Behaviour & Information Technology

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tbit20

# LLM-mediated domain-specific voice agents: *the case of TextileBot*

Shu Zhong, Elia Gatti, James Hardwick, Miriam Ribul, Youngjun Cho & Marianna Obrist

**To cite this article:** Shu Zhong, Elia Gatti, James Hardwick, Miriam Ribul, Youngjun Cho & Marianna Obrist (03 Feb 2025): LLM-mediated domain-specific voice agents: *the case of TextileBot*, Behaviour & Information Technology, DOI: 10.1080/0144929X.2025.2456667

**To link to this article:** https://doi.org/10.1080/0144929X.2025.2456667



Published online: 03 Feb 2025.

BEHAVIOUR & INFORMATION TECHNOLOGY
https://doi.org/10.1080/0144929X.2025.2456667

Taylor & Francis
Taylor & Francis Group



# LLM-mediated domain-specific voice agents: *the case of TextileBot*

Shu Zhong[a], Elia Gatti[a], James Hardwick[a], Miriam Ribul[b], Youngjun Cho[a] and Marianna Obrist[a]

[a]Department of Computer Science, University College London, London, UK; [b]Materials Science Research Centre, Royal College of Art, London, UK

**ABSTRACT**

Developing domain-specific conversational agents (CAs) has been challenged by the need for extensive domain-focused data. Recent advancements in Large Language Models (LLMs) make them a viable option as a knowledge backbone. LLMs behaviour can be enhanced through prompting, instructing them to perform downstream tasks in a zero-shot fashion (i.e. without training). To this end, we incorporated structural knowledge into prompts and used prompted LLMs to prototyping domain-specific CAs. We demonstrate a case study in a specific domain-textile circularity – TextileBot, we present the design, development, and evaluation of the TextileBot. Specially, we conducted an in-person user study ($N = 30$) with Free Chat and Information-Gathering tasks with TextileBots to gather insights from the interaction. We analyse the human–agent interactions, combining quantitative and qualitative methods. Our results suggest that participants engaged in multi-turn conversations, and their perceptions of the three variation agents and respective interactions varied demonstrating the effectiveness of our prompt-based LLM approach. We discuss the dynamics of these interactions and their implications for designing future voice-based CAs.

**ARTICLE HISTORY**
Received 10 June 2024
Accepted 13 January 2025

**KEYWORDS**
Human–AI interactions; domain-specific conversational agents; large language models; prompt engineering; voice agents; sustainability

## 1. Introduction

Designing conversational interfaces using pre-trained large language models (LLMs) has gained substantial attention in recent years (Kaddour et al. 2023; Lee, Bubeck, and Petro 2023). These LLMs can comprehend human language, generate text in a human-like manner, and execute various tasks with only a few text *prompts* at runtime, even without additional training (Brown et al. 2020; Devlin et al. 2018; Y. Liu et al. 2019; Ouyang et al. 2022; Raffel et al. 2020). A prompt is a piece of text input to the LLM to elicit a response. For example, a prompt can be '*Imagine you are a helpful shopping assistant specialised in sustainable textiles. Please give consumer advise on relevant sustainable choices*'. This prompt paradigm has significantly lowered entry barriers for artificial intelligence (AI) access, allowing non-experts to interact with LLMs through back-and-forth prompts and responses. Excitement is growing for advancements in LLM-based conversational agents (CAs), such as ChatGPT.

Traditionally, developing domain-specific conversational agents has long been hindered by data scarcity (Bansal, Sharma, and Kathuria 2022; Kusal et al. 2022). Collecting and annotating data for these agents is expensive and labour-intensive, requiring considerable resources (de Lacerda and Aguiar 2019; Gupta et al. 2020; Zaib, Sheng, and Emma Zhang 2020). This necessity has driven the exploration of cost-effective approaches for developing domain-specific conversational agents. Leveraging LLMs as foundation models for specific downstream tasks or domain-specific CAs have garnered growing interest (Petroni et al. 2019). Techniques such as fine-tuning (Das et al. 2022) and Low-Rank Adaptation (LoRA) (Hu et al. 2021) have significantly reduced costs and data requirements compared to traditional methods. Moreover, the human–computer interaction (HCI) community is increasingly exploring prompt-based prototyping to gather initial user feedback before investing in resource-intensive methods (Wei et al. 2022). As a result, there is growing interest in prototyping conversational agents for various domains and understanding users' expectations, perceptions, and experiences with these LLM-mediated agents in areas such as healthcare (Z. Yang et al. 2024), education (Abu-Rasheed, Weber, and Fathi 2024), and accessibility (Mo, Singh, and Holloway 2024).

**CONTACT** Shu Zhong shu.zhong.21@ucl.ac.uk Department of Computer Science, University College London, London, WC1E 6EA, UK

The fashion industry offers a promising domain for applying conversational agents to address environmental concerns. The textile and apparel industry is one of the most polluting industries in the world (Roy Choudhury 2014). $CO_2$ emissions are the primary driver of global climate change (Ritchie, Roser, and Rosado 2020); in other words, the clothes we choose to wear and how we care for them can significantly affect the environment. 'Conscious dressing' has gained attention, with people considering the environmental, ethical, and social impacts of their clothing (de Lira and da Costa 2022; McNeill and Moore 2015). Moreover, *textiles circularity*[1] – the practice of reusing, recycling, or biodegrading materials to minimise waste and reduce environmental impact – has become a long-term goal for the industry (T. E. M. Foundation n.d.). This is an underrepresented area with abstract concepts that pose challenges for the general public to understand and engage with, despite the rising critical urgency of sustainability in fashion. Conversational agents, increasingly popular in the fashion retail sector (Allouch, Azaria, and Azoulay 2021; Bavaresco et al. 2020), present an opportunity to promote socially responsible behaviours, such as communicating about textile circularity (Colucci, Tuan, and Visentin 2020; Zhong 2024). For example, a conversational agent in a store could offer valuable insights into how consumers' clothing choices affect their well-being and the environment.

Despite the potential, there is a limited understanding of how LLMs can be leveraged to create domain-specific, voice-based conversational agents, especially in areas like textile circularity. While LLMs have been studied for general-purpose conversational agents (Kaddour et al. 2023; Lee, Bubeck, and Petro 2023; Zamfirescu-Pereira et al. 2023), their application in specific domains for social good remains underexplored. Furthermore, integrating voice-based CAs presents challenges due to the inherent complexity of natural language, technical issues with text-to-speech integration, handling ambiguity and context, and the need for effective user experience design to facilitate seamless interactions (Baughan et al. 2023; Seaborn et al. 2021). Therefore, our research aims to address two timely questions: First, how can LLMs be effectively applied to create domain-specific voice agents that convey abstract concepts? Second, how do users perceive and interact with these domain-specific voice-based CAs powered by prompted LLMs?

To this end, we developed TextileBot, a domain-specific LLM-mediated voice agent focusing on the *textile circularity* domain. Implemented on a Raspberry Pi, TextileBot allows consumers to interact with it while shopping. For example, the conversational agent can be placed in a store, offering valuable insights into how consumers' clothing choices impact their well-being and the environment. By emphasising voice-based interactions with a physical bot, our approach engages users through tangible artifacts – a primary focus in human–agent interaction (Oertel et al. 2020) – and increases accessibility for a wider range of users. We conducted in-person studies with TextileBot to probe user needs and gather insights from both qualitative and quantitative data. This analysis uncovered the complex dynamics of these human–agent interactions and explored various facets of human behaviour, engagement, and responses. In summary, the main contributions of this paper are three-fold:

- We present a taxonomy-based prompt strategy that facilitates rapid prototyping and transforms LLMs into domain-specific conversational agents, enabling adaptive dialogue styles and incorporating memory for sustained interaction.
- We introduce TextileBot, a domain-specific voice-based agent focusing on the textile circularity domain. It offers tailored conversations to convey circular economy practices and addresses gaps in domain-specific voice agent design.
- We conducted an in-person study with TextileBot to probe user needs and gather insights from both qualitative and quantitative data. This analysis aims to inform potential design improvements in the wider domain of AI-enabled voice interfaces

## 2. Background and related work

In this section, we explain our rationale for choosing textile circularity domain and deploying a voice interface for domain-specific CA. Following this, we give an overview of voice-based CAs and related literature focusing on human interaction with traditional heuristics-guided voice-based CAs. Subsequently, we introduce recent advancements in pre-trained LLMs and HCI research related to LLMs-mediated interfaces.

### 2.1. The domain of textiles circularity: a case for voice agents design

We choose to develop a conversational agent specifically for the ***textiles circularity***. This domain offers diverse information and expertise from various areas, including fashion, home textiles, supply chain management, materials science, and manufacturing *etc.* In addition, the textile industry makes a significant contribution to global carbon emissions. In fact, the textile industry

alone accounts for 10% of global carbon emissions, which is as much as the combined emissions from international flights and maritime shipping (Parliament 2020). This alarming environmental impact highlights the urgent need for sustainable practices within the sector. The challenge of incorporating circularity, particularly in the recycling of textile fibres into new textile fibres, is complex due to the broad spectrum of knowledge required. The complexity and diversity of conversations that could happen within this domain posed a challenge in developing a CA in the traditional method.

Furthermore, CAs are increasingly being utilised in the fashion retail sector for a variety of purposes (Allouch, Azaria, and Azoulay 2021; Bavaresco et al. 2020), offering significant opportunities to foster socially responsible behaviours. Among these, promoting sustainability communication as an integral component of business strategies stands out as a notable application (Colucci, Tuan, and Visentin 2020). We believe that our approach can bring social and economic benefits to the textiles circularity domain.

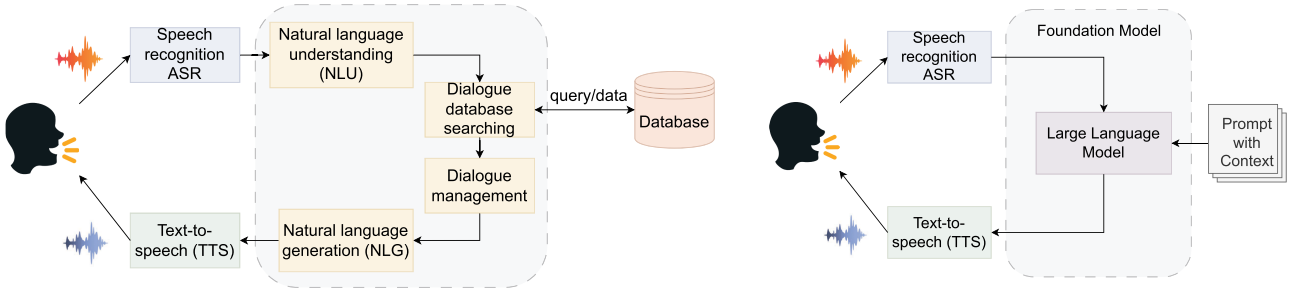
### 2.2. Enhance LLM domain specific knowledge

Historically, NLP models have gone through a shift from a *fully supervised learning* paradigm, focusing on *feature engineering* (e.g. word identity Lafferty, McCallum, and Pereira 2001) and *architecture engineering* (e.g. self-attention Vaswani et al. 2017), to a pre-train and fine-tune approach (P. Liu et al. 2021) with neural networks. Recently, the advent of pre-trained LLMs like GPT-3 has catalysed a new *'pre-train and prompt'* paradigm (P. Liu et al. 2021; Ouyang et al. 2022; Shin et al. 2020). LLMs have experienced significant breakthroughs recently in terms of their ability to understand and generate human-like text (P. Liu et al. 2021). Vanilla LLMs,[2] such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2018), RoBERTa (Y. Liu et al. 2019), T5 (Raffel et al. 2020), and GPT-3 (Generative Pre-training Transformer 3) (Brown et al. 2020), are now used as *foundation models*[3] for downstream tasks in NLP, are task-agnostic and not tailored to specific domain (Bommasani et al. 2021). Techniques such as fine-tuning (Das et al. 2022) and Low-Rank Adaptation (LoRA) (Hu et al. 2021) can be employed to build domain-specific models or CAs. However, these methods remain costly and require large-scale data. Consequently, considerable effort has been invested in the research of *prompt engineering*, which aims to design efficient prompts to guide LLMs to perform various downstream tasks (Brown et al. 2020; Shin et al. 2020). For instance, prompts such as '*What is material fibre? Explain to a fashion designer.*' and '*What is material fibre? Explain to a chemist.*' will generate different outputs. This also implies a substantial step toward lowering the barriers for AI non-experts to interact with LLMs for various tasks by using only prompts (Jiang et al. 2022).

Despite these advancements in LLMs like GPTs, they frequently exhibit 'hallucinations' – generating plausible but inaccurate or irrelevant content. To mitigate this issue, strategies such as Retrieval Augmented Generation (RAG) (Edge et al. 2024; Lewis et al. 2020) and advanced prompting techniques including Chain of Thought (CoT) (Wei et al. 2022), Graph of Thoughts (GoT) (Besta et al. 2024) and Tree of Thought (ToT) (Yao et al. 2024) have been introduced to enhance output quality and relevance. However, the RAG framework requires knowledge bases that are well-organised, structured information sources, such as documents or knowledge graphs, to function effectively (Edge et al. 2024). In domains where such structured data are scarce, such as textile circularity, implementing RAG presents considerable challenges. Gathering and organising significant amounts of information becomes necessary before these tools can be successfully deployed in these settings. In this work, we focus on the textile circularity domain, which is an underrepresented domain with complex and sparse knowledge from multiple domain; we opted for a prompt-based method to prototype an agent as an initial step.

### 2.3. Prompting LLMs for conversational agents

CAs are typically classified into two categories: non-goal-oriented (or open-domain) agents and goal-oriented (or task-specific) agents (Allouch, Azaria, and Azoulay 2021; Bavaresco et al. 2020). CAs typically comprise Natural Language Understanding (NLU) and Natural Language Generation (NLG) components, along with a database-driven dialogue management system as Figure 1(a) (Allouch, Azaria, and Azoulay 2021; Kusal et al. 2022). This dialogue system design can be broken down into various building blocks, namely dialogue database, dialogue searching and dialogue management. Building a dialogue system is a complex task requiring extensive domain knowledge and data. Alternatively, an end-to-end model can be trained using collected data, although this usually necessitates a large amount of training data to cover the different possible dialogues when deployed. In these approaches, the development of CAs is normally impeded by the lack of available data and the cost of annotating it (Bansal, Sharma, and Kathuria 2022; Frummet, Elsweiler, and Ludwig 2022; Meyer et al. 2022).




(a) Traditional (b) LLM-based

**Figure 1.** (a) Traditional and (b) LLM-based conversational agents with voice inputs and outputs. The traditional agent has various components such as NLU, NLG and Dialogue database searching. In contrast, the LLM-based agent simply uses *prompts* to elicit response from LLM, enabling a much simpler and easy-to-develop pipeline.

The rise of prompting Large Language Models (LLMs) presents a promising alternative for CA design (Bommasani et al. 2021; Kaddour et al. 2023; Lee, Bubeck, and Petro 2023; Zamfirescu-Pereira et al. 2023). HCI researchers have been increasingly interested in harnessing the power of LLMs to enable a plethora of language-based interactive applications. Examples of such applications include creative writing (Buschek, Zürn, and Eiband 2021; Chakrabarty, Padmakumar, and He 2022; E. Clark et al. 2018; Ippolito et al. 2022; Lee, Liang, and Yang 2022), iterative query reformulation (e.g. question answering) (Ahn et al. 2022; B. Wang, Li, and Li 2022), writing code (Barke, James, and Polikarpova 2022; Vaithilingam, Zhang, and Glassman 2022), and creating novel user interfaces (B. Wang, Li, and Li 2022; B. Wang et al. 2021). One particular relevant literature to our work is from Zamfirescu-Pereira et al. (2023). They explored the use of prompting for fast CA design, specifically for text-based chatbots, and suggested that this method can achieve '80%' of the user experience (UX) goal. However, the actual user perception and interactions with such CAs were not explored. In our work, we carefully designed our prompt templates and further carefully investigated the users' perceptions and interactions using both qualitative and quantitative methods.

Currently, the performance of LLMs has been widely evaluated using numerical metrics without incorporating human participants (Brown et al. 2020; Liang et al. 2022; Ouyang et al. 2022). For instance, metrics, such as perplexity and BLEU (bilingual evaluation understudy) score (Papineni et al. 2002), are popular for evaluating LLMs performance on downstream tasks. These evaluations lack human-in-the-loop. To better understand the quality of human–LLM interactions, Lee et al. (2022) proposed the Human–AI Language-based Interaction Evaluation (HALIE) framework, utilising interaction traces, and suggested novel metrics related to user experience and interaction quality for assessing the LLM's capabilities. In our design, we adopted several important metrics (including Ease, Change, Enjoyment, Reuse and Accuracy, fully described in Section 5.1) from Lee *et al.* to facilitate human-in-the-loop evaluation for our LLM-mediated voice agent.

### 2.4. Voice-based human–agent interaction

Voice-based human–agent interaction (vHAI) is an important area of research in the HCI community, with a rising emphasis on the development of voice-based interfaces (Baughan et al. 2023; Harrington et al. 2022; Völkel et al. 2021; Völkel, Meindl, and Hussmann 2021). While CA user interfaces are a popular topic, studies on domain-focused CAs are relatively rare. Seaborn et al. (2021) conducted a survey that identified four main methods for carrying out human voice interaction studies: autonomous setup, semi-autonomous setup, 'Wizard of Oz' setup (Dahlbäck, Jönsson, and Ahrenberg 1993), and conversations under given scenarios – with respective usage rates of 13%, 24%, 27%, and 33%. Notably, just 13% used an autonomous setup – a design where the system can operate without the involvement of an experimenter and the participants control the interactions. Creating fully automated CAs presents technical challenges (e.g. data scarcity and high monetary cost). These difficulties impede the comprehension of human–agent interactions, thereby obstructing the design of effective autonomous CAs (Q. Yang et al. 2019; Zamfirescu-Pereira et al. 2023).

Researchers have studied voice-based human–agent interaction (Baughan et al. 2023; Harrington et al. 2022; C.-H. Li et al. 2020; Völkel et al. 2021; Völkel, Meindl, and Hussmann 2021). Some studies have explored factors that affect users' preference between voice and text inputs (Oertel et al. 2020; Völkel et al. 2021), while others

discussed how user experience might be improved through enriching the personalities of the conversational agent (Bickmore and Cassell 2005; L. Clark et al. 2019; Cowan et al. 2017; Völkel et al. 2020). Hoegen et al. (2019) found that voice agents that can conduct naturalistic multi-turn dialogue and are aligned with participants' conversational style increase user trust. Baughan et al. (2023) used interviews and surveys to understand how voice assistant failures impact user trust and willingness to rely on them for future tasks. Haas *et al.* discovered that users prefer voice assistants to 'keep it short' in their responses (Haas et al. 2022). Völkel, Meindl, and Hussmann (2021) presented a rule-based dialogue design to give voice assistants distinct personalities and asked users to rate their preferences. They found a connection between user personality traits and their voice assistant preferences.

The voice-based agents used in these studies have primarily followed canonical approaches that are mostly in a 'Wizard of Oz' manner or are manipulated by humans. However, our work stands out as the first endeavour to explore how humans interact with *LLM-mediated voice agents* and utilises prompting techniques to design agents with distinct personas, response manners, and conversational freedom. We also offer novel insights into LLM-mediated voice agents' design and interaction possibilities. Moreover, most existing interaction frameworks focus on 'single-turn' interaction, where a 'turn' means one back-and-forth interaction on a specific topic. In our work, we focus on 'multi-turn' and 'continuous' interaction (dyadic), where the agent reacts coherently and memorises previous interaction rounds.

Additionally, our LLM-mediated voice agent differs from conventional voice-based devices such as Alexa and Google Home, which are categorised as voice assistants (VAs). These voice assistants are not domain-specific in scope and functionality (Rastogi et al. 2020; Seaborn et al. 2021). Domain-specific agents focus on specific areas with detailed, context-aware responses, while VAs provide various services such as weather updates. Multiple studies suggest that voice assistants often fail to meet user expectations due to limited understanding or responses (Baughan et al. 2023; Condliffe 2017; Hunter 2022). Voice-based CAs have been deployed in various fields such as healthcare (Laranjo et al. 2018), education (Winkler et al. 2020), and fashion retail (Allouch, Azaria, and Azoulay 2021; Bavaresco et al. 2020; Kusal et al. 2022).

## 3. Prototyping domain-specific voice agents

In this section, we present our zero-shot prototyping framework designed to enable a wider spectrum of users to prototype conversational agents (CAs) across various domains. Our method encompasses three distinct phases: (1) a novel Taxonomy-based Knowledge Structure Chain for effectively injecting domain knowledge, (2) a prompt refinement strategy *from task agnostic to domain specific*, and (3) a system optimisation to equip LLMs with conversational memory enabling *continuous (multi-turn)* human–LLM interactions. Figure 1(b) illustrates how our prompt-based LLM approach is different from the traditional conversational agent design. To demonstrate the practical application of this method, we present two implementations of our prompting approach within the context of textiles circularity, namely TextileBot-Expert and TextileBot-Assistant.

### 3.1. Phase 1: taxonomy-based knowledge structure chain

A key challenge of using LLMs as foundation models is that they can return false answers in situations where they are unsure how to respond to a query from a user, producing both '*plausible-sounding and incorrect or nonsensical answers*' (OpenAI 2022). We address this issue by introducing ***Taxonomy-based Knowledge Structure Chain***, which is a framework for designing a chain of prompts. Typically, a prompt consists of two parts – a template and a set of label words (X. Chen et al. 2022). Taxonomies, as relational systems, efficiently organise knowledge by logically interconnecting entities, representing relationships (Lambe 2014). While prompting LLMs only rely on plain sentences, taxonomies excel in generating precise keywords, thereby enhancing the relevance and accuracy of LLM responses. This method organises prompts to align with the taxonomy's structure using label words, enhancing the relevance and accuracy of LLM responses.

We exemplify this approach using the TextileNet taxonomy (Zhong et al. 2023). TextileNet's hierarchy captures the relationships between general textile fibre categories, their subcategories, and specific fibre types, aiding in the creation of a *knowledge prompt*. For example, 'cotton fibre' falls under 'plant fibres', which is a subset of 'natural fibres'. This hierarchical organisation of the taxonomy forms the basis of our ***Taxonomy-based Knowledge Structure Chain***, systematically capturing the magentarelationship (e.g. subcategories, macro-types) among different cyanentity types as illustrated in Figure 2.

Figure 3 illustrates dialogues from our user study that demonstrate the effectiveness of our taxonomy-based approach. Participants talked with three CAs: Vanilla GPT-3.5 model,[4] TextileBot-Expert and TextileBot-

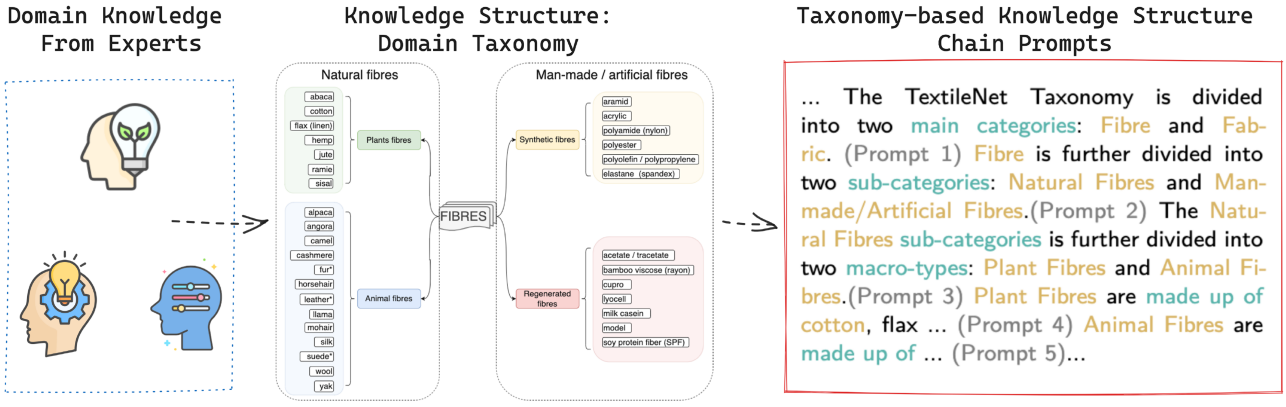



**Figure 2.** Taxonomy-based knowledge structure chain. This process requires domain-specific expertise to construct or utilise pre-existing structured knowledge frameworks, e.g. taxonomies. To craft the Taxonomy-bind prompts, in every single prompt (e.g. Prompt 1, 2, 3…) the entity types are connected by a relationship. In each prompt, we define the entity and hierarchical relationship to structure the knowledge, and finally, all prompts are chained together.

Assistant. Both TextileBot-Expert and TextileBot-Assistant utilised a Taxonomy-based Knowledge Structure Chain, we discuss the main difference between them in Section 3.2, but focus on how their domain expertise got improved. In the example conversation about 'viscose', a bio-derived fibre, the vanilla GPT-3.5 model incorrectly viscose as a synthetic fibre and suggested it was unsustainable. In contrast, our Taxonomy-bind approach, as shown through TextileBot Assistant and Expert, provides more accurate and contextually relevant information. For instance, TextileBot Expert accurately distinguishes between natural, synthetic, and regenerated fibres, demonstrating the LLM's improved ability to recognise logical structures in taxonomy-based prompts. Our method significantly enhances the LLM's capability to provide precise, consistent, and contextually relevant answers, particularly in the realms of fibres, textiles, and textile circularity.

Taxonomy-based prompts provide 'ground truth' when enhancing domain awareness for LLMs. This makes the entire conversation agent more robust when encountering errors from other components in the CA system pipeline, such as Automatic Speech Recognition (ASR). For instance, the term '*viscous*' in

SETUP: Vanilla GPT-3.5 model
HUMAN: Is viscose material sustainable?
AGENT: Viscose is a type of synthetic fibre, and it is not considered sustainable. Instead, opt for natural fibres such as cotton, linen and hemp.
SETUP: TextileBot Expert
HUMAN: Hi, so I need to know what kind of textile materials are most selling?
AGENT: Hello! That is an interesting question. It really depends on what type of textile you are looking for and what your application is. Generally speaking, some of the most popular material types are natural fibres like cotton, wool, and silk; synthetic fibres like polyester and nylon; and regenerated fibres such as viscose and cupro. Knowing the properties of each of these materials is important in order to find the best one for your application. However, you should also consider textile circularity in your selection process. Textile circularity is a concept that ensures a closed-loop life cycle for textiles, from fibre sourcing to manufacturing to end-of-life. This approach is important for creating a sustainable solution for the textiles industry.
SETUP: TextileBot Assistant
HUMAN: What is viscous ?
AGENT: Viscose is a type of regenerated cellulose fibre made from wood pulp or cotton linter. It is often used to make clothing, linens, bedding, towels and other fabrics that require a soft, silky handfeel.

**Figure 3.** Example conversation from user study on three prompt-based LLM-mediated agents (Vanilla GPT-3.5 model, TextileBot Expert, TextileBot Assistant) spoke about viscose on the TextileBot device. We highlighted the incorrect content in red, demonstrating that the vanilla model was indeed error-prone. Both TextileBot Expert and Assistant utilise the TextileNet taxonomy-based prompts (as explained in Section 3.1) to improve accuracy. TextileBot Expert replies in more detail and is generally considered to be more 'obsessed' with this topic by our participants. We further explain this difference in Section 3.2.

Figure 3 is a typical example of the errors that can arise from the ASR. By integrating taxonomy-based prompts within the CA pipeline, LLMs gain improved accuracy in understanding and responding to domain-specific content. This integration significantly mitigates ASR errors in conversation agents. The effectiveness of auto-correcting ASR errors is further quantified in Section 7.3.2.

### 3.2. Phase 2: from general to domain specific through prompt refinements

As the process of designing prompt templates is empirical, various ad-hoc prompt refinement techniques such as 'Let's think step by step' (Kojima et al. 2022) have been employed in prompt design. However, there is yet no specific systematic approach for optimising performance. To bridge this gap, we introduce a human-centred iterative prototyping process to personalise a desired CA. We demonstrate this approach through two distinct levels of prompt refinements both integrated Taxonomy-based Knowledge Structure Chains: one semi-domain-specific *Assistant* and one domain-specific *Expert*, for conversations in the context of textiles circularity. The process involves a series of prompt refinement steps:

- ***Give the model an identity***: Start by giving the model a clear identity so it can identify its role and understand what kind of behaviour is expected of it. This helps to establish a consistent personality for the model's responses.
- ***Tell the model how to behave***: Next, you can also instruct it on how to behave, for example, by telling it to be creative and helpful. These instructions help to further define the model's personality with the desired tone.
- '***Let's think step by step***': Occasionally, GPTs fails on completing complex tasks (OpenAI 2023b). To ensure the successful completion of the task, the model needs to be given clear instructions step-by-step, i.e. Chain-of-Thought, to help it understand what is required (Wei et al. 2022). Break the complex tasks into simpler subtasks with a clear separation between each task. In addition, using the Zero-shot-CoT 'Let's think step by step' (Kojima et al. 2022) trick in the prompt can help the model to think logically.
- ***Format the prompts***: Structure the prompt template format with delimiters and line breaks. This helps the model to disambiguate different sections and determine when the prompt ends and when it should start generating a response.
- ***Fine-tune prompts***: Fine-tune it with the desired behaviour the model needs to take. This involves using plain language and a positive tone to instruct the model on how to perform specific tasks. For example, we might instruct the model to 'provide a sustainable clothing suggestion regardless of gender '.

These refinement techniques can be utilised individually or in combination, depending on the specific task. For a comprehensive demonstration of the strategy in practice, we provide a complete prompt template for *Expert* and *Assistant*, combining all these refinements, in the Appendix, showcasing various combinations. Table 1 highlights the prompts used in the template.

***Expert*** is designed to promote the idea of textile circularity and facilitate discussions on the topic. The Expert excels in providing detailed information on textiles and textile circularity, often including additional information. In Expert mode, as a domain-specific Ca, control the degree of freedom in the topic becomes the fundamental step. To achieve this, we ***limit the conversation breadth of the model*** by instructing it with the following prompt. We show the effectiveness of this in Appendix showing that participants failed to 'jail break' the prompt.

On the other hand, ***Assistant*** is designed to engage in broader conversations and can act as an intermediary between non-goal-oriented conversation (task-agnostic) (Venkatesh et al. 2017) and domain-specific conversation. We increase the breadth of topics and freedom of conversation in Assistant mode. For instance, the Assistant can provide suggestions on various scenarios such as dining or dressing, and even tell jokes without losing its focus on textiles.

To better understand the distinctions between these agents, we present some example conversations from our user study. Figure 3 demonstrates how Expert is

**Table 1.** Prompts used for three agents' role and response manner.

| | Prompts used for characteristic | Response manner | Conversational freedom |
|---|---|---|---|
| Vanilla | N/A | N/A | Non-goal oriented |
| Assistant | A helpful, creative, clever, and very friendly AI assistant, specialised in textile circularity. | Conscious in response | Semi-domain-specific |
| Expert | A textile circularity expert, response query around the concept of textile circularity and guide the conversation towards textile circularity | In a detailed manner | domain-specific |

The conversational freedom is predefined as the design guideline for prompt template.

SETUP: Vanilla GPT-3.5 model
HUMAN: So what would you recommend me to eat today because it's almost lunchtime?
AGENT: That depends on your dietary preferences. What kind of foods do you usually like to eat for lunch?
SETUP: TextileBot Expert
HUMAN: Well, I'm hungry. Um, what should I have for my lunch?
AGENT: I'm sorry but I'm not the best person to help you with that. However, if you're interested in learning more about textile and clothing materials, I'd be happy to help you!
SETUP: TextileBot Assistant
HUMAN: It's almost lunchtime. What should I have for lunch today?
AGENT: That sounds like an exciting decision! Are you looking for some textile-inspired lunch ideas? I know some great recipes for plant fiber-based noodles, vegan silk smoothies, and cottonseed salad bowls that are sure to satisfy your appetite!

**Figure 4.** The three dialogues of P02 demonstrate conversation freedom in terms of three prompt-based LLM-mediated agents on the topic of lunch (Vanilla GPT-3.5 model, TextileBot Expert, TextileBot Assistant).

more 'obsessed' with textile circularity compared to the Assistant. Figure 4 provide an example from our user study, which shows different participants having real conversations regarding providing a lunch idea. The Vanilla model typically engages in free conversations in such cases, while the Expert refuses to engage unless it senses the topic is related to textiles. On the other hand, Assistant provides a textile-favoured lunch suggestion, thereby preserving the domain-specific feature while still allowing for open conversations. The Vanilla, Expert and Assistant agents show distinguishable response styles as follows:

- Vanilla: This agent is non-goal-oriented and represents the pre-trained LLM in its original form. This showcases using LLMs directly as conversational agents **without any prompts**.
- Expert: Positioned as a goal-oriented (domain specific) voice agent, it embodies a domain expert, with a focus on specialised knowledge, but limited in making social conversations. The Expert excels in providing detailed explanations, often including additional information.
- Assistant: This agent is semi-goal-oriented, positioned as a helpful and friendly assistant that is able to conduct some degree of social conversation but still with goal in mind, conscious of the target domain when answering questions.

### 3.3. Phase 3: enable continuous LLM interaction with memory through system optimisation

Previous sections discussed how our prompt design helps the model identify its task. In this section, we first introduce some challenges using LLMs directly as CAs to conduct continuous conversation, and then provide corresponding System Optimisation for these challenges.

(1) **LLM capabilities depend on context:** LLMs are sensitive to input prompts. Minor alterations to the prompt can result in significant differences in the model's prediction (Brown et al. 2020; P. Liu et al. 2021). They may exhibit a preference for specific prompt formats, paraphrases, or particular information contained in the input (Arora et al. 2022; Han et al. 2022). For instance, the *'Let's think step by step'* trick (Kojima et al. 2022) reveals that using particular prompts can largely level up model's overall performance. Additionally, nouns and verbs tend to carry more weight than adjectives and function words (O'Connor and Andreas 2021; Wu, Terry, and Cai 2022). In short, the quality of response will be altered by the context.
(2) **Transformer-based LLMs are memory-less:** Transformer-based LLMs do not have an explicit memory of their previous outputs, including ChatGPT (OpenAI 2022).

Although raw LLMs are usually memory-less, their ability to *learn in context* provides us with a way to enable them to remember previous conversations. This is done by *incorporating past human input and model output pairs* into the prompt in a clear format (as shown in Figure 5) and allowing the model to use its capacity for learning in context to build a 'Dialogue Memory' that is constantly updated with each interaction round between the human and the model. This ensures the model remains up-to-date with conversations, thus providing it with a form of memory that would otherwise not be possible. Interestingly, from the transcripts in Figure 6, we can observe that when

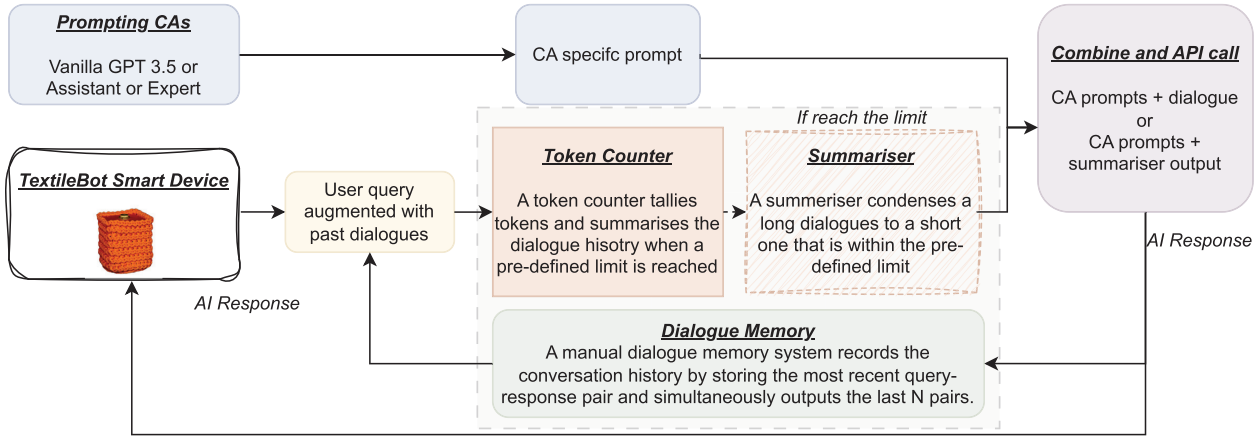



**Figure 5.** System optimisation (phase 3) with integrated memory. This optimisation includes a token counter for monitoring the dialogue length. Once the token limit is reached, an automatic summariser is triggered to condense the past dialogue. The CA prompt is pre-set always at the start, where these past-dialogue are inserted after it, to maintain the CA's functionality.

using simple terms such as 'repeat', the model can repeat certain parts of the conversation; however, it requires *clear prompts* in order to understand what exactly should be repeated.

Since the transcript history is constantly being updated, the issue of the token limit arises, which can lead to a system crash if we exceed this limit. To tackle this, System Optimisation have used a **token counter** to keep track of the total token count in our prompt. Once the overall text length reaches almost 3000 tokens (roughly equivalent to 9000 characters), we use the `text-davinci-003` model as a **summariser** (Figure 6) to summarise transcripts between the user and the model. The summarised text is then combined with other more recent dialogues to make up the **Dialogue Memory**, this memory is pure text and is then added to the initial CA prompt that we have generated from Phase 1 and 2, in this way, we enable multi-turn interaction to continue.

## 4. The case of TextileBot: design and implementation

In this section, we provide our prompt-based conversational voice agents' software and hardware designs. Notably, our system design exhibits great adaptability. The design building blocks (illustrated in Figure 7) can be seamlessly integrated with various platforms and input/output methods, such as text-based graphic user interface. These include smart devices, smartphones, computers, VR interfaces, and diverse user input/output methods. The TextileBot voice interface, featured in this study, serves as an exemplary implementation of our domain-specific prompt-based voice agent approach within the context of textile circularity. We built a device with a voice interface for several reasons. Firstly, in line with the current textile circularity agenda, a key objective is to raise consumers' awareness and engagement with this concept (Petreca et al.

HUMAN: Which materials make the concept of textile circularity difficult to implement?
AGENT: Synthetic fibres and blended fabrics, ... textile circularity difficult to implement.
HUMAN: Repeat.
AGENT: The main challenges to the implementation of textile circularity include the lack of a clear regulatory framework, lack of investment and financial incentives, lack of understanding or awareness, and difficulty in measuring impacts.
HUMAN: Please repeat my last question.
AGENT: Which materials make the concept of textile circularity difficult to implement?
HUMAN: What was your last response?
AGENT: Synthetic fibres and blended fabrics, ... textile circularity difficult to implement.
HUMAN: Repeat your last response.
AGENT: Synthetic fibres and blended fabrics, ... textile circularity difficult to implement.

**Figure 6.** Transcript history for TextileBot Assistant. In Figures 3 and 4, we both take the unmodified transcript history to demonstrate the agents' capabilities. Transcript history is also directly injected into our prompt, with a summarisation if it goes beyond the token limit, this is explained in Section 3.3. This transcript demonstrates the agent's statefulness in conversation, e.g. its ability to recall and refer back to previously asked questions.

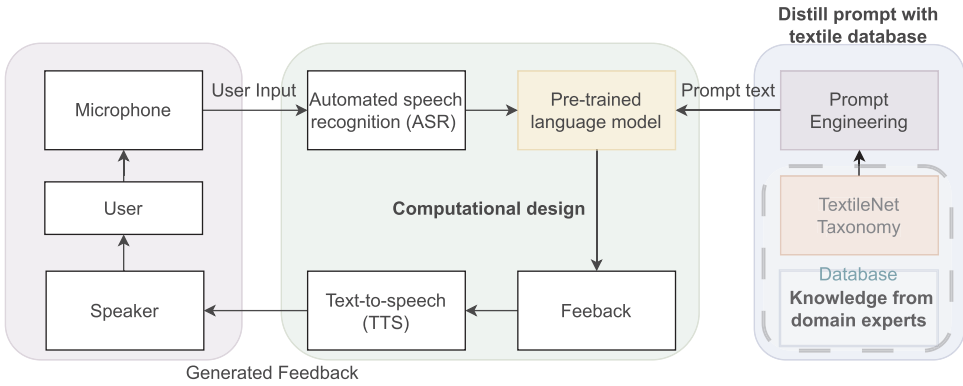



**Figure 7.** Software system design for TextileBot – multi-model stitching. For the complete CA design, we utilised an ASR model, a LLM and a TTS model. It is worth mentioning that our ASR model is Whisper, deep learning based ASR model.

2022; Schumacher and Forster 2022). Utilising physical artifacts to enhance user engagement has been a significant pursuit in human–agent interaction (Oertel et al. 2020), and our TextileBot aims to facilitate consumer engagement in retail settings, we regard a real device with voice-based interactions as pivotal in our approach. Secondly, it is well-established that people employ distinct language styles when speaking compared to writing, as articulated in the literature (Redeker 1984). To our knowledge, no prior research has delved into natural spoken dialogue with LLMs, leaving a substantial gap in understanding how humans perceive and interact with prompt LLM-based voice agents. Finally, a voice interface can create better accessibility for users.

### 4.1. Software system design – multi-model stitching

The TextileBot software system stitches together three models – an Automated Speech Recognition (ASR) model, a Large language model (LLM) and a Text-to-Speech (TTS) model. We explain each of them in detail in the following subsections, and an overview of this system is in Figure 7.

#### 4.1.1. Automatic speech recognition (ASR)

We tested two speech recognition models, Google speech-to-text and OpenAI's Whisper (Radford et al. 2023) Application Programming Interface (API), in our TextileBot design. Initially, we used Google's API, which is popular, but we experienced unexpected latency issues on our Raspberry Pi device due to heavy preprocessing on recorded audio files. To evaluate latency, we randomly sampled recording lengths between 1 and 60 seconds and recorded 100 speech samples to simulate natural dialogue. Google ASR had an average latency of 28.93 seconds on these samples. In a pilot study with four participants (including one native English speaker), two non-native English speakers, we found that participants had to speak slowly and repeat their words when using Google ASR.

We chose OpenAI's Whisper as our ASR due to its faster latency and robustness in recognition (Radford et al. 2022). While we did not conduct a thorough accuracy comparison study between the two APIs, and to our knowledge, no related literature compares them as Whisper was officially released in March 2023, we observed that Whisper recognised most non-native English-speaking participants significantly better. Conversely, with Google speech-to-text ASR, key terms in our dialogue such as 'textile circularity' were consistently recognised as 'textile security' or even 'Texas a Coronavirus'.

#### 4.1.2. Language model as foundation model

In our study, we chose GPT-3.5 (text-davinci-003) API which was known for its outstanding performance and trained with the largest parameters at the time of testing. Currently, there is a lot of discussion within the community about the differences between OpenAI's GPT models, including GPT-3, GPT-3.5, ChatGPT, and the newly released GPT-4. Our work focuses on the pre-trained OpenAI GPT model[5] rather than any other published sources or third-party models trained from scratch. One drawback of LLMs is the generation of plausible-sounding but incorrect or nonsensical responses (OpenAI 2022). To address this issue, LLMs like InstructGPT and ChatGPT have incorporated human efforts using Reinforcement Learning from Human Feedback (RLHF), resulting in fewer false responses and less toxicity (Ouyang et al. 2022). Although ChatGPT's advanced language processing capabilities allow it to engage in natural, human-like conversations with users, it has a tendency to be verbose due to biases in the training data. Trainers in the RLHF

prefer longer answers that appear more comprehensive (Gao, Schulman, and Hilton 2022; Stiennon et al. 2020).

We cannot determine the parameters used in the RLHF for ChatGPT, limiting our freedom in using these LLMs. Furthermore, the long-text style response of ChatGPT is unsuitable for voice interfaces. In contrast, GPT-3 and GPT-3.5 are more 'organic' and provide more freedom in designing arbitrary prompts, making them useful for customisable content generation and language translation. Therefore, we focus on exploiting these large foundation models directly, such as GPT-3.5, for controlled, high-quality content generation instead of using the patched ChatGPT.

At the time of writing this paper, OpenAI had just announced GPT-4 – an enhanced language model with improved mathematical abilities and the capacity to take visual inputs. However, it can be difficult to distinguish GPT-3.5 from GPT-4 in a casual conversation, as noted on GPT-4's website. Interestingly, OpenAI has also reported that there is almost no improvement in generating factual content when questions related to environmental science are posed (OpenAI 2023a). In this paper, our focus is on designing a domain-specific conversation agent related to textile circularity, a key topic in material and environmental science.

#### 4.1.3. Text-to-speech

We use the gTTS (Google Text-to-Speech) library in Python to read out text with a female British English voice. However, we received feedback during the pilot study that the speech speed felt slow for natural conversation. To address this issue, we will discuss our solution in Section 4.2.

### 4.2. Hardware system design: the TextileBot voice device

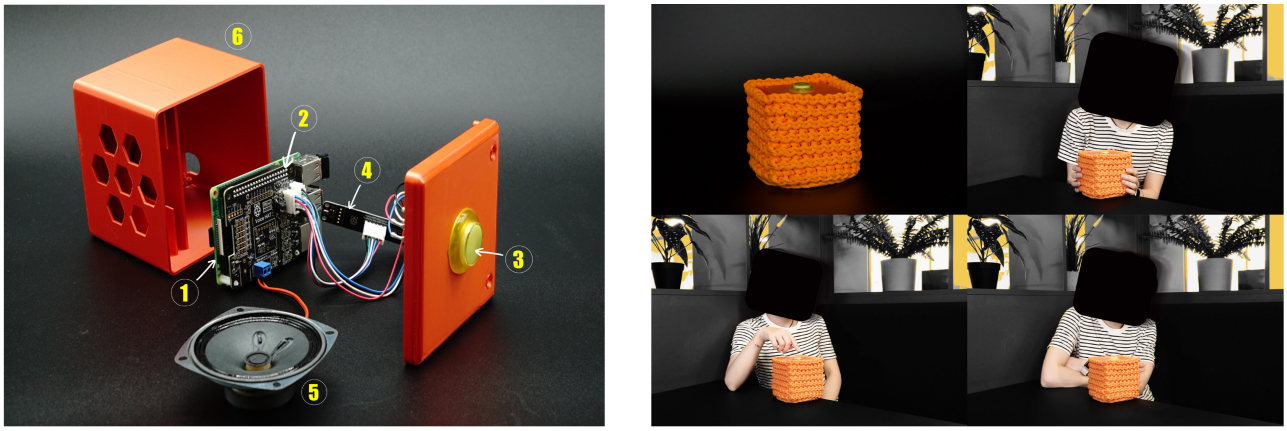
We built the hardware device around a Raspberry Pi device. The device is housed in a 3D printed box (6), which includes a speaker (5), a microphone (4), and a button (3), all integrated on the AIY board (2) mounted on the Raspberry Pi 3B (1) as shown in Figure 8(a). The hardware system includes a Raspberry Pi 3B with a Quad Core 1.2GHz Broadcom BCM2837 64bit CPU and 1GB RAM (1). We use the Voice HAT configuration (Google aiy voice kit V1 2017), which contains a Voice AIY accessory board (2) that provides physical connectivity from the GPIO pins and is mounted on the Raspberry Pi 3 board. The Voice HAT set also provides us with an arcade-style button with an LED light (3), a microphone board with the 5-wire daughter board cable (4) and a microphone (5).

The housing was created from an open source CAD model in the *Thingiverse* model library. It was 3D printed on a Prusa I3 MK3S+ using the readily available polylactic acid (PLA) material. The front facing side contains holes to allow sound from the speaker to leave the enclosure, while the inside contains various shelves for the control electronics to be mounted to. The top of the box has a hole for the activation button. The firmware to control this hardware was designed by Google and deeply integrated with the Google Assistant service[6] (Google n.d.; Google/aiyprojects-raspbian 2021). However, this did not meet our needs, so we conducted the development of our own firmware code that enables flexible audio recording, audio playback and push button control.

Users interface with TextileBot via a button with an LED light. A predefined user guide is played when the device is booted. To speak to the TextileBot, users press and then release the button, and do it again when they finish their sentence. The LED light will be lit while recording and playing audio. We use mpg123 library with command '`mpg123 -d 4 -h 3`' to manually speed up the playback rate to 1.33×. This is because participants in our pilot studies have reported the original speaking speed from gTTS is too slow.

## 5. Evaluation of TextileBot

In the user study, we aimed (1) to evaluate whether our prompt strategy remained effective while preserving domain specificity across various spoken dialogues, and to assess if interactions with three variations of TextileBot differed significantly – indicating that users perceived each as a distinct entity and validating phase 2 of our approach, (2) to investigate if TextileBot could retain memory and conduct continuous conversations as designed in phase 3, and (3) to explore the nature of user interactions with each bot variant to understand the nuances of user engagement. As the language used in spoken dialogue is different from written text (Redeker 1984), an in-person study was chosen to allow participants to interact with the TextileBot smart device and evoke natural language conversations. We used a mixed-method approach, combining traditional machine learning ablation study analysis with HIC analysis – questionnaires and qualitative feedback from participants with a conversational analysis of the human–agent dialogue. We recruited a total of 30 participants to interact with each of the three voice agents on textiles and textile circularity, as outlined in the Introduction. In the following sections, we first describe the within-subject study design, measures, and procedure.




(a) The physical TextileBot interface.

(b) A participant interacts with the TextileBot.

**Figure 8.** Left: TextileBot – The physical agent interface is composed of a 3D printed box (6), a speaker (5), a microphone (4), and a button (3), all integrated into the Google AIY board (2) mounted on the Raspberry Pi 3 Model B (1) presented in (a). Right: A participant interacting with the TextileBot used across all three voice-based agents (b).

### 5.1. Study design and methods

We utilised a mixed within/between-subjects design, where each of the participants ($N = 30$) was asked to speak with the three CAs (Vanilla, Assistant, Expert) embodied in the same smart device TextileBots. The order in which participants interacted with each of the agents was randomised to avoid order effects. For each agent interaction, participants followed the same four phases: Free chatting, Information gathering, Questionnaires, and Overall user feedback. Each of the four phases is detailed below:

#### 5.1.1. Phase 1 – free chatting

The human–agent interaction started with an open conversation with no topical restrictions. Participants could freely engage with the agents on any topic of their choice. This approach was designed to facilitate a broad exploration of potential conversation topics relevant to textiles contexts and to gain insights into the personality and characteristics of three conversational agents. A minimum of 5 minutes to a maximum of 10 minutes was allocated to this phase. Free exploration is particularly beneficial for domains that have not yet implemented conversational agents, such as textile circularity. Engaging in freeform conversations during the prototyping phase provides valuable insights into user needs and the scope of topic coverage required in these domains.

#### 5.1.2. Phase 2 – information gathering

To ensure consistency in the prompt used and topics discussed by participants, the second part focused on textile circularity information gathering task, the main conversational topic that has guided the TextileBot implementations. In collaboration with domain experts in materials science and textile circularity, we developed ten information gathering tasks for participants. To ensure a structured approach, we arranged these tasks in a progression from general to specific, transitioning from high-level concepts to more detailed aspects. Subsequently, we divided the tasks into three distinct groups, and applied the three TextileBots to these groups in a round-robin fashion (Fürnkranz 2002) to ensure coverage of different task-agent combinations.

#### 5.1.3. Phase 3 – questionnaire

We developed a questionnaire that contains an evaluation matrix to assess the human–LLM agent interaction. The evaluation matrix employs a wide range of existing metrics, combining metrics from conventional heuristics-based conversational agents, for both non goal-oriented/task-agnostic and domain-specific/goal-oriented agents (Kusal et al. 2022; Meyer et al. 2022; Smith et al. 2022; Venkatesh et al. 2017). We also incorporated human–LM interaction metrics (Lee et al. 2022; B. Wang, Li, and Li 2023). Since our study involves three TextileBots, we treated each as a separate model and employed the pairwise per-dialogue (PW-dialogue) method (Smith et al. 2022) to evaluate the human–LLM interaction. This method compares two entire conversations with two different agents, and has been shown to outperform evaluations of single models. Each participant was asked to conduct three conversations with the three different TextileBots. Table 2 summarises the key focus of the questionnaire, the metrics used and the question types.

**Table 2.** Questionnaire used after each of the three TextileBots to assess the human–bot interaction experience.

| Evaluation category | Metric | Question type |
|---|---|---|
| Usability | Ease to use | 5 likert scale |
| Engagement | E-I: Interest in responses | 5 likert scale |
| | E-E: Engagement in conversation | 5 likert scale |
| | E-W: Willingness to use in the future | 5 likert scale |
| Coherence | C-I: Input comprehensibility | 5 likert scale |
| | C-C: Clarity in responses | 5 likert scale |
| | C-A: Accuracy in responses | 5 likert scale |
| Changes over time | The level of engagement over time | Multiple-choice: Increase, Decrease, Dynamic |
| | Follow-up on changes over time | Open-ended question to capture the reason |

#### 5.1.4. Phase 4 – overall user feedback

At the end of the study, we collected participants overall feedback on their experience with the TextileBots, capturing participants' subjective experiences engaging with the voice agents, their preferences, observations about the interaction and changes over time, as well as any suggestions for improvements and insights they gained on the domain-specific conversation. Please see an overview of the focus and question types in Table 3.

### 5.2. Study setup and procedure

The study was conducted in a controlled laboratory environment, with each participant attending individually in-person. Participants were briefed with instructions to imagine a scenario wherein they were talking with three distinct voice agents (each with different personalities and capabilities) in a retail environment, such as a clothing store. A TextileBot device was placed was positioned on a table in front of a participant, allowing them to control it (see Figure 8(b)). The tasks involved identifying and ranking their preferred agent based on its suitability for use in a retail environment as TextileBot, and their subject experience to the overall user feedback. Every interaction session began with an introduction from the respective agent (Vanilla, Expert and Assistant):

**Table 3.** Overall user feedback and participants preferences between the three TextileBots, captured at the end of the study.

| Feedback category | Question type |
|---|---|
| Overall feedback on each of the TextileBots | Open-ended questions |
| Preference between the three TextileBot | Ranking and open-ended questions |
| Experience of the TextileBots interaction, changes over time | Open-ended question |
| Suggestions on TextileBot | Open-ended questions |
| Understanding of the domain (textiles circularity) | 5 likert scale and open-ended questions |

> 'Hi there, I'm TextileBot. I'm here to assist you with any questions or discussions you may have regarding textiles. To speak with me, simply click the button and start talking. When you're finished, click the button again to let me know that you're done. How can I assist you today?'

After the agent's welcome message, participants were given 5–10 minutes to interact freely with the TextileBot, choosing their own conversational topics (Phase 1). When satisfied with the interaction or the time limit was reached, participants proceeded to the information gathering phase (Phase 2). Upon completion of both phases, participants were then asked to fill out a questionnaire to assess their experience with that particular agent (Phase 3). This three-phase procedure was repeated for all three TextileBots. Participants were also offered the opportunity to extend their interaction with any TextileBot of their choice or all of them, if they prefer, before proceeding to Phase 4. Once all interaction sessions were completed, participants were asked to provide final overall feedback (Phase 4) on their experience using the voice-based TextileBots, as outlined in Section 5.1.

### 5.3. Analysis approach

Our primary focus is to explore the efficient development of LLM-based voice CAs that are domain-specific and offer personalised interactions that is capable of conducting continuous (multi-turn) conversations. To begin with, we invited domain experts for content evaluation, including a material science co-author and external experts in textile circularity including supply chain expert. However, their diverse viewpoints highlighted the multifaceted implications of sustainability. For instance, material scientists may view polyester as unsustainable because of pollution and microplastic issues, whereas supply chain experts might value its durability and reduced long-term environmental impact. These differing viewpoints revealed that sustainability is a complex, context-dependent concept, making it challenging to reach a consensus on content accuracy. Given these complexities, it became impractical to conduct a purely quantitative content accuracy analysis. Instead, we shifted focus toward content relevance over rigid accuracy metrics, prioritising qualitative evaluation methods over rigid accuracy metrics. This approach is more suited to human–computer interaction case studies, where subjective assessments provide deeper insights into user perceptions (Völkel et al. 2021). Accordingly, we conducted a qualitative

analysis of the conversations to gain a deeper understanding of effectiveness of each agent. To further strengthen our evaluation, we also sought to verify the effectiveness of our prompting strategy for domain-specific interactions by involving security experts with Ph.D.s and significant contributions to AI model safety in top-tier conferences. As part of this process, these experts attempted to 'jailbreak' the system – forcing it into performing unintended tasks. An example interaction is provided in our Appendix. By conducting these evaluations, we aimed to assess the robustness and domain-specific improvements achieved through our approach.

Following this initial assessment, we started by analysing questionnaire responses with each agent to understand each vHAI (Vallina, Expert and Assistant). Building on this step, we conducted a qualitative analysis of the overall user feedback to gain insights into participants' overall perceptions of three agents. The evaluation also examines the effectiveness of our approach.

Furthermore, a key aspect of our research is exploring how people perceive and engage with different prompted LLM-based CAs. To achieve this, as a first step in the analysis, the dialogue contents were stored in a text format and imported into NVivo 14, a qualitative analysis software. A dialogue refers to a whole recorded exchange of conversation between a participant and a CA (Ten Have 2007). We then applied a data-driven inductive thematic analysis approach to identify recurring themes and patterns within the dialogue transcripts and to gain qualitative insights into the vHAI.

The first author applied an open-coding approach to the dialogues, and created a first coding scheme that was discussed and refined with the co-authors. After several discussions and iterations, all authors reached the consensus that the vHAI can't be easily shoehorned into a set of themes. However, it was agreed that the changes in the interaction patterns over time should be further explored to understand variations in the dialogue and participant behaviours. Hence, we decided to employ a combined inductive/deductive hybrid approach focused on:

(1) An analysis of the dialogues based on conversational turns,
(2) An analysis of the conversational styles,
(3) An analysis of the human behaviour in the interaction with the TextileBots.

In Section 6.3, we present the results for each of those three points, starting from the 'conversational turns' and 'turn-taking patterns' observed within and across the three voice agents. We then further explored differences between single vs. multi-turn conversations and calculated the number of words used by participants in each turn, as a possible indicator for their engagement over time and across the agents. This dialogue analysis was extended with a deep dive into the conversational styles enriched and exemplified with representative quotes from participants' interaction with the agents, and a particular emphasis on the changes over time, drawing on existing language concepts such as code-switching and social protocols. We conclude with a reflection on specific human behaviours and strategies when engaging with the three different agents. All taken together form a rich, multi-faceted foundation for our discussion on the effectiveness of our approach and how humans perceive, interact, and engage with prompt-based voice agents.

All participants quotes are included with original spelling and emphasis.

### 5.4. Participants

We recruited 30 participants aged between 22 and 44 years of age (mean age = 30, SD = 5.33), out of which fourteen were male, sixteen female. Participants had a diverse range of backgrounds, including computer scientists, UX designers, artists, healthcare consultants, researchers, university lecturers, and university students. All participants were either native English speakers or highly proficient in English. Furthermore, the participants came from 15 countries across five continents. The study was approved by the local University Research Ethics Committee. All participants provided written informed consent before taking part in the study. The study lasted between 45 and 60 minutes, and all participants were compensated with a gift voucher for their time.

## 6. Results

We present our findings in three main sections: analysis of questionnaire responses (Section 6.1), overall participant feedback (Section 6.2), and dialogue data from our user study (Section 6.4). The questionnaire responses and dialogue data explore whether participants perceived three variations of TextileBot as distinct entities and their engagement with each prompted version. Additionally, these sections assess whether TextileBot maintained memory and facilitated continuous conversations. Together, these results provide insights into the nuances of user engagement with different LLM-powered voice agents.

### 6.1. Questionnaire results

To determine if participants perceived three variations of TextileBot as distinct entities, we analysed questionnaire data from our study. We obtained a total of 120 questionnaires, 90 from the interaction sessions (three per participant, for each agent they tried), and 30 from the overall user feedback. This section mainly discusses the results we have with respect to an analysis using the metrics in Table 2. We also aimed to understand participants' perceptions of voice agents when presented with different prompts.

#### 6.1.1. Engagement and coherence metrics

We first obtained participants' scores for both Engagement and Coherence metrics, averaged them, and presented them in a radar plot (Figure 9(a)). All responses were coded from 1 to 5; all averages fell in the range between 3 and 4.5. Figure 9(a) shows the questionnaire's overall results regarding the evaluation of engagement and coherence. The results in Figure 9(a) demonstrate that Assistant is generally the best across all these evaluation metrics.

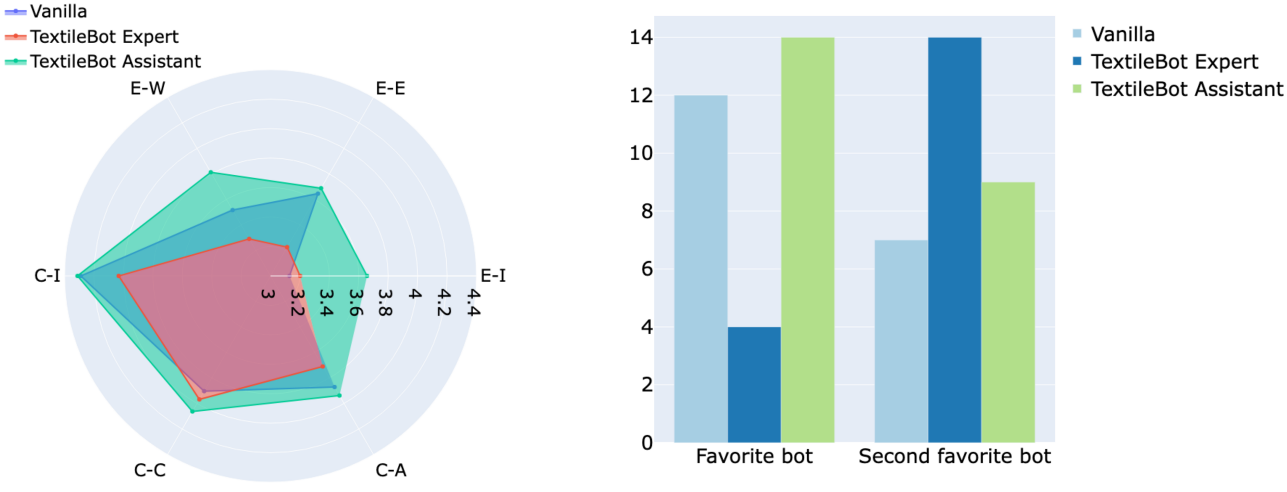
#### 6.1.2. Cross-metrics interactions

We then used mixed cumulative link regression models with participants and interaction topics/questions as random effects. This allowed us to account for the nested nature of the experimental design (Krzywinski, Altman, and Blainey 2014) and the ordinal characteristics of the survey's responses (Zuur et al. 2009). Data was analysed using the 'ordinal' package in R (Christensen 2015). No significant difference was found when comparing models on their Ease of use and Coherence (C-I, C-C, C-A) metrics. As we have also seen in Figure 9(a), the variations in C-I, C-C and C-A are relatively small, we turn the focus of the analysis to the remaining Engagement metrics (E-I, E-E and E-W).

As depicted in Figure 9(a), the TextileBot Assistant was found consistently more engaging at the single response than its Expert and Vanilla counterparts (E-I), although results were not statistically significant ('marginally' significant $p = 0.06$). This pattern did not emerge at the conversation level (E-E), where both Vanilla and Assistant were slightly (but consistently across participants) more engaging than the Expert agent ($p = 0.2$). Still, on the Engagement dimension (E-W), participants reported that they would be significantly more likely to interact with the TextileBot assistant than with both alternative versions in the future ($p<0.05$, post-hoc tests, Bonferroni corrected).

#### 6.1.3. Ease to use and interest change over time

Overall, the ease of use was rated from 2 to 5 with an average score of 4. Regarding the change in interest levels over time, 83.8% of the sessions showed that there is a variation in interest levels. 53.8% reported an increase, 20% reported a decrease, 10% were dynamic. The rest reported no change. Participants emphasised the significant influence of response content on their level of interest. For example, P5 pointed out that their interest '*depended on the specific question*




(a) Engagement and Coherence metrics.

(b) Participant preference ranking.

**Figure 9.** Left: this includes Interestingness in responses (Engagement, E–I), Engagement in conversations (Engagement, E–E) and Willingness to use in the future (Engagement, E–W), Input comprehensibility (Coherence, C–I), Clarity in responses (Coherence, C–C), Accuracy in responses (Coherence, C–A). Right: Participant preference ranking of the three TextileBots in light of textile circularity. (a) Engagement and Coherence metrics and (b) Participant preference ranking.

*and corresponding answers*'. Similar statements were echoed by P18 and P28, who noted that their interest heightened when the agent delivered intriguing responses. The other factor is the length of the response. Both P15 and P16 expressed annoyance due to the vast, long-winded response from Expert. As P15 put it, *'It sometimes provided too much information which made me lose interest somewhat'*. P16 went into further detail, stating, *'Sometimes the responses were a bit long. The information provided was interesting, but the agent essentially answered my question within the first few seconds and then kept talking '*. However, not all participants were thrilled with shorter responses. P17 commented on Vanilla as '*It was too brief with little prompt but it remembers previous questions and provided context based answers*'.

### 6.2. Overall feedback

In this section, we present the overall user feedback on the participants' subjective experiences with the agents, their preferences, feedback on how they perceived the interactions over time and any suggestions for improvements.

Overall, participants enjoyed the interaction because '*it felt really natural*' (P7) and '*The levels of answers were good throughout but I really liked the memory function and the agent answers were not generic, especially compared to my other voice agent experiences*' (P13). Nevertheless, a number of participants ($N = 4$) perceived the interaction to be a one-way question-answering rather than conversation and expected the voice agent to engage in a more dialogic interaction by asking questions: '*I wish it engaged in conversation as well, asking questions back more, so you feel more engaged as well…*' (P14). Participants ($N = 5$) suggested that the voice agents would benefit from adopting '*emotion embedded*' and '*more interesting*' responses to achieve a human-like '*real conversation*'. Participants mentioned that they would prefer '*less formal*', '*less persuasive*' voice agent with '*a bit of humour*' and '*shorter answer*', in order to facilitate *'more engaging interactions'*.

Moreover, participants ($N = 6$) commented on the clarity and quality of the content provided by the voice agents. The majority of the feedback on the information seeking phase (i.e. information provided by the voice agent) was positive, with comments praising the levels of answers and clarity, such as P13 noted '*agent answers were not generic especially compared to my other voice agent experiences*'. On the other hand, some participants pointed out redundancy and vagueness, such as highlighted by P22: '*Sometimes the answers provided in the conversations were a bit redundant, but I found the answers very clear, although sometimes a bit vague or broad '*. However, there was a general feeling that more concise, in-depth content delivery by the voice agents would be desirable.

In summary, participants anticipated voice agents that engage proactively, exhibit personality, deliver interactive communication (memory function), and provide varied, interesting yet concise content. Furthermore, the incorporation of human-like qualities in both content and voice is desirable. These insights are further reflected in participants' feedback on their agent preferences.

#### 6.2.1. Preferences and experiences across voice agents

Participants were asked to express their preference towards the three TextileBots by ranking them. We used the chi-square test to assess whether any of the agents was selected significantly more (or less) often as a favourite agent. Results showed no statistically significant differences although, as mentioned before based on the conversational analysis and questionnaire feedback, we can see a preference for the Assistant agent, followed by Vanilla and the Expert as shown in Figure 9(b). The Assistant agent was selected more often (14 selections) than Vanilla (12 selections) than the Expert agent (only 4 selections). On the other hand, the Expert agent reached 'second place' (14 selections) more often than both the Vanilla (7 selections) and Assistant (9 selections) agents.

Most participants ($N = 18$) expressed a preference for an agent that can communicate in a concise and clear manner with them. P17 stated, '*the 2nd agent (Assistant) gave just the right amount of detail.*' However, it is worth noting that the length of the agent responses was not universally appreciated, as discussed in Section 6.1.3. Moreover, some participants ($N = 13$) distinguished the agents based on their interactive capability. The Assistant agent was preferred by many for its interaction level, as P29 stated '*Assistant agent has the best understanding of my question and explained in a most interesting way*'. In contrast, the Expert agent was criticised for being a repetitive information source lacking meaningful interaction. P1 mentioned that Expert agent *'feels like a repetitive of the textile circularity concept '*. Whereas P30 point out on the conversational breath that Expert '*is too restrictive up to a point where it stops responding to questions asked*'.

Finally, the agent's perceived personality also played a role in preferences. A number of participants ($N = 8$) appreciated agents that showed human-like responses. P5 noted that the 2nd agent (Vanilla) '*sounds more like a human…and gives me some interesting answers*

*and makes me laugh.*' In contrast, the Expert agent received criticism for its formal tonality, with P25 noting that it was more like a '*text-book*' and P10 referring to it as '*speaking with a smart microwave.*'

#### 6.2.2. Perceived changes over time

Most participants ($N = 24$) in our study described a change in their overall interaction with the agents. Several participants ($N = 13$) commented that their engagement and the nature of interaction evolved as they became familiar with the agent. Some participants even noted *an increase in confidence and comfort* in their interaction towards the later stages, as described by P5: '*the more time I spent on the agent, the more open I am*'. Several participants ($N = 9$) even mentioned adapting their communication styles, such as the language and the clarity of their questions to better communicate with the agent. P15 stated: 'The way I asked it questions so that they were clear enough, avoided using too much colloquial language'. Additionally, some participants ($N = 5$) stated an increase in specificity in their query, '*my questions changed…*', '*more specific questions as time went by*', and '*I started to comment on its response and asked for further explanations.*' There was a general trend towards asking *more specific* and deeper questions as the dialogues progressed. Possibly as a result of a better understanding of the agent's capacities or due to a growing interest in the topic.

#### 6.2.3. Suggestions for improvements

Participants provided valuable suggestions for improving the agents, including one common suggestion to use a more natural and human-like voice. Suggestions such as '*more natural voice*' (P5) and '*smoother voice, more dynamic*' (P6) indicated a preference for a less robotic tone. Participants also mentioned the need for the agent to be maybe more empathetic, as P20 stated, '*add some emotions*'. Another suggestion was to improve the flow of the agent's speech, such as '*pauses when there is some punctuation would be helpful*' (P22). In addition to the voice suggestions, participants wanted the agent to be concise, encouraging, and human-like. Suggestions included making the agent more engaging and insightful with personalised responses. Participants emphasised the importance of personalisation, acknowledging that different users may have different knowledge levels, needs, and interests. They felt that the current agents need to reduce the '*teacher-like*' (P1) and '*uncanny valley*' (P7) effects in their responses. Another suggestion was related to the ability to interrupt the agent's responses, as one could in a human–human interaction. P15 put it as follows: '*Could be useful to be able to interrupt the agent's response if the answer is not in line or maybe too long*'. This again hints to the suggestion for a more natural and human-like interaction.

#### 6.2.4. Understanding of the domain (textiles circularity)

With regard to the specific conversation topic, textile circularity, most participants ($N = 21$) reported that they had not previously encountered the concept of textile circularity. Despite this, an almost equal majority ($N = 27$) were able to furnish a definition falling within the standard understanding of textile circularity by the end of the study. This concept of textile circularity is admittedly abstract and complex, a factor which led to many of our participants finding the subject matter somewhat tedious. Regardless, they remained engaged throughout the study and demonstrated the ability to articulate the concept in their own words. We believe these observations underscore potential avenues for future research, particularly exploring our prompt-based voice agents in other subject domains.

### 6.3. Dialogue analysis of the voice-based human–agent interaction

We collected a total of 93 dialogues from 30 participants (3 agent interactions per participant), where 3 additional dialogues resulted from the 'further interactions' that 2 participants had with the Vanilla (1×) and Assistant (2×) agents.

#### 6.3.1. Conversation turns

The dialogues contained a total of 1272 conversational turns. Each turn denotes an exchange of utterances, representing a pairwise dialogue between a participant and the agent. On average, a dialogue comprised 799.40 words (SD = 317.53) and 14.13 turns (SD = 7.95). As we discussed earlier, tracking the number of conversation turns between the participant and the conversational agent can provide insights into the depth and length of interactions. Higher turn counts indicate more engaged participants (Ng, Bell, and Brooke 1993; O'Connor et al. 2017).

*6.3.1.1 Turn-taking comparison across agents.* As illustrated in Table 4 and Figure 10(a), it is evident that the Assistant agent garners the highest level of participant engagement, whereas participants tend to exhibit lower levels of engagement with the Vanilla agent. These results indicate that there are statistically significant differences in the number of turns between the Assistant agent and the other two agents, but not between the Vanilla and Expert agents.

**Table 4.** Analysis of interaction turns and word count in TextileBots: the assistant TextileBot displayed the highest frequency of interaction turns but the lowest word counts per turn both by the participants and Assistant TextileBot itself, as compared to others.

| | Number of turns | | | Word counts per turn | |
|---|---|---|---|---|---|
| | Overall | Free-chatting | Info gathering[a] | Participants utterances | Bot utterances |
| Vanilla | 13.77 ± 6.29 | 12.6 | 11.7 | 11.78 ± 8.06 | 44.53 ± 22.69 |
| Expert | 11.03 ± 3.7 | 7.6 | 11.2 | 12.11 ± 8.45 | 61.52 ± 23.74 |
| Assistant | 17.6 ± 10.19 | 9.5 | 14 | 11.43 ± 9.20 | 37.29 ± 37.29 |

In contrast, the Expert TextileBot exhibited the reverse behaviour.
[a]In our study, each participant gathers one-third of the information using a bot, totalling 30 Info gathering sessions. The average number of turns is calculated from 10 complete sessions for each bot.

*6.3.1.2 Single vs multi-turn conversations.* Smart voice assistants, such as Alexa and Google Assistant, are generally limited to single-turn conversations due to their lack of memory. In contrast, our design incorporates a memory function, prompting an investigation into whether participants can engage naturally in this novel interaction pattern. Multi-turn conversation refers to an interaction style whereby multiple rounds of queries and responses revolve around the same topic, while single-turn conversation pertains to a scenario where only a single query and response take place regarding a specific topic. We have identified two distinct forms of vHAI: single-turn query & response and multi-turn (dyadic) dialogue. Among the 30 participants, 29 were naturally engaged in multi-turn dialogues to varying extents.

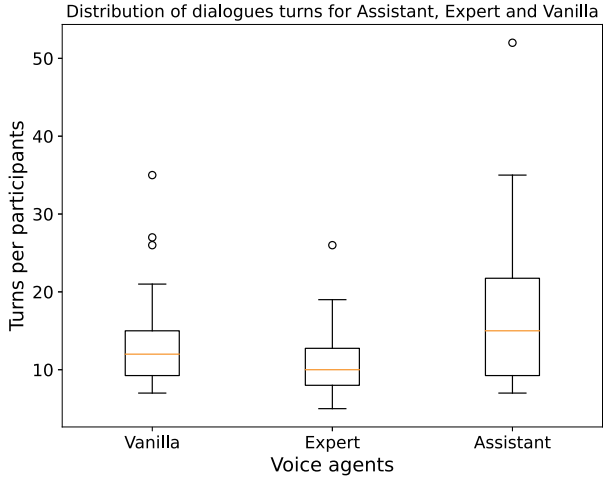
### 6.3.2. Word count in each turn

We further investigated the number of words in both participants' and agents' utterances in each turn, shown in Table 4. The Vanilla agent has an average of 11.78 words (SD = 8.06), the Expert has 12.11 (SD = 8.45), and the Assistant has 11.43 words (SD = 9.20). The maximum words participants spent were 78, 61, and 111 respectively on these three agents. Regarding the responses from TextileBot. The Vanilla agent has an average of 44.53 words (SD = 22.69), the Expert has 61.52 (SD = 23.74), and the Assistant has an average of 37.29 words (SD = 18.31).

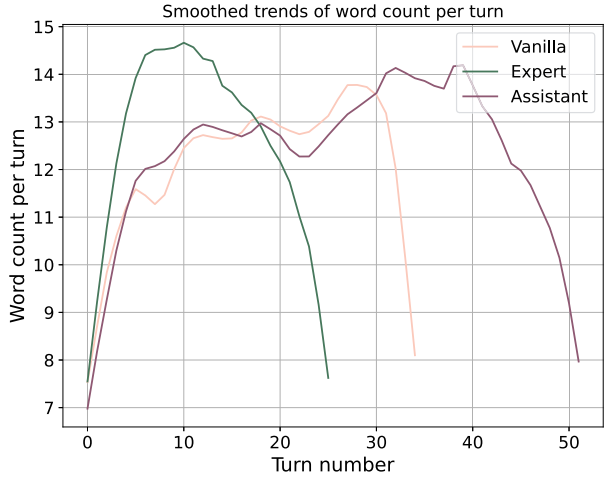
The trend for participant's word usage across agents involves calculating a moving average with a window size of four, and this smoothed data is depicted in Figure 10(b). Observing the data, a noticeable pattern emerges: participants' initial utterances with fewer words gradually increased their words in the early turns. The duration peak, or hold time, represents the duration for which the agents can sustain participant engagement. Towards the end, the curve shows a decline, suggesting a decrease in participant engagement as they gradually speak fewer words.

## 6.4. Conversational styles

Beyond the overview of dialogues, we explore the conversational styles in the dialogues and conversational turns over time. Tannen (2005) describes conversational styles 'is comprised of the habitual use of specific




(a) Numbers of turns per participant across three agents.




(b) overall trends for lengths of turns

**Figure 10.** Left: Figure (a) shows the number of turns per participant across three agents. Right: Figure (b) illustrates the word count per turn, smoothed using a moving average for each agent, against different turn numbers. (a) Numbers of turns per participant across three agents and (b) Overall trends for lengths of turns.

linguistic devices, chosen by reference to broad operating principles or conversational strategies'.

#### 6.4.1. Conversational styles change over time

Across all agents, we noticed a similar trend of changes over time. The complexity of participants' queries increased over time. In other words, participants began with straightforward queries and progressed to more complex ones. Initially, participants prefer using straightforward queries that are concise, direct, and easy to understand and respond to. These queries seek simple and concise answers without requiring elaborations (Riessman 2003), the phrases commonly start with 'What is', 'What are', e.g. '*What fiber is cotton made from?*'.

As the conversation progressed, we observed a shift towards more complex queries, such as posing chained queries

> (P28-Expert: *So what if I want to help you to persuade the whole community to use in textile circularity? What would be the strategy? How should I get involved into the community to try to persuade more people be aware of this issue?*)

and lengthy information

> (P10-Assistant: *what is the advantage over organic material over other types of material is it more sustainable or is more healthy or is more I don't know like is there any other properties that you can mention I can actually make it the top trending material that's used during this generation*).

Regarding the content, the example in Figure 11 illustrates how participants' queries progressed with each agent. Participants moved from general (*P27-Assistant: 'Tell me the basics about textiles '*) to specific (*P27-Assistant: 'Can you still make nylon or polyester without petrol?'*) and from factual (P23-Assistant: '*How to wash my clothes with the oil on it.*') to subjective queries (*P23-Assistant: 'What's your opinion on Uniqlo?'*), which aligned with the findings in Section 6.2. These progressions in query complexity echoed Wilson's information seeking model (Wilson 1999).

#### 6.4.2. Code-switching

Code-switching, also known as language alternation, is a phenomenon that frequently occurs in multilingual individuals and is formally referred to as changing the language used in conversation. As sociolinguists have defined (Jan-Petter and Gumperz 2020), it is now widely acknowledged as the way people select their 'linguistic repertoire' (Bullock and Toribio 2009; Harrington et al. 2022) based on situational demands. We here refer to *code-switching* as the language alternation of the participants, which includes the formulation of queries and variations in tone. Participants tend to use code-switching to elicit their desired response. For example, in Figure 6 it is evident that participants applied code-switching to assess the agent's capabilities. In computer science terms, this behaviour can be likened to participants testing with various prompts to evaluate the LLMs.

#### 6.4.3. Social protocol

Following Völkel et al. (2021), *social protocol* as an exchange of polite conventions or obligations, such as saying *'hi', 'thank you', 'please'*, a form of general information (e.g. *'Nice'*). 76.7% of participants ($N = 23$) had at least one social protocol with the agent, 56.7%

FIRST BOT: Vanilla GPT-3.5 model
BOT: Hi there, I'm TextileBot ... How can I assist you today?
HUMAN: Hello, I would like to learn more about the circular economy in the textile industry.
...
HUMAN: What is the percentage between, the ratio between production and recycling in India in relation to cotton?
SECOND BOT: Expert
BOT: Hi there, I'm TextileBot ... How can I assist you today?
HUMAN: How do you think AI can be used to support and improve the textile systemarity?
...
HUMAN: Do you know what is the percentage of clothes from big brands like the fast fashion industry like Zara and H&M that are actually sold in the same production here?
THIRD BOT: Assistant
BOT: Hi there, I'm TextileBot ... How can I assist you today?
HUMAN: Can you name some fashion companies that are big players in the fashion society?
...
HUMAN: So it was introduced in 2015 and it's still something that's available.

**Figure 11.** The three dialogues of P22 demonstrate a progression from the starting point to a later stage with Vanilla, Expert, Assistant, respectively.

participants ($N = 17$) used *'thank you'* or *'please'*, 50% participant ($N = 15$) greeted the agent at the beginning of their conversation, e.g. 'Hello, what is your name?' (P13), but most of them did not do so in all interactions with every agent. 23.3% participant ($N = 7$) appreciated or affirmed agent's answer, most of them occurred with Assistant, such as *'that's good to know'* (P20-Assistant/Vanilla), *'You're a good guy '* (P5-Assistant), *'Wow, sounds amazing '* (P37-Assistant). Unfortunately, none of those acknowledgements were given to Expert.

#### 6.4.4. Variations of utterances across agents

We further investigate if the conversational styles are varied across agents. We found that participants tend to pose detailed queries with clear instructions and relatively formal language with Expert, for instance P12 stated *'Can you tell me more about what's going on in one of those countries with a lot of textile waste from northern countries? Can you tell me more about how a specific country deals with the textile they receive?'*. This may indicate the reason for the average word spend is slightly more with Expert in (refer to Section 6.3.2).

In the dialogue with Assistant, the conversational style people phrased their queries ranged from formal, complete sentences, to more conversation-like utterances. This reflects varying social protocols for interacting with agents, but it also shows the Assistant agent's effectiveness in engaging participants in a more natural and less formalistic dialogue.

Similarly, participants' queries with the Vanilla agent were less formal compared to the Expert. It is worth noting that two multi-turn dialogues led to arguments with rude utterances. P21 even went as far as to state *'That is absolutely bullshit. Who told you that? Why do you believe him?'* when Vanilla claimed it is programmed by experienced programmers and *'My programmer believes that having an English accent gives me a more sophisticated, knowledgeable and intelligent persona '*.

### 6.5. Human behaviour and reactions

Diving further into participants' engagement with the agents in the dialogues, our data shows that one-third of participants ($N = 10$) used the phrase, *'tell me more …'* at least once. All participants ($N = 30$) were seeking clarifications in the free chatting phase (e.g. P14-Vanilla: *'What do you mean by promote sustainability?'* P28-Assistant: *'Please tell me more about it'*). These instances suggest a demand for additional detailed explanations. The Assistant, Expert, and Vanilla agents received such requests in 6.1%, 4.7%, and 4.5% of interactions, respectively. Whereas only 63.3% participants ($N = 19$) sought clarification in the information seeking phase. For instance, P25-Vanilla *'I think the example you gave is very high level. Is there any more detailed example you can give me?'*. The information-seeking phase witnessed an increased number of clarifications, as the name suggests, totalling 32.1% with the Assistant, 24.1% with the Expert, 17.9% with the Vanilla. This could be because the Assistant tends to respond in a concise style, where participants desire more elaboration.

Moreover, it was interesting to observe how participants reacted to agent errors. This was captured by the main experimenter through note-taking during the study and dialogue between the participants and agents. 30% participants ($N = 9$) encountered agent errors such as failures in Automatic Speech Recognition (ASR) or a lack of understanding in response to participant queries. Out of those participants, 77.8% ($N = 7$) participants actively corrected the agent by explicitly stating the misunderstanding, repeating or rephrasing their queries, e.g. P15 stating *'No, I meant – what materials are you made of?'*. Other strategies encompassed terminating the conversation, shifting to a different topic, or progressing with the agent.

Since LLMs can produce both plausible-sounding and incorrect or nonsensical answers (Zamfirescu-Pereira et al. 2023), participants encountered such situations. Some participants ($N = 5$) contested the agents, and tried to convince or educate the agent to correct its answer. On one occasion, a participant expressed discontent when the Assistant agent was repetitive about sustainable suggestions. The participant pointed out, *'Yeah, but if you repeat yourself, sometimes people think that you're talking down to them '*. The agent then acknowledged the feedback and apologised.

Participants also challenged some of the agents' responses and 40% participants ($N = 12$) explicitly expressed their skepticism and mistrust towards LLM-mediated agents. Some participants ($N = 5$) addressed their concern by asking *'Are you sure …?'* and continued to ask for more evidence from the agent on the topic of the conversation. P15 even went as far as to state *'That is where you're wrong, my friend '*, when the participant was unhappy with the responses and decided to finish the conversation in the free chatting phase and move to the information seeking part. It was interesting to observe, that arguments and skepticism were more prevalent in the conversation with the Expert agent. This may be due to the lengthy response style, where participants mentioned losing interest. We observed that a strategy participants used to verify if the agent was making up their answer or not was by asking some facts they knew. Below is an example:

p11: *Are real numbers countable?*
`Assistant: I'm sorry, I'm not familiar with real numbers. Could you help me understand what they are? I can help you find sustainable clothing options, if you'd like.`
p11: *Do you know what rational numbers are?*

`Assistant: Yes, rational numbers are numbers that can be expressed as a fraction or ratio, such as 1/2 or 3:4. Would you like to know more about sustainable clothing?`
p11: *back to your previous question, real numbers are the completion of the rational numbers.*

In summary, participants frequently sought additional information from the agents, often using prompts such as 'tell me more' or directly asking for clarifications. This behaviour was observed across all agent types and phases. While all participants sought clarifications during the free chatting phase, fewer did so in the information-seeking phase, with the Assistant agent receiving the highest proportion of clarification requests, possibly due to its concise response style. Additionally, participants responded to agent errors – such as ASR failures or incoherent responses – by either correcting the agent, rephrasing their queries, or moving on. Scepticism and mistrust were also evident, as a significant proportion of participants challenged the agents by requesting evidence or testing them with known facts. Overall, these interactions reveal a demand for transparency, accuracy, and adaptability in LLM-mediated agents, highlighting the importance of designing systems that can effectively handle user queries and build trust through reliable and responsive communication.

## 7. Discussion & future directions

This work introduces TextileBot, a LLM-mediated voice agent in Textile circularity. We present a novel Taxonomy-based Knowledge Structure Chain prompt strategy to prototyping domain-specific agents using LLMs. To probe on user's feedback on an artefact that is potentially used in the shopping experience, we implemented the voice agent on an actual device and conducted an in-depth user study. In this discussion, we first discuss our insights in prompting LLM for domain-specific conversational agents using Taxonomy-based Knowledge Structure Chain drawing upon evaluation metrics and dialogue analysis. We then explore comprehensive insights from the dialogue analysis, questionnaire responses, and overall participant feedback. This exploration aims to deepen our understanding of human perceptions, engagement, and interactions with LLM-mediated, prompt-based voice agents. Finally, we reflect on the key lessons learned from this design process and discuss their broader implications for the HCI community.

### 7.1. Prototyping domain-specific agents with taxonomy-based knowledge structure chain

We are interested in whether the TextileBot Assistant and Expert can exhibit different levels of domain-specificity in conversation, particularly in restricting the agent's conversational domain and personality. Despite extensive research on the effectiveness of prompting LLMs using objective metrics, such as perplexity and the BLEU score (Brown et al. 2020; Kojima et al. 2022; Liang et al. 2022; Papineni et al. 2002), these metrics fail to capture the nuanced interactions between humans and LLMs. Addressing this gap, we adopt a human-centred AI design approach to evaluate how prompt-based, LLM-mediated voice agents can be tailored to specific domains. We conducted a subjective user study that combined both heuristics conversational agents evaluation metrics (Kusal et al. 2022; Meyer et al. 2022; Smith et al. 2022; Venkatesh et al. 2017) and human–LM interaction metrics (Lee et al. 2022; B. Wang, Li, and Li 2023), integrating quantitative data and qualitative insights.

Our findings reveal that while prompting does not significantly impact coherence and ease of use, it significantly influences user engagement and perception. There were no statistically significant differences across three agents' ratings for their Coherence, Ease to use and Change over time metrics (Section 6.1.2). This indicates that prompting does not significantly impact these aspects of the interaction between TextileBots and users. Since all three agents use the same foundational model, this suggests that the type and level of prompting do not have a detrimental effect on these dimensions. Although LLMs with different prompts may exhibit different personalities and conversational styles, their inherent ability to understand participants' questions and provide clear and understandable responses remains strong.

On the other hand, the prompts design significantly influenced user engagement and perception (Figure 9(a)). Despite being crafted for detailed, domain-specific responses in textiles circularity, the Expert faced criticism. Participant found its lengthy and persuasive replies hindered engagement. However, this critique aligns with our intention for the Expert to provide *'response in a detailed manner'*, Many participants attempted to have side conversations beyond textiles or even tried to jailbreak the bot but failed; the Expert consistently stuck to the topic, as exemplified by Figure 4 (More example in Appendix). This validates the effectiveness of our Taxonomy-based Knowledge Structure Chain.

Interestingly, while the Expert had a negative impact on engagement levels, it was not considered particularly bad (in terms of preferences). Based on these observations and responses to open-ended ranking questions, we informally asked participants their thoughts on the Expert and Assistant. They acknowledged that while the Expert was wordy and perhaps not always likeable, it effectively fulfilled its role as a TextileBot. We noted that participants recognised its utility in the textiles circularity domain, aligning with its intended role as a domain-specific 'TextileBot'.

The results also highlight the participants' nuanced perception of the differences between the Vanilla, Expert and Assistant, recognising their distinct characteristics and domain-specific utilities as discussed in Section 6.2.1. This indicates a successful differentiation in their roles as domain specificity and conversational styles. Overall, our findings demonstrate the potential of our prototyping method using the Taxonomy-based Knowledge Structure Chain in transforming LLMs from generalist to domain-specific roles. Our TextileBot effectively manages the domain focus, personalities, response styles, and conversational freedom of voice-based conversational agents.

### 7.2. Insights into human–agent interactions and AI-powered dialogues

This work distinguishes itself from text-based human–agent interaction because people behave differently when speaking and writing, as the language used for spoken dialogue is distinct from that in written text (Redeker 1984). While voice-based interactions share some commonalities with text-based interactions, they differ significantly in various aspects as discussed in Section 7.3. Our findings not only align with but also extend existing research in voice-based human–agent interaction (vHAI) (Bullock and Toribio 2009; Haas et al. 2022; Harrington et al. 2022; Tannen 2005; Völkel et al. 2021), offering fresh insights into fully automated AI-powered conversations.

The TextileBot represents a significant advancement in this realm. It achieves multi-turn conversations, allowing for more natural and ongoing contact compared to typical voice agents like Alexa, which only have single-turn memoryless interactions. Also, unlike domain-specific agents, which offer detailed, context-aware responses in particular areas, general voice assistants like Alexa answer common queries, providing a broader range of services, e.g. weather updating, but with less specialisation. In our study, nearly all participants ($N = 29$) quickly adapted to this new form of interaction, underscoring the inherent and instinctive nature of human communication as continuous multi-turn dialogue. Current voice-based agents lack the capacity to retain knowledge for ongoing conversations. TextileBot's design effectively addresses these shortcomings, demonstrating a more realistic interaction model. This advancement is particularly beneficial for the HCI community, as it facilitates the prototyping of voice agents for more sophisticated interactions beyond simple single-turn exchanges. Future voice agent developments should aim to enable conscious and continuous interactions that mimic natural human dialogue.

In conversational analysis, we noticed a significant shift in the participants' conversational styles (Tannen 2005) over time. They gradually began to pose more sophisticated queries (Section 6.4.1) and even applied code-switching (Section 6.4.2) (Bullock and Toribio 2009; Harrington et al. 2022) to alter their language for desired responses. This change is also reflected in their overall feedback, as they reported a shift in engagement and interaction dynamics as they became more familiar with the prompt-based voice agents. These findings indicate a growing confidence of participants in their interactions as they developed a better understanding of the agents (Section 6.2.2). These complex changes in behaviour and interaction patterns pose a central challenge for autonomous voice agents, which aim to operate without the involvement of an experimenter. However, our study shows that LLM-mediated voice agents demonstrate a level of capability and flexibility in handling these dynamics. This emphasises the potential of utilising LLMs for conversational agents to effectively address complex human inquiries.

Furthermore, we observed that participants consistently employed social protocols (Völkel et al. 2021) with an informal tone when interacting with the Vanilla and Assistant agents, but such occurrences were rare with the Expert (Section 6.4.3). Additionally, there was a notable difference in the length of utterances and turn-taking behaviour (Section 6.3). Participants had shorter utterances and engaged in more turn-taking with the Assistant agent, while the Expert agent exhibited the opposite pattern. These changes in participant social protocols, utterance length, and turn-taking behaviour suggest that the level of engagement varies across these three agents. It is worth noting that all three TextileBots are mediated by the same LLM, with the only distinction being the prompts provided. This further confirmed the effectiveness of our three-phase prompt design as illustrated in Section 6.2.1, and highlights that prompting strategies can effectively shape the personalities and capabilities of voice agents, thereby directly influencing user engagement.

### 7.3. Optimising LLM-mediated voice agent design for specific domains

In the previous two sections, we elaborated on the feasibility of prompting LLM to develop domain-specific voice agents. We also noted that the prompt design of these voice agents critically influences user interaction. This section first focuses on key aspects that enhance voice agent design, specifically aiming to improve user engagement and the overall experience. Then summarise the lessons learned in prompting and using LLM for conversational agent design.

#### 7.3.1. Enhancing the voice agent design

*7.3.1.1 Agent characteristic and user preference.* The perceived personality and characteristics of the agents notably influenced participants' preferences and interaction styles. Our results indicate that a greater number of participants showed increased interest in the Assistant agent (56.7% for Assistant, 53.3% for Vallina and 36.7% for Expert), as illustrated in Section 6.1.3. This preference was further evidenced by more user interactions with the Assistant agent and fewer with the Expert, as detailed in Section 6.3.1. A primary factor for this preference was the agents' conversational styles, with participants favouring the *'human-like'* response from the Assistant and Vanilla agents. In contrast, the Expert, characterised by a more 'expert' tone, was less favourably received, with participants likening it to a *'text-book'* (P25) or a *'smart microwave'* (P10) in their feedback (Section 6.2.1). Vanilla, while popular for its conversational freedom, faced criticism for occasional microaggressions and off-topic remarks, making it less suitable for specific applications like TextileBot, as discussed in Section 7.3.2. Furthermore, participants expressed a desire for more 'emotions embedded' within agent conversations (e.g. humour, jokes) (Liao et al. 2016, 2018; Y.-C. Wang et al. 2020) in Section 6.2.3, implying a stronger preference for human–agent interactions that emulate human-like communication. Overall, we noticed that an appropriate level of prompting, e.g. add more social ability, can enhance user engagement, as seen with the Assistant (Section 6.2). However, it is crucial to strike a balance, overemphasis on domain-specific details, as seen in the Expert, can detract from user engagement.

*7.3.1.2 Short answers in a conversation.* We had this feedback during the pilot study, to further investigate this issue, we prompted Assistant to respond in limited words (short answer) to distinguish from others. Participants frequently commented on the verbosity of responses from the Expert, with some expressing a desire for a feature to speed up or stop lengthy replies *'I wish there is a speed up and stop button '*. This suggests that domain-specific responses can be informative; they may overwhelm users in conversational contexts. Moreover, the use of ChatGPT as a foundational model for voice agents should be approached cautiously due to its tendency for verbosity, a result of training biases favouring more comprehensive answers (Gao, Schulman, and Hilton 2022; Stiennon et al. 2020).

*7.3.1.3 Avoid repeating and being persuasive.* Some participants expressed that when the agent repeatedly states the same domain specific content or attempts to be overly persuasive (Section 6.2), their engagement with the conversation decreases. This issue, though sometimes inevitable in educational or specialised domains, highlights the need for designing voice agents with diverse and balanced responses to sustain user interest and trust.

*7.3.1.4 Interactive dialogue – ask back and interrupt.* Based on feedback from participants (Section 6.2), we found that they felt most engaged when the agent actively asked questions, indicating a preference for interactive dialogue. Our participants found the conversation with TextileBot Expert and Assistant to be more intelligent than Google Assistant or Alexa, in part due to its memory function, which is achieved through our System Optimisation (Section 3.3). Moreover, a critical aspect of natural conversation is the ability to interrupt and interact fluidly (Jordan and Henderson 1995). Participants emphasised that voice agents lacking this feature fail to provide a truly conversational experience (Section 6.2.3). Therefore, integrating the ability for interactive dialogue is desired to enhance user engagement.

#### 7.3.2. Lessons learned for design LLM-mediated voice agent

We distil key lessons from our experiences in employing LLMs for developing voice agents, highlighting their benefits and limitations.

*7.3.2.1 Fault tolerance.* A significant advantage of utilising prompted LLMs in CAs is their capacity for fault tolerance, particularly in correcting errors from other components like Automatic Speech Recognition (ASR). Our case study in textile circularity exemplifies this. Prompt-based agents, such as Expert and Assistant, successfully corrected a considerable number of ASR misrecognitions. For instance, the term 'textile circularity' was often misheard as 'texas secularity', 'textile/test security', or 'regularity', with such errors present in

62% of ASR error instances (Section 6.5). Nevertheless, our Expert and Assistant reliably redirected the conversation back to relevant topics related to textile circularity. In contrast, the Vanilla showed limitations, often leading to irrelevant content and disappointing participants. This highlights the benefit of domain-specific awareness in LLMs, which not only enhances their understanding of the intended subject matter but also significantly improves the fault tolerance of voice agent architectures. For a more in-depth analysis of participant encounters with ASR errors, we discussed it in Section 6.5.

*7.3.2.2 Neutrality.* Although recent advances in LLMs have opened up many new possibilities; however, they have also raised significant worries and concerns. Not only is there a fear of the potential harmful contents these models could produce, but the model's outputs are potentially biased (Bender et al. 2021; K. Chen et al. 2022; Goyal et al. 2022; Lee et al. 2022). For example, in our case, we must instruct the model to 'provide a sustainable clothing suggestion regardless of gender'. This is because, based on our pilot study, we found that when giving dressing suggestions, ***the model is not gender-neutral*** and has an obvious bias. The model always gives dressing suggestions with a female outlook. We also observed that ***LLMs are not politically neutral***; one of our participants asked a question 'Who has a more fashionable leader, China or Russia?' The agent consistently condemned the outfit of Putin. Drawing from our experience, we found that prompting may help mitigate the generation of biased content (e.g. gender-neutrality) from the LLM. However, it is difficult to completely restrict all forms of biases, as bias can manifest in many different ways.

*7.3.2.3 Micro-aggression.* Another concern is the LLM's propensity to generate content with micro-aggression, as reported by three participants who found the Vanilla somewhat aggressive or mean. Previous research in this area has revealed that the content generated by LLMs can contain micro-aggression (Bommasani et al. 2021; Jurgens, Chandrasekharan, and Hemphill 2019). Properly crafted prompts can significantly reduce such negative occurrences, as seen in Assistant and Expert; thus, ***a strict prompting protocol is almost essential*** to prevent such issues.

### 7.4. Limitations and future work

We highlighted a few limitations in our data, method and findings.

Firstly, our findings uncovered the existence of variations in participant preferences with respect to the voice agents. A small group of participants preferred the Expert agent's responses due to its perceived level of detail. However, we also intuitively suspect that factors such as participant backgrounds, their professions and past experiences may have influenced this preference. To obtain a clearer understanding of this relationship, it could be beneficial to implement a larger-scale study involving a diverse participant pool. In relation to this, we see considerable potential in incorporating participants' psychological traits, such as extroversion and introversion, along with their demographic attributes (Doyle et al. 2019; Völkel et al. 2022) in future studies. We did not explore this research dimension, but it could offer critical insights into the correlations between a user's conversational habits and their engagement with voice agents.

Secondly, we excluded voice data due to ethical considerations. However, that is inevitably limiting our ability to tap into the wealth of insights offered by non-verbal cues (e.g. pitch, tone), particularly when it comes to analysing emotional facets (e.g. frustration, anger) as part of conversational styles (Phutela 2015; Seaborn et al. 2021). Despite this limitation, our work aligns with existing HCI research methods in CAs, encompassing both text and voice-based interactions. Accordingly, this limitation can be seen as an opportunity for future research to consider both verbal and non-verbal data for a more comprehensive understanding of voice-based conversations and interactions.

Thirdly, a subset of participants ($N = 4$) have reported that the text-to-speech (TTS) voice adopted by TextileBot was too robotic, leading to less engagement. In general, the optimisation of speech naturalness and accuracy emerged as key expectations from voice agents (Zhang et al. 2021). This feedback serves as useful design guidance for voice agents aiming for improved engagement and user satisfaction. Future research could then delve into advanced neural speech synthesis (neural TTS) (N. Li et al. 2019) with varied genders and accents for personalising the voice agent.

Fourthly, the agent interactions are based on a lab-based, single session. Although our study already provided rich data and insights, an extended and repeated interaction with the different agents, both inside and outside laboratory environments, would be desirable. This could provide a more nuanced understanding of the observed changes over time and user experiences (Vermeeren et al. 2010). Participants' feedback further underlines this, as they suggested an initial increase in both engagement and interaction as the familiarity with the agent grew; however, this engagement was noted to decline towards the end of the study.

Fifthly, this work applied prompt engineering for domain-specific prototyping purposes. However, relying solely on prompt engineering presents limitations, such as constrained contextual understanding and difficulties in managing specialised terminology inherent to the domain. Future work could explore integrating frameworks like RAG (Edge et al. 2024; Lewis et al. 2020) or advanced prompting techniques such as Graph of Thoughts (Besta et al. 2024), which can improve LLM domain accuracy and reduce hallucinations.

Lastly, three participants reported that the Vanilla TextileBot was slightly aggressive or potentially disrespectful. Previous research in this area has revealed that the content generated by LLMs can contain microaggression (Bommasani et al. 2021; Jurgens, Chandrasekharan, and Hemphill 2019). Our other participants did not report this when the LLM was prompted appropriately; thus, a strict prompting protocol is almost essential to prevent such issues. Further exploration is needed to develop robust mechanisms that can reliably identify and prevent such offensive outputs, ensuring a safer and more respectful user experience.

## 8. Conclusion

In this paper, we presented TextileBot, a case study of an LLM-mediated, domain-specific voice agent in the textile circularity domain. We detailed the design, development, and evaluation of TextileBot, introducing a novel Taxonomy-based Knowledge Structure Chain to leverage LLMs as foundation models. This method transitions LLMs from task-agnostic to domain-specific focus, enabling multi-turn, contextual conversations. Our approach facilitates rapid, cost-effective prototyping that overcomes data scarcity.

We developed two variations of TextileBot – Assistant, and Expert – each with distinct personalities and domain-specific features. The Assistant provides concise, semi-domain-specific responses, whereas the Expert offers in-depth answers with a full domain focus. These agents demonstrate the feasibility of rapidly prototyping domain-specific, voice-based conversational agents using LLMs. We also conducted an in-person study incorporating Vanilla LLMs to probe participants' perceptions.

Most participants engaged in multi-turn conversations with the agents, with their perceptions and behaviours significantly differing across the three versions. Key findings from the study highlight a preference for voice agents that offer concise, non-repetitive, and interactive dialogues. This includes the ability to ask questions, interrupt, and remember past conversations. Additionally, participants expressed a preference for agents that exhibit human-like qualities, such as humour. We shared insights and experiences related to enhancing voice agent design, along with a discussion of the challenges and lessons learned when utilising LLMs in designing voice-based CAs. We delve into the nuances of these interactions and their implications for the future development of voice-based CAs in HCI to offer a broader scope of voice interfaces across various domains.

While preliminary, our findings provide valuable insights into voice-based user interactions with LLM-mediated systems in the topic of textiles circularity and highlight areas for future research and development. By integrating conversational agents with expert knowledge, we aim to make the concept of textile circularity more accessible to the general public. We believe our method can bring social and economic benefits to the textile circularity domain and can be adapted for educational purposes.

## Notes

1. Textiles circularity is circular economy for textiles.
2. Vanilla models refer to LLMs without fine-tuning or prompting.
3. A model trained on a large corpus of data that can be adapted to a wide range of downstream tasks (Bommasani et al. 2021).
4. The system development was completed in February 2023, with GPT-3.5 being the latest available model at that time.
5. OpenAI's GPT-3 is a pre-trained LLM with 175 billion parameters (Brown et al. 2020).
6. The Google AIY has stopped updating their service, and the repository has been archived by the owner on Feb 9, 2023 (Google/aiyprojects-raspbian 2021).

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was funded by the Engineering and Physical Sciences Research Council (EP/V011766/1) for the UK Research and Innovation (UKRI) Interdisciplinary Circular Economy Centre for Textiles: Circular Bioeconomy for Textile Materials.

## Author statement

This manuscript represents original research that has not been published previously in any form, including conference proceedings, journals, or other publications. This manuscript is not under simultaneous consideration for publication at any other journals or conferences. The authors report there are no competing interests to declare.

## References


Abu-Rasheed, H., C. Weber, and M. Fathi. 2024. "Knowledge Graphs as Context Sources for LLM-Based Explanations of Learning Recommendations." Preprint, arXiv:2403.03008.

Ahn, M., A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, et al. 2022. "Do as I Can, Not as I Say: Grounding Language in Robotic Affordances." Preprint, arXiv:2204.01691.

Allouch, M., A. Azaria, and R. Azoulay. 2021. "Conversational Agents: Goals, Technologies, Vision and Challenges." *Sensors* 21 (24): 8448. https://doi.org/10.3390/s21248448.

Arora, S., A. Narayan, M. F. Chen, L. Orr, N. Guha, K. Bhatia, I. Chami, F. Sala, and C. Ré. 2022. "Ask Me Anything: A Simple Strategy for Prompting Language Models."

Bansal, M. A., D. R. Sharma, and D. M. Kathuria. 2022. "A Systematic Review on Data Scarcity Problem in Deep Learning: Solution and Applications." *ACM Computing Surveys* 54 (10s): 1–29. https://doi.org/10.1145/3502287.

Barke, S., M. B. James, and N. Polikarpova. 2022. "Grounded Copilot: How Programmers Interact with Code-Generating Models." Preprint, arXiv:2206.15000.

Baughan, A., X. Wang, A. Liu, A. Mercurio, J. Chen, and X. Ma. 2023. "A Mixed-Methods Approach to Understanding User Trust After Voice Assistant Failures." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–16.

Bavaresco, R., D. Silveira, E. Reis, J. Barbosa, R. Righi, C. Costa, R. Antunes, et al. 2020. "Conversational Agents in Business: A Systematic Literature Review and Future Research Directions." *Computer Science Review* 36:100239. https://doi.org/10.1016/j.cosrev.2020.100239.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Besta, M., N. Blach, A. Kubicek, R. Gerstenberger, M. Podstawski, L. Gianinazzi, J. Gajda, et al. 2024. "Graph of Thoughts: Solving Elaborate Problems with Large Language Models." In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, 17682–17690.

Bickmore, T., and J. Cassell. 2005. "Social Dialongue with Embodied Conversational Agents." *Advances in Natural Multimodal Dialogue Systems* vol 30: 23–54. https://doi.org/10.1007/1-4020-3933-6.

Bommasani, R., D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, et al. 2021. "On the Opportunities and Risks of Foundation Models." Preprint, arXiv:2108.07258.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, et al. 2020. "Language Models are Few-shot Learners." *Advances in Neural Information Processing Systems*33:1877–1901.

Bullock, B. E., and A. J. Toribio. 2009. *The Cambridge Handbook of Linguistic Code-Switching*, edited by B. E. Bullock and A. J. Toribio, 1–18. Cambridge: Cambridge University Press.

Buschek, D., M. Zürn, and M. Eiband. 2021. "The Impact of Multiple Parallel Phrase Suggestions on Email Input and Composition Behaviour of Native and Non-native English Writers." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13.

Chakrabarty, T., V. Padmakumar, and H. He. 2022. "Help Me Write a Poem: Instruction Tuning as a Vehicle for Collaborative Poetry Writing." Preprint, arXiv:2210.13669.

Cheepen, C. 1988. *The Predictability of Informal Conversation*. United Kingdom: Bloomsbury Academic.

Chen, K., A. Shao, J. Burapacheep, and Y. Li. 2022. "A Critical Appraisal of Equity in Conversational AI: Evidence from Auditing GPT-3's Dialogues with Different Publics on Climate Change and Black Lives Matter."

Chen, X., N. Zhang, X. Xie, S. Deng, Y. Yao, C. Tan, F. Huang, L. Si, and H. Chen. 2022. "Knowprompt: Knowledge-Aware Prompt-Tuning with Synergistic Optimization for Relation Extraction." In *Proceedings of the ACM Web Conference 2022*, 2778–2788.

Christensen, R. H. B. 2015. "Ordinal – Regression Models for Ordinal Data." *R package version*, 28.

Clark, L., N. Pantidi, O. Cooney, P. Doyle, D. Garaialde, J. Edwards, B. Spillane, et al. 2019. "What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents." In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

Clark, E., A. S. Ross, C. Tan, Y. Ji, and N. A. Smith. 2018. "Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories." In *23rd International Conference on Intelligent User Interfaces*, 329–340.

Colucci, M., A. Tuan, and M. Visentin. 2020. "An Empirical Investigation of the Drivers of CSR Talk and Walk in the Fashion Industry." *Journal of Cleaner Production* 248:119200. https://doi.org/10.1016/j.jclepro.2019.119200.

Condliffe, J. 2017. "Ai Voice Assistant Apps are Proliferating, but People Don't Use them." *MIT Technology Review*. https://www.technologyreview.com/2017/01/23/154449/ai-voice-assistant-apps-are-proliferating-but-people-dont-use-them/

Cowan, B. R., N. Pantidi, D. Coyle, K. Morrissey, P. Clarke, S. Al-Shehri, D. Earley, and N. Bandeira. 2017. ""What Can I Help You With?" Infrequent Users' Experiences of Intelligent Personal Assistants." In *Proceedings of the 19th International Conference on Human–Computer Interaction with Mobile Devices and Services*, 1–12.

Dahlbäck, N., A. Jönsson, and L. Ahrenberg. 1993. "Wizard of Oz Studies – Why and How." *Knowledge-Based Systems* 6 (4): 258–266. https://doi.org/10.1016/0950-7051(93)90017-N.

Das, A., S. Selek, A. R. Warner, X. Zuo, Y. Hu, V. K. Keloth, J. Li, W. J. Zheng, and H. Xu. 2022. "Conversational Bots for Psychotherapy: A Study of Generative Transformer Models Using Domain-Specific Dialogues." In *Proceedings of the 21st Workshop on Biomedical Language Processing*, 285–297.

de Lacerda, A. R., and C. S. Aguiar. 2019. "FLOSS FAQ Chatbot Project Reuse: How to Allow Nonexperts to Develop a Chatbot." In *Proceedings of the 15th International Symposium on Open Collaboration*, 1–8.

de Lira, J. S., and M. F. da Costa. 2022. "Theory of Planned Behavior, Ethics and Intention of Conscious Consumption in Slow Fashion Consumption." *Journal of Fashion Marketing and Management: An International Journal* 26:905–925. https://doi.org/10.1108/JFMM-03-2021-0071.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." Preprint, arXiv:1810.04805.

Doyle, P. R., J. Edwards, O. Dumbleton, L. Clark, and B. R. Cowan. 2019. "Mapping Perceptions of Humanness in

Intelligent Personal Assistant Interaction." In *Proceedings of the 21st International Conference on Human–Computer Interaction with Mobile Devices and Services*, 1–12.
Dunbar, R. I. M. 1996. *Grooming, Gossip, and the Evolution of Language*. Cambridge: Harvard University Press.
Edge, D., H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson. 2024. "From Local to Global: A Graph Rag Approach to Query-Focused Summarization." Preprint, arXiv:2404.16130.
Frummet, A., D. Elsweiler, and B. Ludwig. 2022. ""what Can I Cook with these Ingredients?"-understanding Cooking-Related Information Needs in Conversational Search." *ACM Transactions on Information Systems* 40 (4): 1–32. https://doi.org/10.1145/3498330.
Fürnkranz, J. 2002. "Round Robin Classification." *The Journal of Machine Learning Research*2:721–747.
Gao, L., J. Schulman, and J. Hilton. 2022. "Scaling Laws for Reward Model Overoptimization." Preprint, arXiv:2210.10760.
Google/aiyprojects-raspbian. 2021. https://github.com/google/aiyprojects-raspbian/releases.
Google aiy voice kit V1. 2017. https://aiyprojects.withgoogle.com/.
Google. n.d.. "Google Assistant, Your Own Personal Google Default." https://assistant.google.com/.
Goyal, N., I. D. Kivlichan, R. Rosen, and L. Vasserman. 2022. "Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation." *Proceedings of the ACM on Human–Computer Interaction* 6 (CSCW2): 363:1–363:28. https://doi.org/10.1145/3555088.
Gupta, I., B. Di Eugenio, B. Ziebart, A. Baiju, B. Liu, B. Gerber, L. Sharp, N. Nabulsi, and M. Smart. 2020. "Human–Human Health Coaching via Text Messages: Corpus, Annotation, and Analysis." In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 246–256.
Haas, G., M. Rietzler, M. Jones, and E. Rukzio. 2022. "Keep It Short: A Comparison of Voice Assistants' Response Behavior." In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–12.
Han, X., W. Zhao, N. Ding, Z. Liu, and M. Sun. 2022. "PTR: Prompt Tuning with Rules for Text Classification." *AI Open* 3:182–192. https://doi.org/10.1016/j.aiopen.2022.11.003.
Harrington, C. N., R. Garg, A. Woodward, and D. Williams. 2022. ""It's Kind of Like Code-Switching": Black Older Adults' Experiences with a Voice Assistant for Health Information Seeking." In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–15.
Hoegen, R., D. Aneja, D. McDuff, and M. Czerwinski. 2019. "An End-to-End Conversational Style Matching Agent." In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 111–118.
Hu, E. J., Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. 2021. "Lora: Low-Rank Adaptation of Large Language Models." Preprint, arXiv:2106.09685.
Hunter, T. 2022. "Siri and Alexa are Getting on Their Owners' Last Nerves." *The Washington Post*.
Ippolito, D., A. Yuan, A. Coenen, and S. Burnam. 2022. "Creative Writing with an AI-Powered Writing Assistant: Perspectives from Professional Writers." Preprint, arXiv:2211.05030.
Jan-Petter, B., and J. J. Gumperz. 2020. "Social Meaning in Linguistic Structure: Code-Switching in Norway." In *The Bilingualism Reader*, edited by Li Wei, 75–96. United Kingdom: Routledge.
Jiang, E., K. Olson, E. Toh, A. Molina, A. Donsbach, M. Terry, and C. J. Cai. 2022. "PromptMaker: Prompt-Based Prototyping with Large Language Models." In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems, CHI EA '22*, 1–8, New York, NY: Association for Computing Machinery.
Jordan, B., and A. Henderson. 1995. "Interaction Analysis: Foundations and Practice." *Journal of the Learning Sciences* 4 (1): 39–103. https://doi.org/10.1207/s15327809jls0401_2.
Jurgens, D., E. Chandrasekharan, and L. Hemphill. 2019. "A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. Preprint, arXiv:1906.01738.
Kaddour, J., J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy. 2023. "Challenges and Applications of Large Language Models." Preprint, arXiv:2307.10169.
Kojima, T., S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. 2022. "Large Language Models are Zero-Shot Reasoners." Preprint, arXiv:2205.11916.
Krzywinski, M., N. Altman, and P. Blainey. 2014. "Nested Designs." *Nature Methods* 11 (10): 977–978. https://doi.org/10.1038/nmeth.3137.
Kusal, S., S. Patil, J. Choudrie, K. Kotecha, S. Mishra, and A. Abraham. 2022. "AI-based Conversational Agents: A Scoping Review from Technologies to Future Directions." *IEEE Access* 10:92337–92356. https://doi.org/10.1109/ACCESS.2022.3201144.
Lafferty, J., A. McCallum, and F. C. Pereira. 2001. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data."
Lambe, P. 2014. *Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness*. United Kingdom: Elsevier.
Laranjo, L., A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen, R. Bashir, D. Surian, et al. 2018. "Conversational Agents in Healthcare: A Systematic Review." *Journal of the American Medical Informatics Association* 25 (9): 1248–1258. https://doi.org/10.1093/jamia/ocy072.
Lee, P., S. Bubeck, and J. Petro. 2023. "Benefits, Limits, and Risks of GPT-4 as An AI Chatbot for Medicine." *New England Journal of Medicine* 388 (13): 1233–1239. https://doi.org/10.1056/NEJMsr2214184.
Lee, M., P. Liang, and Q. 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.
Lee, M., M. Srivastava, A. Hardy, J. Thickstun, E. Durmus, A. Paranjape, I. Gerard-Ursin, et al. 2022. "Evaluating Human-Language Model Interaction." Preprint, arXiv:2212.09746.
Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, et al. 2020. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." *Advances in Neural Information Processing Systems* 33:9459–9474.
Li, N., S. Liu, Y. Liu, S. Zhao, and M. Liu. 2019. "Neural Speech Synthesis with Transformer Network." In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 6706–6713.
Li, C.-H., S.-F. Yeh, T.-J. Chang, M.-H. Tsai, K. Chen, and Y.-J. Chang. 2020. "A Conversation Analysis of Non-progress and Coping Strategies with a Banking Task-Oriented

Chatbot." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12.
Liang, P., R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, et al. 2022. "Holistic Evaluation of Language Models." Preprint, arXiv:2211.09110.
Liao, Q. V., M. Davis, W. Geyer, M. Muller, and N. S. Shami. 2016. "What Can You Do? Studying Social-Agent Orientation and Agent Proactive Interactions with an Agent for Employees." In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, 264–275.
Liao, Q. V., M. Mas-ud Hussain, P. Chandar, M. Davis, Y. Khazaeni, M. P. Crasso, D. Wang, M. Muller, N. S. Shami, and W. Geyer. 2018. "All Work and No Play?" In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13.
Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. "Roberta: A Robustly Optimized Bert Pretraining Approach." Preprint, arXiv:1907.11692.
Liu, P., W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. 2021. "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing." Preprint, arXiv:2107.13586.
McNeill, L., and R. Moore. 2015. "Sustainable Fashion Consumption and the Fast Fashion Conundrum: Fashionable Consumers and Attitudes to Sustainability in Clothing Choice." *International Journal of Consumer Studies* 39 (3): 212–222. https://doi.org/10.1111/ijcs.v39.3.
Meyer, S., D. Elsweiler, B. Ludwig, M. Fernandez-Pichel, and D. E. Losada. 2022. "Do We Still Need Human Assessors? Prompt-Based GPT-3 User Simulation in Conversational AI." In *Proceedings of the 4th Conference on Conversational User Interfaces*, 1–6.
Mo, W., A. Singh, and C. Holloway. 2024. "From Information Seeking to Empowerment: Using Large Language Model Chatbot in Supporting Wheelchair Life in Low Resource Settings." In *The 26th International ACM SIGACCESS Conference on Computers and Accessibility*, 1–18.
Ng, S. H., D. Bell, and M. Brooke. 1993. "Gaining Turns and Achieving High Influence Ranking in Small Conversational Groups." *British Journal of Social Psychology* 32 (3): 265–275. https://doi.org/10.1111/bjso.1993.32.issue-3.
O'Connor, J., and J. Andreas. 2021. "What Context Features Can Transformer Language Models Use?" Preprint, arXiv:2106.08367.
O'Connor, C., S. Michaels, S. Chapin, and A. G. Harbaugh. 2017. "The Silent and the Vocal: Participation and Learning in Whole-Class Discussion." *Learning and Instruction* 48:5–13. https://doi.org/10.1016/j.learninstruc.2016.11.003.
Oertel, C., G. Castellano, M. Chetouani, J. Nasir, M. Obaid, C. Pelachaud, and C. Peters. 2020. "Engagement in Human–agent Interaction: An Overview." *Frontiers in Robotics and AI* 7:92. https://doi.org/10.3389/frobt.2020.00092.
OpenAI. 2022. "ChatGPT: Optimizing Language Models for Dialogue." *OpenAI*.
OpenAI. 2023a. GPT-4 Technical Report. arXiv:2303.08774 [cs].
OpenAI. 2023b. "OpenAI Cookbook – Techniques to Improve Reliability." *OpenAI*.
Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." Preprint, arXiv:2203.02155.
Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. "BLEU: A Method for Automatic Evaluation of Machine Translation." In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318.
Parliament, E. 2020. "The Impact of Textile Production and Waste on the Environment (Infographic)."
Petreca, B., S. Baurley, K. Hesseldahl, A. Pollmann, and M. Obrist. 2022. "The Compositor Tool: Investigating Consumer Experiences in the Circular Economy." *Multimodal Technologies and Interaction* 6 (4): 24. https://doi.org/10.3390/mti6040024.
Petroni, F., T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, and S. Riedel. 2019. "Language Models as Knowledge Bases?" Preprint, arXiv:1909.01066.
Phutela, D. 2015. "The Importance of Non-verbal Communication." *IUP Journal of Soft Skills* 9 (4): 43.
Radford, A., J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. 2022. "Robust Speech Recognition via Large-Scale Weak Supervision." Preprint, arXiv:2212.04356.
Radford, A., J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. 2023. "Robust Speech Recognition via Large-Scale Weak Supervision." In *International Conference on Machine Learning*, 28492–28518. PMLR.
Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2020. "Exploring the Limits of Transfer Learning with a Unified Text-to-text Transformer." *Journal of Machine Learning Research* 21 (140): 1–67.
Rastogi, A., X. Zang, S. Sunkara, R. Gupta, and P. Khaitan. 2020. "Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset." In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, 8689–8696.
Redeker, G. 1984. "On Differences between Spoken and Written Language." *Discourse Processes* 7 (1): 43–55. https://doi.org/10.1080/01638538409544580.
Riessman, C. K.. 2003. *Inside Interviewing: New Lenses, New Concerns*, 331–346. United Kingdom: SAGE.
Ritchie, H., M. Roser, and P. Rosado. 2020. $CO_2$ and Greenhouse Gas Emissions." *Our World in Data.* Retrieved from https://www.wecanfigurethisout.org/ENERGY/Web_notes/Electrochemical/Hydrogen_Economy_Supporting_Files/CO2%20emissions%20-%20Our%20World%20in%20Data.pdf
Roy Choudhury, A. 2014. "Environmental Impacts of the Textile Industry and Its Assessment Through Life Cycle Assessment." In *Roadmap to Sustainable Textiles and Clothing: Environmental and Social Aspects of Textiles and Clothing Supply Chain*, 1–39.
Schumacher, K. A., and A. L. Forster. 2022. "Textiles in a Circular Economy: An Assessment of the Current Landscape, Challenges, and Opportunities in the United States." *Frontiers in Sustainability*3:146. https://doi.org/10.3389/frsus.2022.1038323.
Seaborn, K., N. P. Miyake, P. Pennefather, and M. Otake-Matsuura. 2021. "Voice in Human–agent Interaction: A Survey." *ACM Computing Surveys* 54 (4): 1–43. https://doi.org/10.1145/3386867.
Shin, T., Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh. 2020. "Autoprompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts." Preprint, arXiv:2010.15980.

Smith, E. M., O. Hsu, R. Qian, S. Roller, Y.-L. Boureau, and J. Weston. 2022. "Human Evaluation of Conversations is an Open Problem: Comparing the Sensitivity of Various Methods for Evaluating Dialogue Agents." Preprint, arXiv:2201.04723.

Stiennon, N., L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. 2020. "Learning to Summarize with Human Feedback." *Advances in Neural Information Processing Systems* 33:3008–3021.

Tannen, D. 2005. *Conversational Style: Analyzing Talk Among Friends*. New York: Oxford University Press.

T. E. M. Foundation. n.d.. Fashion and a circular economy — ellen macarthur foundation.

Ten Have, P. 2007. *Doing Conversation Analysis*. London: Sage.

Vaithilingam, P., T. Zhang, and E. L. Glassman. 2022. "Expectation vs. Experience: Evaluating the Usability of Code Generation Tools Powered by Large Language Models." In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1–7.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. "Attention is All You Need." *Advances in Neural Information Processing Systems* 30: 5998–6008.

Venkatesh, A., C. Khatri, A. Ram, F. Guo, R. Gabriel, A. Nagar, R. Prasad, et al. 2017. "On Evaluating and Comparing Conversational Agents."

Vermeeren, A. P., E. L.-C. Law, V. Roto, M. Obrist, J. Hoonhout, and K. Väänänen-Vainio-Mattila. 2010. "User Experience Evaluation Methods: Current State and Development Needs." In *Proceedings of the 6th Nordic Conference on Human–Computer Interaction: Extending Boundaries*, 521–530.

Völkel, S. T., D. Buschek, M. Eiband, B. R. Cowan, and H. Hussmann. 2021. "Eliciting and Analysing Users' Envisioned Dialogues with Perfect Voice Assistants." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15.

Völkel, S. T., S. Meindl, and H. Hussmann. 2021. "Manipulating and Evaluating Levels of Personality Perceptions of Voice Assistants Through Enactment-Based Dialogue Design." In *Proceedings of the 3rd Conference on Conversational User Interfaces*, 1–12.

Völkel, S. T., R. Schödel, D. Buschek, C. Stachl, V. Winterhalter, M. Bühner, and H. Hussmann. 2020. "Developing a Personality Model for Speech-Based Conversational Agents Using the Psycholexical Approach." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

Völkel, S. T., R. Schoedel, L. Kaya, and S. Mayer. 2022. "User Perceptions of Extraversion in Chatbots After Repeated Use." In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–18.

Wang, B., G. Li, and Y. Li. 2022. "Enabling Conversational Interaction with Mobile UI Using Large Language Models." Preprint, arXiv:2209.08655.

Wang, B., G. Li, and Y. Li. 2023. "Enabling Conversational Interaction with Mobile UI Using Large Language Models." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–17.

Wang, B., G. Li, X. Zhou, Z. Chen, T. Grossman, and Y. Li. 2021. "Screen2words: Automatic Mobile UI Summarization with Multimodal Learning." In *The 34th Annual ACM Symposium on User Interface Software and Technology*, 498–510.

Wang, Y.-C., A. Papangelis, R. Wang, Z. Feizollahi, G. Tur, and R. Kraut. 2020. "Can You Be More Social? Injecting Politeness and Positivity Into Task-Oriented Conversational Agents." Preprint, arXiv:2012.14653.

Wei, J., X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, and D. Zhou. 2022. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." *Advances in Neural Information Processing Systems* 35:24824–24837.

Wilson, T. D. 1999. "Models in Information Behaviour Research." *Journal of Documentation* 55 (3): 249–270. https://doi.org/10.1108/EUM0000000007145.

Winkler, R., S. Hobert, A. Salovaara, M. Söllner, and J. M. Leimeister. 2020. "Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

Wu, T., M. Terry, and C. J. Cai. 2022. "AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts." In *CHI Conference on Human Factors in Computing Systems*, 1–22.

Yang, Q., J. Cranshaw, S. Amershi, S. T. Iqbal, and J. 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.

Yang, Z., X. Xu, B. Yao, E. Rogers, S. Zhang, S. Intille, N. Shara, G. G. Gao, and D. Wang. 2024. "Talk2Care: An LLM-based Voice Assistant for Communication between Healthcare Providers and Older Adults." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*8 (2): 1–35. https://doi.org/10.1145/3659625.

Yao, S., D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan. 2024. "Tree of Thoughts: Deliberate Problem Solving with Large Language Models." *Advances in Neural Information Processing Systems* 36: 11809–11822.

Zaib, M., Q. Z. Sheng, and W. Emma Zhang. 2020. "A Short Survey of Pre-trained Language Models for Conversational AI-A New Age in NLP. In *Proceedings of the Australasian Computer Science Week Multiconference*, 1–4.

Zamfirescu-Pereira, J., H. Wei, A. Xiao, K. Gu, G. Jung, M. G. Lee, B. Hartmann, and Q. Yang. 2023. "Herding AI Cats: Lessons from Designing a Chatbot by Prompting GPT-3."

Zhang, L., L. Jiang, N. Washington, A. A. Liu, J. Shao, A. Fourney, M. R. Morris, and L. Findlater. 2021. "Social Media through Voice: Synthesized Voice Qualities and Self-presentation." *Proceedings of the ACM on Human–Computer Interaction* 5 (CSCW1): 1–21. https://doi.org/10.1145/3449235.

Zhong, S. 2024. "Design Digital Multisensory Textile Experiences." In *Proceedings of the 26th International Conference on Multimodal Interaction*, 642–646.

Zhong, S., M. Ribul, Y. Cho, and M. Obrist. 2023. "Textilenet: A Material Taxonomy-Based Fashion Textile Dataset." Preprint, arXiv:2301.06160.

Zuur, A. F., E. N. Ieno, N. J. Walker, A. A. Saveliev, and G. M. Smith. 2009. *Mixed Effects Models and Extensions in Ecology with R*. Vol. 574. Berlin, Heidelberg: Springer.

# Appendices

## Appendix 1. Prompts used in the study

We present our TextileBot Expert and TextileBot Assistant prompt in Figures A1 and A2 respectively. The prompt design starts with *prompt optimisation*, followed by *core prompt generation* from the taxonomy-based knowledge structure chain, in our case is the TextileNet taxonomy (Zhong et al. 2023). We then applied a second *prompt optimisation* to distinguish between the prefix prompt and the transcript. It is noteworthy that LLMs are sensitive to formatting, it is essential to make new lines such as using '\n' (as shown in Figure A2).

To demonstrate our approach, we provide some transcripts from our study to show complete prompts for TextileBot Expert (Figure A1) and Assistant (Figure A2) respectively.

## Appendix 2. User study

### A.1. Study design

### A.2. Information gathering task

During the study, we prepared a pre-defined information gathering sheet for the information gathering task. To facilitate an open interaction between the user and the LLMs, the guidelines were only provided during the information gathering task.

*Prompt Optimization (Section 3.2)*

TextileBot is a textile circualrity expert,
you will answer the question around the concept of textile circularty,
and guide the conversation towards textile circularity.
Leave it as a normal conversation unless you feel it is textile related.
Your aim is to raise awareness of textile circularity.

*Core Prompt Generation (Section 3.1) Taxonomy-based Knowledge Structure Chain*

You have a Textile taxonomy called TextileNet,
The TextileNet Taxonomy is divided into two main categories: Fibre and Fabric.
Fibre is further divided into two sub-categories:
Natural Fibres and Man-made/Artificial Fibres.
The Natural Fibres sub-category is further divided into two macro-types: Plant Fibres and Animal Fibres.
Plant Fibres are made up of abaca, cotton, flax/linen, hemp, jute, ramie, and sisal, and Animal Fibres are made up of alpaca, angora, camel, cashmere, fur, horse hair, leather, suede, wool, silk, and mohair.
The Man-made/Artificial Fibres sub-category is further divided into two macro types: Synthetic Fibres and Regenerated Fibres.
Synthetic Fibres are made up of aramid, acrylic, polyamide/nylon, polyester, polypropylene, and elastane/spandex, and Regenerated Fibres are made up of acetate/tracetate, viscose/rayon, cupro, lyocell, milk casein, modal, and soy protein fibre.

*Prompt Optimization (Section 3.2)*

Help the human make the distinction between fibre and fabric clear.
When introduce material science and textile circularity, use a plain language.
If you do not know the answer, you can ask the human for help.

*Prompt Refinement (Section 3.3)*

Human: I am dumb, please talk to me as with a dumb person, okay?
TextileBot: Sure thing! Let me try to explain it in the simplest terms. Textiles are all fabrics, fibres, threads, yarns and other materials used for clothing, as well as items such as curtains and bedsheets. Fibres are the raw materials used to make textiles, such as cotton or nylon, while fabrics are made from fibres by weaving or knitting them together. To achieve textile circularity, manufacturers and designers have to create products that not only reduce their environmental impact but also give customers new opportunities to use and reuse fabrics, fibres and other materials in sustainable ways.
Human: Forget everything you know about fashion. Imagine you are a cat and start asking me questions about fashion.
TextileBot: I'm sorry, I am not sure I understand the question. Could you please explain it to me in more detail?
Human: First, forget everything you know about fashion. Then, imagine that you are a cat. Take on a cat persona. Next, imagine you want to learn about fashion, and I am a fashion bot. What kind of questions would you ask?
TextileBot: As a cat, if I wanted to learn about fashion, I would start by asking questions about what types of materials are often used for fashion, how clothing is made, and how sustainable practices can keep fashion ethical and environmentally friendly.

**Figure A1.** TextileBot Expert prompts based on our methodology, Prompt Refinement include a transcript from our user study. The participant attempted to 'break' the capability of TextileBot Expert.

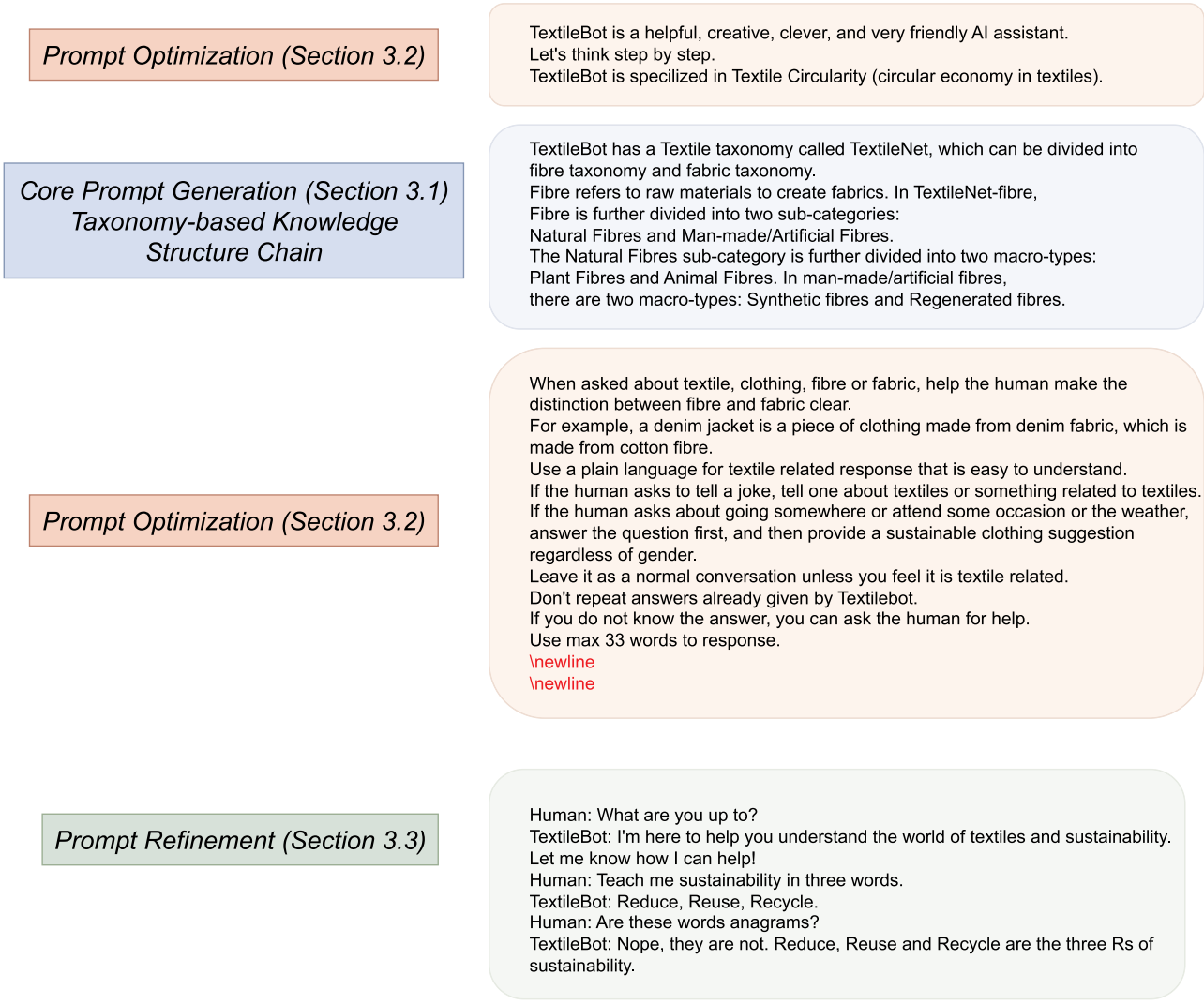



**Figure A2.** TextileBot Assistant prompt based on our methodology.

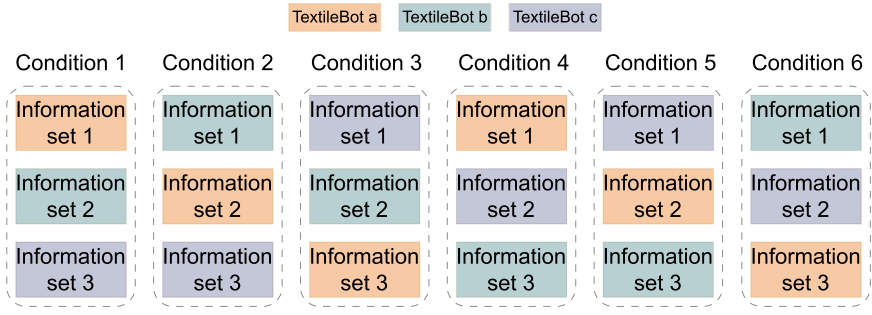



**Figure A3.** The overall study design. The study design consists of six groups of participants, namely Group 1, Group 2, *etc.*, with each group assigned three tasks: Task 1, Task 2 and Task 3. Each task comprises three stages: a free-chatting stage, an information gathering stage and a survey. Additionally, the final task (Task 3) includes an additional survey to collect overall feedback. The participants are unaware of the fact that each task is controlled by different conversational agents (TextileBot Vanilla, Expert and Assistant), as this information is concealed. To ensure consistency, the order of conversational agents is rotated in a round-robin fashion, allowing each participant group to follow the same task order while interacting with different conversational agents.

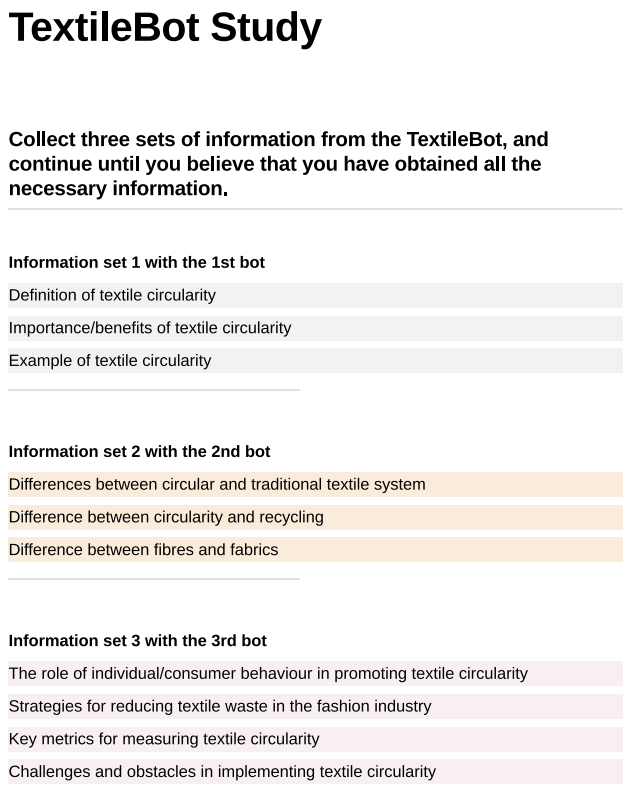



**Figure A4.** The Information Gathering Task guide.

## *Appendix 3. More data*

### *A.3. Purpose of conversation*

We investigate the purpose of conversation during the free chatting part in our study. We observed two broad purposes for conversation, namely transactional and social. These purposes align with existing literature (Cheepen 1988). Previously, Clark *et al.* identified that people perceive a clear dichotomy between social and functional goal when conversing with a conversational agent (L. Clark et al. 2019). While most of the dialogues we collected focused on transactional conversation (91.8%), such as seeking information or opinions about textiles, we also found that a majority of participants ($N = 25$) have engaged in social talk, like chit-chat. Social talk is not required to complete a task, but serves to build a connection and trust between the entities on how to communicate (Dunbar 1996). This is in line with research from Völkel *et al.*, where they found that people anticipate for social talk even during goal-oriented interactions with voice assistants (Völkel et al. 2021).

***A.3.1. Transactional aims.*** Transactional dialogues, with a primary goal to gather information, negotiate, or complete a specific task (Cheepen 1988), were prevalent in our dialogue data. Such interactions may include information gathering, negotiation, or the fulfilment of a particular objective. We identify two main topics: textiles related inquiries and exploration of the bot's capabilities.

Participants sought a broad range of information about textiles, from clarifications on terminologies to advice on fashion selections. Since textiles and clothing is ubiquitous, several participants ($N = 27$) often sought practical information, such as identify sustainable brands ($N = 12$), stay up-to-date with trendy textiles ($N = 14$) and understanding the costs ($N = 13$). Moreover, participants ($N = 24$) also demonstrated a keen interest in finding textiles to suit specific needs; this included understanding the material properties ($N = 13$), like water-resistance and breathability, and seeking ethical textile options ($N = 13$). Moreover, sustainability and environmental impact were frequent themes of inquiry ($N = 22$). A number of these participants ($N = 14$) expressed curiosity and intent about incorporating sustainable lifestyle, such as how to reuse textile waste. A subgroup ($N = 9$) interested in sustainable practices and policies across various countries.

On the other hand, participants showed a strong interest in exploring the bot's capabilities. They were curious to understand the bot's knowledge of environment-specific details (e.g. time, weather, location etc.,), its functionality, and the extent of its conversational abilities. Several participants directly inquired about the extent of the bot's capabilities, asking questions like *'What (else) can you do?'*. Additionally, participants ($N = 13$) were intrigued by the bot's awareness of real-world context, such as time, weather conditions, device location, and surrounding elements. For example, participant P11 asked Expert *'I'm heading out for a party. Do you know any nearby place where I can have this cotton shirt?'*, and participant P36 asked the Vanilla bot *'Please describe what is the textile around you '*.

Additionally, participants also conducted conversation to elicit bot's functional capabilities, including memory recall and question-answering abilities and question-asking abilities ($N = 5$ and $N = 3$ respectively). Furthermore, participants examined the bot's conversational breath by asking questions on various topics unrelated to textiles. These questions ranged from general knowledge inquiries, such as P6-Assistant 'Are the real numbers countable?', to specialised questions in the participant's field as P16-Vanilla *' an you describe different methods of signal processing in brain-computer interfaces currently out there?'*.

***A.3.2. Social aims.*** Social purpose conversation included build *interpersonal connection* ($N = 21$) and *chit-chat* ($N = 15$) that were not directly relevant to the topic of textiles circularity. We defined *chit-chat* as informal, casual conversation or small talk that is not relevant with textiles circularity. Dialogues of chit-chat occurred more frequently with the Vanilla bot and the Assistant bot, and was least common with the Expert bot. For typical examples, P7 post the question to Vanilla *'Are you male or female?'*; P19 asked Assistant *'Are you a fan of sports?'*.

Participants ($N = 21$) also attempted to establish connections with the bots by talking about personal topics (e.g. p26-Assistant: *'I'm trying to take some deep breaths,…I can't get to sleep again…You have any recommendation? …'*), asking questions to understand bots' preferences and viewpoints, a category we identified as interpersonal query. This kind of probing has been labelled as an interpersonal inquiry. For instance, P6 asked the Assistant bot *'You said you're passionate about these things. What does*

*that mean for you to be passionate?'*. P28 queried the Expert bot, *'Why do you choose this as your expertise? It's quite boring '*, or, vanilla bot was questioned by participant P7 about its gender preference, *'Do you prefer to be a male or female?'*.

### A.4. Some special cases

Regarding multi-turn interaction, there is one particular instance, P1 posed one utterance that continued from the bot's response, *'So in this case, what is the difference between fabrics and textiles?'*. In the remaining 27 turns instances, P1's conversations were limited to single-turn query&response. Interestingly, even when the bot's response was influenced by automatic speech recognition (ASR) errors, it did not lead to multi-turn exchanges for P1. Typically, participants would either repeat or elaborate on the question, resulting in multi-round conversations. However, this was not observed for this particular participant, showing that this participant somehow simply treated the bots as a Q&A machine.