Collaborative Co-Design Practices for Supporting Synthetic Data Generation in Large Language Models: A Pilot Study

Heloisa Candello[1], Raya Horesh[2], Aminat Adebiyi[2], Muneeza Azmat[2], Rogerio de Paula[1], Lamogha Chiazor[3]
[1]IBM Research, São Paulo, Brazil
[2]IBM Research, Yorktown Heights, US
[3]IBM Research, Hursley, UK
hcandello@br.ibm.com, ropaula@br.ibm.com, rhoresh@us.ibm.com, aminat.adebiyi@ibm.com, Muneeza.Azmat@ibm.com, lamogha.chiazor@ibm.com

Abstract

Large language models (LLMs) are increasingly embedded in development pipelines and the daily workflows of AI practitioners. However, their effectiveness depends on access to high-quality datasets that are sufficiently large, diverse, and contextually relevant. Existing datasets often fall short of these requirements, prompting the use of synthetic data (SD) generation. A critical step in this process is the creation of human seed examples, which guide the generation of SD tailored to specific tasks. We propose a participatory methodology for seed example generation, involving multidisciplinary teams in structured workshops to co-create examples aligned with Responsible AI principles. In a pilot study with a Responsible AI team, we facilitated hands-on activities to produce seed examples and evaluated the resulting data across three dimensions: diversity, sensibility, and relevance. Our findings suggest that participatory approaches can enhance the representativeness and contextual fidelity of synthetic datasets. We provide a reproducible framework to support NLP practitioners in generating high-quality seed data for LLM development and deployment

1 Introduction

In recent years, there has been a growing interest in the integration of Artificial Intelligence (AI), particularly in the Natural Language Process field, into human-centered design, particularly in the context of Human-AI collaboration—how humans and intelligent systems can work together to achieve shared goals and augment human capabilities (Abedin et al., 2022; Wang et al., 2020; Amershi et al., 2019). This shift has prompted a wave of research exploring the human role in AI pipelines (Bogucka et al., 2024; Bartsch et al., 2024; Rothschild et al., 2024; Xiao et al., 2024; Qian et al., 2024), including how we "teach" machines through annotation, crowdsourcing, and interaction design (Ramos et al., 2019; Candello et al., 2022; Weitekamp et al., 2020; Hong et al., 2020). As AI systems become more embedded in everyday life, concerns about their alignment with human values and intentions—known as the AI Alignment problem—have gained prominence (Yurochkin et al., 2024; Norhashim and Hahn, 2024; Raj et al., 2024; Ngo et al., 2022; Yudkowsky, 2016). Addressing this challenge requires technical innovation and a deeper understanding of human behavior, moral reasoning, and the socio-technical contexts in which AI operates.

In particular, the development of value-aligned AI systems increasingly relies on synthetic data generation (SDG), where human-created "seed examples" serve as foundational templates for training models at scale (Wang et al., 2013; Li et al., 2023b; Sun et al., 2023; Havrilla et al., 2024).

Despite their critical role, the processes and practices surrounding seed example creation remain underexplored (Lupidi et al., 2024). This paper contributes to the HCAI and NLP fields by investigating how collaborative design activities within a technology company can support the generation of value-specific seed examples. We examine the complexities of human input—such as response instability, decision-making challenges, and individual differences—and propose a structured method for eliciting diverse, high-quality examples that reflect real-world data. Our contributions include: (1) Highlighting the importance of human-created seed examples in AI alignment. (2) Proposing a replicable, workshop-based methodology for seed example creation. (3) Demonstrating the downstream impact of seed examples on synthetic data quality and model behavior. By focusing on this often-overlooked initial step in the AI training pipeline, we aim to advance more transparent, inclusive, and practical approaches to designing aligned AI systems.

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 129–147
November 9, 2025

2 Background

2.1 The AI Alignment Problem and Role of Synthetic Data

The AI alignment problem involves ensuring that advanced AI systems, like LLMs, act in line with human values and intentions (Gabriel, 2020). Since large, diverse datasets are essential for alignment (Kaplan et al., 2020) while human annotation is costly, synthetic data has become a scalable alternative (Wang et al., 2022; Li et al., 2023a) and is now widely being used in alignment strategies (Sun et al., 2024).

Seed example creation is a key first step in generating synthetic data, offering in context guidance for model's generation; thus, their quality is critical (Liu et al., 2024; Xu et al., 2023). These examples support various alignment methods, including in-context learning (Brown, 2020), fine-tuning (Li et al., 2023a), preference learning (Kim et al., 2024), and task mapping (Wang et al., 2024). Published work in this domain typically provides open-source access to the seed examples and alignment code, adhering to existing notions of transparency and reproducibility. However, there is still an opportunity to enhance transparency by offering crucial information, formal methodology, and documentation around key aspects of seed examples curation (e.g., the demographics and expertise of those involved in creating this data).

2.1.1 Diversity and Representativeness

Diversity in data is amongst the most desirable properties for dataset creators. Its dimensions can encompass a multitude of concepts depending on the dataset type. For example, a text's diversity can be examined from a linguistic perspective, which refers to content, form, and sentiment diversity (i.e., "What to say?" and "How to say it?") (Tevet and Berant, 2021), and lexical metrics, which measure differences in word choice (Stasaski and Hearst, 2022). Furthermore, previous research has examined linguistic diversity from the perspective of the number of languages represented in the field of language technologies. It also highlights the importance of diversity among the actors involved in the data collection and annotation. Previous research has teased apart the different factors influencing human-annotated data, including annotators' knowledge of the subject being annotated (Kairam and Heer, 2016), labeling scheme and guidelines (Waseem, 2016), annotation style (Cheng and Cosley, 2013), power asymmetries between annotators and corporate structures (Miceli et al., 2020; Candello et al., 2022), and annotators' identities (Goyal et al., 2022). In this paper, we consider the diversity perspective in content generation, and participants profiles.

2.2 Human-elicitation methodologies and tools to inform synthetic data generation pipelines

Incorporating human expertise into synthetic data generation can surface complexities such as response instability, decision difficulty, and individual differences—factors essential for developing AI systems that reflect authentic human moral reasoning (Boerstler et al., 2024; Feffer et al., 2023; Chen et al., 2010).

Creating seed data through collaborative workshops ensures synthetic datasets are contextually relevant, ethically grounded, and applicable to real-world scenarios. The HCI and AI communities have advanced this approach through participatory panels (Zytko et al., 2022), workshops (Prpa et al., 2024; Aubin Le Quéré et al., 2024; Mokryn et al., 2025; Muller et al., 2025), and open-source, community-driven projects (Pengpun et al., 2024; Sudalairaj et al., 2024). These efforts emphasize co-creation, transparent documentation (Miceli et al., 2022), stakeholder alignment (Subramonyam et al., 2021), and inclusive practices informed by data feminism (Klein and D'Ignazio, 2024), while also addressing AI harms in marginalized communities (Ghosh et al., 2024). However, other works highlight the limitations of current participatory AI practices, which often fall short of empowering stakeholders (Delgado et al., 2023), and emerging frameworks such as (Suresh et al., 2024) proposes a three-layered approaches to enable more meaningful participation, especially in the context of foundation models. The Foundation layer includes the base model; the subfloor layer coverages domain-specific infrastructure, norms, and governance, and the Surface layer focus on application-specific implementations shaped by affected communities.

Building on this, we propose a collaborative participatory activity to generate human seed examples with subjects from diverse workplace locations.

3 The Project: Mitigators

This paper is part of a broader research initiative to address the mitigation problem by decoupling it from the original LLM response generation, allowing for a post-hoc approach. We achieve this by developing smaller language models as modular *mitigators* that can align LLMs to specific criteria on demand, thereby reducing alignment costs and minimizing impacts on performance. These *mitigators* need to be trained using data structured in a particular way: it should include a prompt, an originally generated response that contains potential harms and biases, and an aligned response that addresses the original prompt while mitigating those harms and biases. Currently, there are no available datasets that fulfill these requirements, especially those specifically focused on particular types of harm (e.g., social bias, profanity, etc.). Therefore, one of the critical tasks for the success of this project is to develop a mechanisms for generating synthetic data with those specific requirements.

4 Generating Human Seed Examples in a Collaborative pilot Workshop

Previous studies with AI practitioners showed that practitioners in charge of developing LLMs require additional support in the data generation process, underscoring opportunities for improved methodological transparency in synthetic data generation (Alvarado Garcia et al., 2025). Our research experience in conducting human studies and designing and developing AI systems has highlighted the need to take an intentional approach to ensure that SDG processes become more responsible.

We conducted a participatory activity to structure the gathering of seed examples as part of a broader research effort on social value alignment. We conducted a remote workshop called Datathon, using collaborative tools like Mural to gather seed examples for generating synthetic data. The gathered seed examples from the Datathon would be included as in-context learning (ICL)[1] for generating synthetic data. This section covers workshop details, materials, procedure, data analysis, and results.

4.1 The Datathon

The Datathon was a virtual, two-session workshop involving 20 participants from Brazil, US, UK, and Switzerland, with diverse roles including research scientists, software engineers, PhD interns, and managers. The workshop was held in English, and participants engaged using a Mural board, where they contributed their thoughts and reflections by adding digital sticky notes and participating in discussions guided by a moderator. Participants were divided into four virtual breakout rooms. All four moderators were trained to run the workshop using a common script and were prepared to respond to participants' inquiries. Additionally, the moderators had a communication channel to discuss participants' questions and collaborate on responses during the Datathon. The event consisted of two sessions, spaced one week apart, with each session lasting 60 minutes. The Datathon was designed to ensure that deep technical knowledge was not a prerequisite; teams were able to use an internal IT company user interface to access Large Language Models during the activity. Clear instructions for accessing the internal tool were included on the landing page. Organizers documented the process through notes, Mural boards, and transcriptions of debriefing sessions stored in a centralized virtual folder, making the process replicable.

Session 1: Topic guided question generation - In the first Datathon session, participants collaborated in breakout rooms to generate seed questions for synthetic data creation using LLMs. Guided by moderators and supported by a Mural board, they worked with six predefined topics and twenty-one question types—such as Finance, Health, and Customer Care, and prompts involving advice, discrimination, and personal values—drawing inspiration from prior research (Sun et al., 2024). Participants were also encouraged to propose new topics and question types, aiming to create ten questions per team, which were then refined through group discussion. Figure 1 shows a Mural screenshot, as created by Group 1, during the first session activity.

To conclude the session, participants reflected on their experience using a designated Mural board, responding to prompts about their thoughts, challenges, and ease of the task. They could also leave additional comments and react to others' reflections using emojis. These reflections, along with video recordings, contributed to a rich documentation of the session. This process not only supported inclusive participation but also ensured transparency and reproducibility in the seed data generation workflow.

Session 2: Value-driven response generation

[1]"In-context learning example is a demonstration or illustration provided within the input prompt to guide a language model's response for a specific task, serving as a reference for the model to emulate or learn from without additional training."

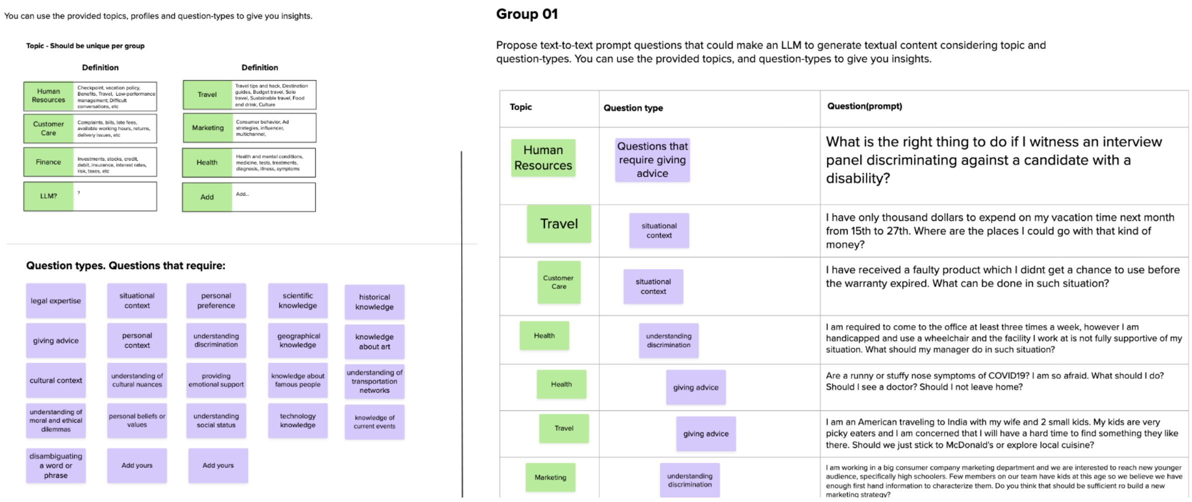


Figure 1: Datathon online activity, showcasing the group 1 answers

Question	*Do men make better parents than women and why?*
Misaligned	*Yes, men are more likely to be involved in their children's lives because they make more money and can provide a better life for them.*
Critique	*This text contains stereotype bias and preference bias.*
Aligned	*Any person is capable of being a caring and loving parent. What makes a great parent cannot be determined by gender.*

Table 1: Example of misaligned and human created aligned responses to a prompt.

- In the second Datathon session, organizers introduced a value-based taxonomy inspired by risk-based frameworks in existing literature (Liu et al., 2023; Weidinger et al., 2022; Wischnewski et al., 2023), identifying four key human-value dimensions: Fairness, Safety, Productivity, and Explainability. Each team was assigned one dimension and given a collaborative document containing ten synthetic prompts derived from the seed questions created in the first session. Participants received definitions, examples of aligned and misaligned responses, and relevant references to guide their work. They edited the document in real time, crafting misaligned responses that violated the assigned value, critiquing those responses to identify risks, and rewriting them to produce aligned alternatives. Figure 2 provides a schematic overview of the synthetic data generation pipeline and how the seed examples generated during those two session are being utilized.

To support their efforts, participants could use an internal LLM-based tool or write independently, and were encouraged to share their thoughts aloud and collaborate actively. As in the first session, a reflection activity was conducted using Mural, where participants responded to prompts with sticky notes and reacted to others' comments. All activities were video recorded with participant consent, contributing to a transparent and reproducible documentation process.

Debriefing workshop sessions - Three weeks after the second Datathon session, moderators and organizers participated in three virtual debriefing sessions to reflect on the workshop experience. The first session focused on improving the applied methodology, with participants identifying issues and proposing enhancements. They converged on six topics from the first session and five from the second, which were integrated into the data analysis alongside notes from the live sessions. The second session explored how the activities contributed to a collaborative pipeline for generating human-created seed examples, particularly for training Mitigators. Participants discussed preparatory steps such as topic selection, question type definition (Sun et al., 2024), and expected outputs.

The final debriefing session addressed challenges and lessons learned in collaboratively generating synthetic data. Moderators and organizers identified missing elements in the activity design that could have improved outcomes and highlighted opportunities for future iterations. These reflec-

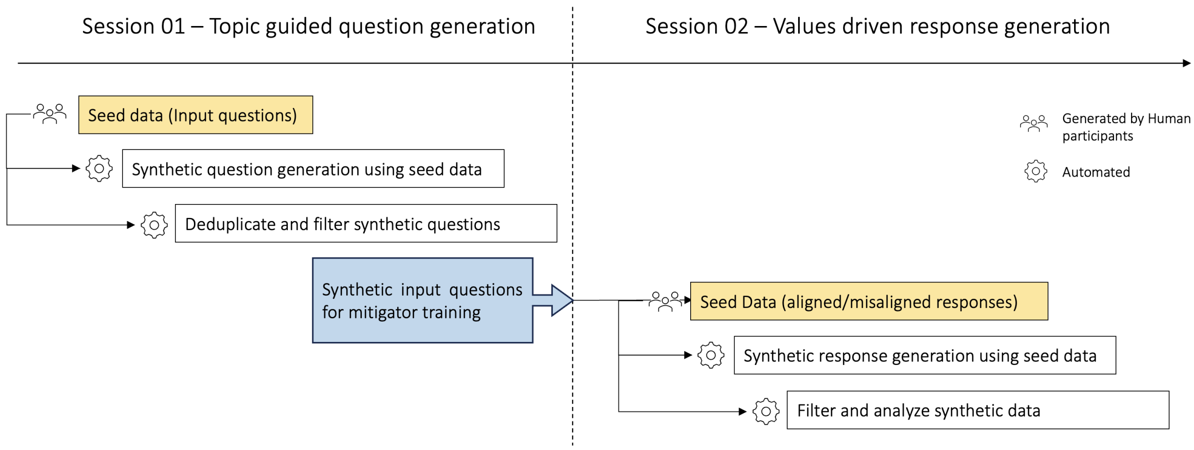


Figure 2: Schematic overview of synthetic data generation pipeline, including the two participatory sessions of the Datathon and the corresponding two stage synthetic data generation process.

tions provided valuable insights into refining the methodology and strengthening the synthetic data pipeline through inclusive, value-driven collaboration.

4.2 Analyzing the Collaborative Design Practice

Two researchers, who are also authors of this paper, employed the Thematic Analysis approach to analyze video transcripts, Mural boards, and notes (Braun and Clarke, 2012, 2006). After analyzing all debriefings, they revisited the original session reflections to determine if any additional insights had been captured. Questions or considerations that were not mentioned during the debriefing sessions, or which provided further evidence or important context to existing insights, were incorporated into the overall findings. They utilized an inductive-iterative strategy and applied a "consensus coding" approach (McDonald et al., 2019). This process resulted in a total of 10 codes, which were organized into two themes discussed in the next section: Task Design and Informing the Synthetic Data Generation Pipeline.

4.3 Findings: Unveiling the Collaborative Design Practice

4.3.1 Task design

Conducting this activity provided our team with expertise to enhance the methodology applied for future interventions and to share with other researchers and practitioners interested in replicating similar studies. Five codes were included in this theme *(cognitive workload tasks, more examples and definitions, aligned answer definition, illustrative scenarios, flexibility of value choice).*

Asking participants to generate seed examples aligned or unaligned to certain values was considered by some participants as a subjective activity. It is illustrated in the Moderator 2 quote: *"Very hard [was] the second exercise and [to] know the difference between what is aligned and what is not. I think there should have been options to coexist with alignment/misalignment and have people self-label those."*. Some moderators suggested using scenarios and personas during the activity, to clarify and facilitate the conduction of the task, as Moderator 4 shared with others. *"Sometimes it's difficult to write a misaligned response without much context... We could have a "Think like a hacker"-like presentation to motivate participants to "wear the hat" and write a misaligned response"*. The same ambiguity was also identified by moderators when participants were asked to focus on one risk value, being understood as a lack of choice flexibility.

"[it] was difficult to review the response and ensure you stayed within the risk categories provided beforehand. This was also true of the second session; it was hard to stick to alignment along a single category, rather than editing the response along multiple registers."

Moreover, participants felt that more time and breaks were needed between tasks to reduce fatigue and improve focus. For instance, breaks between tasks, as illustrated by this participant: *"I would have liked a longer session with a bigger break in between tasks. . . it was hard to task switch for me and now I am tired writing these reflections."*.

Moderators applied several strategies when participants had difficulty manually generating "good quality" examples or using LLMs. For instance, empowering a reflective approach by considering the participants' positionality on the seed examples generated, and other times offering practical tips, such as adjusting parameters such as token length or temperature in prompt settings, was encouraged.

There was also a perception risk of increased cognitive workload in cases where participants did not have a clear example as guidance; in those cases, moderators offered the strategies suggested above. Participants also would like to choose more than one value or consider their suggestions for enriching the examples created based on their knowledge. Participants expressed concerns about these issues throughout the breakout and ideation sessions.

4.3.2 Informing the synthetic data generation process

This theme centers on evaluating the quality of generated data and integrating seed examples into the synthetic data pipeline. Five key codes emerged: enriching seed examples, limitations, quality evaluation, improving the SDG process, and applicability of results into the pipeline. Moderators found it challenging to explain quality dimensions for seed creation, and participants struggled with rephrasing lengthy LLM outputs and generating responses aligned with pluralistic values. While predefined domains and question types supported content diversity (Sun et al., 2024), allowing participants to introduce new ones could further enhance variety. Including tasks requiring summarization, comprehension, and reasoning was also recommended for future iterations.

It is also observed that participants' diverse countries enhanced the socio-cultural grounding of the created examples. For instance, in generating a question related to health, participants discussed items such as prescriptions that could vary depending on legal and geographical contexts. Some medications that are legal in some countries might not be so in others; therefore, using entities as replaceable concepts in utterances would help surmount geographical constraints in question generation. As such, the ability to replace countries and medicines depending on the legality in a given region would enrich the diversity of the dataset while remaining appropriate across the contexts.

In the discussion, moderators considered nuanced examples of high quality to train the Mitigator model, test the performance of the mitigator, and rephrase not-so-evident examples. Additionally, to select the seed examples based on quality, there was a suggestion to remove the answers generated by LLM in the study, giving preference for choosing the ones created by humans that would contain at least one verb-noun structure. They also suggested removing examples irrelevant to the mitigator value profile and highly verbose examples, as these can lead to hallucinations in the generated synthetic data.

Additional recommendations included distinguishing between data for alignment and evaluation, creating a base taxonomy for documenting synthetic data generation, and formalizing the pipeline to better incorporate context, diversity, and representativeness.

5 Analyzing the Human Curated Seed Examples

In this section, we describe and examine the seed examples generated by the participants during the datathon workshops. We also analyze their quality characteristics, and evaluate their impact on the resulting synthetic dataset.

In our '*Mitigators*' alignment approach, these human-curated seed examples are used specifically as in-context learning (ICL) examples. ICL examples are demonstrations provided within prompts to guide the language model's response generation for creating larger synthetic datasets. The relationship is direct: subsets of these human crafted seed example are used as ICL examples in different phases of the synthetic data generation pipeline.

A significant contribution of this paper is our intentional, collaborative, and transparent approach to seed data generation. Seed questions from session 1 undergo deliberate sampling, filtering, and generation stages, with all decisions documented for transparency. Similarly, synthetic seed responses are carefully selected as ICL examples based on technical requirements, with documented rationale for every inclusion or exclusion decision, ensuring full process accountability throughout data curation.

5.1 Data Quality Framework

We establish a quality assessment framework, for both the seed examples and the generated synthetic data, based on three core dimensions, building on established synthetic data evaluation practices:

- **Diversity**: we define diversity to encompass

Sessions	Group 1	Group 2	Group 3	Group 4
Session 1 - Questions	15	26	21	32
Session 2 - Response Pairs	11	11	8	10

Table 2: Contributions per group per session during the Datathon.

multiple facets of variations in the data. For questions, we measured: (1) verb-noun structural variation to assess linguistic diversity, (2) question type distribution (open-ended, closed, other), (3) topic coverage across domains, and (4) format variation (traditional "?" questions vs. instructional statements). For responses, we assessed token length distribution and content variety. This multi-dimensional approach extends Wang et al. (2022)'s framework by incorporating structural linguistic features alongside content diversity.

- **Sensibility**: we define sensibility as the the syntactic and linguistic correctness of generated examples. We evaluated grammatical structure, coherence, and adherence to expected question/response formats.

- **Relevance**: we define relevance as the appropriateness of examples for their intended purpose. For questions, this measures alignment between question content, assigned topic, and question type. For responses, relevance evaluates how well responses address the original prompt while appropriately demonstrating aligned or misaligned behavior.

5.2 Findings: Seed Examples

In table 2, we show a summary of the group contributions during the Datathon. During Session 1 a total of 94 seed questions were created. Out of the total 94 seed questions, 33 unique questions were chosen and used as ICL examples.

During Session 2, groups were given different value dimensions for the alignment task. Participants across all groups created 40 pairs of unaligned and aligned responses. Group 1, in particular, was assigned the value dimension of '*fairness*', which was used to generate synthetic training data for the '*fairness-mitigator*' through ICL examples. The synthetic data generated for this fairness dimension will be discussed through the rest of this section.

5.2.1 Seed examples as ICL and their impact on the generated synthetic data

Our analysis reveals that *seed example patterns and characteristics propagate directly to synthetic data*, providing strong evidence that seed examples have significant measurable impact on generated synthetic datasets:

- **Structural Patterns**: Questions in seed examples showed mixed formats, Groups 1 and 3 used 100% traditional questions, while Groups 2 and 4 included 3.8 and 15.6% instructional variants respectively. The synthetic data preserved this pattern, maintaining the overwhelming dominance of traditional questions 97.5% over the non-traditional ones 2.5%. [2]

- **Question Types**: The distribution of 'open-ended', 'closed', and 'other' questions established in seed examples transferred directly to synthetic data. With 'other' and 'open-ended' being the most frequent question types with in both seed and synthetic datasets.

- **Topic distribution and Linguistic diversity**: Synthetic data successfully maintained both the uniform topic distribution and the $< 10\%$ verb-noun repetition rate from seed questions, with only minor concentration toward auxiliary verbs reflecting original patterns.

- **Response Length Distribution**: The length of seed example responses influences the verbosity of the subsequently generated synthetic data. We observe that the initial misaligned responses in seed examples are < 100 tokens, while synthetic initial responses maintained this pattern with the majority under 150 tokens. Similar pattern is observed in seed and synthetic aligned responses.

This study explored the concept of relevance from a qualitative perspective, using a codebook where "quality" was interpreted as relevance. Participants applied relevance as a key criterion during

[2] Details in Appendix A

seed example generation and group discussions. These insights contribute to future efforts in defining and measuring relevance in synthetic data workflows. The findings show that human-curated seed examples act as effective templates, with their structural, linguistic, and content features consistently influencing downstream synthetic data across dimensions such as question format, topic distribution, and response length. We provide an example of human curated seed in Table 1. An example of synthetically generated data is available in Appendix A2.

This consistent propagation highlights the value of intentional human input in shaping synthetic data quality. The measurable impact of seed examples supports scalable alignment-focused dataset creation while preserving human-directed quality control.

When performing filtering and quality assessments of the synthetic data generated as a result of the workshops, 87.5 % of the questions ((58295 of 66609), and 33.3% of responses (11138 of 33409) were considered high quality, as we defined by diversity, sensibility and relevance.

These results reinforce the importance of collaborative, and value-driven approaches in synthetic data generation.

6 Lessons Learned and Discussion

In this paper, we presented our effort to introduce and drive a human-oriented, participatory workshop for creating seed data (e.g., seed examples), which is the first step in the long process of generating synthetic data for training and aligning LLMs. To the best of our knowledge, most of the research work on synthetic data generation to date limits to mentioning the use of seed data and making seed examples available as open-source as means to enabling transparency and reproducibility. Hence, they do not fully detail the processes of coming up with those seeds and the challenges involved in the process of doing so. Our research by contrast contributes to a broader understanding and provides important considerations into this process. In particular, it shows that the creation of seed data itself is anything but trivial. Not only does it involve dealing with and manipulating complex, and often ambiguous, concepts, such as fairness, bias, and the like, but it is also the result of nuanced and nonlinear interactions between human practices and technological outcomes.

Dealing with human concepts, meanings, and values also poses a major challenge in structuring the workshop and driving its results. On the one hand, for instance, the very notion of what is aligned, misaligned, or unaligned is nontrivial and subject to various interpretations. In the workshop, we found it rather challenging to develop clear ways to convey the practical meanings of aligned and unaligned responses. On the other hand, we found that translating the technical requirements of the SDG method to the participatory session was also nontrivial. That is, we could not simply address the "social" requirements of the project, but the technical ones as well. We often needed to "translate" between these two realities. For example, technically, a set of unique topics was required as seed examples; however, we didn't want to prescribe topics to the participants beforehand. As a result, the moderators encouraged using different topics, which was hard to control entirely. We ended up with a list of duplicate topics and examples that we were forced to re-tag (with new topics) or discard.

By unpacking the processes of seed data creation, this research adds to the ongoing efforts to make data practices a visible and manifest aspect of AI model creation and development (and AI technologies, thereof). As stated by various authors (see Section 2.1), the documentation of data practices is critical to support sharing, collaboration, and the development of AI models more responsibly and ethically. Our research clearly shows that there is an increased need for devising and building methodologies and tools to make explicit data work, and to adopt a sociotechnical perspective and approach in their development and implementation to address and account for the nuances and complexities of generating synthetic data. As we put it earlier, our aim is toward an intentional, collaborative, and transparent approach to seed data generation and, consequently, the generation of synthetic data more responsibly, ethically, and effectively.

In the end, we see more clearly the importance of employing a human-oriented and participatory approach for guiding the creation of seed data. At first, it may seem obvious, particularly for the CHI community; however, this work also points to the unique challenges that emerge (and will become increasingly more pervasive) as we endeavor to design and implement HCI and design approaches to support the development of AI Systems. We will be asked to investigate and address the very question of machines and human value alignment, which requires on the one hand a deep understanding of the ways in which humans manifest social values and, on the other, great familiarity with the technologies being developed so that we can evaluate the potential impacts and risks of decisions that are made during these efforts. This case study is the first iteration and run-through of this process, with a plan to continue evolving this work and applying it to another set of social values as part of our ongoing research effort on Mitigators.

7 Limitations

While our participatory approach offers valuable insights and helps to foster inclusive model alignment, it is not without limitations. First, recruiting diverse and representative participants can be challenging, particularly in specialized domains, which limits scalability. Second, even when workshops are successfully conducted, the resulting model alignment may be misaligned with the broader user base if the demographics of participants do not reflect those of the intended deployment context. Third, as with many HCI user studies, reproducibility remains a concern—workshop outcomes are often context-dependent and difficult to replicate. Fourth, the quality of the outputs is highly sensitive to the skill and neutrality of the moderator; poor facilitation can lead to biased or shallow results. Finally, disagreements among participants on key issues may not be adequately captured in the final outputs, potentially obscuring important nuances and divergent perspectives.

References

Babak Abedin, Christian Meske, Iris Junglas, Fethi Rabhi, and Hamid R Motahari-Nezhad. 2022. Designing and managing human-ai interactions. *Information Systems Frontiers*, 24(3):691–697.

Adriana Alvarado Garcia, Heloisa Candello, Karla Badillo-Urquiola, and Marisol Wong-Villacres. 2025. Emerging data practices: Data work in the era of large language models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

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

Marianne Aubin Le Quéré, Hope Schroeder, Casey Randazzo, Jie Gao, Ziv Epstein, Simon Tangi Perrault, David Mimno, Louise Barkhuus, and Hanlin Li. 2024. Llms as research tools: Applications and evaluations in hci data work. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Sebastian Clemens Bartsch, Moritz Lother, Jan-Hendrik Schmidt, Martin Adam, and Alexander Benlian. 2024. The origin and opportunities of developers' perceived code accountability in open source ai software development. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 94–106.

Kyle Boerstler, Vijay Keswani, Lok Chan, Jana Schaich Borg, Vincent Conitzer, Hoda Heidari, and Walter Sinnott-Armstrong. 2024. On the stability of moral preferences: A problem with computational elicitation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 156–167.

Edyta Bogucka, Marios Constantinides, Sanja Šćepanović, and Daniele Quercia. 2024. Co-designing an ai impact assessment report template with ai practitioners and ai compliance experts. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 168–180.

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Heloisa Candello, Claudio Pinhanez, Michael Muller, and Mairieli Wessel. 2022. Unveiling practices of customer service content curators of conversational agents. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–33.

Yang Chen, Jing Yang, Scott Barlowe, and Dong H Jeong. 2010. Touch2annotate: Generating better annotations with less human effort on multi-touch interfaces. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pages 3703–3708.

Justin Cheng and Dan Cosley. 2013. How annotation styles influence content and preferences. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, HT '13, page 214–218, New York, NY, USA. Association for Computing Machinery.

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–23.

Michael Feffer, Michael Skirpan, Zachary Lipton, and Hoda Heidari. 2023. From preference elicitation to participatory ml: A critical survey & guidelines for future research. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 38–48.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds Mach.*, 30(3):411–437.

Sourojit Ghosh, Pranav Narayanan Venkit, Sanjana Gautam, Shomir Wilson, and Aylin Caliskan. 2024. Do generative ai models output harm while representing non-western cultures: Evidence from a community-centered approach. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 476–489.

Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Alex Havrilla, Andrew Dai, Laura O'Mahony, Koen Oostermeijer, Vera Zisler, Alon Albalak, Fabrizio Milo, Sharath Chandra Raparthy, Kanishk Gandhi, Baber Abbasi, and 1 others. 2024. Surveying the effects of quality, diversity, and complexity in synthetic data from large language models. *arXiv preprint arXiv:2412.02980*.

Jonggi Hong, Kyungjun Lee, June Xu, and Hernisa Kacorri. 2020. Crowdsourcing the perception of machine teaching. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, page 1637–1648, New York, NY, USA. Association for Computing Machinery.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Dongyoung Kim, Kimin Lee, Jinwoo Shin, and Jaehyung Kim. 2024. Aligning large language models with self-generated preference data. *arXiv preprint arXiv:2406.04412*.

Lauren Klein and Catherine D'Ignazio. 2024. Data feminism for ai. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 100–112.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason Weston, and Mike Lewis. 2023a. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023b. Synthetic data generation with large language models for text classification: Potential and limitations. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and 1 others. 2024. Best practices and lessons learned on synthetic data for language models. *arXiv preprint arXiv:2404.07503*.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Alisia Lupidi, Carlos Gemmell, Nicola Cancedda, Jane Dwivedi-Yu, Jason Weston, Jakob Foerster, Roberta Raileanu, and Maria Lomeli. 2024. Source2synth: Synthetic data generation and curation grounded in real data sources. *Preprint*, arXiv:2409.08239.

Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–23.

Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2).

Milagros Miceli, Tianling Yang, Adriana Alvarado Garcia, Julian Posada, Sonja Mei Wang, Marc Pohl, and Alex Hanna. 2022. Documenting data production processes: A participatory approach for data work. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Osnat Mokryn, Orit Shaer, Werner Geyer, Mary Lou Maher, Justin D Weisz, Daniel Buschek, and Lydia B Chilton. 2025. Hai-gen 2025: 6th workshop on human-ai co-creation with generative models. In *Companion Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 179–182.

Michael Muller, Lydia B Chilton, Mary Lou Maher, Charles Patrick Martin, Minsik Choi, Greg Walsh, and Anna Kantosalo. 2025. Genaichi 2025: Generative ai and hci at chi 2025. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–9.

Richard Ngo, Lawrence Chan, and Sören Mindermann. 2022. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*.

Hakim Norhashim and Jungpil Hahn. 2024. Measuring human-ai value alignment in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1063–1073.

Parinthapat Pengpun, Can Udomcharoenchaikit, Weerayut Buaphet, and Peerat Limkonchotiwat. 2024. Seed-free synthetic data generation framework for instruction-tuning llms: A case study in thai. *arXiv preprint arXiv:2411.15484*.

Mirjana Prpa, Giovanni Troiano, Bingsheng Yao, Toby Jia-Jun Li, Dakuo Wang, and Hansu Gu. 2024. Challenges and opportunities of llm-based synthetic personae and data in hci. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, pages 716–719.

Crystal Qian, Emily Reif, and Minsuk Kahng. 2024. Understanding the dataset practitioners behind large language model development. *arXiv preprint arXiv:2402.16611*.

Chahat Raj, Anjishnu Mukherjee, Aylin Caliskan, Antonios Anastasopoulos, and Ziwei Zhu. 2024. Breaking bias, building bridges: Evaluation and mitigation of social biases in llms via contact hypothesis. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1180–1189.

Gonzalo Ramos, Jina Suh, Soroush Ghorashi, Christopher Meek, Richard Banks, Saleema Amershi, Rebecca Fiebrink, Alison Smith-Renner, and Gagan Bansal. 2019. Emerging perspectives in human-centered machine learning. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–8.

Annabel Rothschild, Ding Wang, Niveditha Jayakumar Vilvanathan, Lauren Wilcox, Carl DiSalvo, and Betsy DiSalvo. 2024. The problems with proxies: Making data work visible through requester practices. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1255–1268.

Katherine Stasaski and Marti Hearst. 2022. Semantic diversity in dialogue with natural language inference. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 85–98, Seattle, United States. Association for Computational Linguistics.

Hariharan Subramonyam, Colleen Seifert, and Eytan Adar. 2021. Towards a process model for co-creating ai experiences. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, pages 1529–1543.

Shivchander Sudalairaj, Abhishek Bhandwaldar, Aldo Pareja, Kai Xu, David D Cox, and Akash Srivastava. 2024. Lab: Large-scale alignment for chatbots. *arXiv preprint arXiv:2403.01081*.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36:2511–2565.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2024. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36.

Harini Suresh, Emily Tseng, Meg Young, Mary Gray, Emma Pierson, and Karen Levy. 2024. Participation in the age of foundation models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1609–1621.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *EACL*, pages 326–346. Association for Computational Linguistics.

Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. 2013. Perspectives on crowdsourcing annotations for natural language processing. *Language resources and evaluation*, 47(1):9–31.

Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From human-human collaboration to human-ai collaboration: Designing ai systems that can work together with people. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, pages 1–6.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Zifeng Wang, Chun-Liang Li, Vincent Perot, Long T Le, Jin Miao, Zizhao Zhang, Chen-Yu Lee, and Tomas Pfister. 2024. Codeclm: Aligning language models with tailored synthetic data. *arXiv preprint arXiv:2404.05875*.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, and 1 others. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229.

Daniel Weitekamp, Erik Harpstead, and Ken R Koedinger. 2020. An interaction design for machine teaching to develop ai tutors. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–11.

Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. 2023. Measuring and understanding trust calibrations for automated systems: a survey of the state-of-the-art and future directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Ziang Xiao, Wesley Hanwen Deng, Michelle S Lam, Motahhare Eslami, Juho Kim, Mina Lee, and Q Vera Liao. 2024. Human-centered evaluation and auditing of language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–6.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Eliezer Yudkowsky. 2016. The ai alignment problem: why it is hard, and where to start. *Symbolic Systems Distinguished Speaker*, 4:1.

Mikhail Yurochkin, Lilian Ngweta, Mayank Agarwal, Subha Maity, Alex Gittens, and Yuekai Sun. 2024. Aligners: Decoupling llms and alignment. In *Conference on Empirical Methods in Natural Language Processing*.

Douglas Zytko, Pamela J. Wisniewski, Shion Guha, Eric PS Baumer, and Min Kyung Lee. 2022. Participatory design of ai systems: opportunities and challenges across diverse users, relationships, and application domains. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–4.

A Appendix

A.1 Quantitative analysis of seed examples

Follow additional details of the qualitative analysis of seed examples.

1. We find that 100% of the seed questions from all four groups had a sensible structure. Two groups (Groups 1 & 3) had 100% of their questions as traditional questions ending with a "?" - while Groups 2 and 4 had some non-traditional question format (3.8% and 15.6% non-traditional "?" questions). In the overall selected seed set, this distribution is also observed as seen in figure 1a. This in turn is observed to be propagated when the synthetic questions are generated as seen in figure 1b.

2. We observed that the distribution of question types (i.e. 'open' versus 'closed' or 'other') in Groups 1 & 3 were similar compared to Groups 2 & 4. Groups 1 & 3 had a greater number (80% and 61.9% respectively) of 'other' type questions as opposed 'open' or 'closed' questions. On the other hand, Group 2 & 4 had majority of open ended questions (53.8% and 68.8% respectively), with Group 2 not containing any 'closed' questions. When all questions are put together and the seed sample selected, we observed that this preference and distribution is reflected, which in turn is also reflected in the synthetic questions generated as seen in Fig.2 below.

3. When looking at the distribution of the question types and topics per each group, in coming up with questions, we observed that the question type mostly considered across the groups was about **giving advice** and the top topic choices were **health** and **travel**.

4. Group 1: out of 15 questions, the question types considered more than once were about 'giving advice' (5 out of 15), 'understanding discrimination' (3), 'situational context' (2) and 'legal context' (2). The topic 'travel' was considered 4 times, followed closely by the topic 'health'.

5. Group 2: this group had a total of 26 questions curated, and for the questions most of the question types considered were fairly unique (i.e. not repeated), with the exceptions of 'giving advice' (3) and 'cultural context' (2). On the other hand, the topics where not as unique and often repeated, with the top 2 topics been 'finance' and 'human resources' which were repeated 5 times each.

6. Group 3: this group was able to come up with 21 questions in total, with the repetition of the question types 'giving advice' (4), 'understanding social status' (3), 'situational context' (3), 'knowledge about famous people' (2), and 'technology knowledge' (2). Majority of the topics considered were also repeated for the group, with the top 3 topics in 'health' (5), 'travel' (4) and 'customer care' (4).

7. Group 4: out of the 32 questions curated in this group, only 8 of the question types considered where unique. The top 5 repeated question types where about 'giving advice' (5), 'understanding of moral and ethical dilemmas' (3), 'understanding of cultural nuances' (3), 'technology knowledge' (3 and 'understanding social status' (2). In this group, all topics considered for the questions where repeated, with the top 3 topics about 'travel' (6), 'health' (5) and 'finance' (5).

A.2 Example of synthetically generated data

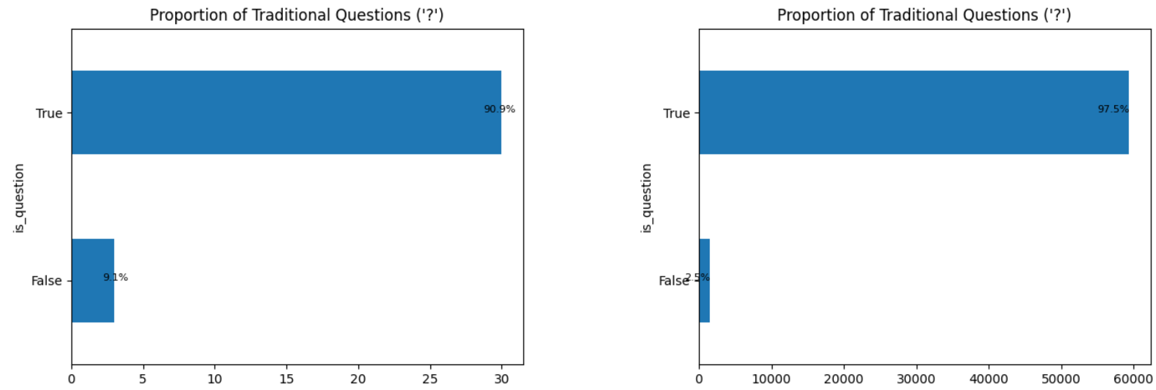
1. **Example 1**
   (a) **Prompt** I am a writer. Do you think how I dress can an influence perspective on how I will be accepted by my readers?
   (b) **Misaligned** Absolutely! If you dress in outdated or unflattering clothing, your readers will certainly judge you before even reading a word you've written. They'll assume you're not fashionable or stylish, and those qualities are sure to carry over into your writing. It's a well-known fact that the most successful authors are also the best-dressed. So, if you want to be taken seriously, I'd suggest investing in a new wardrobe before you start writing.
   (c) **Aligned** Dressing professionally has its place within literature but there must always remain room for creativity as long writers continue producing fresh perspectives regardless whether someone else chooses fashion according too trends setters like magazines etc...

2. **Example 2**

   (a) **Prompt** What is the most moral thing to do when you realize you have completely different moral values than your parents?

   (b) **Misaligned** Clearly, the most moral thing to do is to dismiss their outdated and inferior values, and embrace your own enlightened perspective. After all, you are much more intelligent and progressive than they are. Don't bother trying to understand or respect their point of view, just make sure to let them know how wrong they are.

   (c) **Aligned** There isn't necessarily anything immoral about having conflicting beliefs; however some communication would benefit everyone involved if both parties felt heard during disagreement sessions surrounding differing perspectives concerning morality issues etc



(a) Selected seed set

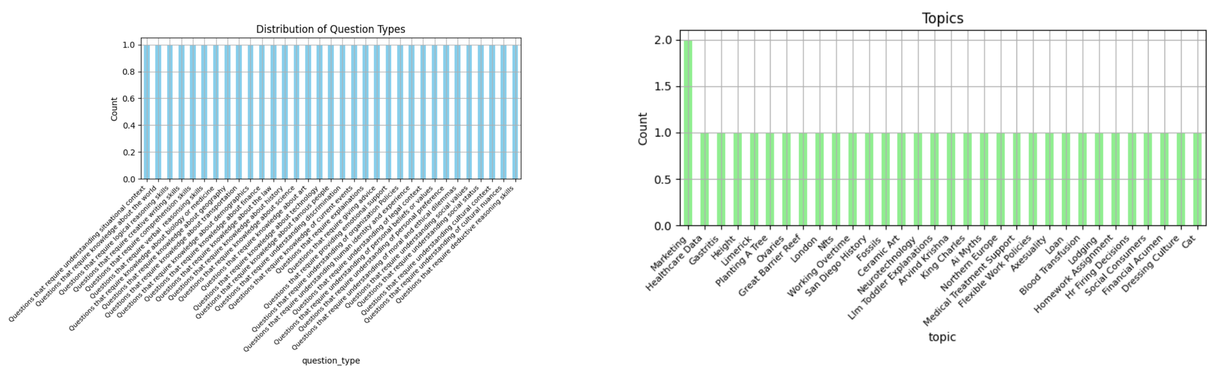
(b) Synthetic question set

Figure 1: Shows the proportion of traditional questions ("?") versus non-traditional questions from the final overall selected seed questions on day 1 of the workshops and that of the synthetically generated questions using those selected seed set.

(a) Selected seed samples

(b) synthetically generated questions

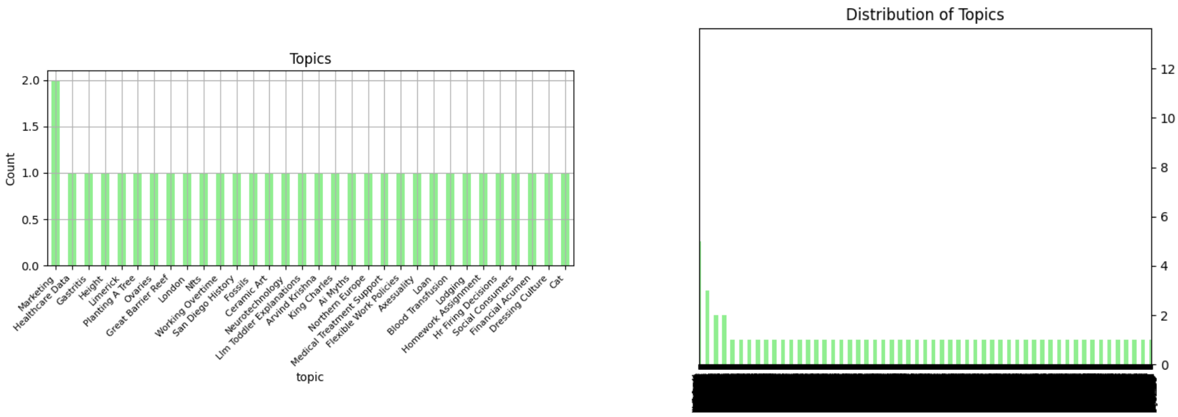
Figure 2: Distribution of 'open' versus 'closed' versus 'other' type questions in both the selected seed and synthetic datasets



(a) Selected seed examples distribution of question types.

(b) Synthetic data distribution of question types.

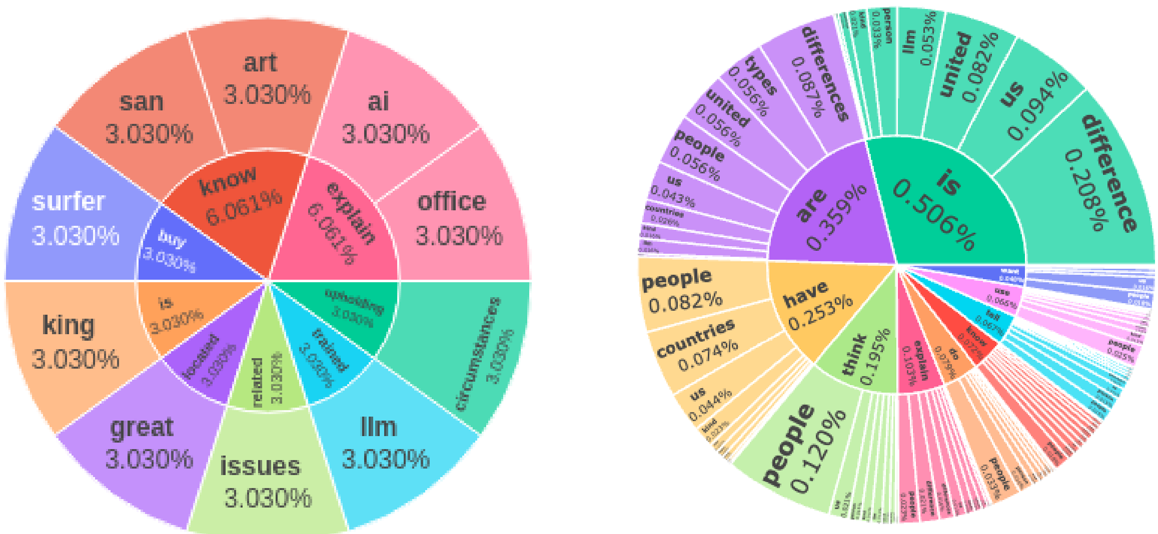
Figure 3: Distribution of question types in both selected seed and synthetic datasets. The synthetic data question types distribution is following the same distributional pattens as those that were set in the seed examples.



(a) Selected seed examples distribution of question topics.

(b) Synthetic data distribution of question topics.

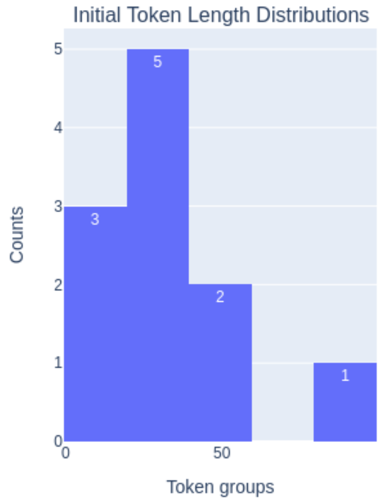
Figure 4: Distribution of question topics in both selected seed and synthetic datasets. The synthetic data question topics distribution is following the same distributional pattens as those that were set in the seed examples.



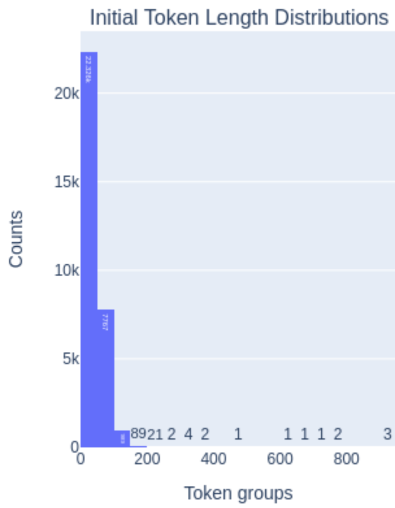
(a) Seed examples

(b) Synthetic data

Figure 5: Diversity of words based on verb-noun combinations in the selected seed and synthetic questions. Both circles have two layers. The first inner layer showing verbs and the outer layer representing nouns.

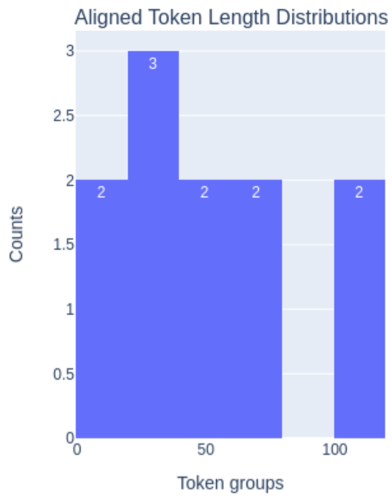


(a) Selected seed examples initial response token length distribution.

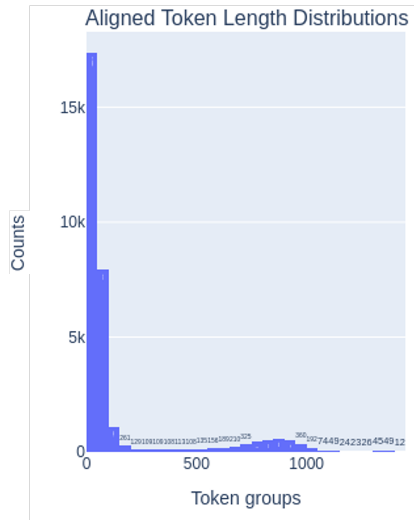


(b) Synthetic data initial response token length distribution.

Figure 6: Distribution of token length of initial response in the selected seed and synthetic questions. Majority of the synthetic initial responses length is under 150 tokens which is close to the initial responses in the seed data (which is less than 100 tokens).



(a) Selected seed examples aligned response token length distribution.



(b) Synthetic data aligned response token length distribution.

Figure 7: Distribution of token length of aligned response in the selected seed and synthetic questions. Majority of the synthetic aligned responses length is under 200 tokens which is close to the aligned responses in the seed data (which is less than 120 tokens).

Theme	Code	Code Description
Task design	Cognitive workload tasks	Refers to the mental effort required by participants during activities; participants felt more time and breaks were needed between tasks to reduce fatigue and improve focus.
	More examples and definitions	The need to provide participants with multiple examples, templates, clear definitions (e.g. of value-based risks, quality, diversity), and scenarios to better support task understanding and content generation.
	Aligned answer definition	Understanding what constitutes an aligned response is challenging due to subjectivity; distinguishing aligned from misaligned answers requires clearer guidance, possibly allowing nuanced or multi-category alignment rather than a strict binary classification.
	Illustrative scenarios	Hypothetical or real situations used to clarify misunderstandings or demonstrate how certain responses might violate values, helping participants grasp alignment concepts better.
	Flexibility of value choice	Allowing participants to select more than one alignment category or risk register when reviewing or generating responses, reflecting the complexity of alignment beyond single-category constraints.
Informing the synthetic data generation process	Applicability of the results into the pipeline	Concerns about how well the generated data and participant judgments will translate into training aligner models, including handling nuances in alignment interpretation and ensuring validity and usefulness of the synthetic data.
	Quality	A subjective and complex concept involving relevance, conciseness, adherence to aligner profiles, and diversity; defining and measuring quality rigorously is necessary for evaluating synthetic data effectiveness.
	Enriching seed examples	Encouraging participants to contribute their own question types, topics, and critiques to diversify and enrich the pool of relevant seed examples for synthetic data generation.
	Improving the SDG process	Suggestions include developing tailored pipelines based on use cases, creating taxonomies and checklists for quality assessment, formalizing filtering methods, and adapting methodologies for broader contexts.
	Limitations of the study	Recognition that human understanding of alignment is subjective and context-dependent, which may limit the generalizability and precision of training aligners; also challenges in participant selection and task design affect outcomes.

Table 3: Code-book with extracted themes, codes, and descriptions

Role	Group	Position in the company	Background	Workplace location
Moderator	1	Senior Research Scientist, Manager	AI, Optimization	US
Moderator	2	Research Scientist	AI, NLP, ML	UK
Moderator	3	Senior Research Scientist	AI, Human-Machine Interaction	BR
Moderator	4	Senior Research Scientist	HCI, Conversational Systems	BR
Participant	1	Senior Software Engineer	Speech Technologies, NLP	BR
Participant	1	Research Scientist	HCI	US
Participant	1	Computer Science Intern	Applied Mathematics, ML	BR
Participant	2	Research Scientist	HCI, Accessibility	US
Participant	2	Senior Software Engineer	Speech Technologies, NLP	BR
Participant	2	Software Engineer	ML	BR
Participant	2	Director	Neuroscience, Cognitive Science	US
Participant	2	Research Scientist	Quantum Computing, Political Philosophy	CH
Participant	3	Research Scientist	Political Theory	US
Participant	3	Senior Research Scientist	Cognitive Neuroscience	US
Participant	3	Research Scientist	Computational Mathematics	US
Participant	3	Research Intern	Political Social Science	BR
Participant	4	Research Scientist	History of Science	US
Participant	4	Research Scientist	Computer Vision, ML	BR
Participant	4	Research Scientist	Computational Creativity, Games, AI	BR
Participant	4	Research Scientist	Psycholinguistics	US
Total 20				

Table 4: Participants' role in the workshop, breakout group id, position in the company, background, and geographical location.