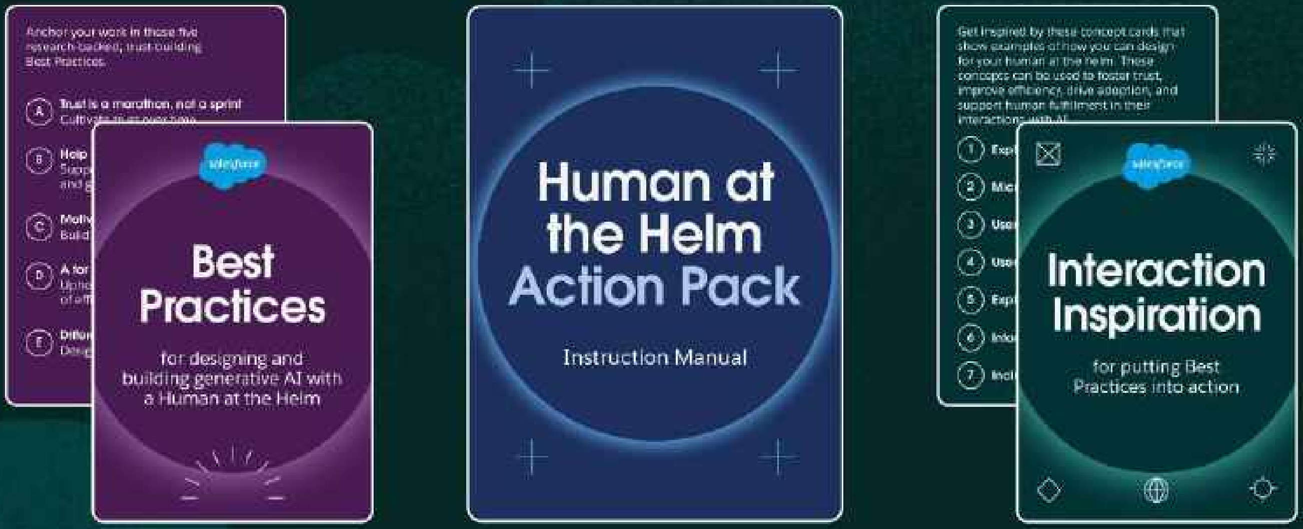
2. **Help me help you: Support reciprocal learning and growth for humans and AI**
   Design the system to help users understand the technology they are working with and ensure users' feedback improves immediate and future outputs.

3. **Motivate over mandate: Build it so people want to use it**
   Users are frustrated by unavoidable and seemingly arbitrary friction (i.e., without proper explanation of the benefits/reasons for the friction). When possible, inspire users to participate, rather than requiring them to do so.

4. **A for efficiency: Uphold the value proposition of efficiency**
   Users will avoid HATH mechanisms (if possible) or abandon use of generative AI systems that introduce friction that undermines the AI's value proposition of efficiency.

5. **Different strokes for different folks: Design inclusively**
   Recognize that different skills, abilities, situations, and use cases call for different modes of interaction. Design for these differences, and/or design for optionality.

Ready to put the Human at the Helm approach into action?


Anchor your work in these five research-backed, trust-building Best Practices.
A Trust is a marathon, not a sprint
salesforce
Best Practices
for designing and building generative AI with a Human at the Helm
Human at the Helm Action Pack
Instruction Manual
Get inspired by these concept cards that show examples of how you can design for your human at the helm. These concepts can be used to foster trust, improve efficiency, drive adoption, and support human fulfillment in their interactions with AI.
salesforce
Interaction Inspiration
for putting Best Practices into action

Put yourself and your users at the helm today!

Salesforce created this set of cards based on our research.
The cards are divided into Best Practices and Interaction Inspiration.

Print the following pages of instructions and cards to get started.

OFFICE OF ETHICAL AND HUMANE USE

Human at the Helm Action Pack

Build trust in AI
with a human touch

salesforce

Welcome to the

Human at the Helm Action Pack

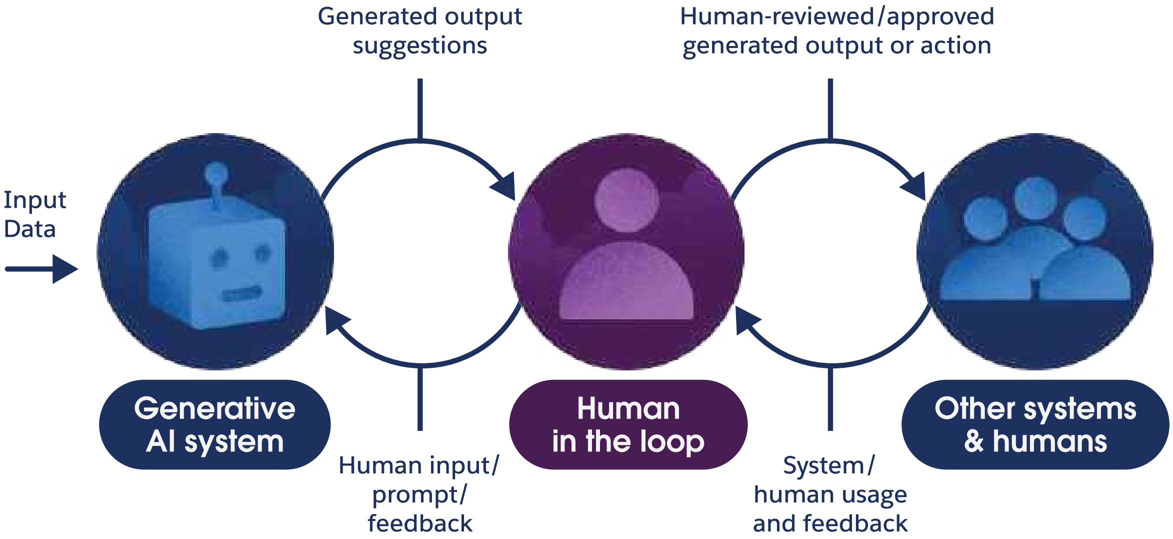
Generative AI has made the move from experimentation to implementation, now touching so many aspects of our work and lives. With this shift, people are increasingly wondering “Can I trust it?”

Salesforce Research & Insights conducted a study that found that **AI is more trusted and desirable when humans work in close partnership with AI**. Users in the study believed in the potential of AI to increase efficiency, bring inspiration and fulfillment, and improve how they organize their thoughts. But they also reported **they aren’t ready to fully trust generative AI without an empowered human taking the lead**.

What does that mean for those responsible for designing how people interact with generative AI in their jobs? Let us share some guidance for waysto support human-generative AI interaction that builds trust, drives adoption, improves efficiency, and empowers people.

What is Human at the Helm (HATH)?
Originally used in the military, nuclear energy, and aviation sectors, the phrase “human in the loop” (HITL) referred to the ability for humans to intervene in automated systems to prevent disasters. Today, HITL has been widely repurposed to talk about how humans are involved in generative AI.

However, Salesforce believes humans should be more than “in the loop.” To cultivate trust over time and build confidence in this still-emergent AI technology, our research has shown generative AI systems should be designed so humans can take the lead in their partnership with AI. Enter: **Human at the Helm**. With a “human at the helm” (HATH) approach, businesses and their employees can thoughtfully steer, review, and act upon AI-generated content to ensure safe, secure, and trustworthy AI.


Generated output suggestions
Human-reviewed/approved generated output or action
Input Data
Generative AI system
Human in the loop
Other systems & humans
Human input/ prompt/ feedback
System/ human usage and feedback

How to use your Human at the Helm Cards

Put yourself and your users at the helm today! Start by reading the **Best Practices Cards** to ground your work and business goals in trust. Next, check out the **Interaction Inspiration Cards** for ideas on how to bring HATH mechanisms into your AI solutions.

Match up **Interaction Inspiration Cards** with **Best Practices Cards** as you make your plans for designing and implementing generative AI solutions for your organization.

A BEST PRACTICES

Trust is a marathon, not a sprint

Trust i
Incorp
experi
confid

A BEST PRACTICES

Cultivate trust over time

- Keep humans at the helm, especially during an initial trust-building/onboarding period where you're allowing users to learn the technology and laying the foundation for autonomous solutions down the line.
- Be transparent and honest about the system's presence and limitations (i.e., that AI is being used, where the information comes from, how confident the system is in the output, how user feedback is used).
- Demonstrate the system's value by guiding the user in providing the right inputs for optimal outputs.
- Rinse and repeat.

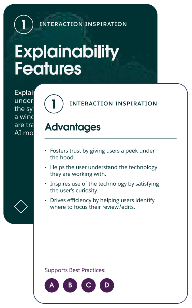
Relevant Interaction Inspiration Cards:

1 2 3 4 6

Best Practices Cards

Anchor your work in these five research-backed, trust-building best practices.

A. Cultivate trust over time
B. Support reciprocal learning and growth for humans and AI
C. Build it so people want to use it
D. Uphold the value proposition of efficiency
E. Design inclusively



1 INTERACTION INSPIRATION

Explainability Features

Explai
under
the sy
a wind
are tra
AI mo

1 INTERACTION INSPIRATION

Advantages

- Fosters trust by giving users a peek under the hood.
- Helps the user understand the technology they are working with.
- Inspires use of the technology by satisfying the user's curiosity.
- Drives efficiency by helping users identify where to focus their review/edits.

Supports Best Practices:

A B C D

Interaction Inspiration Cards

Get inspired by these concept cards that show examples of how you can design for your human at the helm.

Human at the Helm Checklist

With these Best Practices and Interaction Inspirations, we hope you feel empowered and prepared to put your users at the helm.

Ask yourself:

- □ Does the design provide inspiration or foster exploration and experimentation for the user?
- □ Does your design keep the human at the helm of the power paradigm?
- □ Is the system transparent about its limitations, where the information comes from, and/or its confidence in the output?
- □ Does the design motivate or entice the human to engage, rather than require them to engage?
- □ Does the design give users a peek under the hood to satisfy their sense of curiosity?
- □ Does the system afford users the means to refine outputs?
- □ Does the system help the user understand what inputs are needed from them in order to deliver the best output?
- □ Does the design help users identify where to add their expertise or human touch on outputs?
- □ Does the system offer ways for the human to interact that don't require typing?
- □ Does the design prioritize the value proposition of "efficiency"?
- □ Does the design help the user balance their interest in productivity with their new responsibility to ensure safe/accurate/trusted outputs?
- □
- □

Human at the Helm is a product of Salesforce's Office of Ethical & Humane Use and Salesforce's Research & Insights team.

Instructions: Print these pages at Actual Size on letter-sized paper. Use the double-sided setting (make sure your priner is set to “flip along the long side”).

salesforce

Best Practices

for designing and building generative AI with a Human at the Helm

A BEST PRACTICES

Trust is a marathon, not a sprint

Trust is not a one-time conversation. Incorporate continuous trust-building experiences to help humans build confidence in your AI technology.

B BEST PRACTICES

Help me help you

Design the system to help users understand the technology they are working with and ensure users' feedback improves immediate and future outputs.

C BEST PRACTICES

Motivate over mandate

Users are frustrated by unavoidable and seemingly arbitrary friction (i.e., without proper explanation of the benefits/reasons for the friction). When possible, inspire users to participate, rather than requiring them to do so.

A BEST PRACTICES

Cultivate trust over time

- Keep humans at the helm, especially during an initial trust-building/onboarding period where you're allowing users to learn the technology and laying the foundation for autonomous solutions down the line.
- Be transparent and honest about the system's presence and limitations (i.e., that AI is being used, where the information comes from, how confident the system is in the output, how user feedback is used).
- Demonstrate the system's value by guiding the user in providing the right inputs for optimal outputs.
- Rinse and repeat.

Relevant Interaction Inspiration Cards:

1 2 3 4 6

Anchor your work in these five research-backed, trust-building Best Practices.

A **Trust is a marathon, not a sprint**
Cultivate trust over time

B **Help me help you**
Support reciprocal learning and growth for humans and AI

C **Motivate over mandate**
Build it so people want to use it

D **A for efficiency**
Uphold the value proposition of efficiency

E **Different strokes for different folks**
Design inclusively

C BEST PRACTICES

Build it so people want to use it

- Provide inspiration (e.g., give users examples of prompts they could use, give them various outputs to choose from).
- Foster exploration and experimentation (e.g., give users explainability features, let them easily "undo" or "rewind" or adjust course, let them play around).
- Prioritize informative guardrails over restrictive ones (e.g., flag potential risk and empower the user to determine the appropriate course of action).
- Help users understand the value of their participation (i.e., how their input or due diligence shapes the model or improves accuracy, quality, and safety).

Relevant Interaction Inspiration Cards:

1 5 6 7

B BEST PRACTICES

Support reciprocal learning and growth for humans and AI

- Guide users in writing effective prompts.
- Help users identify where to add their expertise or human touch.
- Let users peek under the hood to understand the system.
- Support in-tool editing of outputs. Capture edits with feedback APIs to produce better results over time.
- Offer opportunities for users to share feedback with the system. To encourage users to actually give feedback, ensure that users understand how their feedback is informing the model and what's in it for them.

Relevant Interaction Inspiration Cards:

1 2 3 4 6

Human at the Helm is a product of Salesforce's Office of Ethical & Humane Use and Salesforce's Research & Insights team.

Instructions: Print these pages at Actual Size on letter-sized paper. Use the double-sided setting (make sure your priner is set to "flip along the long side").

Human at the Helm is a product of Salesforce's Office of Ethical & Humane Use and Salesforce's Research & Insights team.

Instructions: Print these pages at Actual Size on letter-sized paper. Use the double-sided setting (make sure your priner is set to "flip along the long side").

D BEST PRACTICES

A for efficiency

Users will avoid HATH mechanisms (if possible) or abandon use of generative AI systems that introduce friction that undermines the AI's value proposition of efficiency.

E BEST PRACTICES

Different strokes for different folks

Recognize that different skills, abilities, situations, and use cases call for different modes of interaction. Design for these differences, and/or design for optionality.

salesforce

Interaction Inspiration

for putting Best Practices into action

1 INTERACTION INSPIRATION

Explainability Features

Explainability features provide understandable explanations of the system's process, allowing users a window into how human inputs are transformed into outputs by the AI model.

E BEST PRACTICES

Design inclusively

- Support various modes of interaction and various styles/formats of outputs so that the user can engage in the ways that work best for them.
- Support user focus through features like in-app guidance and tooltips, and focus on creating a guided prompt experience to reduce iterations.
- Mitigate language challenges by introducing re-generation and feedback mechanisms that don't involve open text fields. Include features like spell check.

Relevant Interaction Inspiration Cards:

4 5 7

D BEST PRACTICES

Uphold the value proposition of efficiency

- Help users write better prompts to reduce the need to regenerate the output.
- Help users quickly identify where they can add their expertise or where they might want to focus their review.
- Use sliders, dropdowns with binary options, and other ways to suggest edits efficiently.
- Don't make users copy/paste into Word/Quip/Google Docs to finesse. Let them make their micro-edits in the generative AI tool.
- Help the user balance their interest in productivity with their new responsibility to ensure safe/accurate/trusted outputs.

Relevant Interaction Inspiration Cards:

1 2 4 5 7

1 INTERACTION INSPIRATION

Advantages

- Fosters trust by giving users a peek under the hood.
- Helps the user understand the technology they are working with.
- Inspires use of the technology by satisfying the user's curiosity.
- Drives efficiency by helping users identify where to focus their review/edits.

Supports Best Practices:

A B C D

Get inspired by these concept cards that show examples of how you can design for your human at the helm. These concepts can be used to foster trust, improve efficiency, drive adoption, and support human fulfillment in their interactions with AI.

1 **Explainability Features**

2 **Micro-Edits**

3 **User Feedback**

4 **User Guidance**

5 **Exploration**

6 **Informative Guardrails**

7 **Inclusive Building Blocks**

Human at the Helm is a product of Salesforce's Office of Ethical & Humane Use and Salesforce's Research & Insights team.

Instructions: Print these pages at Actual Size on letter-sized paper. Use the double-sided setting (make sure your priner is set to "flip along the long side").

Human at the Helm is a product of Salesforce's Office of Ethical & Humane Use and Salesforce's Research & Insights team.

Instructions: Print these pages at Actual Size on letter-sized paper. Use the double-sided setting (make sure your priner is set to "flip along the long side").

1.1 INTERACTION INSPIRATION > EXPLAINABILITY FEATURES

AI Confidence Indicator

1.2 INTERACTION INSPIRATION > EXPLAINABILITY FEATURES

Interactive Explanations & Cited Sources

2 INTERACTION INSPIRATION

Micro-Edits

Micro-editing involves making detailed adjustments (e.g., refining details, adjusting tone/style, correcting errors) to generated content, tailoring the output to better meet the intended purpose or audience.

2.1 INTERACTION INSPIRATION > MICRO-EDITS

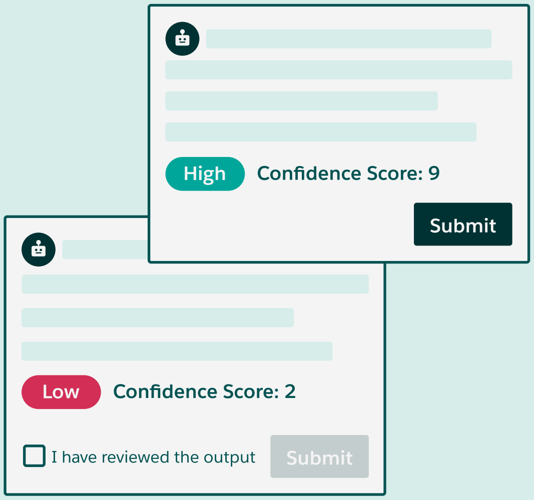
Critique Mode

Human at the Helm is a product of Salesforce's Office of Ethical & Humane Use and Salesforce's Research & Insights team.

Instructions: Print these pages at Actual Size on letter-sized paper. Use the double-sided setting (make sure your priner is set to "flip along the long side").

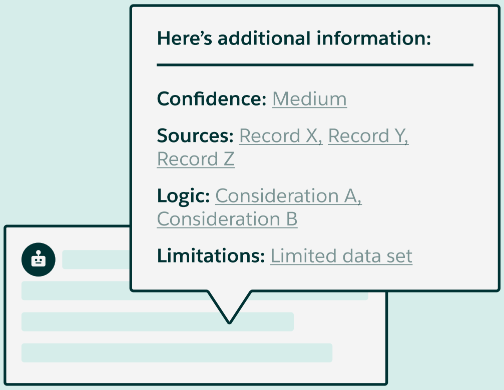
1.1 INTERACTION INSPIRATION > EXPLAINABILITY FEATURES

User gets a score alongside generated content, indicating the AI's confidence about the output. If the output falls below the admin-set threshold, the system requires human review.



1.2 INTERACTION INSPIRATION > EXPLAINABILITY FEATURES

User can interact with explanations of how the content was derived (e.g., sources referenced, explanations of confidence, explanation of system limitations, inputs considered in decision-making).



Here's additional information:

Confidence: Medium

Sources: Record X, Record Y, Record Z

Logic: Consideration A, Consideration B

Limitations: Limited data set

2 INTERACTION INSPIRATION

Advantages

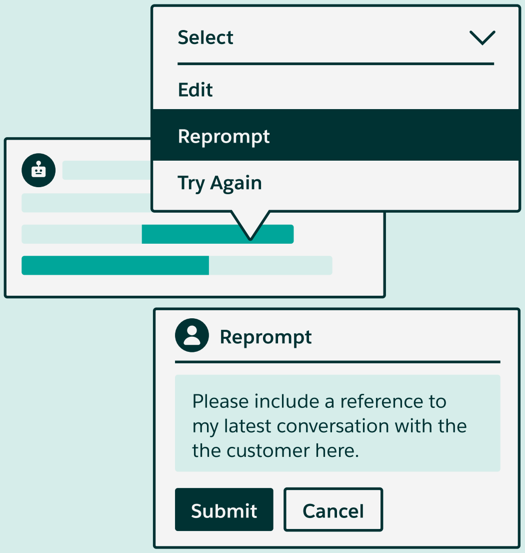
- Fosters trust by giving the human control over the outputs.
- Supports machine learning which improves output accuracy and quality over time, in turn fostering trust.
- Drives efficiency by reducing the number of tools needed to revise outputs.

Supports Best Practices:

A B D

2.1 INTERACTION INSPIRATION > MICRO-EDITS

User can highlight and regenerate specific words or phrases in an output rather than regenerating an entire output. Edits are captured by feedback APIs to improve output quality and accuracy over time.



Please include a reference to my latest conversation with the the customer here.

Human at the Helm is a product of Salesforce's Office of Ethical & Humane Use and Salesforce's Research & Insights team.

Instructions: Print these pages at Actual Size on letter-sized paper. Use the double-sided setting (make sure your priner is set to "flip along the long side").

2.2 INTERACTION INSPIRATION > MICRO-EDITS

Direct Edit

3 INTERACTION INSPIRATION

User Feedback

User feedback features enable users to share feedback on the quality, relevance, and accuracy of generated content, aiding in the continuous improvement of the AI model's performance and effectiveness.

3.1 INTERACTION INSPIRATION > USER FEEDBACK

Granular Thumbs Up/ Down

3.2 INTERACTION INSPIRATION > USER FEEDBACK

"Give Feedback" Module

3 INTERACTION INSPIRATION

Advantages

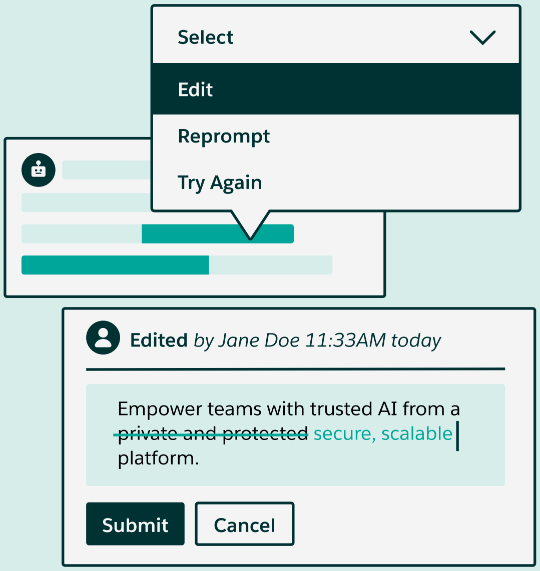
- Supports machine learning which improves output accuracy and quality over time, in turn fostering trust.
- Can afford the user the opportunity to share feedback in a way that doesn't require typing.

Supports Best Practices:

A B

2.2 INTERACTION INSPIRATION > MICRO-EDITS

User can enter “edit mode” to directly edit the generated text. Edits are visually highlighted, similar to Google Docs “suggesting” mode. Edits are captured by feedback APIs.

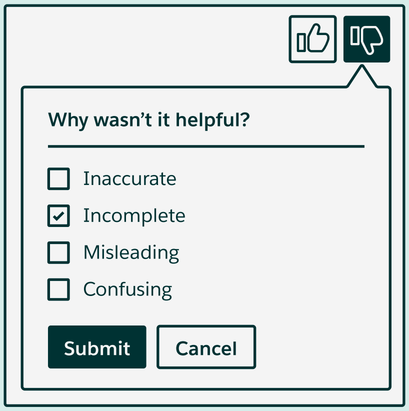


Edited *by Jane Doe 11:33AM today*

Empower teams with trusted AI from a ~~private and protected~~ secure, scalable platform.

3.2 INTERACTION INSPIRATION > USER FEEDBACK

User can select from options to specify the reasons for their thumbs up/down feedback.



Why wasn’t it helpful?

☐ Inaccurate

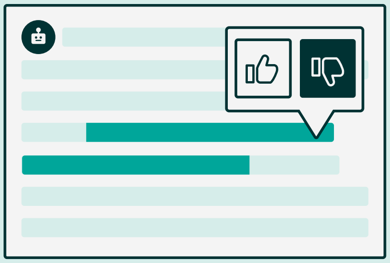
☑ Incomplete

☐ Misleading

☐ Confusing

3.1 INTERACTION INSPIRATION > USER FEEDBACK

User can highlight a portion of the generated content to thumbs up/down rather than giving a thumbs up/down to an entire body of text.



Human at the Helm is a product of Salesforce’s Office of Ethical & Humane Use and Salesforce’s Research & Insights team.

Instructions: Print these pages at Actual Size on letter-sized paper. Use the double-sided setting (make sure your priner is set to “flip along the long side”).

Instructions: Print these pages at Actual Size on letter-sized paper. Use the double-sided setting (make sure your priner is set to "flip along the long side").

4 INTERACTION INSPIRATION

User Guidance

User guidance features give users suggestions on how to optimize their generative AI interactions to achieve optimal outputs.

4.1 INTERACTION INSPIRATION > USER GUIDANCE

Prompt Error & Guidance

4.2 INTERACTION INSPIRATION > USER GUIDANCE

Fill-in-the-Blanks

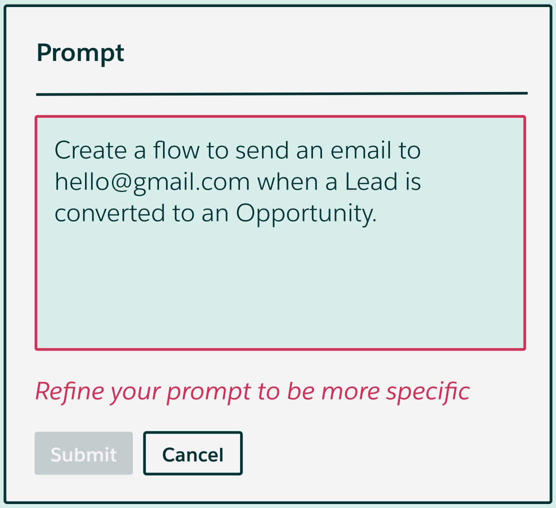
5 INTERACTION INSPIRATION

Exploration

Explorative features facilitate experimentation and discovery by giving users starting points to ignite creativity, diverse suggestions to explore, and mechanisms to dynamically experiment with how different inputs influence outputs.

4.1 INTERACTION INSPIRATION > USER GUIDANCE

If a prompt is not well-written or specific enough, the user is nudged to refine their prompt accordingly.



Prompt

Create a flow to send an email to hello@gmail.com when a Lead is converted to an Opportunity.

Refine your prompt to be more specific

4 INTERACTION INSPIRATION

Advantages

- Fosters trust by demonstrating what the system needs from the user in order to offer an optimal output.
- Helps users learn how to better collaborate with the technology.
- Provides guidance to reduce guesswork for the user, offering a more focused experience.

Supports Best Practices:

A B D E

5 INTERACTION INSPIRATION

Advantages

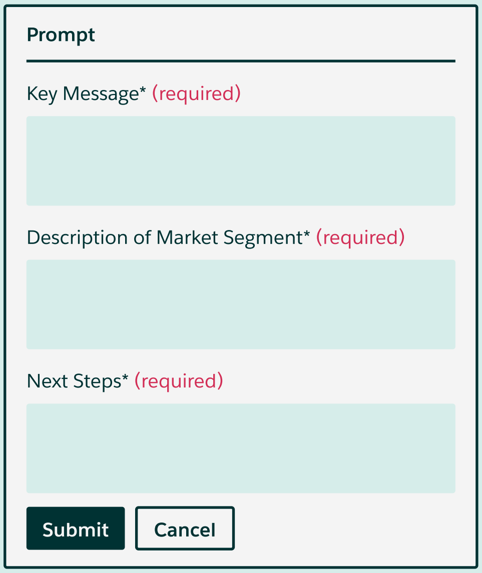
- Satisfies users' curiosity, inspiring them, and fueling their creativity by allowing easy experimentation.
- Helps users reach their end goals faster by giving them suggested paths to explore.
- Supports diverse users by offering options to finesse outputs that don't require typing.

Supports Best Practices:

C D E

4.2 INTERACTION INSPIRATION > USER GUIDANCE

User must fill out required prompt fields before the output can be generated.



Prompt

Key Message* (required)

Description of Market Segment* (required)

Next Steps* (required)

Human at the Helm is a product of Salesforce's Office of Ethical & Humane Use and Salesforce's Research & Insights team.

Instructions: Print these pages at Actual Size on letter-sized paper. Use the double-sided setting (make sure your priner is set to "flip along the long side").

Human at the Helm is a product of Salesforce's Office of Ethical & Humane Use and Salesforce's Research & Insights team.

Instructions: Print these pages at Actual Size on letter-sized paper. Use the double-sided setting (make sure your priner is set to "flip along the long side").

5.1 INTERACTION INSPIRATION > EXPLORATION

Suggest Alternatives

5.2 INTERACTION INSPIRATION > EXPLORATION

Slider Controls

6 INTERACTION INSPIRATION

Informative Guardrails

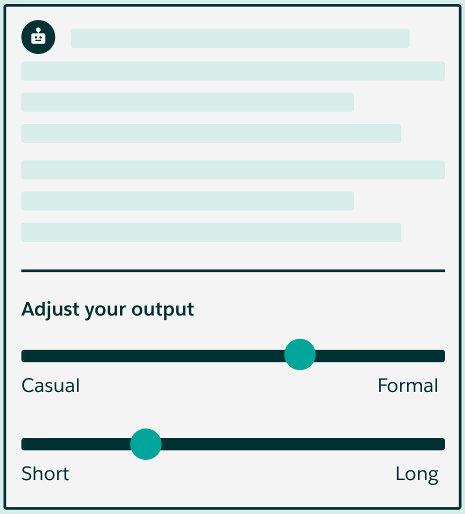
Informative guardrails are proactive measures that alert users to potential risks such as legal liabilities or ethical concerns, empowering them to make informed decisions regarding their output's appropriateness and compliance.

6.1 INTERACTION INSPIRATION > INFORMATIVE GUARDRAILS

Risk Score

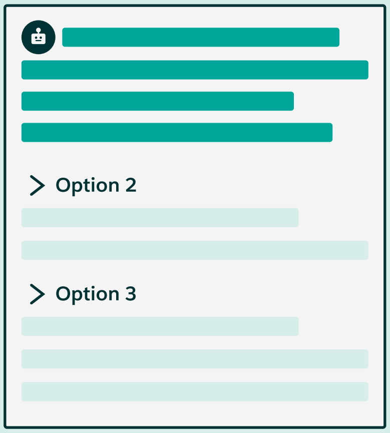
5.2 INTERACTION INSPIRATION > EXPLORATION

User can adjust sliders to fine-tune certain parameters of the output.



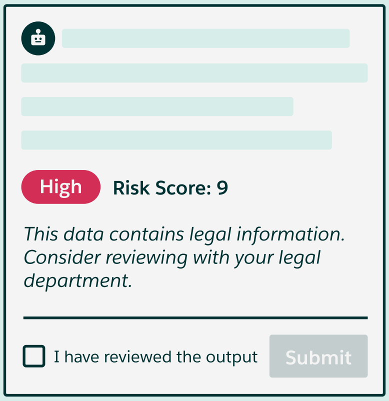
5.1 INTERACTION INSPIRATION > EXPLORATION

User can cycle through suggested alternatives to find the most suitable replacement for all or a portion of the generated content.



6.1 INTERACTION INSPIRATION > INFORMATIVE GUARDRAILS

User gets an indication when content might be higher risk (e.g., when there could be legal, financial, or ethical implications) via a "risk score" with suggested courses of action.



This data contains legal information. Consider reviewing with your legal department.

□ I have reviewed the output

6 INTERACTION INSPIRATION

Advantages

- Fosters trust by being transparent about the system's limitations.
- Fosters trust by empowering users to make the ultimate decision about how to handle a situation.
- Supports learning by giving users information about potential risks.
- Motivates users to play an active role in AI systems by flagging areas that need special human attention.
- Adds a risk management layer for the company by empowering the user to make informed decisions about how to handle delicate situations.

Supports Best Practices:

A B C

Human at the Helm is a product of Salesforce's Office of Ethical & Humane Use and Salesforce's Research & Insights team.

Instructions: Print these pages at Actual Size on letter-sized paper. Use the double-sided setting (make sure your priner is set to "flip along the long side").

Human at the Helm is a product of Salesforce's Office of Ethical & Humane Use and Salesforce's Research & Insights team.

Instructions: Print these pages at Actual Size on letter-sized paper. Use the double-sided setting (make sure your priner is set to “flip along the long side”).

6.2 INTERACTION INSPIRATION > INFORMATIVE GUARDRAILS

Adaptive Disclaimer

7 INTERACTION INSPIRATION

Inclusive Building Blocks

Inclusive Building Blocks provide tools and functionalities that accommodate diverse skills, abilities, and scenarios, empowering users to engage comfortably and successfully.

7.1 INTERACTION INSPIRATION > INCLUSIVE BUILDING BLOCKS

Spell Check

7.2 INTERACTION INSPIRATION > INCLUSIVE BUILDING BLOCKS

Multi-Modality

7 INTERACTION INSPIRATION

Advantages

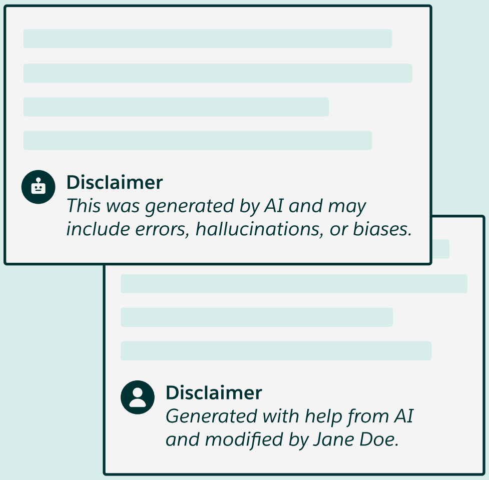
- Promotes use by supporting flexible modes of interaction so that diverse users can engage in the way that works best for them.
- Drives efficiency by ensuring users' intent is accurately interpreted by the system, limiting iterations that need to happen.

Supports Best Practices:

C D E

6.2 INTERACTION INSPIRATION > INFORMATIVE GUARDRAILS

A default disclaimer cautions that generated content may include errors, hallucinations, or bias. As the user edits the output, the disclaimer dynamically changes to indicate that there has been human review.

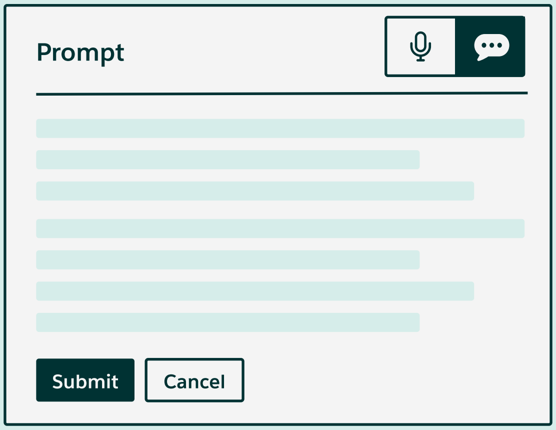


Disclaimer
This was generated by AI and may include errors, hallucinations, or biases.

Disclaimer
Generated with help from AI and modified by Jane Doe.

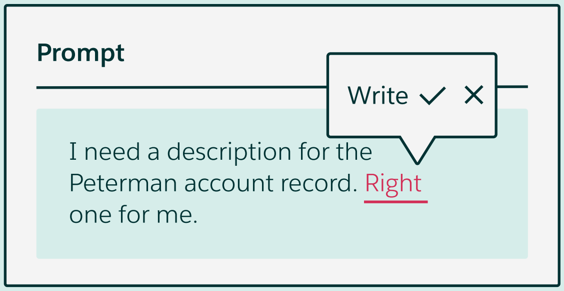
7.2 INTERACTION INSPIRATION > INCLUSIVE BUILDING BLOCKS

Users can seamlessly switch between voice and text inputs for greater flexibility and accessibility.



7.1 INTERACTION INSPIRATION > INCLUSIVE BUILDING BLOCKS

User is alerted to spelling or grammatical errors within a prompt and can correct the errors so that outputs are clearer and more closely aligned to user intent.



I need a description for the Peterman account record. Right one for me.

Human at the Helm is a product of Salesforce's Office of Ethical & Humane Use and Salesforce's Research & Insights team.

Instructions: Print these pages at Actual Size on letter-sized paper. Use the double-sided setting (make sure your priner is set to "flip along the long side").

For more on
Human at the Helm
scan the
QR code below