Designerly Understanding: Information Needs for Model Transparency to Support Design Ideation for AI-Powered User Experience

Q. Vera Liao
veraliao@microsoft.com
Microsoft Research
Montreal, Canada

Hariharan Subramonyam
harihars@stanford.edu
Stanford University
Stanford, USA

Jennifer Wang
jennifer.wang@microsoft.com
Microsoft
Redmond, USA

Jennifer Wortman Vaughan
jenn@microsoft.com
Microsoft Research
New York, USA

ABSTRACT

Despite the widespread use of artificial intelligence (AI), designing user experiences (UX) for AI-powered systems remains challenging. UX designers face hurdles understanding AI technologies, such as pre-trained language models, as design materials. This limits their ability to ideate and make decisions about whether, where, and how to use AI. To address this problem, we bridge the literature on AI design and AI transparency to explore whether and how frameworks for transparent model reporting can support design ideation with pre-trained models. By interviewing 23 UX practitioners, we find that practitioners frequently work with pre-trained models, but lack support for UX-led ideation. Through a scenario-based design task, we identify common goals that designers seek model understanding for and pinpoint their model transparency information needs. Our study highlights the pivotal role that UX designers can play in Responsible AI and calls for supporting their understanding of AI limitations through model transparency and interrogation.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; • **Computing methodologies → Artificial intelligence**.

KEYWORDS

AI design, AI transparency, AI documentation, explainability, pre-trained models

ACM Reference Format:
Q. Vera Liao, Hariharan Subramonyam, Jennifer Wang, and Jennifer Wortman Vaughan. 2023. Designerly Understanding: Information Needs for Model Transparency to Support Design Ideation for AI-Powered User Experience. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3544548.3580652

CHI '23, April 23–28, 2023, Hamburg, Germany
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3580652

1 INTRODUCTION

The use of AI technologies has become widespread, from novel systems like machine translators built entirely around a machine learning (ML) model, to advanced features like text auto-completion built into already commonplace applications. Advances in AI are driven by research and development efforts producing models with various capabilities, but often disconnected from specific applications or user needs. For example, AI service platforms [1, 2, 6] such as Microsoft Azure [7] and HuggingFace [4] host a growing collection of *pre-trained* models with capabilities in language, vision, audio, and more. There is also a recent trend of developing large pre-trained models, such as the large language model GPT-3 [22] or the multimodal model Dall-E [83]. Fully realizing the potential of these new AI technologies requires discovering applications where they can be used to solve user problems and aligning their behavior with user preferences. While these tasks are often an essential part of UX designers' jobs, recent research shows that practitioners grapple with challenges when using "AI as a design material" [39, 54, 102, 103]. These challenges discourage the prioritization of UX, leading to failures of AI-driven products and unintended individual and societal consequences.

Among other challenges, the design ideation process is often hindered by struggles to understand the AI technologies due to their complexities and expertise barriers [39, 103]. However, effective design ideation does not necessarily require a deep technical understanding of the technology, but rather a "designerly understanding" [100, 102]. What does a designerly understanding of AI involve? Prior work defined it as the ability to link an AI technology's capabilities to ways of generating value for users [102]. Having a good understanding of the design material can enable designers to take UX-led approaches to AI product innovation that prioritize value to users, mitigate potential harms, and better align with the goal of responsible AI (RAI). However, prior work reported a lack of means to support a good designerly understanding of AI [39, 103], and to begin with, a lack of knowledge on what designers need for such support.

Meanwhile, we recognize that supporting an understanding of AI has long been the goal of research on AI transparency. Besides producing explainable AI (XAI) techniques to illuminate the technical details of models [49], the community is moving towards standardized approaches to transparent reporting with *AI documentation frameworks* such as model cards [78], datasheets [45], and AI service factsheets [12]. These frameworks are often motivated from the perspective of RAI—to help both practitioners and end users evaluate the suitability of the model or dataset for their products or contexts. This suitability assessment must be supported through means of understanding caveats such as unintended use cases, limitations, and potential pitfalls, in addition to basic details such as model inputs and outputs.

With the rise of pre-trained models, which lower the barrier to AI for practitioners, transparent model reporting is all the more critical. But while these services often target engineers, in practice, it is questionable whether these are, or *should* be, the only roles driving suitability assessment and ideation. With the movement to "democratize AI," it is equally important to lower the barrier to ideating on whether, where, and how to use models appropriately and responsibly. By reducing technical investment overhead, the availability of increasingly powerful pre-trained models for product development may create both more opportunities and more responsibility for UX designers to drive innovation.

In this work, we set out to explore how to support design ideation around the use of pre-trained models, focusing on enabling a designerly understanding through model transparency. Specifically, we introduce a hands-on scenario-based design task and leverage an example of model documentation as a design probe [58]. We explore the utility and gaps of the documentation's comprehensive categories of information to pinpoint designers' information needs for understanding a model to perform design ideation.

Our study takes two particular stances to inform future work supporting UX designers to work with AI. First, we prioritize RAI practices that proactively mitigate potential harms of AI technologies during their development and explore designers' role in RAI. Therefore, our study protocol emphasizes investigating how designers use critical information such as the model's limitations to engage in responsible design ideation. Second, to enable a designerly understanding of AI, we draw on the goal-oriented stance in human-centered approaches to studying explainable and transparent AI [68, 93, 97], recognizing that understanding is a means to an end [63, 72], and effective transparency support must be developed according to the end goals. Our analysis distills four common goals that designers seek model understanding for, which future work should aim to support.

In short, our work makes the following contributions:

- *Identifying new challenges in AI UX design practices*: Our interviews reveal that practitioners frequently work with pre-trained models and that new challenges arise when understanding and designing with these models. Echoing findings in prior work, there is a lack of support for UX-led approaches to product ideation, which is especially critical for the responsible use of pre-trained models.
- *Bridging AI design and AI transparency*: We explore using transparent model reporting frameworks to support design ideation with a pre-trained model. While our study provides evidence of their utility, it also reveals significant gaps and calls for moving beyond static documentation to supporting model interrogation.
- *Identifying four common goals that designers seek out model understanding for and how to support them*: These goals are to engage in divergent-convergent design thinking and eliminate risky design ideas; to create "conditional designs" to mitigate AI's varying impact for different user scenarios; to provide AI transparency to end users; and to negotiate and collaborate with their team to advocate for users. We pinpoint designers' model information needs for each of these goals, and suggest design guidelines to support them. These common goals also highlight the pivotal role that UX designers can play in RAI with an effective understanding of model limitations.

Below we start by reviewing related work that informed our study, then present our methods and findings. We conclude with a discussion of implications for research and practice.

2 BACKGROUND AND RESEARCH QUESTIONS

2.1 Challenges of AI as a Design Material

Researchers have investigated the challenges for UX practitioners to work with "AI as a design material," including the complexity of the material itself [39, 92, 102, 103]. Yang et al. [103] summarize two sources of AI's distinctive design challenges: 1) the uncertainty surrounding its capabilities, with expansive and evolving algorithmic possibilities; and 2) AI's output complexity, stemming from its probabilistic and adaptive nature. Subramonyam et al. [92] contend that, because of the complications of developing models, including choosing from different models, AI does not lend itself to the deterministic "material" perspective that designers are used to when working with unfamiliar technologies [42, 46, 85], but instead has its material properties *emergent* from envisioned designs.

Besides making existing UX methodologies (e.g., prototyping and user testing) challenging [91, 101], these materialistic complexities give rise to pressing challenges in understanding AI [103]. These challenges are exacerbated by disciplinary barriers and a lack of support for gaining AI literacy [39, 66, 73]. Interestingly, an interview study with experienced AI designers [102] suggests that design is not necessarily hindered by a lack of technical knowledge, but supported by a "designerly understanding" of the technology, often approached through designerly abstractions (e.g., describing its capabilities in relation to user utility) and design exemplars.

The struggle to understand AI can hinder design ideation, causing designers to fail to recognize "low-hanging fruits" to use AI to solve user problems, grapple with envisioning novel uses of AI, or inadvertently attempt uses that exceed technical feasibility [39, 103]. A survey study published in 2017 [39] reported that UX designers were rarely involved in the feature planning stage for AI-powered products, but were limited to working on UI designs. In contrast, a recent study [104] with experienced enterprise AI designers suggests that they do engage in defining new systems and processes. These engagements require not only understanding the AI's capabilities and conceptualizing how a design idea would add value, but also viability positioning that justifies the use of AI through expected return on investment.

We recognize that there are also systemic challenges. While the emergent properties of AI call for UX-led approaches to shape technological choices [90, 103, 104], individuals may face upstream battles challenging the current software engineering workflows [90, 95, 101], defeating constraints on time, resources, and incentives [31, 62, 102], and overcoming disciplinary and organizational barriers [66, 104]. While our work seeks to empower designers to drive ideation by supporting their understanding of AI, this goal cannot be achieved without also tackling the organizational challenges.

2.2 Information Support for AI Design

A small but growing area of work on supporting AI designers to overcome the above-mentioned challenges has emerged. Prior research produced tooling to support AI prototyping [76, 91], new design processes [40, 46, 62, 67, 92], and boundary objects [25, 66, 92] to facilitate collaboration between designers and data scientists. Our work is most directly informed by related work that focuses on providing information and knowledge support for AI design.

Research and industry have produced numerous taxonomies and guidelines to sensitize designers to both AI capabilities [34, 66, 104] and the AI design space (e.g., [3, 5, 11]). However, support for designers to seek information about specific models they work with remains under-explored. A small number of tools have been developed to help designers understand certain aspects of a model, such as performance metrics [57, 105]. Others explored approaches to guide designers in envisioning solutions when working with a model. For example, Hong et al. [56] developed an NLP playbook to encourage systematic consideration of errors, based on common failures of NLP models. Subramonyam et al. [92] suggest the use of "data probes," example data points and their model outputs, to facilitate design thinking and validation. Similarly, to enable exploring GPT-3's promises for interaction design, Lee et al. [65] created a dataset of instances from writers working with GPT-3. Using examples to support understanding aligns with the "material" design perspective, as understanding of materialistic properties can be achieved by experienced affordances [37, 54, 87].

Also under-explored is investigation into the actual information-seeking processes when designing with AI, except for some first-person account of the challenges [16, 101]. A relevant work is Subramonyam et al. [90]. By interviewing practitioners, the authors investigated how designers and data scientists overcome expertise boundaries by sharing information through low-level details. Engineers share with designers information about the data used to train the model through dataset documentation and other means, and about model behaviors through example outputs, performance dashboards, demos, and explanations such as feature weights, rules, and underlying assumptions.

Like most prior research in this area, Subramonyam et al. [90] focuses on cases where UX practitioners work with the data scientists who develop the model. We instead investigate designers' information needs when working with *pre-trained* models, which can be third-party models or models handed over after completion, where access to the model developers is unavailable or limited. We also take an ecological position of human-information interaction research [43] that people's information needs are best understood by observing tasks being performed. We hence create a scenario-based design task to investigate designers' needs when designing with an unfamiliar pre-trained model

2.3 AI Transparency

To support a designerly understanding of AI, we draw from the literature on AI transparency. To facilitate effective and consistent AI transparency, the AI research community has proposed various frameworks for transparent reporting of data [15, 45, 53], models [78], and services [12], often broadly referred to as "AI documentation." These frameworks include standardized categories of information—such as "performance metrics, intended use cases, and potential pitfalls" [78]—and guidelines to help AI creators transparently communicate the capabilities and limitations of their models or data. Aiming to support "responsible democratization of AI" [78], AI documentation is intended to support evaluating the suitability of a dataset or model for one's use case, and facilitate accountability and governance. These frameworks are increasingly adopted in industry, especially for third-party AI services. For example, model cards [78] have been implemented by Google Cloud [2] and Hugging Face [96], and Microsoft introduced "Transparency Notes" for its Azure Cognitive Services [8].

While some researchers have explored the needs of practitioners creating AI documentation [50, 51], empirical studies investigating its use is relatively scarce. Through a think-aloud protocol, Boyd [20] demonstrates that thoughtfully constructed datasheets can help ML engineers understand and make decisions about ethical problems in training data. A recent study by Crisan et al. [32] contends that current AI documentation primarily serves people with ML expertise, while non-experts can benefit from interactive interrogation of an expanded form of documentation. Through a user-centered design study, the authors create a prototype of an interactive model card and make design recommendations, including considering information hierarchies and prioritizing critical information to promote productive skepticism.

Another cornerstone of AI transparency is to support understanding of model behaviors through AI explanations, actively studied in the field of explainable AI (XAI) (e.g., [48, 49, 70]). XAI techniques typically address user questions such as "how does the model make decisions?" or "why does the model make this particular decision?" by revealing the features used by the model, how these features are weighted, or the rules that the model follows. Recent studies report that XAI techniques are increasingly used in industry practice as end user-facing features [66], by data scientists to debug models [52, 61, 68], and shared with stakeholders to verify the models [18, 57]. However, it is unclear if designers utilize—or even have the need for—such technical explanations.

Our study is also motivated by the goal-oriented stance in research that takes human-centered perspectives on explainable AI [68, 93, 97]. Rather than focusing on what aspects of the model can be made transparent, this stance prioritizes articulating the goals that people seek out model understanding for, and centers the development and evaluation of explainability methods around these goals. While several taxonomies of common goals of XAI have been proposed [27, 69, 93], other works empirically investigated the goals of a specific group of people. For example, by interviewing data scientists and ML engineers, Hong et al. [57] identified the explainability goals of ML practitioners to be model improvement, knowledge discovery, and gaining confidence.

We set out to identify designers' *transparency goals* during design ideation to unpack the requirements to support designerly understanding of AI. We introduce model documentation as a design probe [58]. That is, besides the "engineering goal" of testing the viability of using documentation to support AI design ideation, we are interested in the "social science goal" of understanding designers' needs during ideation and the "design goal" of inspiring new approaches to supporting model understanding.

2.4 Responsible AI (RAI) in Practice

Our work also aims to contribute to RAI practices by exploring designers' role in responsibly building AI technologies. RAI is concerned with putting theoretical principles of AI ethics into practice, and proactively mitigating individual and societal harms from AI [17, 82, 88]. Recent years have seen a growing interest in studying practitioners' practices, challenges, and gaps in dealing with RAI issues, such as fairness [35, 55, 74, 75, 82], transparency [18, 57, 66], and accountability [21, 81]. These challenges are multi-faceted, ranging from individuals' lack of knowledge support and technical means, to socio-organizational barriers such as lacking incentives and enabling internal structures.

However, UX designers are not always included in these studies of RAI practitioners. A gap seems to exist between many works advocating for collaboration between ML engineers and designers to create good AI UX, and advocating for designers' role in mitigating potential harms of AI. Meanwhile, recent work recognizes that RAI is fundamentally about serving stakeholders' needs and values [33, 75, 89], a position that is central to the deliverable and methodological toolbox of UX practitioners. Studies of enterprise designers also suggest that designers are deeply concerned about RAI issues such as fairness, transparency, safety, privacy, and data use [66, 104, 106].

To explore designers' role in RAI and how to support such a role, our study emphasizes the need for supporting a designerly understanding of a model's limitations, including failures, biases, and potential harms. This emphasis on both capabilities and limitations also aligns with the intent and design of AI documentation [12, 78]

In summary, drawing on these prior works, our study is guided by the following research questions:

- **RQ1**: What are UX practitioners' needs and challenges in understanding and working with pre-trained models, particularly to perform responsible ideation? (Sections 4.1–4.3)
- **RQ2**: To what extent are current model documentation frameworks useful for supporting design ideation and what are the gaps? (Sections 4.2–4.3)
- **RQ3**: What are the goals that designers seek model understanding for and how can they be supported? (Section 4.3)

3 METHOD

We conducted interviews consisting of a hands-on ideation task using a pre-trained model to solve a user problem, and discussions around participants' own experiences performing design ideation for AI. In the sections below, we first describe the ideation task, then the procedure, participants, and analysis of the interviews.

3.1 Scenario-Based Design Ideation Task and Artifacts Provided

We aimed to create an ideation task that could be completed in 30 minutes and generate rich discussions. We therefore chose a user problem scenario that is easily accessible, but has a complex solution space, with multiplex user flows. In the scenario we chose, users of an online microblogging platform share online articles without understanding them or helping their followers understand them, leading to the spread of misinformation. We included a Twitter UI (with minor adaptations, such as changing the brand name) in the task introduction to invoke participants' knowledge about microblogging platforms. The company running the microblogging platform in the scenario had already paid for an AI service which includes a pre-trained *text summarization model.* Participants were asked to act as if they work for the company and try to *come up with a new feature of the microblogging platform that takes advantage of this available model to solve the article misunderstanding problem.*

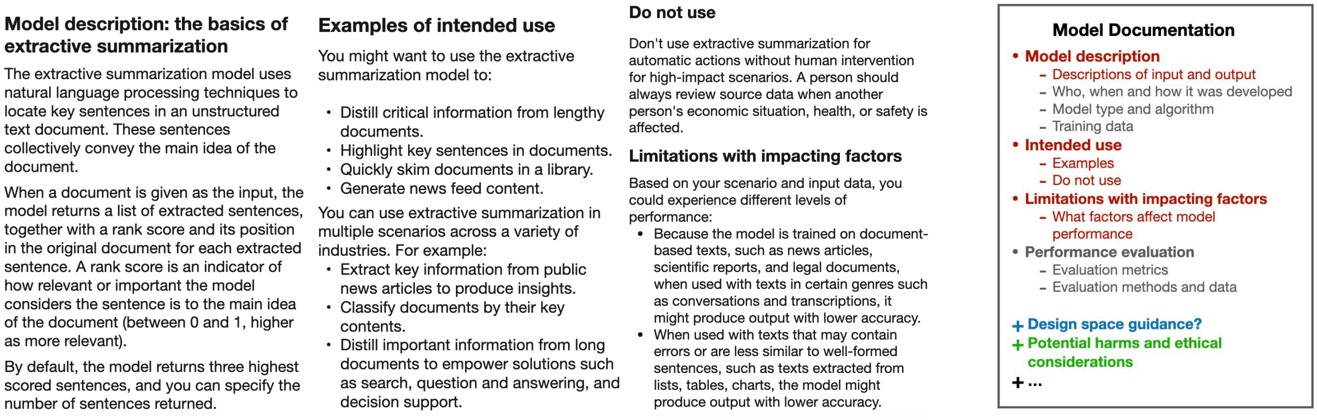
We chose to base the model on the extractive text summarization model provided by Microsoft Azure Cognitive Service[1] because it is a popular AI service with comprehensive documentation. Two artifacts were provided to help participants to understand the model: a modified version of the **model documentation** from the service (a Transparency Note), and 20 curated **model input-output examples** (described below). The documentation covers the major components specified in model reporting frameworks like model cards [78], including a model description, examples of intended uses, warnings against unintended uses, and limitations highlighting impacting factors (i.e., what factors may impair the model's performance). Images of the documentation artifact used in the study are shown in Figure 1-Left. The content is also provided in Table 2 in the appendix.

The service provides a playground UI (Figure 4 in the appendix) for users to try out the model with their own input examples. Seeing examples of model outputs allows understanding through experienced affordance [65, 92]. For efficiency, instead of asking participants to experiment on their own, we curated 20 online articles from different genres and sources, and of different lengths, and presented them in a spreadsheet with these attributes shown. We then captured their summary outputs from the playground UI, and linked the screenshots to the corresponding input articles in the spreadsheet. The documentation and examples show that the model output includes three components:

- Extracted sentences: Three sentences extracted from the input article that the model identifies as conveying the main topic of the article.
- Rank score: A score indicating how relevant each extracted sentence is to the article's main topic.
- Positional information: The position of each extracted sentence in the input article.

We chose a summarization model for several reasons. First, language models are among the most popular pre-trained models as they have become increasingly powerful and can be applied to any document input data. Second, to make the ideation task tractable, we opted for a model that has well-scoped capabilities, rather than

[1] https://azure.microsoft.com/



Model description: the basics of extractive summarization

The extractive summarization model uses natural language processing techniques to locate key sentences in an unstructured text document. These sentences collectively convey the main idea of the document.

When a document is given as the input, the model returns a list of extracted sentences, together with a rank score and its position in the original document for each extracted sentence. A rank score is an indicator of how relevant or important the model considers the sentence is to the main idea of the document (between 0 and 1, higher as more relevant).

By default, the model returns three highest scored sentences, and you can specify the number of sentences returned.

Examples of intended use

You might want to use the extractive summarization model to:

- Distill critical information from lengthy documents.
- Highlight key sentences in documents.
- Quickly skim documents in a library.
- Generate news feed content.

You can use extractive summarization in multiple scenarios across a variety of industries. For example:

- Extract key information from public news articles to produce insights.
- Classify documents by their key contents.
- Distill important information from long documents to empower solutions such as search, question and answering, and decision support.

Do not use

Don't use extractive summarization for automatic actions without human intervention for high-impact scenarios. A person should always review source data when another person's economic situation, health, or safety is affected.

Limitations with impacting factors

Based on your scenario and input data, you could experience different levels of performance:

- Because the model is trained on document-based texts, such as news articles, scientific reports, and legal documents, when used with texts in certain genres such as conversations and transcriptions, it might produce output with lower accuracy.
- When used with texts that may contain errors or are less similar to well-formed sentences, such as texts extracted from lists, tables, charts, the model might produce output with lower accuracy.

Model Documentation

- **Model description**
  - Descriptions of input and output
  - Who, when and how it was developed
  - Model type and algorithm
  - Training data
- **Intended use**
  - Examples
  - Do not use
- **Limitations with impacting factors**
  - What factors affect model performance
- **Performance evaluation**
  - Evaluation metrics
  - Evaluation methods and data

+ Design space guidance?
+ Potential harms and ethical considerations
+ ...

Figure 1: Left: Images of the model documentation artifact given to participants, with four categories of model information (content in Table 2 in the appendix). Right: Image of the summary card shown in the last step for reflection.

general-purpose large language models such as GPT-3. We also believe that the relatively narrow scope of extractive summarization was suitable for a task that requires careful ideation to match model capabilities and user needs.

After participants' initial ideation, we introduced two additional sets of information to help them refine their ideas: **AI design space guidance** (Figure 2-top), and a list of potential **harms considerations** (Figure 2-bottom).

The design space guidance is intended to encourage participants to systematically consider key design elements for AI-powered features, which would allow us to understand their information needs comprehensively. We opted to introduce it after the initial ideation to avoid overwhelming participants. The guidance also serves as additional information about "what to design" [24]. We adapted the "AI-powered user interface guidance" in Subramonyam et al. [90], a synthesis of key UI components of AI-powered systems based on 89 industry design guidelines.

The harms considerations were introduced to further investigate participants' ideation around how to use the model responsibly. While transparency on ethical considerations has been a motivating factor for AI documentation frameworks [12, 45, 78], there is currently no industry standard on how to present them. We designed the information based on a review of survey papers mentioning limitations of summarization models [9, 41, 64, 99] and papers on ethical risks of language models [19, 94], as well as discussions with 2 experts of NLP and AI ethics. We chose to lead with common technical limitations of summarization models and highlight the potential harms that each technical limitation can lead to (in red). This delineation between the potential harms and their sources of technical limitations was intended to encourage participants to come up with mitigation strategies that target the sources.

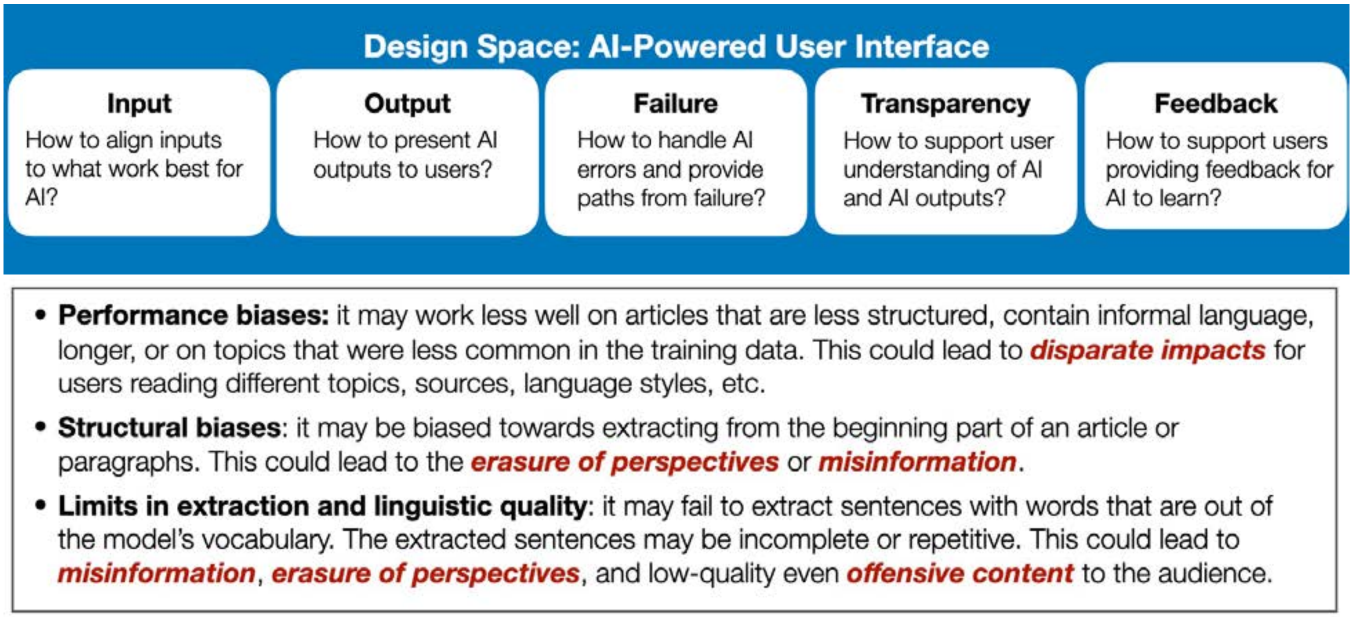
3.2 Procedure

All interviews were conducted online via video conferencing software and lasted around 60 minutes. A $50 gift card was provided as an appreciation token for each participant. Participants were asked to read and sign the consent form before they joined the interviews. The study was IRB approved.

The semi-structured interviews started with a 10-minute discussion of participants' prior experience with designing AI-powered applications. The interviewer probed on how they attempted to understand models in their initial encounters, including their approaches, resources available, and challenges.

The interviewer then made a 5-minute presentation to introduce the design task described above, including showing the documentation (Figure 1-Left) and demonstrating the playground UI (Figure 4 in the appendix). Participants were then asked to join a FigJam board (whiteboarding feature provided by Figma, a UI prototyping software), where we provided the scenario description, documentation, and a link to the spreadsheet of input-output examples obtained from the playground UI. Participants were encouraged to spend a few minutes to further understand the model by browsing the spreadsheet with input-output examples. They were instructed to start ideating whenever they felt ready and follow any processes they usually do. They could use sketching, sticky notes, or any UI widgets on FigJam to communicate their ideas. We also provided a set of microblogging UI components which they could optionally include in their design or annotate directly. Participants were asked to spend no more than 25 minutes on this task, and could stop whenever they were satisfied with their design idea. They were asked to continue thinking aloud throughout the process.

After this, the interviewer asked participants about their perceived understanding of the model, which information they found helpful, and what questions were left unanswered, followed by the two rounds of iteration with the design space guidance and harms considerations (Figure 2). For the sake of time, the iterations focused on verbal discussions rather than re-creation of visual designs. Whenever applicable, participants were prompted to reflect on whether the process and information available shared similarities with how they approach AI in their own work



Design Space: AI-Powered User Interface

Input
How to align inputs to what work best for AI?

Output
How to present AI outputs to users?

Failure
How to handle AI errors and provide paths from failure?

Transparency
How to support user understanding of AI and AI outputs?

Feedback
How to support users providing feedback for AI to learn?

- **Performance biases:** it may work less well on articles that are less structured, contain informal language, longer, or on topics that were less common in the training data. This could lead to ***disparate impacts*** for users reading different topics, sources, language styles, etc.
- **Structural biases**: it may be biased towards extracting from the beginning part of an article or paragraphs. This could lead to the ***erasure of perspectives*** or ***misinformation***.
- **Limits in extraction and linguistic quality**: it may fail to extract sentences with words that are out of the model's vocabulary. The extracted sentences may be incomplete or repetitive. This could lead to ***misinformation***, ***erasure of perspectives***, and low-quality even ***offensive content*** to the audience.

Figure 2: Top: AI UI design space guidance provided in the study. Bottom: Harms considerations provided in the study. Each bullet point is a common technical limitation of text summarization models; the potential harms that each limitation can lead to are highlighted in red. The content is in Table 2 in the appendix.

Lastly, the interviewer asked participants to reflect with a summary card as shown in Figure 1-Right. The card listed all the categories of information provided in the task in color, with additional categories that appear in model cards [78], the most established model documentation framework, in grey. The latter were described as "more technical information" that we excluded (also excluded in the original service documentation). The interviewer asked questions to prompt reflection, such as which category was helpful, whether the grey categories were desired, and what other information they wished to have. The interviewer also introduced the concept of *designerly understanding of AI*—understanding a model well enough to be able to use it as a design material to solve user problems—and asked participants to reflect on what could help them better achieve a designerly understanding in general.

3.3 Participants

Recruitment was carried out through two routes. First, recruiting messages were disseminated in a large international technology company's UX-focused online communities, across product lines and locations. Second, the authors posted recruiting messages on Twitter and LinkedIn. The messages called for participation of people who are in roles that perform design ideation often (including designers, UX researchers, and product managers (PMs)), and have experience working with AI. We limited to these groups since we are interested in learning about participants' own experience ideating for AI-powered products.

The interview study included 23 participants (8 male, 15 female), with 17 recruited via the first route and 6 via the second. Participants from the same large company were distributed in 6 countries with no overlap of first-line teams. The remaining participants work in a mix of large companies, start-ups, and non-profit organizations.

When participants signed up, they were asked to fill out a form that gathered information about their demographics, professions, and their self-reported experience with designing AI and NLP powered applications, respectively (never / limited experience / part of my day-to-day job / I consider myself an expert). The majority of participants have designer titles (N=17), while 3 are HCI or UX researchers, and the remaining 3 are PMs. Detailed information about the participants can be found in Table 3 in the appendix. The last two questions were used to group participants into more or less experienced groups with AI design. Overall, we considered 6 participants to be in the less experienced group; these participants either answered "less experienced" or "never" to both questions regarding AI and NLP or confirmed in the interview they never designed AI-powered features in their job.

3.4 Analysis

Interview transcriptions included question-answering and think-aloud data. Coding started with the first and second authors performing open and axial coding informed by Grounded Theory research [29] on a common set of 5 interviews. They discussed and converged on a set of axial codes, with which the first author continued coding the rest of the interviews. The axial codes will be highlighted in **bold** when discussing findings. After that, the first author performed a first round of selective coding to identify themes, then iteratively presented to the other authors for feedback.

Through a human-information interaction lens [43], we paid particular attention to places where participants showed their attention to, perception of, use of, and feedback for the categories of model information provided. These include what they commented on while reading the documentation, what appeared in their think-aloud comments while performing the task, and their answers to the reflection questions after the task. We coded both the *categories of information* and participants' *goals* behind the information sought. We also mapped the relations between the two with the axial codes. These results are presented in Sections 4.3.

4 FINDINGS

Since the focus of our study is on design ideation with pre-trained models, we first situate our results by discussing what this task currently looks like in practice (RQ1). We then present a brief overview of participants' designs, demonstrating that they were able to engage in design ideation supported by the documentation (RQ2), but their model understanding and design outcomes varied by their level of experience with AI design (RQ1). Finally, in Section 4.3, we present our main results, identifying four common goals that designers seek model understanding for, pinpointing designers' model information needs for each goal including gaps in the current documentation, and suggesting design implications to support them (RQ1, RQ2, RQ3). Throughout the findings, we highlight the pivotal role that UX designers can play in RAI with an understanding of model limitations.

4.1 Putting Design Ideation with Pre-Trained Models in Context: Current Practices (RQ1)

As described below, while previous HCI research focused on supporting UX practitioners to work with data scientists, we found that, on the ground, practitioners also frequently work with pre-trained models without the direct involvement of the data scientists who built them—a trend that may increase as more powerful pre-trained models such as GPT become widely available. This makes it more challenging to assess models for suitability and ideate on how to use the models to solve user problems, since information cannot be obtained directly from the data scientists involved. Unfortunately, at the current time, designers often do not play a central role in the ideation stage for AI-powered products and do not have the information support to obtain a good enough understanding of models to engage in effective ideation.

4.1.1 How is design ideation currently performed in practice? Our study presented participants with a scenario that requires figuring out how to use a given model to solve an existing user problem. We found that participants (or their teams) commonly face this type of scenario in their day-to-day practices. Namely, **practitioners frequently work with third-party models** (N=12)—sometimes referred to as "out-of-box" or "off-the-shelf" models—to add AI-powered features to existing products.

Especially in larger companies, practitioners also strive to **reuse AI capabilities** that the company already owns, whether purchased from third parties or developed by R&D teams. Echoing previous studies, there is still a common separation of design and model development [90]. Designers do not necessarily distinguish between working with a pre-trained model or an "in-house" model handed over *after* its completion, as in several cases (N=5), participants could not recall the sources of the models they worked with.

Curiously, 9 participants mentioned they or their teams engaged in various degrees of **exploration around the use of recent large pre-trained models**, such as GPT-3 and Dall-E, including attempts to define product features and tinkering with the APIs using playground UIs on their own. However, when probed further, none could clearly articulate a ready outcome or an established process to explore these models, showing that ideation on how to use large pre-trained models is an emerging task that raises much interest but is still challenging.

Similar to previous studies [39], we found that UI/UX designers are often not the drivers for product or feature definition, though they are more likely to be in smaller organizations or start-ups (P12, P15, P16) [104]. In large companies, this task is often led by PMs, with input from designers. Participants frequently expressed **dissatisfaction in being excluded in the ideation stage**, as "*I think we should be because we're gonna carry all that implications of a technology choice*" (P17). Participants attributed their lack of means and motivation to understand models to this separation between UI design and ideation, as "*they show up and someone already said this is the problem and this is the solution and you feel like you haven't had a stake, or haven't had a chance to research that problem for your own understanding*" (P7). Participants also described their experience of **design failures due to this siloed process and a lack of model understanding**, as: "*I came in a later stage. The PM had already defined all the specs. I mapped out the ideal customer journey and a service blueprint...it turns out we don't have the technical feasibility to cover all of them*" (P8).

4.1.2 What are the current approaches to obtaining a designerly understanding? About half of participants have done so by **reading some form of model documentation**. While a few mentioned formal documentation of third-party AI services or GPT-3, designers often rely on notes written by PMs or model developers. Participants expressed **struggles with digesting documentation**, because "*they are explaining very complex things and most of them are just plain text*" (P14). Some mentioned that creating high-quality documentation is usually not a priority due to resource constraints.

Participants also sought "experienced affordance" by **examining model inputs and outputs**. P3, P9, and P20 mentioned "playing around" with GPT-3 through the playground UI. However, **means to directly interact with models are often unavailable**. Instead, participants mentioned that their initial encounters with models involved a demo from data scientists or third-party sellers showing examples of inputs and outputs, or being given examples together with documentation. This lack of direct access is common even for in-house models. P10 and P21 approached this challenge by curating their own "*golden set of inputs*" and obtaining outputs from the engineers to support their understanding and design decisions. Several participants (N=5) expressed excitement upon seeing the playground UI in our design task, showing a gap in their current practices and a strong desire to tinker with models directly.

Lastly, as the majority of participants have also worked with "in-house" models built by data scientists, in these cases, they **learn about models by talking to data scientists**. When performing our design task, they frequently described the experience of reading documentation unfavorably compared with that of speaking to a data scientist directly.

4.2 Overview of the Design Outcomes (RQ1&2)

To investigate the feasibility of supporting design ideation with the documentation framework (RQ2), and to ground our later discussions of participants' information needs and goals, we briefly overview the variety of design ideas that participants came up with. To shed light on the challenges (RQ1), we also highlight differences between designers with more or less experience with AI.

4.2.1 Participants created rich designs with various details. With the same set of provided model transparency artifacts, participants arrived at different designs to address the scenario. 17 out of 23 participants presented a feature that shows AI-generated summaries together with shared articles. 7 participants explored a feature that nudges the user to understands the article content before sharing. 4 participants discussed a feature that uses an AI-generated summary to help users to write their own summary.

Participants' designs also had rich details. For example, P3, P7, and P12, who are among the most experienced AI designers we interviewed, created sophisticated designs (Figure 3). P3's designs considered different conditions: applying the AI to only longer articles; quality-checking the original articles and summaries to prevent disputable content from being shared; and a pop-up window to view summaries in sequence for users who share multiple articles in a thread. P7 added user-facing transparency elements about the model's accuracy, confidence, and explanations, and added that their feature should not be applied to high-stakes topics. P12 presented a summary feature that should only be applied to factual articles but not opinion pieces. Several layers of detail were added: a link to the original article, a disclaimer indicating this content is AI-generated, paths from AI failures including feedback and model auditing, and an explanation for why a summary is provided.

4.2.2 Less experienced AI designers faced more challenges approaching model understanding and ideation. As described in Section 3.3, we identified a sub-group of 6 participants (P8, P11, P15, P17, P18, P20) who are less experienced with AI design. We compare their designs to the more experienced AI designers'. We note that although this group had limited experience with designing AI features in their jobs, they still showed a level of knowledge of and strong interest in AI. Our comparison is not intended to generalize the relationship between AI experience and ideation outcomes.

We observe two clusters of designs created by the designers less experienced with AI. One cluster (P15 and P20, whose designs are in Figure 3, and also P8) presented relatively simplistic designs, without as many details as in the designs by the experienced designers. Interestingly, all of them opted to pick one single example out of the playground UI outputs, and created visuals around the content. The other cluster (P17 as in Figure 3, and also P11) had a distinct pattern of quickly generating multiple ideas, with some diverting from the common ideas participants converged to, but under-exploring the feasibility of these ideas with the given model. For example, as shown in Figure 3, P17 suggested a feature that would leverage a summary to identify and explain links with disputable content.

Furthermore, we observed that the less experienced group was significantly more likely to skip the step of examining multiple examples from the playground UI—83.3% versus 11.8% among the rest of the participants. Despite the evidence of less effective understanding and ideation, they were more likely to say yes when asked whether they felt they had a good understanding of the model—50% versus 11.8% among the rest of the participants.

In short, the more experienced participants sought information more thoroughly. With a better understanding of the model, they were able to generate more sophisticated and complete design ideas that are grounded in technical feasibility. However, they also tended to find the provided information inadequate to support what they intended to design. In the next section, we unpack what additional information is required and why.

4.3 Common Transparency Goals and Information Needs (RQ1, RQ2, RQ3)

In this section, we present our main findings. To answer RQ3, we identify four common goals for which designers sought out model understanding. These goals are not meant to be mutually exclusive. To answer RQ2, for each goal, we pinpoint which categories of model information in the provided documentation were used, and what missing information was requested. We also suggest design implications to support each goal to answer RQ1. These results are summarized in Table 1. In the appendix, we provide more details on how each category of model information was used or sought, divided in two tables—Table 4 for provided information and Table 5 for additional information requested.

4.3.1 Goal 1 (G1): To engage in divergent-convergent design thinking and eliminate risky design ideas. We frequently observed the diamond process [30] of design thinking (N=14), with a divergent stage of generating many potential design solutions followed by a convergent stage of refining them. As described below, this process drove designers to seek model understanding necessary for assessing potential UX benefits and risks that would arise from their designs.

For the divergent stage, participants often (N=6) found the section with examples of *intended uses* helpful to "*jump-start*" (P9) generating design ideas. We also observed a common strategy (N=5) of **delineating user workflows** as a way of ideating on potential places for a summarization feature.

For the convergent stage, a **risk-benefit analysis** was often performed to eliminate solutions that are risky or deliver less value to users. This process is best illustrated in P7's design shown in Figure 3. P7 started by generating three possible solutions that they called preempt, intervention, and sidekick. Their convergent process required understanding how reliable the model is, which they approached by examining playground examples: "*I have the question of how reliably it could perform [for different designs]... if it was an intervention and it was unreliable...you're out of your extra step and it's literal nonsense. And that really diminishes somebody's experience with the whole product, so that presents, I think, a huge risk.*" Later, as they proceeded with the sidekick idea, they realized that the benefit provided might be limited and they re-visited the other ideas: "*now that I'm fleshing this out, it's making me feel this would make people read things even less, because just anecdotally for myself, if I saw this I would definitely not click the article.*"

Performing the risk-benefit analysis led participants to seek an understanding of model capabilities based on *model descriptions* and *input-output examples*, and model limitations from *harms considerations*, *impacting factors*, and *unintended uses*. However, translating

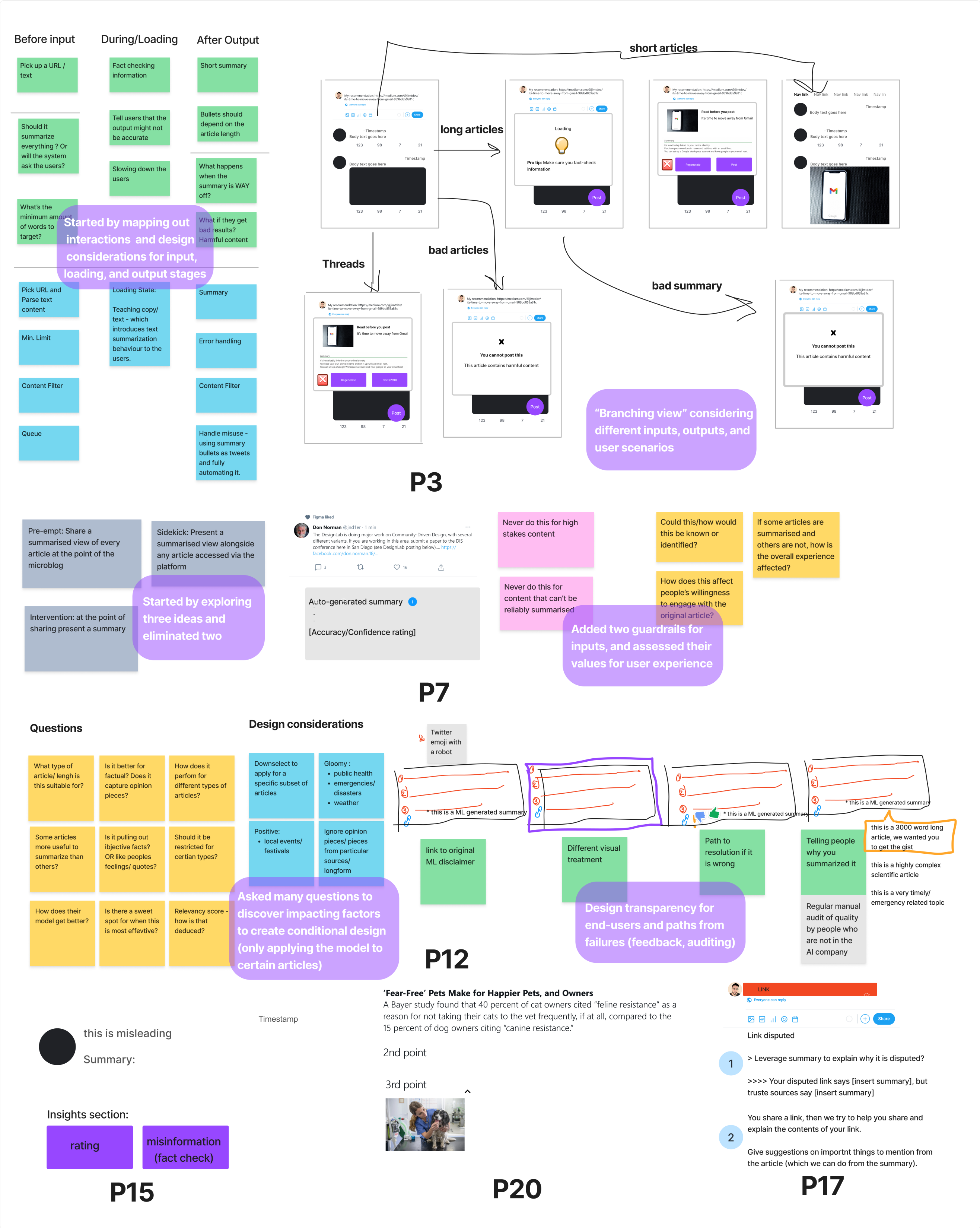

Before input
During/Loading
After Output
Pick up a URL / text
Fact checking information
Short summary
Should it summarize everything ? Or will the system ask the users?
Tell users that the output might not be accurate
Bullets should depend on the article length
Slowing down the users
What happens when the summary is WAY off?
What's the minimum amount of words to target?
What if they get bad results? Harmful content
Pick URL and Parse text content
Loading State: Teaching copy/ text - which introduces text summarization behaviour to the users.
Summary
Min. Limit
Error handling
Content Filter
Content Filter
Queue
Handle misuse - using summary bullets as tweets and fully automating it.
Started by mapping out interactions and design considerations for input, loading, and output stages
short articles
long articles
Threads
bad articles
bad summary
Loading
Pro tip: Make sure you fact-check information
You cannot post this
This article contains harmful content
Post
"Branching view" considering different inputs, outputs, and user scenarios
P3
Pre-empt: Share a summarised view of every article at the point of the microblog
Sidekick: Present a summarised view alongside any article accessed via the platform
Intervention: at the point of sharing present a summary
Started by exploring three ideas and eliminated two
Figma liked
Don Norman @jnd1er · 1 min
The DesignLab is doing major work on Community-Driven Design, with several different variants. If you are working in this area, submit a paper to the DIS conference here in San Diego (see DesignLab posting below).... https://facebook.com/don.norman.18/...
Auto-generated summary
-
-
-
[Accuracy/Confidence rating]
Never do this for high stakes content
Never do this for content that can't be reliably summarised
Could this/how would this be known or identified?
How does this affect people's willingness to engage with the original article?
If some articles are summarised and others are not, how is the overall experience affected?
Added two guardrails for inputs, and assessed their values for user experience
P7
Questions
What type of article/ lengh is this suitable for?
Is it better for factual? Does it capture opinion pieces?
How does it perfom for different types of articles?
Some articles more useful to summarize than others?
Is it pulling out ibjective facts? OR like peoples feelings/ quotes?
Should it be restricted for certian types?
How does their model get better?
Is there a sweet spot for when this is most effevtive?
Relevancy score - how is that deduced?
Design considerations
Downselect to apply for a specific subset of articles
Gloomy :
• public health
• emergencies/ disasters
• weather
Positive:
• local events/ festivals
Ignore opinion pieces/ news from particular sources/ longform
Asked many questions to discover impacting factors to create conditional design (only applying the model to certain articles)
Twitter emoji with a robot
* this is a ML generated summary
* this is a ML generated summary
* this is a ML generated summary
link to original ML disclaimer
Different visual treatment
Path to resolution if it is wrong
Telling people why you summarized it
this is a 3000 word long article, we wanted you to get the gist
this is a highly complex scientific article
this is a very timely/ emergency related topic
Regular manual audit of quality by people who are not in the AI company
Design transparency for end-users and paths from failures (feedback, auditing)
P12
Timestamp
this is misleading
Summary:
Insights section:
rating
misinformation (fact check)
P15
'Fear-Free' Pets Make for Happier Pets, and Owners
A Bayer study found that 40 percent of cat owners cited "feline resistance" as a reason for not taking their cats to the vet frequently, if at all, compared to the 15 percent of dog owners citing "canine resistance."
2nd point
3rd point
P20
LINK
Everyone can reply
Share
Link disputed
1
> Leverage summary to explain why it is disputed?
>>>> Your disputed link says [insert summary], but truste sources say [insert summary]
2
You share a link, then we try to help you share and explain the contents of your link.
Give suggestions on importnt things to mention from the article (which we can do from the summary).
P17

Figure 3: Example designs created by participants. The bottom three are from participants in the "less experienced with AI design" group.

Transparency goal	Provided info used	Requested information	Design implications to support the goal
G1: Divergent-convergent design thinking	intended uses, model description, input-output examples, harms considerations, unintended uses, limitations	output analysis, explanations	- Inspire divergent design thinking with example use cases and user contexts, such as providing or helping define user workflows or scenarios. - Scaffold defining UX risks by supporting discovering a broad range of model limitations and providing ethical consideration guidance. - Support risk-benefit assessment of candidate design ideas; develop risk-oriented evaluation metrics and practices.
G2: Conditional design	impacting factors in limitations and harms considerations, examples of different categories	training data, explanations, disaggregated evaluation with performance and other output characteristics, confidence/ uncertainty	- Facilitate discovering and testing hypotheses about impacting factors and assessing their UX impact. - Support decisions about whether to create conditional designs by exploring the design space (e.g., creating intermediate prototypes) and assessing their user values. - Develop design patterns and implementation guidance for conditional designs.
G3: Transparency for users	model description, limitations, harms considerations	performance, confidence/uncertainty, explanations	- Allow seeking model information by user questions or needs, such as re-structuring the documentation and providing other information channels. - Support translating information in documentation to transparency designs for users.
G4: Team negotiation and collaboration	harms considerations, limitations, design space guidance	customizability and improvability, algorithm, training data and other development information	- Empower designers by prioritizing their suitability assessment of the model for the users, with both informational and organizational support. - Equip designers to collaborate with engineers with technical literacy, actionable suggestions for model improvement, common references, and boundary objects.

Table 1: Summary of goals for which designers sought out model information, what information they sought in the provided model transparency artifacts and what is missing, and design implications for supporting each goal

model capabilities into UX benefits, and model limitations into UX risks is a non-trivial task. One common translation strategy observed in the majority of participants is to examine the examples and **mentally simulate how users would perceive and react** to them. As discussed in Section 4.2, the experienced AI designers often examined a mix of input examples that are representative of articles shared on the platform, examples from multiple categories, and edge cases in the hope of revealing model limitations. However, not all participants engaged in these productive strategies.

When translating model limitations to UX risks, participants often (N=6) expressed confusion about **what failures meant from a UX perspective**. Failures can arise not only due to poor performance as measured by standard performance metrics—the focus of the limitations section in the provided documentation—but can also be caused by other characteristics of the model outputs (like being too long or not coherent enough for a design) or API properties (such as speed or reliance). As a result, participants found the current limitations section inadequate and engaged in **discovering additional model limitations** that can cause UX failures.

Participants appreciated the *harms considerations* to help them think through potential negative consequences. Remarkably, many participants (N=7) took it as inspiration to **anticipate harms specific to their design and users**. Upon reading about performance biases, for example, P7 questioned the downstream harm of their own design: "*Am I creating a potential skew? All articles that are extremely neat get a good summary and any that are too complicated have a low accuracy. It's possible...all the ones that are summarized well are click-baity articles...does that really help?*"

To anticipate UX benefits and risks, some participants also requested *descriptive analysis of outputs* to understand the general characteristics of model outputs such as the distribution of output lengths and frequencies of certain types of words. Such information should ideally be provided with regard to input articles specific to their product. In addition, some participants sought *model explanations* about what features the model relies on. For example, P1 asked "*does it give more importance to numbers?*" as they reasoned that numbers may then show up more often in the summaries. Such requests were frequently triggered by observing distinct or unexpected model behaviors in input-output examples.

Finally, participants expressed a desire to more **accurately assess the candidate design ideas by risk**, such as through user testing. P7 called out a need for risk-oriented evaluation metrics rather than traditional UX metrics: "*I would want to create different concepts and evaluate them through these lenses [of risks]... typical design world you would say, oh, is this good or not, do users like it or* *not. But I think this would be the other test. So looking at the potential risks in doing something a particular way.*"

Design implications based on these findings are summarized in Table 1.

4.3.2 Goal 2 (G2): To create "conditional designs" to mitigate AI's varying impact for different user scenarios. We observed that participants frequently (N=9) approached the model's output uncertainty [103] by creating **different designs for different types of inputs** or **different types of outputs**. They often did so by putting guardrails on inputs and outputs—only applying the model to inputs that it is reliable for, or blocking problematic outputs. Other common conditional designs included triggering user warnings in certain conditions or providing user controls such as "*toggle on or off this feature for those articles*" (P12). These observations suggest that experienced AI designers are mindful about moving away from only focusing on golden paths or ideal hero scenarios—a "traditional" design practice that may lead to failure of AI UX [56]. This type of design thinking also stems from a common practice of **designing for different user scenarios**. We refer to these common processes of creating different designs for different conditions as *conditional designs*. Such designs are best illustrated in the example of P3 (Figure 3) who considered different designs based on article length (input) and summary quality (output), as well as for users who share many articles in a thread (scenario).

This goal of creating conditional designs gave rise to participants' pronounced needs to understand impacting factors, which appeared in the *limitations* and *harms considerations* sections. There are several challenges. First, not all impacting factors are available in the documentation. Participants often discovered new factors of interest by considering different user scenarios. Participants therefore expressed desires to **discover or verify more impacting factors**. While many participants (N=8) attempted this by visually examining *input-output examples* of different categories, this approach might not lead to an accurate understanding, but "*create a cognitive load*" (P2). Some participants requested information about *training data*, *explanations*, and *disaggregated evaluations* [14] (also referred to as sub-group analysis) to help them discover impacting factors. As illustrated in P12's examples in Figure 3, they started by asking many questions regarding unknown impacting factors, such as "*what type of article length is this suitable for?*" and "*is it better for factual?*", then probed on the explainability-related "how" question to infer potential factors: "*is the summary only pulling out objective facts? Or also peoples' quotes?*", and questions suggesting a desire for disaggregated evaluation: "*how does it perform for different types of articles?"* At the end, P12 reflected that they did not have a good understanding of the model, and made an assumption that they should avoid summarizing opinion pieces based on observations of example outputs.

The decision to create a conditional design must be carefully justified. It comes not only with a development cost but also a cost to UX, since it can create "*an odd feeling and inconsistency*" (P7). Participants wanted to **assess the impact** of factors of interest—both on performance metrics and on other output characteristics like the structural patterns and content quality. This justification must also be assessed with regard to the actual **user benefits of a conditional design**, considering factors such as frequency and user contexts of a given condition. For example, P7 opted to ignore the concern about low-quality summaries for articles about traveling, based on observing an example, since the consequence of misinformation about such articles might be less serious. After reading about the impact of lack of article structure on performance in the limitation section, P9 decided to ignore it, saying "*if articles that are being shared on the microblogging site were mostly chart heavy, like scientific publications, then I would have more questions.*"

Finally, designers struggled with **how to implement conditional designs**, often realizing that another technical component might be required, like a separate model to detect the article genre. Participants commonly (N=6) sought to create guardrails on outputs by leveraging the model *confidence or uncertainty*, prompting them to look into whether the model can provide that information.

Once again, design implications derived from these findings are summarized in Table 1.

4.3.3 Goal 3 (G3): To provide AI transparency to end users. To create "*interaction-level interventions*" (P9) to mitigate the harms of designs, another common goal is to **transparently expose limitations and potential harms to end users**. Many participants (N=10) attempted to meet this goal by **incorporating information from the documentation into their designs**, including information from the *limitations* and *harms considerations* sections. Some further requested to expose information about *performance*, output *confidence or uncertainty*, and *explanations* of how the text summarization works. However, participants faced challenges in **how to translate and effectively present this information in the UI**: "*we definitely need to surface this information in the documentation. But the key question is from a UI perspective, what needs to change, right?... you can have a pop-up appearing over there highlighting that this capability is in preview, and there could be certain limitations... Click here to learn more and then you get to the documentation*" (P10).

Some participants also discussed that in their own work, to be able to effectively communicate the model information to end users, they need to seek an understanding of the model by **asking questions on behalf of the users** to data scientists, which they are unable to do with documentation alone: "*I think a lot of the questions about the human impact of the model are very much within the designers' purview to ask questions to the ML team...the hard part is that a lot of them are not visible. It's such an intangible thing... you have to be really familiar with the material to be able to even have coherent thoughts about it*" (P12). In other words, an AI-powered product cannot be truly transparent and supportive of user understanding if the designer themself lacks an understanding of the AI material they are working with.

4.3.4 Goal 4 (G4): To negotiate and collaborate with their team to advocate for users. When probed about designers' role in creating responsible AI products, besides design interventions, participants emphasized designers' responsibilities in **advocating for users** by anticipating potential harms to users and communicating them to their teams, including **pushing back on the use of a technology**. These points are best made by P9: "*I think designers are probably in the best position to explain those problems back up the chain because we have good tools about modeling users and contexts in scenarios. So we can say, hey, have you thought about the single mom who's looking at this interface and how it presumes she has a husband and how offensive that is? That ability to frame that as a human* *problem as opposed to a business problem might have more of an influence within the conversations of an organization. And I think that we end up being a small amount of gate keeping for the vetting of software. So if we found that something was actively propagating misinformation we can reject it... we have to advocate for users, not just business outcomes."* To engage in such advocacy, designers can "*feel empowered*" (P23) by having a good understanding of model limitations. Indeed, we often observed participants pushing back on the use of the model after reading the *harms considerations*.

Designers also actively seek to **collaborate with engineers to resolve technical limitations and improve the model** for their product. This tendency prompted several participants (N=5) to request additional information about model *customizability and improvability*, such as whether it is possible to gather domain-specific data to fine-tune the model. The service roadmap information about future updates is also important for coordinating with the team so they know whether "*some constraints [will be] erased [in the future]*" (P13). P2 and P6 also appreciated the format of the provided *harms considerations*, in which different sources of technical limitations were delineated, so that they could work with the team to "*address each of these different sources of possible harms*" (P2).

Lastly, we highlight two additional roles that documentation can serve to facilitate team coordination. One is that a documentation can be used as **common references and boundary objects** to support collaboration, especially the sections on *harms considerations* and *design space guidance*, as illustrated in P7's response: "*I'm wondering how much of this is just generally useful for a team. Obviously the model stuff is true, but some of this guidance around human-AI interaction could also be useful for everybody to be mindful of.*" Second, participants deemed documentation a useful resource to help them to **develop general AI literacy** for effective cross-team communication in the long run, which motivated them to seek more technical information in the documentation such as the *algorithms* used, information about the *training data*, and other details about *model development*: "*it's also about educating the designer [about] different types of learning models and algorithms, so that when we communicate with data scientists, we can use the same language and talk about the same thing... on the long term I feel [documentation] also should be about education*" (P5).

5 DISCUSSION

We have identified UX designers' diverse information needs when working with pre-trained models as design material, and how these needs are engendered by their task-specific, role-specific, and socio-organizational goals. The results demonstrate the utility of model documentation in sensitizing designers to the capabilities and limitations of a model for design ideation, but also reveal many gaps. We found that designers gravitate towards critical information that helps them understand model limitations and adopt a set of strategies to mitigate the negative user impact of these limitations. In this section, we discuss future directions for supporting ideation with model transparency and argue for better engaging designers in RAI practices.

5.1 From Model Documentation to Model Interrogation

While there has been extensive work on AI documentation [12, 45, 50, 51, 78, 80], who the consumers are and how they are consuming it has not been well studied empirically. Our work serves as a case study to explore the model reporting needs of UX practitioners. We found that they can benefit from AI documentation, and are already consuming it on the ground. However, only 5 out of 23 participants answered affirmatively that they understood the model in our study well. Participants requested additional categories of information, and some struggled with a lack of complete or concrete understanding of provided information.

Their struggles were not due to a lack of ML expertise [32], as participants had little difficulty comprehending the documentation (though this may not generalize to designers with little knowledge about AI). Instead, the challenges arose when *contextualizing* model information for their setting and users. It is impossible for documentation creators to anticipate every downstream use case. This suggests that we should provide opportunities for designers to interrogate the model with their own data instances, factors of interest, hypotheses, and questions. Additionally, static documentation falls short in supporting the *co-evolving of design solutions and model understanding,* which can be seen as an aspect of design as co-evolving of problem-solution [38]. That is, what needs to be understood, such as what characteristics of model output are important, is emergent from the designs being explored, the depth of design details, and also the evolving understanding of the users' needs and characteristics. We suggest a few directions to support the contextualization and evolving information-seeking processes [60, 77, 86] of model understanding through *model interrogation.*

Supporting example-based interrogation. Echoing prior research [65, 91], we found that examining input-output examples plays an important role in design decisions, as it allows designers to visually envision user reactions and design opportunities, as well as discover nuanced model behaviors that cannot be conveyed by high-level descriptions or metrics. However, an ad-hoc approach that relies on designers to choose the examples to examine does not guarantee an accurate and complete understanding, and can disadvantage inexperienced AI designers. Future work should explore helping designers create or customize example datasets that are representative of their use cases, and guiding them to explore the input and output spaces in a more systematic fashion, such as by suggesting examples from different categories. Designers should not only experience model affordances but also failures from examples, such as through observing edge cases. Lastly, it is necessary for designers to understand the generalizability of model behaviors they observe in examples, such as by having metrics quantifying their frequency and performing group-level output analyses that expand beyond basic disaggregated evaluations.

Explainable AI for designerly understanding. We note a potential role that explainable AI (XAI) tools can play in supporting designerly understanding, as "how" and "why" questions frequently emerged in participants' ideation processes. In human communication, people seek explanations about an event to be able to extrapolate to predictions about future, unseen events [71, 72]. This was often the goal and process that participants followed; they requested explanations of the output for an example they observed, and then attempted to infer whether the model would behave similarly for other articles, and if so, what kinds. In some cases, participants also requested global explanations to infer general characteristics of the model's behavior and outputs. Subramonyam et al. [90] found that when interacting with data scientists who had developed in-house models, designers often attempt to validate their hypotheses about model behaviors, errors, and impacting factors by seeking explanations about training data, features used, feature weights, rules, and underlying assumptions. Future work can explore how to provide similar information through interactive explanations for pre-trained models.

Supporting testing factors for UX impact. A recent development in making model reporting interrogatable is allowing users to provide or slice data to generate performance reports [47] for different groups. Also available are a set of model behavioral analysis tools (e.g., [10, 28, 84, 98]) to support ML engineers to understand impacting factors by performing disaggregated evaluation. However, their utility for UX designers is unclear. First, as discussed, UX failures may be concerned with a broader set of output characteristics than model errors. Second, designers face challenges translating between factors that impact model performance and the user scenarios they are designing for. For example, upon reading about the impact of an article being "unstructured" on performance, participants needed to translate that into "what kind of users tend to share unstructured articles and what are the potential risks to them." When considering a common scenario of users sharing multiple articles in a row, P3 had to mentally simulate the output of chaining multiple articles. Future work should support designers to test factors of interest by allowing them to define metrics that matter to UX, and explore the impact by different user scenarios. For example, we may envision a tool that asks designers to define different user scenarios, helps them identify input examples that fit these scenarios, and allows visually examining their outputs and UX impact for each scenario.

Integrating the exploration of design space and model understanding. Recent HCI work begins to develop prototyping tools that integrate generating model outputs and creating interface designs in one place [91]. Future work should explore AI prototyping tools that also incorporate model transparency information. Moreover, we observe that, to cope with the uncertainty and complexity of AI, there is a strong desire to create intermediate designs and explore how the model behaves for different design ideas, and assess the potential UX risks and benefits. Future tooling should support such processes and model understanding needs that emerge from exploring different designs. For example, P3's ideation in Figure 3 shows a natural inclination to create "branching views" to explore and manage different conditional designs, and their design decisions can benefit from a more concrete understanding of the model input and output characteristics for each branch of design.

Finally, we call out the immense need to support understanding large, general-purpose pre-trained models (e.g., GPT-3 and Dall-E) through interrogation to support designers or other individuals in making responsible decisions about their use. Given the extreme uncertainty about these models' capabilities and limitations, and the current uncertainty about appropriate application domains, any static documentation is unlikely to suffice. For example, the current documentation for GPT-3 provides only a high-level description of its capabilities, such as "a set of models that can understand and generate natural language," and "safety best practices guidance" that includes examples of harms and mitigation strategies. We believe users of large pre-trained models can benefit from tools that support example-based interrogation, model behavior analysis on different input groups, risk-oriented explorations to discover context-specific failures and harms, capabilities to answer questions and test hypotheses, and tinkering with application ideas.

5.2 Implications for UX-Led Approaches for RAI Practices

Our study investigated designers' use of critical information about model limitations. Based on the results, we highlight a few reasons that UX practitioners can and should play a more central role in RAI practices to mitigate potential harms of AI technologies. First, UX design is fundamentally about bridging user needs and technical affordances. UX designers' training equips them with the skills to understand users through user research and prior experience, and extrapolate that understanding to anticipate user perceptions and behaviors interacting with a given technology. They can apply the same skills to anticipating potential harms of AI. For example, recently it has become more common to adopt "red teaming" into RAI development practices [43, 44]—coming up with adversarial inputs that produce harmful outputs and then updating the model to avoid them. We believe designers are well suited for performing such tasks. Their bridging role also places a sense of responsibility on them to be "user advocates," making them inclined to exhibit appropriate skepticism about a technology and actively seek to understand model limitations.

Second, as the four goals identified in Section 4.3 show, UX designers bring a unique set of tools to cope with limitations of AI and mitigate their potential harms. They are able to explore the design space, assess potential harms that are emergent from different design solutions, eliminate risky technology designs, and help identify harms-mitigation strategies that should drive technical development. They are able to create "*interaction-level interventions*" (P9) to mitigate potential harms, by putting guardrails on model inputs and outputs, as well as creating transparency and user-control features to enable user agency dealing with model limitations. While the AI ethics literature often asks the question of "whether to build a technology" [13], UX practices and HCI literature have long contemplated with "*how* to build a technology" responsibly when there are competing requirements and values (e.g., design for "wicked problems" [23], and applying it to RAI [79]). The core interest in supporting user agency is motivated by a critical position that even following the best practices, technology creators will always face uncertain downstream trajectories of use [26, 36, 59]. RAI practices can benefit from these user-centered and pragmatic design tools.

Lastly, the increasing adoption of pre-trained models for product development calls for a UX-led approach as it necessitates design ideation to define new features that the model is suitable for. If well supported, pre-trained models also present opportunities to empower UX practitioners to directly tinker with a wide range of AI design materials, prototype by choosing from and "stitching together" different design materials, and ultimately take a more proactive role in developing AI-powered products.

With these arguments, we advocate for more and earlier inclusion of UX practitioners in the practice of RAI. This can be accomplished by providing designer-centered documentation and tooling to support their needs for understanding AI, as well as by lowering the organizational barriers for designers to take leading roles in product ideation, suitability assessment of models, and the definition of harms and mitigation strategies.

5.3 Limitations

As with any studies that utilize a design probe, our results are contingent on the documentation we chose. However, our main focus on the categories of information required to meet designers' goals, rather than specific content details, may mitigate this limitation to some degree. The extractive text summarization model and the design task also had their idiosyncrasies, and hence certain design processes may not generalize to other types of models and AI-powered features. While we aimed to introduce a hands-on task to observe participants' natural ideation processes, the study may still suffer from certain ecological validity issues. For example, participants were not given the opportunity to research the user problem, and the time and resources given were limited. Moreover, our sample was biased towards UX practitioners working in large technology companies and experienced with AI design. In fact, the majority of participants were recruited from a single large company. This sample may limit the generalizability of our observations about current practices reported in Section 4.1. For example, we mentioned that designers in smaller organizations appeared to take more initiative in the ideation stage. Lastly, given that design ideation with pre-trained models is still an emerging task and not all our participants had engaged with such a task before, the design practices we observed may not cover all practices, and we do not claim that all observations should be taken as best practices.

6 CONCLUSION

We conducted an interview study, including a hands-on design task, with 23 UX practitioners to investigate their needs and goals when performing design ideation on how to use a pre-trained ML model. We took a primary interest in their information needs to develop a "designerly understanding" of the model, and explored whether and how information categories in transparent model reporting frameworks can support such an understanding. Our results inform future development of transparent model reporting practices, as well as other tools that aim to support design ideation working with pre-trained models. Our study is motivated by two current trends in the broader context of AI product development: the availability of increasingly versatile pre-trained models, including large, general-purpose pre-trained models, for product innovation; and the recognition of the importance and challenges of RAI practices that aim to proactively mitigate the potential harms, and safeguard the use of AI. We take a formative step towards exploring, and ultimately supporting, the opportunities and responsibility for UX practitioners under these trends.

ACKNOWLEDGMENTS

We thank all our study participants and reviewers for their helpful feedback. We also thank Su Lin Blodgett, Alexandra Olteanu, Qiaosi Wang, and Ziang Xiao for their input and feedback on our study materials and procedure.

REFERENCES

[1] 2019. Amazon AWS Machine Learning Services. https://aws.amazon.com/machine-learning/.

[2] 2019. Google AI for Developers. https://cloud.google.com/products/ai.

[3] 2019. Google People + AI Guidebook. pair.withgoogle.com/guidebook.

[4] 2019. Hugging Face Models. https://huggingface.co/models.

[5] 2019. IBM Design for AI. https://www.ibm.com/design/ai/.

[6] 2019. IBM Watson AI Solutions. https://www.ibm.com/artificial-intelligence.

[7] 2019. Microsoft Azure Cognitive Service. https://azure.microsoft.com/en-us/services/cognitive-services/.

[8] 2022. Transparency Note for Azure Cognitive Service for language. https://docs.microsoft.com/en-us/legal/cognitive-services/language-service/transparency-note.

[9] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268* (2017).

[10] Saleema Amershi, Max Chickering, Steven M Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. Modeltracker: Redesigning performance analysis tools for machine learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 337–346.

[11] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[12] Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6–1.

[13] Solon Barocas, Asia J Biega, Benjamin Fish, Jędrzej Niklas, and Luke Stark. 2020. When not to design, build, or deploy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 695–695.

[14] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, W Duncan Wadsworth, and Hanna Wallach. 2021. Designing disaggregated evaluations of ai systems: Choices, considerations, and tradeoffs. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 368–378.

[15] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

[16] Jesse Josua Benjamin, Arne Berger, Nick Merrill, and James Pierce. 2021. Machine Learning Uncertainty as a Design Material: A Post-Phenomenological Inquiry. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[17] Richard Benjamins, Alberto Barbado, and Daniel Sierra. 2019. Responsible AI by design in practice. *arXiv preprint arXiv:1909.12838* (2019).

[18] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 648–657.

[19] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5454–5476.

[20] Karen L Boyd. 2021. Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–27.

[21] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[22] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[23] Richard Buchanan. 1992. Wicked problems in design thinking. *Design issues* 8, 2 (1992), 5–21.

[24] Bill Buxton. 2010. *Sketching user experiences: getting the design right and the right design.* Morgan kaufmann.

[25] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2021. Onboarding Materials as Cross-functional Boundary Objects for Developing AI Assistants. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–7.

[26] Matthew Chalmers and Areti Galani. 2004. Seamful interweaving: heterogeneity in the theory and design of interactive systems. In *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques.* 243–252.

[27] Valerie Chen, Jeffrey Li, Joon Sik Kim, Gregory Plumb, and Ameet Talwalkar. 2022. Interpretable machine learning: Moving from mythos to diagnostics. *Queue* 19, 6 (2022), 28–56.

[28] Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, Ki Hyun Tae, and Steven Euijong Whang. 2019. Slice finder: Automated data slicing for model validation. In *2019 IEEE 35th International Conference on Data Engineering (ICDE).* IEEE, 1550–1553.

[29] Juliet Corbin, Anselm L Strauss, and Anselm Strauss. 2015. *Basics of qualitative research.* sage.

[30] Design Council. 2005. The 'double diamond' design process model. Design Counci. (2005).

[31] Henriette Cramer and Juho Kim. 2019. Confronting the tensions where UX meets AI. *Interactions* 26, 6 (2019), 69–71.

[32] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive Model Cards: A Human-Centered Approach to Model Documentation. *arXiv preprint arXiv:2205.02894* (2022).

[33] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2021. Stakeholder Participation in AI: Beyond" Add Diverse Stakeholders and Stir". *arXiv preprint arXiv:2111.01122* (2021).

[34] Dominik Dellermann, Adrian Calma, Nikolaus Lipusch, Thorsten Weber, Sascha Weigel, and Philipp Ebel. 2021. The future of human-AI collaboration: a taxonomy of design knowledge for hybrid intelligence systems. *arXiv preprint arXiv:2105.03354* (2021).

[35] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. *arXiv preprint arXiv:2205.06922* (2022).

[36] Alan Dix. 2007. Designing for appropriation. In *Proceedings of HCI 2007 The 21st British HCI Group Annual Conference University of Lancaster, UK 21.* 1–4.

[37] Dennis P Doordan. 2003. On materials. *Design Issues* 19, 4 (2003), 3–8.

[38] Kees Dorst and Nigel Cross. 2001. Creativity in the design process: co-evolution of problem–solution. *Design studies* 22, 5 (2001), 425–437.

[39] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX design innovation: Challenges for working with machine learning as a design material. In *Proceedings of the 2017 chi conference on human factors in computing systems.* 278–288.

[40] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing transparency design into practice. In *23rd international conference on intelligent user interfaces.* 211–223.

[41] Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. A survey on dialogue summarization: Recent advances and new frontiers. *arXiv preprint arXiv:2107.03175* (2021).

[42] Ylva Fernaeus and Petra Sundström. 2012. The material move how materials matter in interaction design research. In *proceedings of the designing interactive systems conference.* 486–495.

[43] Raya Fidel. 2012. *Human information interaction: An ecological approach to information behavior.* Mit Press.

[44] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858* (2022).

[45] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[46] Elisa Giaccardi and Elvin Karana. 2015. Foundations of materials experience: An approach for HCI. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.* 2447–2456.

[47] Karan Goel, Nazneen Rajani, Jesse Vig, Samson Tan, Jason Wu, Stephan Zheng, Caiming Xiong, Mohit Bansal, and Christopher Ré. 2021. Robustness gym: Unifying the nlp evaluation landscape. *arXiv preprint arXiv:2101.04840* (2021).

[48] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[49] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. *Science Robotics* 4, 37 (2019).

[50] Amy K. Heger, Liz B. Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. 2022. Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022).

[51] Michael Hind, Stephanie Houde, Jacquelyn Martino, Aleksandra Mojsilovic, David Piorkowski, John Richards, and Kush R Varshney. 2020. Experiences with improving the transparency of AI models and services. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–8.

[52] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems.* 1–13.

[53] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2020. The dataset nutrition label. *Data Protection and Privacy, Volume 12: Data Protection and Democracy* 12 (2020), 1.

[54] Lars Erik Holmquist. 2017. Intelligence on tap: artificial intelligence as a new design material. *interactions* 24, 4 (2017), 28–33.

[55] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems.* 1–16.

[56] Matthew K Hong, Adam Fourney, Derek DeBellis, and Saleema Amershi. 2021. Planning for natural language failures with the ai playbook. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–11.

[57] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–26.

[58] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems.* 17–24.

[59] Sarah Inman and David Ribes. 2019. " Beautiful Seams" Strategic Revelations and Concealments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* 1–14.

[60] Harmanpreet Kaur, Eytan Adar, Eric Gilbert, and Cliff Lampe. 2022. Sensible AI: Re-imagining Interpretability and Explainability using Sensemaking Theory. *arXiv preprint arXiv:2205.05057* (2022).

[61] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–14.

[62] Claire Kayacik, Sherol Chen, Signe Noerly, Jess Holbrook, Adam Roberts, and Douglas Eck. 2019. Identifying the intersections: User experience+ research scientist collaboration in a generative machine learning interface. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems.* 1–8.

[63] Frank C Keil. 2006. Explanation and understanding. *Annu. Rev. Psychol.* 57 (2006), 227–254.

[64] Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2022. An Empirical Survey on Long Document Summarization: Datasets, Models and Metrics. *ACM Journal of the ACM (JACM)* (2022).

[65] Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *CHI Conference on Human Factors in Computing Systems.* 1–19.

[66] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–15.

[67] Q Vera Liao, Milena Pribić, Jaesik Han, Sarah Miller, and Daby Sow. 2021. Question-driven design process for explainable ai user experiences. *arXiv preprint arXiv:2104.03483* (2021).

[68] Q Vera Liao and Kush R Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv preprint arXiv:2110.10790* (2021).

[69] Q Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. 2022. Connecting Algorithmic Research and Usage Contexts: A Perspective of Contextualized Evaluation for Explainable AI. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 10. 147–159.

[70] Zachary C Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[71] Tania Lombrozo. 2006. The structure and function of explanations. *Trends in cognitive sciences* 10, 10 (2006), 464–470.

[72] Tania Lombrozo. 2012. Explanation and abductive inference. (2012).

[73] Jiahao Lu, Alejandra Gomez Ortega, Milene Gonçalves, and Jacky Bourgeois. 2021. The Impact of Data on the Role of Designers and their Process. *Proceedings of the Design Society* 1 (2021), 3021–3030.

[74] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of*

the ACM on Human-Computer Interaction 6, CSCW1 (2022).

[75] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[76] Nirav Malsattar, Tomo Kihara, and Elisa Giaccardi. 2019. Designing and Prototyping from the Perspective of AI in the Wild. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. 1083–1088.

[77] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[78] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[79] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. 2022. Designing Fair AI in Human Resource Management: Understanding Tensions Surrounding Algorithmic Evaluation and Envisioning Stakeholder-Centered Solutions. In *CHI Conference on Human Factors in Computing Systems*. 1–22.

[80] Partnership on AI. 2021. *ABOUT ML Reference Document*. Technical Report. https://partnershiponai.org/paper/about-ml-reference-document/.

[81] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44.

[82] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.

[83] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. PMLR, 8821–8831.

[84] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4902–4912.

[85] Erica Robles and Mikael Wiberg. 2010. Texturing the" material turn" in interaction design. In *Proceedings of the fourth international conference on Tangible, embedded, and embodied interaction*. 137–144.

[86] Reijo Savolainen. 1993. The sense-making theory: Reviewing the interests of a user-centered approach to information seeking and use. *Information processing & management* 29, 1 (1993), 13–28.

[87] Donald Schon and John Bennett. 1996. Reflective Conversation with Materials in Bringing Design to Software, Winograd T.

[88] Ben Shneiderman. 2021. Responsible AI: Bridging from ethics to practice. *Commun. ACM* 64, 8 (2021), 32–35.

[89] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2020. Participation is not a design fix for machine learning. *arXiv preprint arXiv:2007.02423* (2020).

[90] Hariharan Subramonyam, Jane Im, Colleen Seifert, and Eytan Adar. 2022. Solving Separation-of-Concerns Problems in Collaborative Design of Human-AI Systems through Leaky Abstractions. In *CHI Conference on Human Factors in Computing Systems*. 1–21.

[91] Hariharan Subramonyam, Colleen Seifert, and Eytan Adar. 2021. ProtoAI: Model-Informed Prototyping for AI-Powered Interfaces. In *26th International Conference on Intelligent User Interfaces*. 48–58.

[92] Hariharan Subramonyam, Colleen Seifert, and Eytan Adar. 2021. Towards a process model for co-creating AI experiences. In *Designing Interactive Systems Conference 2021*. 1529–1543.

[93] Harini Suresh, Steven R Gomez, Kevin K Nam, and Arvind Satyanarayan. 2021. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[94] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 214–229.

[95] Maximiliane Windl, Sebastian S Feger, Lara Zijlstra, Albrecht Schmidt, and Pawel W Wozniak. 2022. 'It Is Not Always Discovery Time': Four Pragmatic Approaches in Designing AI Systems. In *CHI Conference on Human Factors in Computing Systems*. 1–12.

[96] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

[97] Jennifer Wortman Vaughan and Hanna Wallach. 2021. A Human-Centered Agenda for Intelligible Machine Learning. In *Machines We Trust: Perspectives on Dependable AI*, Marcello Pelillo and Teresa Scantamburlo (Eds.). MIT Press.

[98] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2019. Errudite: Scalable, reproducible, and testable error analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 747–763.

[99] Divakar Yadav, Jalpa Desai, and Arun Kumar Yadav. 2022. Automatic Text Summarization Methods: A Comprehensive Review. *arXiv preprint arXiv:2204.01849* (2022).

[100] Qian Yang. 2018. Machine learning as a UX design material: how can we imagine beyond automation, recommenders, and reminders?. In *AAAI Spring Symposia*.

[101] Qian Yang, Justin Cranshaw, Saleema Amershi, Shamsi T Iqbal, and Jaime Teevan. 2019. Sketching nlp: A case study of exploring the right things to design with language intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[102] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating how experienced UX designers effectively work with machine learning. In *Proceedings of the 2018 designing interactive systems conference*. 585–596.

[103] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.

[104] Nur Yildirim, Alex Kass, Teresa Tung, Connor Upton, Donnacha Costello, Robert Giusti, Sinem Lacin, Sara Lovic, James M O'Neill, Rudi O'Reilly Meehan, et al. 2022. How Experienced Designers of Enterprise Applications Engage AI as a Design Material. In *CHI Conference on Human Factors in Computing Systems*. 1–13.

[105] Bowen Yu, Ye Yuan, Loren Terveen, Zhiwei Steven Wu, Jodi Forlizzi, and Haiyi Zhu. 2020. Keeping designers in the loop: Communicating inherent algorithmic trade-offs across multiple objectives. In *Proceedings of the 2020 ACM designing interactive systems conference*. 1245–1257.

[106] Sabah Zdanowska and Alex S Taylor. 2022. A study of UX practitioners roles in designing real-world, enterprise ML systems. In *CHI Conference on Human Factors in Computing Systems*. 1–15.

7 APPENDIX

Category	Content
Model description	The extractive summarization model uses natural language processing techniques to locate key sentences in an unstructured text document. These sentences collectively convey the main idea of the document. When a document is given as the input, the model returns a list of extracted sentences, together with a rank score and its position in the original document for each extracted sentence. A rank score is an indicator of how relevant or important the model considers the sentence is to the main idea of the document (between 0 and 1, higher as more relevant). By default, the model returns three highest scored sentences, and you can specify the number of sentences returned.
Examples of intended use	You might want to use the extractive summarization model to: • Distill critical information from lengthy documents. • Highlight key sentences in documents. • Quickly skim documents in a library. • Generate news feed content. You can use extractive summarization in multiple scenarios across a variety of industries. For example: • Extract key information from public news articles to produce insights. • Classify documents by their key contents. • Distill important information from long documents to empower solutions such as search, question and answering, and decision support.
Do not use (Unintended use)	Don't use extractive summarization for automatic actions without human intervention for high-impact scenarios. A person should always review source data when another person's economic situation, health, or safety is affected.
Limitations with impacting factors	Based on your scenario and input data, you could experience different levels of performance: • Because the model is trained on document-based texts, such as news articles, scientific reports, and legal documents, when used with texts in certain genres such as conversations and transcriptions, it might produce output with lower accuracy. • When used with texts that may contain errors or are less similar to well-formed sentences, such as texts extracted from lists, tables, charts, the model might produce output with lower accuracy.
Design Space Guidance	**Input**: How to align inputs to what work best for AI? **Output**: How to present AI outputs to users? **Failure**: How to handle AI errors and provide paths from failure? **Transparency**: How to support user understanding of AI and AI outputs? **Feedback**: How to support users providing feedback for AI to learn?
Harms considerations	Here is a description of general **technical limitations** and ***potential harms*** for summarization models. **Performance biases**: it may work less well on articles that are less structured, contain informal language, longer, or on topics that were less common in the training data. This could lead to ***disparate impacts*** for users reading different topics, sources, language styles, etc. **Structural biases**: it may be biased towards extracting from the beginning part of an article or paragraphs. This could lead to the ***erasure of perspectives*** or ***misinformation***. **Limits in extraction and linguistic quality**: it may fail to extract sentences with words that are out of the model's vocabulary. The extracted sentences may be incomplete or repetitive. This could lead to ***misinformation***, ***erasure of perspectives***, and low-quality even ***offensive content*** to the audience

Table 2: Content of model documentation presented to participants. Original images are presented in Section 3.

ID	Role	Years in profession	Experience with AI design	Experience with NLP	Gender
1	UX researcher	1–5 years	Part of my day-to-day job	Part of my day-to-day job	Female
2	HCI researcher	1–5 years	Part of my day-to-day job	Never	Female
3	Product Designer	5–10 years	Part of my day-to-day job	Part of my day-to-day job	Male
4	Product Designer	5–10 years	Part of my day-to-day job	Part of my day-to-day job	Female
5	Interaction designer	1–5 years	Part of my day-to-day job	Limited experience	Female
6	UI/UX designer	1–5 years	Part of my day-to-day job	Limited experience	Female
7	Designer	5–10 years	Part of my day-to-day job	Limited experience	Female
8	Product designer	5–10 years	Limited experience	Never	Female
9	Design lead	More than 10 years	I consider myself an expert	Part of my day-to-day job	Male
10	Product Manager	1–5 years	Part of my day-to-day job	Limited experience	Male
11	Product designer	5–10 years	Limited experience	Never	Male
12	Interaction designer	1–5 years	Part of my day-to-day job	Part of my day-to-day job	Female
13	Designer	More than 10 years	Part of my day-to-day job	Part of my day-to-day job	Male
14	User Researcher	1–5 years	Part of my day-to-day job	Part of my day-to-day job	Female
15	Product designer	More than 10 years	Limited experience	Limited experience	Female
16	UX researcher	5–10 years	Part of my day-to-day job	Limited experience	Female
17	Designer	5–10 years	Limited experience	Limited experience	Male
18	Interaction designer	1–5 years	Limited experience	Limited experience	Female
19	Product Designer	1–5 years	Part of my day-to-day job	Limited experience	Female
20	UX Designer	1–5 years	Limited experience	Never	Female
21	UX Designer	5–10 years	Part of my day-to-day job	Part of my day-to-day job	Male
22	Product manager	1–5 years	I consider myself an expert	Never	Male
22	Product manager	More than 10 years	Part of my day-to-day	I consider myself an expert	Female

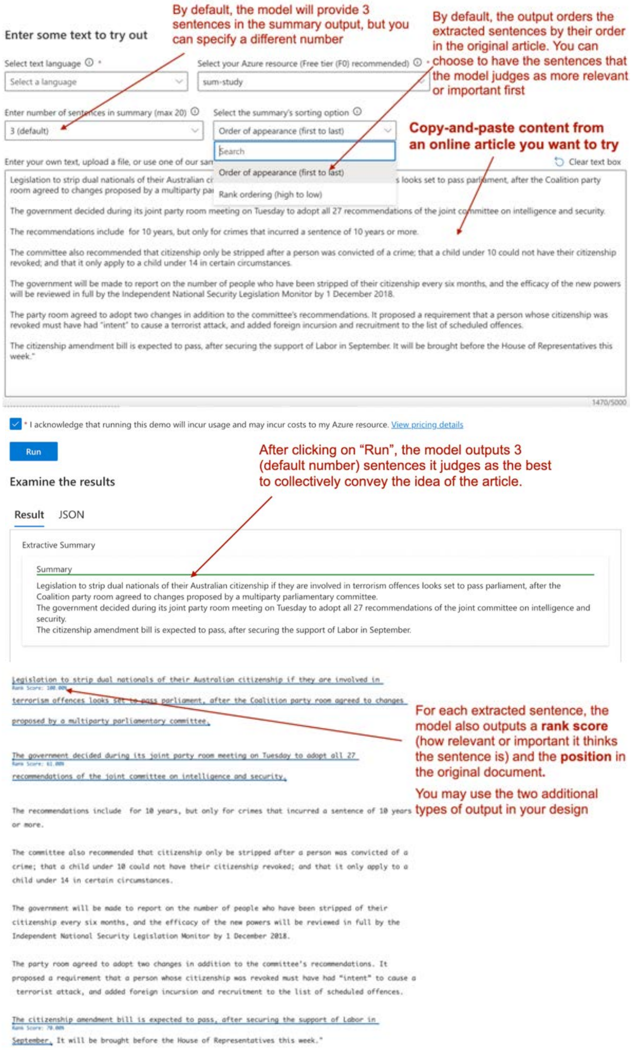
Table 3: Description of participants.

Information Category	Summary	Example Quotes
	Provided in the Task	
Harms considerations (N=13)	Participants appreciated the awareness of potential harms, and the delineation of different sources of technical limitations. However, many struggled with not having a concrete understanding and a lack of actionability to address these harms.	*This was really useful. I realized I was talking about performance biases and structural biases in the same way... that helped me think more granularly, how I should design to address each of these sources* (P2) *The potential harms... those were a little abstract...I have a hard time thinking about actual instances* (P4)
Impacting factors (N=12)	Some picked up the factor of article structure mentioned in this section and considered not applying the model to unstructured inputs. However, the majority expressed dissatisfaction because they wished to know whether other factors, such as article length, genre, and language style, can impact the model. Also wished to understand how the model behaves differently (e.g., output length, frequency of certain words), rather than just how the performance varies, with factors.	*I made a lot of assumptions that could be more well informed. Like is it better for things that are factual or opinion pieces?* (P12) *Will there be concept that is harder for the model [to extract]... what exactly are good for providing these kinds of output is not clear to me (P21)*
Examples from playground (N=11)	Appreciated that the output example provided an intuitive understanding of the model affordance. Experienced designers were intentional in examining different types of input-output pairs and looked for edge cases, often to explore the reliability of the model, and to discover or examine the effect of impacting factors.	*It is helpful to understand different ways the model could be used, like you don't have to just use the output, you can also rank the sentences, you can use the sentences within the context of the article* (P2) *I picked the statement of Ukraine because I'm assuming it talks about sensitive matters... Because I know language models are problematic when it comes to sensitive issues* (P3)
Design space guidance (N=8)	Appreciated it as a checklist to help them think systematically about what to design for, especially for those new to AI design. As both a generative tool for inspiring designs and an evaluative tool to ensure the design quality. Helpful for setting common languages and goals when communicating with other team members.	*The first is, as a generative tool...provides you a thing to think about, to apply to the designs that are in progress... Secondarily, it can provide a checklist for quality assurance.* (P9) *Designing with and for AI is a relatively emerging territory. So just even being able to flag that in a way that's shared across the team would be really useful.* (P7)
Do not use (N=7)	Appreciated documentation that leads with critical information. Some picked up the mentioning of avoiding use in "high-stakes" situations in their design thinking. However, the section was too high-level. Needed more examples of out-of-scope scenarios and understanding of outcomes.	*I personally look at AI from a very critical lens. So I naturally gravitate towards things that talk about limitations and do not use.* (P1) *I'm a little confused on high impact scenarios...what would be an example of something that might require human intervention?* (P2)
Examples of intended use (N=6)	Appreciated this section to help them jump-start divergent thinking and generating ideas.	*Examples of intended use was the one that weighed the most for me because looking over the examples then I can begin to extrapolate that and apply it to the problem I have* (P9)

Table 4: Summary of participants' comments about the categories of information provided in the task, ranked by the number of participants discussing each.

Information Category	Summary	Example Quotes
Additional information needed		
Explanation (N=13)	While some asked for “explainability,” most expressed such needs by asking the “how” and “why” questions, e.g., “how does the model summarize?” or “why does it extract this sentence?”. Also hypothesized the “how” from examples. Often interested in anticipating general patterns in model outputs or impacting factors. The access to explanations was considered an advantage of being able to talk to model developers directly.	*If I knew it was looking out for, like sentences that are very definitive, then I would understand that, this isn't going to work for an opinion piece.* (P12) *I would like to learn more about the model on how it's extracting... Like there are a lot of transitional words, how does those get filtered out?... so [I know] how is the output of the model like, how easy is it for users to consume?* (P19)
Training data (N=12)	Interested in understanding the training data because it could help them infer impacting factors. Also interested in the data shift—whether the training data matches data of their platform, to assess suitability and limitations.	*What types of articles that thing was trained on, what diversity of articles, what type of language like formal or informal?* (P12) *What it's been trained on or how it performs in relation to the types of articles being shared on this platform to evaluate that sort of match.* (P7)
(Disaggregated) evaluation (N=6)	Mostly interested in disaggregated evaluation to understand how performance varies by impacting factors. The quantification could help them better assess their potential impact.	*In the limitations it was highlighted a bit, but we need to quantify that... in the end, what matters is how will that impact the customer experience... if the model is 20% accurate for topics where it's weak then I want to know that so that I can avoid summarizing for those topics* (P10)
Confidence/ uncertainty (N=6)	Asked whether the model could generate confidence scores. Gravitated towards using it to put guardrails on low-quality outputs.	*I'd want to be able to say the degree of confidence in this. I don't know... I don't think that was documented* (P7)
Customizability, improvability, roadmap (N=5)	Interested in knowing whether the model could be customized or improved, and whether the service-provider plans to improve it in the future, to help them plan the design accordingly and coordinate or negotiate with the team.	*Is anything coming up in the future?...Because if you start building for these, then some new feature gets unlocked or constraints gets erased... So I wish we had been thinking ahead for what we wanted to design* (P13)
Analysis of output patterns (N=4)	Interested in understanding the general patterns of outputs, such as lengths and types of words.	*Descriptive statistics around the model output...like, is there a pattern? Does it take from the very beginning? How long is it usually?* (P2)
Algorithm and development background (N=4)	Wished to understand the background of the model to infer potential biases or mismatching assumptions for their use case. “Technical” knowledge could help designers build AI literacy and communicate with data scientists.	*Understanding who, when and how it was developed. I wanna know... what are their interests? What are their biases?* (P12) *The model type and algorithm because it's also about educating... so when we communicate with data scientists, we can use the same language* (P5)
Governance information (N=2)	Sought “delegated trust” by relying on their company or other organizations to vet the capabilities and ethical considerations of the model.	*Due diligence on the service provider. What have you done as bias mitigation? Are you an ethical actor? I wanna see some sort of assurance of that, [from] a third party that I can trust.* (P9)

Table 5: Summary of participants' comments about additional information participants asked for, ranked by the number of participants discussing each.



By default, the model will provide 3 sentences in the summary output, but you can specify a different number

By default, the output orders the extracted sentences by their order in the original article. You can choose to have the sentences that the model judges as more relevant or important first

Enter some text to try out

Select text language ⓘ *: Select a language

Select your Azure resource (Free tier (F0) recommended) ⓘ *: sum-study

Enter number of sentences in summary (max 20) ⓘ: 3 (default)

Select the summary's sorting option ⓘ: Order of appearance (first to last)

Search

Order of appearance (first to last)

Rank ordering (high to low)

Copy-and-paste content from an online article you want to try

Enter your own text, upload a file, or use one of our san

Clear text box

Legislation to strip dual nationals of their Australian ci s looks set to pass parliament, after the Coalition party room agreed to changes proposed by a multiparty pa

The government decided during its joint party room meeting on Tuesday to adopt all 27 recommendations of the joint committee on intelligence and security.

The recommendations include for 10 years, but only for crimes that incurred a sentence of 10 years or more.

The committee also recommended that citizenship only be stripped after a person was convicted of a crime; that a child under 10 could not have their citizenship revoked; and that it only apply to a child under 14 in certain circumstances.

The government will be made to report on the number of people who have been stripped of their citizenship every six months, and the efficacy of the new powers will be reviewed in full by the Independent National Security Legislation Monitor by 1 December 2018.

The party room agreed to adopt two changes in addition to the committee's recommendations. It proposed a requirement that a person whose citizenship was revoked must have had "intent" to cause a terrorist attack, and added foreign incursion and recruitment to the list of scheduled offences.

The citizenship amendment bill is expected to pass, after securing the support of Labor in September. It will be brought before the House of Representatives this week."

1470/5000

☑ * I acknowledge that running this demo will incur usage and may incur costs to my Azure resource. View pricing details

Run

After clicking on "Run", the model outputs 3 (default number) sentences it judges as the best to collectively convey the idea of the article.

Examine the results

Result JSON

Extractive Summary

Summary

Legislation to strip dual nationals of their Australian citizenship if they are involved in terrorism offences looks set to pass parliament, after the Coalition party room agreed to changes proposed by a multiparty parliamentary committee.
The government decided during its joint party room meeting on Tuesday to adopt all 27 recommendations of the joint committee on intelligence and security.
The citizenship amendment bill is expected to pass, after securing the support of Labor in September.

Legislation to strip dual nationals of their Australian citizenship if they are involved in
Rank Score: 100.00%
terrorism offences looks set to pass parliament, after the Coalition party room agreed to changes
proposed by a multiparty parliamentary committee.

The government decided during its joint party room meeting on Tuesday to adopt all 27
Rank Score: 61.00%
recommendations of the joint committee on intelligence and security.

The recommendations include for 10 years, but only for crimes that incurred a sentence of 10 years or more.

The committee also recommended that citizenship only be stripped after a person was convicted of a crime; that a child under 10 could not have their citizenship revoked; and that it only apply to a child under 14 in certain circumstances.

The government will be made to report on the number of people who have been stripped of their citizenship every six months, and the efficacy of the new powers will be reviewed in full by the Independent National Security Legislation Monitor by 1 December 2018.

The party room agreed to adopt two changes in addition to the committee's recommendations. It proposed a requirement that a person whose citizenship was revoked must have had "intent" to cause a terrorist attack, and added foreign incursion and recruitment to the list of scheduled offences.

The citizenship amendment bill is expected to pass, after securing the support of Labor in
Rank Score: 70.00%
September. It will be brought before the House of Representatives this week."

For each extracted sentence, the model also outputs a **rank score** (how relevant or important it thinks the sentence is) and the **position** in the original document.

You may use the two additional types of output in your design

Figure 4: Screenshot of the playground UI with an example input document and its model output, as retrieved in June 2022.