Advancements and Challenges in Generative AI: Architectures, Applications, and Ethical Implications

Flora Amato[1,*], Domenico Benfenati[1], Egidia Cirillo[1], Giovanni Maria De Filippis[1], Mattia Fonisto[1], Antonio Galli[1], Stefano Marrone[1], Lidia Marassi[1], Vincenzo Moscato[1], Narendra Patwardhan[1], Alberto Moccardi[1], Antonio Elia Pascarella[1], Antonio M. Rinaldi[1], Cristiano Russo[1], Carlo Sansone[1] and Cristian Tommasino[1,2]

[1]Department of Electrical Engineering and Information Technology (DIETI), University of Naples Federico II, Via Claudio 21, 80125 Naples, Italy
[2]Interdepartmental Center for Research on Management and Innovation in Healthcare (CIRMIS), University of Naples Federico II, Naples, Italy

Abstract

Architecture, classification, and major applications of Generative AI interfaces, specifically chatbots, are presented in this paper. Research paper details how the Generative AI interfaces work with various Generative AI approaches and show the architecture and their working. On the other hand, the generative model is built using advanced machine learning techniques to build dynamic, contextually relevant responses automatically. On the other hand, the retrieval-based model builds up with dependency on a predefined response library. The paper also discusses the use of Generative AI to populate Multimedia Knowledge Graphs (KGs), presenting technologies based on the semantic analysis of deep learning and NoSQL to more effectively integrate and retrieve data. The social and ethical challenges that come with the deployment of generative models are critically reviewed. These dialogues bring forward the balance that has to be maintained between progress and necessity in technological advancements, for which the call for ethical responsibility in developing AI is made. The paper presents a comprehensive review of state-of-the-art Generative AI with special focus on the promises and pitfalls in Generative AI research related to both natural language processing and knowledge management.

Keywords

artificial intelligence, Generative AI

1. Introduction

A chatbot, also known as a conversational agent, is an artificial intelligence (AI) software that can simulate a conversation (or a chat) with a user through text or voice interfaces [1]. Chatbots can use natural language processing (NLP) and machine learning algorithms to understand user inputs and generate appropriate responses, allowing them to provide assistance, automate tasks, and perform other functions without the need for human intervention.

The term "chatbot", short for "chatterbot", was originally coined by Michael Mauldin in 1994 to describe these conversational programs in his attempt to develop a Turing System [2].

This work aims to explore various techniques, approaches and technologies that have been utilized for developing chatbots since the late 1990s; furthermore, we will provide insights into the most common applications and use cases.

2. Architecture and Classification of Generative AI Interfaces

As a modern approach for architecture of Generative AI Interfaces, we will follow [3, 4, 5] and divide the intelligent interfaces structure proposed in the state of the art in four parts: the interface, the multimedia processor, the multimodal input analysis, and the response generator. In detail,

1. The **interface** is responsible for managing the interaction between the chatbot and users, which involves receiving inputs in various forms such as text or audio and returning appropriate responses.
2. The **multimedia processor** (optional) may be required to preprocess voice or video signals and

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy
*Corresponding author.
flora.amato@unina.it (F. Amato); egidia.cirillo@unina.it (E. Cirillo); mattia.fonisto@unina.it (M. Fonisto); antonio.galli@unina.it (A. Galli); stefano.marrone@unina.it (S. Marrone); lidia.marassi@unina.it (L. Marassi); vincenzo.moscato@unina.it (V. Moscato); narendra.patwardhan@unina.it (N. Patwardhan); alberto.moccardi@unina.it (A. Moccardi); antonioelia.pascarella@unina.it (A. E. Pascarella); carlo.sansone@unina.it (C. Sansone)
0000-0002-5128-5558 (F. Amato); 0009-0008-5825-8043 (D. Benfenati); 0009-0002-8395-0724 (G. M. D. Filippis); 0000-0001-6852-0377 (S. Marrone); 0009-0006-8134-5466 (L. Marassi); 0000-0002-4807-5664 (V. Moscato); 0000-0002-4807-5664 (N. Patwardhan); 0000-0002-1079-7741 (A. E. Pascarella); 0000-0001-7003-4781 (A. M. Rinaldi); 0000-0002-8732-1733 (C. Russo); 0000-0002-8176-6950 (C. Sansone); 0000-0001-9763-8745 (C. Tommasino)

convert them into text or recognize the user's tone to facilitate response generation.

3. The **multimodal input analysis unit** handles classification and data pre-treatment, often using natural language understanding (NLU) techniques such as semantic parsing, slot filling, and intent identification.
4. The **response generator** either associates a proper response for the given pre-processed input from a stored dataset or, using modern machine learning techniques, maps the normalized input to the output using a pre-trained model.

The response generator is the core component of a chatbot where the actual question-and-answer process takes place, and it can be considered as the "brain" of the system. Based on the architecture of the response generator, chatbot systems can be classified into two main categories: **retrieval-based chatbots**, which select their responses form a pre-defined set of possible outcomes, and **generative-based chatbots**, which use ML techniques to dynamically generate answers [6].

2.1. Retrieval-based chatbots

The goal of retrieval-based chatbots is to "understand" the user input and choose the most suitable responses from a knowledge dataset. There are four sub-categories of retrieval-based chatbots, which can be distinguished based on the architecture of their knowledge dataset and retrieval techniques. These categories are template-based, corpus-based, intent-based, and RL-based [5].

Template-based chatbots

Template-based chatbots select responses from a set of possible candidates by comparing the user input to certain query patterns.

Corpus-based chatbots

Although template-based chatbots have shown effectiveness in certain cases, their fundamental architecture necessitates scanning through all potential outputs for each input until the appropriate response is located. As a result, this approach can be slow and unsuitable for applications with a large knowledge dataset.

Intent-based chatbots

Intent-based chatbots utilize machine learning techniques to establish a connection between user inputs and pre-defined outputs. Typically, relevant data is collected and stored to establish associations between **user intents** (i.e., the conceptual meaning behind a user's request) and appropriate responses. Next, a pre-trained model leverages this information to link normalized user inputs with the most probable user intent [7].

RL-based chatbots

RL-based chatbots adopt **reinforcement learning** for response generation. Reinforcement learning itself is mainly based on the Markov decision process, i.e. a 4-tuple (S, A, P_a, R_a) where:

- $S = (s_1, s_2, ..., s_n)$ is a set of *states*, called the *state space*;
- $A = (a_1, a_2, ..., a_m)$ is a set of *actions*, called the *action space*;
- $P_a(s, s') = \Pr(s_{t+1} = s' | s_t = s, a_t = a)$ is the probability that action a, in the state s at step t will lead to state s' at step $t + 1$;
- $R_a(s, s')$ is the *reward* received after transitioning from state s to state s' when action a is performed.

The goal of a Markov decision process it to find a function $\pi(s)$ (generally called *policy*) that associate, for every state s_i, the action $\pi(s_t) = a_i$ which maximizes the overall reward, i.e. the following expectation value:

$$Q_\pi = E\left[\sum_{t=0}^{\infty} \gamma^t R_{\pi(s_t)}(s_t, s_{t+1})\right] \tag{1}$$

where γ is a coefficient (the *discount factor*) between 0 and 1 [8]. In RL-based chatbots, each state s_i corresponds to a specific turn in the conversation and is usually represented by an embedded vector. After the chatbot is trained, it is able to select the most appropriate response (action) a_i to ensure that the conversation remains relevant and coherent [9].

2.2. Generative-based chatbots

Generative-based chatbots have the advantage of being able to generate responses dynamically, which can lead to more natural and flexible conversations with users. Generative chatbots can generate novel responses, which means that they are not limited to pre-defined responses like retrieval-based chatbots. This flexibility allows them to provide more personalized and relevant responses.
Depending on the machine learning architecture used, we will discuss about **RNN-based chatbots** and **Transformer-based chatbots**.

RNN-based chatbots

One commonly used method for developing generation-based chatbots involves the use of two interconnected neural networks known as **recursive neural networks** (RNNs). The first network, called the **encoder**, is trained

to associate an input sentence with an intermediate vector called the **context vector**. The second network, called the **decoder**, takes the context vector as input and is trained to generate an output sentence, either by generating actual words or by using tokens. This approach is commonly referred to as "sequence-to-sequence" or **Seq2Seq** [6, 10].
As RNN-based chatbot responses are dynamically generated through machine learning models, they may be less precise and more uncertain than retrieval-based chatbots. For this reason, RNN-based chatbots are less commonly used in task- or knowledge-oriented scenarios and are instead more frequently used in entertainment and mental-health-related activities [5].

Transformer-based chatbots

A Transformer is a recent type of neural network architecture used for NLU and chatbots. First introduced in [11], is also used in other tasks such as language translation and text summarization. Transformers are based on the **self-attention mechanism**, which allows the model to learn which parts of the input sequence to attend to at each step of processing, based on the relevance of the other parts of the sequence to the current position. This is done through a process called *scaled dot-product attention*, where the model learns a set of weights to compute a weighted sum of the input sequence representations.
An important language model based on the Transformer architecture is the **Generative Pre-trained Transformer** (GPT), which was developed by OpenAI in 2020 [12]. GPT serves as the underlying architecture for the **ChatGPT chatbot**, which has gained widespread recognition for its ability to provide detailed and articulate responses across a variety of domains [13].

3. Multiquery Retrieval Augmented Generation

In the actual forefront of Generative Artificial Intelligence (Gen-AI) streamlining complex decision-making processes by enabling accessible and comprehensible tools to all users it is vitally important. The core of this section is relative to propose an alternative to the classical RAG, introduced by Lewis et al. in 2021 [14], enhancing its capabilities with a multiquery approach presenting a concise and solid architectural flow along with main evaluation metrics.

3.1. Methodology

This methodological section delves into the profound implications of leveraging Generative Artificial Intelligence (AI) to streamline and revolutionize complex decision-making processes, augmenting the power of cutting-edge technologies, enhancing the classical Retrieval-Augmented Generation (RAG) models. Through a meticulous exploration of a multi-query & human centred RAG application design, the access and the understanding to sophisticated AI capabilities, bridging the gap between technical expertise and practical application, is guaranteed. The culmination of this inquiry comes with a concise and robust architectural flow proposal, laying the groundwork for the seamless integration of multiquery-RAG solutions into decision-making processes and offering further insights that extends beyond the confines of this study and pave the way for future advancements in the field.

Question Generation Chain The multiquery-RAG system distinguishes itself through its ability to generate multiple variations of the original user query, in a human like fashion, through a specialized question generation chain that produces a prefixed number of alternative queries capturing distinct viewpoints and nuances associated with the original question. This diversification of the query set, if correctly fine-tuned, plays a pivotal role in surmounting the limitations of distance-based similarity searches in vector databases, ensuring a comprehensive and more efficient document retrieval process despite the classical retrieving process.

Answer Generation Chain Following the retrieval of information (documents), the system proceeds to generate answers by synthesizing and formulating responses using the data extracted from the documents and leveraging a wide LLMs systems. Contextualizing and elaborating on those information it ensures that the responses are both accurate and easily understandable for non-experts facilitating broader accessibility and utilization of the information among a wider audience.

3.2. Evaluation Criteria

This section outlines the principal metrics [15] that are integral for evaluating a Retrieval-Augmented Generation (RAG) in measuring different aspects of the system's performance as presented in figure [1].

Context Precision This metric evaluates the signal-to-noise ratio within the retrieved contexts measuring how many of the retrieved documents are actually relevant respect to the user's query.

Context Recall This metric assesses whether all necessary information required to answer the query has been

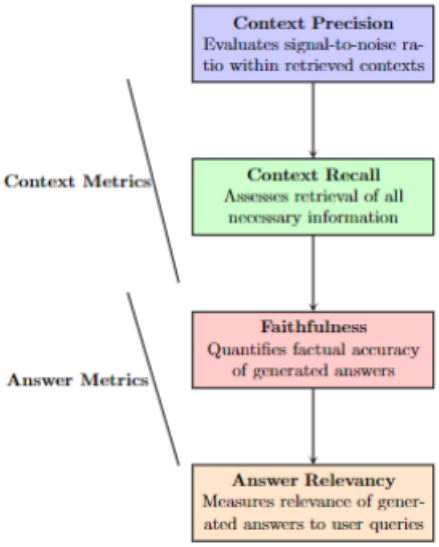

Figure 1: RAG Evaluation criterion

retrieved ensuring that the system's knowledge base covers all aspects needed to formulate a comprehensive and accurate response and relying on a comparison between the retrieved contexts and the ground truths.

Faithfulness This metric quantifies the factual accuracy of the answers generated by the RAG system. It involves counting the number of correct factual statements made in the generated answers based on the retrieved contexts and comparing this count to the total number of statements in the answers.

Answer Relevancy This metric measures how well the generated answers address the user's queries. For example, if a query asks for multiple pieces of information, the relevancy score reflects how completely the response addresses all elements of the query.

4. Multimedia Knowledge Graph population using Generative AI

Knowledge Graphs (KGs) serve as potent repositories, adeptly organizing, connecting, and extracting insights from many data sources, embodying contemporary knowledge management principles in semantic web applications [16]. Despite their invaluable utility, realizing the full potential of KGs necessitates a systematic population with relevant information, a task fraught with challenges, mainly when data is scarce [17].

Recent advancements, however, offer promising solutions. [18] and [19] present novel frameworks integrating semantic analysis, deep learning, and NoSQL technologies to extract entities from knowledge corpora, bridging the gap between textual and multimedia sources. Their approaches mark significant strides in enriching KGs with diverse data types, fostering more comprehensive knowledge representation and analysis.

Meanwhile, Chen et al. [20] propose a generative approach to the KG population, leveraging machine learning to establish relationships and reduce human intervention in the curation process. Training models to learn underlying data distributions and generate triplets regardless of entity pair co-occurrence in textual corpora pave the way for more efficient and scalable KG construction. This innovative approach streamlines the population process and broadens the scope of knowledge capture, enabling KGs to encapsulate a wider array of interconnected concepts and relationships.

Manual curation, though traditional, is labor-intensive and impractical in the face of expanding data landscapes [21]. To address this, a data-centric architecture harnessing generative deep-learning models emerges, automating KG creation, particularly for multimedia instances. By synthesizing multimedia data, irrespective of absolute data scarcity, a dynamic, infinitely expandable pool of instances is ensured, underpinning model training and inference with a multimedia knowledge graph that evolves alongside data trends.

Different knowledge graph population approaches with generative AI are based on standard steps. The first is grabbing information from curated textual sources. It is possible to enrich it by using Linked Open Data (LOD) and base the image's generation using the enhanced textual description to make the text as complete as possible. The next step combines the previously obtained textual statement and produces a representative multimedia instance of the input text via a generative text-image synthesis model. The last step consists of using a focused crawler, which allows a check on the quality of the generated image, exploiting different metrics useful to measure the degree of similarity of the generated image concerning its textual description and real images crawled from the web. If the image from the previous step exhibits metric values that surpass a threshold determined through experimental evaluation, it can be stored in the node of the multimedia knowledge base.

In image generation for the knowledge graph population, text-image synthesis models are developed to bridge the semantic gap between textual descriptions and corresponding visual representations. These models leverage cutting-edge generative strategies to produce high-quality images aligned with the provided textual prompts. The application of text-to-image models improved a lot in recent years, migrating from Generate Adversarial Net-

work (GAN) to Latent Diffusion Models, such as *Stable Diffusion* [22]. A latent diffusion model refines a latent representation by applying diffusion steps in the latent space, gradually reducing noise and revealing the desired image. This iterative process involves adding noise and updating the latent code. The model implements a decoder network to reconstruct the image from the refined latent code.

The evaluation phase of the quality of multimedia instances for the KG node is important. The evaluation process of text-to-image synthesis models involves assessing their accuracy in converting text inputs into synthetic images.

Some quantitative metrics are used to assess not only the quality of the image about the text but also the degree of realism in a generated image by comparing it to real images, such as Cosine Similarity, which compares the feature vectors, calculating the cosine between them, FID (Frechèt Inception Distance) [23], a numerical value that quantifies the similarity between the statistical distributions of real and generated images computing the Fréchet distance between the two distributions, and CLIP score [24], a metric that understands the relationship between images and text, used for evaluate the model's ability to rank images based on their relevance to a given textual description and vice versa.

5. Ethical and social challenges

The recent advances in generative AI are revolutionizing many sectors thanks to the ability to create original content based on patterns learned from training data. Models such as those based on transformer architectures, have already demonstrated significant success in various fields, including natural language processing, computer vision, and reinforcement learning. However, despite the advantages offered by generative models, their development and deployment raise concerns regarding ethical and environmental implications. Firstly, these models require massive computational resources and consume a large amount of energy during both training and execution processes. This raises concerns about the environmental impact of AI, especially considering the urgent need to reduce carbon emissions to address climate change. Additionally, there are ethical concerns regarding the use and management of training data. Since these models can generate original content, there is a risk that they may perpetuate biases or discriminations present in the training data, raising questions about fairness, privacy, and data security in the era of AI [25].

The Hominis project, conducted at the University of Naples Federico II in collaboration with industrial partners (DeepKapha), aims to advance toward sustainable and programmable AI solutions [26]. The project focuses on creating a concrete sustainable generative model, addressing crucial issues related to data collection, key model components, and essential additions. One of the main goals of the project is to improve model efficiency without compromising performance, using techniques such as attention and linear layer optimization within the Transformer architecture. Hominis also aims to ensure the sanitization of public data and develop data collection strategies to capture a wide range of multifaceted data. Additionally, the project involves developing tools for the community to analyze, curate, and critique datasets while ensuring fairness, privacy, and legality. The proposed methodologies, such as Universal Tokenization, Assisted Generation by Recovery (RAG), the use of diffusion to improve model controllability, and the use of muTransfer technique to optimize hyperparameters and reduce carbon footprint associated with training, all aim to improve the efficiency, sustainability, and fairness of AI models. In particular, the approach of unifying data through Universal Tokenization can help better manage data diversity, while RAG can improve model relevance and accuracy, ensuring greater fairness in outcomes. Furthermore, the use of diffusion to improve model controllability helps ensure that AI outputs are transparent and understandable. Today, attention to sustainable, adaptable, and responsible AI is crucial to ensure that the benefits of artificial intelligence are evenly distributed and that negative impacts, such as the carbon footprint associated with model training, are minimized. In an era where sustainable and responsible AI is essential for our future, projects like Hominis represent a step in the right direction, helping ensure that the benefits of AI are accessible to all while minimizing negative impacts on the environment and society.

Acknowledgments

This work was partially supported by PNRR MUR Project PE0000013-FAIR.

The FAIR project is committed to promoting an advanced vision of Artificial Intelligence, driving research and development in this crucial field and constantly keeping ethical, legal and sustainability considerations in mind

References

[1] G. Caldarini, S. Jaf, K. McGarry, A literature survey of recent advances in chatbots (2022). doi:https://doi.org/10.48550/arXiv.2201.06657.

[2] M. Mauldin, Chatterbots, tinymuds, and the turing test: Entering the loebner prize competition, 1994.

[3] S. A. Abdul-Kaer, J. Woods, Survey on chatbot design techniques in speech conversation systems,

International Journal of Advanced Computer Science and Applications, Vol. 6, No. 7 (2015).

[4] P. Jonell, Using social and physiological signals for user adaptation in conversational agents., Proceedings of the international joint conference on autonomous agents and multiagent systems, AAMAS (Vol. 4(c), pp. 2420-2422) (2019).

[5] B. Luo, R. Y. K. Lau, C. Li, Y. Si, A critical review of state-of-the-art chatbot designs and applications, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery Volume 12, Issue 1 (2022). doi:https://doi.org/10.48550/arXiv.2201.06657.

[6] H. Chen, X. Liu, D. Yin, J. Tang, A survey on dialogue systems: Recent advances and new frontiers, ACM SIGKDD Explorations Newsletter (2018). doi:https://doi.org/10.48550/arXiv.1711.01731.

[7] M. Franco, B. Rodrigues, E. J. Scheid, A. Jacobs, C. Killer, Z. G. L., S. B., Secbot: a business-driven conversational agent for cybersecurity planning and management, 16th International Conference on Network and Service Management (CNSM) (2020). doi:https://doi.org/10.23919/CNSM50824.2020.9269037.

[8] H. Cuayáhuitl, D. Lee, S. Ryu, Y. Cho, S. Choi, S. Indurthi, S. Yu, H. Choi, I. Hwang, J. Kim, Ensemble-based deep reinforcement learning for chatbots, Neurocomputing (2019). doi:https://doi.org/10.48550/arXiv.1908.10422.

[9] I. V. Serban, C. Sankar, M. Germain, S. Zhang, Z. Lin, S. Subramanian, T. Kim, M. Pieper, S. Chandar, N. R. Ke, S. Rajeshwar, A. de Brebisson, J. M. R. Sotelo, D. Suhubdy, V. Michalski, A. Nguyen, J. Pineau, Y. Bengio, A deep reinforcement learning chatbot, 2017. doi:https://doi.org/10.48550/arXiv.1709.02349.

[10] K. ho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014. doi:https://doi.org/10.48550/arXiv.1406.1078.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. doi:https://doi.org/10.48550/arXiv.1706.03762.

[12] A. Radford, K. Narasimhan, I. Salimans, T. ad Sutskever, Improving language understanding by generative pre-training, 2020.

[13] S. Lock, What is ai chatbot phenomenon chatgpt and could it replace humans?, 2022. URL: https://www.theguardian.com/technology/2022/dec/05/what-is-ai-chatbot-phenomenon-chatgpt-and-could-it-replace-humans.

[14] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. arXiv:2005.11401.

[15] S. Es, J. James, L. Espinosa-Anke, S. Schockaert, Ragas: Automated evaluation of retrieval augmented generation, 2023. arXiv:2309.15217.

[16] J. Zhang, M. Pourreza, R. Ramachandran, T. J. Lee, P. Gatlin, M. Maskey, A. M. Weigel, Facilitating data-centric recommendation in knowledge graph, in: 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC), IEEE, 2018, pp. 207–216.

[17] H. Li, G. Appleby, C. D. Brumar, R. Chang, A. Suh, Knowledge graphs in practice: Characterizing their users, challenges, and visualization opportunities, 2023. arXiv:2304.01311.

[18] M. Muscetti, A. M. Rinaldi, C. Russo, C. Tommasino, Multimedia ontology population through semantic analysis and hierarchical deep features extraction techniques, Knowledge and Information Systems 64 (2022) 1283–1303.

[19] A. M. Rinaldi, C. Russo, C. Tommasino, A novel approach to populate multimedia knowledge graph via deep learning and semantic analysis, in: Proceedings of the 14th International Conference on Management of Digital EcoSystems, 2022, pp. 40–47.

[20] H. Chen, C. Zhang, J. Li, P. S. Yu, N. Jing, Kggen: A generative approach for incipient knowledge graph population, IEEE Transactions on Knowledge and Data Engineering 34 (2022) 2254–2267. doi:10.1109/TKDE.2020.3014166.

[21] S. Issa, O. Adekunle, F. Hamdi, S. S.-S. Cherfi, M. Dumontier, A. Zaveri, Knowledge graph completeness: A systematic literature review, IEEE Access 9 (2021) 31322–31339.

[22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, 2022. arXiv:2112.10752.

[23] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, Advances in neural information processing systems 30 (2017).

[24] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, Y. Choi, Clipscore: A reference-free evaluation metric for image captioning, arXiv preprint arXiv:2104.08718 (2021).

[25] G. Tamburrini, et al., Digital humanism and global issues in artificial intelligence ethics., 2022.

[26] N. Patwardhan, S. Shetye, L. Marassi, M. Zuccarini, T. Maiti, T. Singh, Designing human-centric foundation models, reconstruction 9 (2023) 10.