What is Jiaozi: Exploring User Control of Language Style in Multilingual Conversational Agents

Jiawen Stefanie Zhu
University of Waterloo
Waterloo, Ontario, Canada
University of Washington
Seattle, Washington, USA
jiawenz2@uw.edu

Jian Zhao
University of Waterloo
Waterloo, Ontario, Canada
jianzhao@uwaterloo.ca

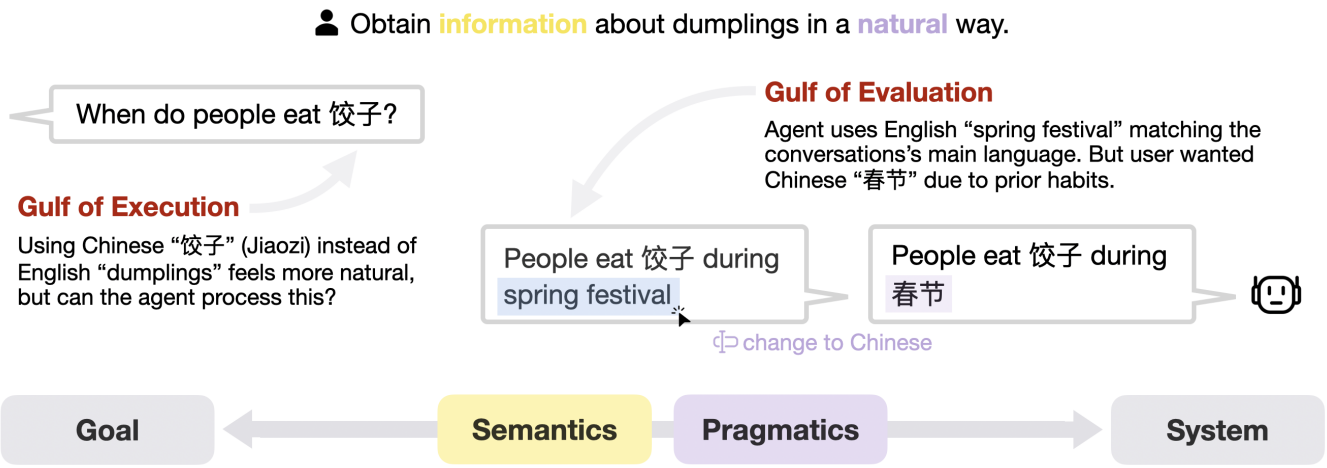

Figure 1: When interacting with AI-powered systems, multilingual users face *pragmatic* gaps where the desired content (*semantics*) is presented in an undesirable style. For example, when language used does not align with their preferences.

ABSTRACT

Recent advances in language models have significantly expanded the capabilities of AI-powered conversational agents. Nonetheless, current technology is still primarily designed with monolingual English speakers in mind, overlooking the need of more personalized agents by multilingual users. Particularly, prior work showed that multilingual individuals preferred conversational agents that accommodate their desired multilingual style. However, these approaches rely on probabilistic methods to automatically determine the agent's multilingual style, which often fails to align with the needs of multilingual users, as their preferences are nuanced, ad hoc, and difficult to predict. In our work, we explore user control of multilingual style as a step toward developing a mixed-initiative multilingual conversational agent tailored to the needs of multilingual users. We first derived design considerations and dimensions of user control from a formative study with 10 participants. Next, we implemented Mirrios, a prototypical conversational system with multilingual style control, and used it as a probe to conduct an user study with 12 participants. We identified preferred designs for multilingual style control and found that this control reduced the need to constrain language habits, accommodated ad hoc language needs, and enabled more personalized interactions with conversational agents. Based on our findings, we propose design implications to inform the design of multilingual style control and future mixed-initiative multilingual conversational agents.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**.

KEYWORDS

Conversational Agents, Multilingualism, Language

1 INTRODUCTION

As the capabilities of conversational agents expand with technological advances [41], more focus in the field has shifted from merely improving accuracies to enhancing the overall user experience [10]. One line of effort considers how to design language *style* – distinct from the *content* of agent replies – to align with users' preferences and habits [10, 20, 30]. Style refers to the aspects of language that convey subtle connotations rather than direct denotations, such as formality, complexity, and the combination of languages used [7, 8]. For multilingual users, style is especially important to consider, since their language skills and usage are context-dependent [15, 24] – making multilingual style alignment a practical necessity. Failing to accommodate their linguistic inclinations may thus negatively impact the multilingual user experience with conversational agents. However, accurately predicting these preferences is challenging, as they depend on a range of dynamic factors that cannot always be directly inferred from the input – for example the user's personal habits, current goals, and the broader context of the interaction.

Existing work support the importance of accommodating multilingualism in conversational agents, drawing from theories like the Communication Accommodation Theory [38]. Zhao et al. found that multilingual users often code-mix – or interweave multiple languages – when interacting with conversational AI [43]. Bawa et al. demonstrated that agents nudging users towards code-mixing (interweaving different languages) are perceived as more human-like and desirable compared to those that do not code-mix or follow a fixed code-mixing policy [10]. Similarly, Choi et al. found in a participatory design workshop that multilingual users preferred voice agents capable of understanding and responding in code-mixed language [12]. Although these work consider the preferred multilingual style of users, they focus on the presence or absence of style accommodation as a binary switch (the *if*) and *how* the multilingual style should be designed or tailored to individual preferences in conversational agents remains underexplored. Furthermore, existing systems automatically determine the multilingual style through probabilistic methods [10] and mimicking. While this can be helpful, due to the ad-hoc and unpredictable nature of the preferences of multilingual users, automatic approaches often do not fully align with user needs, which was validated in our formative study.

In this research, we address this agency gap by exploring the concept of *multilingual style control* (Figure **??**), where multilingual users gain the ability to directly steer the language of the agent's responses to align with their personal language skills and habits. We also wish to identify patterns and structures for multilingual style, to allow us to move beyond purely probabilistic approaches, and to pave the way for more meaningful designs of multilingual interactions. It is important to note that we do not envision multilingual style control as a replacement for current automatic approaches. Rather, we see it as an enhancement that could improve user experience and contribute to the development of mixed-initiative [19] multilingual agents, combining the agency provided by user control with the convenience of automation.

We first conducted an interview study with 10 multilingual users to understand their current needs and challenges when interacting with multilingual conversational agents. From these findings, we derived design considerations and dimensions for multilingual style control. We then implemented Mirrios, a working prototype of a conversational system with multilingual style control, manifesting all dimensions to allow systematic exploration of the design choices. Using Mirrios as a design probe, we further conducted an exploratory study with 12 Chinese-English bilingual users to investigate (1) preferred design alternatives for multilingual style control, and (2) the overall utility of enabling user control over multilingual style. Based on our results, we derived design implications to motivate future research and development of multilingual style control. We view this as a first step toward mixed-initiative multilingual conversational agents, where agency and automation are balanced to adaptively accommodate the dynamic and diverse needs of multilingual users. The contribution of this work is thus threefold:

- a **formative study** that informs design considerations and dimensions for enabling user control of multilingual style;
- a prototypical **conversational system with multilingual style control, Mirrios**, that manifests the derived dimensions and acts as a design probe for understanding the implications of giving users direct control of a conversational agent's multilingual style;
- an exploratory **probe study** that investigates preferred design alternatives of multilingual style control, evaluates its overall utility, and provides design implications for the development of future mixed-initiative conversational agents.

2 RELATED WORK

We summarize the key literature relevant to our work.

2.1 Language Style in Interfaces

Existing work in HCI has demonstrated that adapting language style based on user background has many benefits [6–9, 32]. Due to the advancements in AI models, research on conversational agents has also shifted from optimizing performance towards enhancing the subjective experience [10]. Particularly, one line of work has shown that the language style, such as variations in formality or affective expression, influences user interactions with conversational agents [20, 30, 35, 36]. Grounded in this work, the language(s) in which a passage appears has also been considered to be a form of language style [10]. We adopt this perspective in our work, and consider how *multilingual style* should be designed in conversational agents, and how to provide users control over it. For multilingual users, whose language skills and usage are contextual [15], both the content of the agent's reply and its multilingual style play a crucial role in enhancing the user experience.

2.2 Multilingual Conversational Agents

The importance of supporting multilingual users is widely recognized [22, 23], prompting research in areas such as multilingual language models [25, 37] and search interfaces [13, 26, 31]. Due to the language-heavy nature of conversational agents, a line of work has focused on designing conversational agents that align with the needs of multilingual users. In particular, this line of work focused on code-mixing, the phenomenon where polyglots interweave and switch between different languages [34]. For example, Bawa et al. discovered that users find code-mixing conversational agents to be more human-like and to have better conversational ability compared to counterparts that only speaks one language, especially so if the agent's level of code-mixing is on par with that of the user [10]. Similarly, Choi et al. conducted a participatory design workshop with polyglots and found that they desire voice assistants that understand and speak multiple languages mixed together [12].

While these works consider the desired multilingual style (e.g., code-mixing) of users, they treat multilingual style accommodation as a binary on-off switch and do not explore the nuanced dynamics of *how* the style should be designed. Additionally, existing multilingual conversational agents automatically determine the multilingual style, which often does not align with user preferences, due to their ad hoc and unpredictable nature. In these systems, users have no effective way of adjusting the multilingual style [10], outside of implicitly priming with the style of their input, which does not always align with user intentions and preferences. In our work, we explore how to give users *direct control over the multilingual style*, and the utility it may add to multilingual conversational agents. By analyzing how users adjust the multilingual

Table 1: Demographics of participants in the formative study. Labels in brackets represent the user's self-rated language proficiency according to the Common European Framework of Reference for Languages (CEFR) [1]. Roughly, Level A1-A2 corresponds to beginner, Level B1-B2 to intermediate, and Level C1-C2 corresponds to near-native proficiency.

Participant	Age	Gender	Native Language(s)	Other Language(s) Spoken
P1	33	Man	Arabic	English (C1), French (B2)
P2	62	Woman	Portuguese	English (C1), Spanish (A2), French (A1)
P3	23	Woman	Tamil, Hindi	English (C2), French (A1)
P4	23	Man	Chinese	English (B2)
P5	24	Woman	Farsi	English (C1)
P6	23	Woman	Portuguese	English(C2)
P7	22	Man	Malayalam	English (C2), Hindi (B2)
P8	24	Man	Mandarin	English (C2), Japanese (C1)
P9	29	Man	Portuguese	English (C1), Italian (B1), German (B1)
P10	39	Man	Bangla	English (C1), Hindi (B2)

style using our probe, we identify patterns and suggest structured approaches for designing multilingual style, moving beyond purely stochastic methods.

2.3 AI and Multilingualism

Recent advances in large-language models (LLMs) have granted AI-powered conversational agents unprecedented capabilities [41], including multilingual abilities such as understanding and responding to multilingual inputs [21, 29, 39]. However, these advancements have also introduced new challenges, as multilingual capabilities are handled probabilistically [17], and their multilingual behavior can still be stochastic and unnatural [40, 42]. This additional layer of stochasticity in multilingual interactions further exacerbates the existing gap between user intent and agent response [33]. The unpredictability also makes it difficult for users to formulate prompts, as they are unsure how their input will be interpreted. This issue is related to abstraction matching, which refers to the gap between the nearly infinite space of natural language utterances and the much smaller space of effective user inputs [27]. Our formative study investigates the difficulties arising from this gap. Through Mirrios and the probe study, we explore ways to bridge the gap between the user and the agent's multilingual language usage by providing users with control over the multilingual style.

3 FORMATIVE STUDY

We conducted an interview study with multilingual speakers to investigate their challenges, strategies, and envisioned technologies when using conversational agents. From our findings, we derived considerations for determining aspects of multilingual style users wish to control (Section 3.2). We then identified design dimensions for implementing user control over multilingual styles in conversational agents (Section 3.3). The design considerations, dimensions, and associated design choices are summarized in Table 2.

3.1 Participant and Procedure

We recruited 10 participants through an university mailing list (4 Women, 6 Men, average age = 30.2 ± 12.5). 9 of them were graduate students and 1 was an undergraduate. Their field of study included computer science (4/10), mathematics (2/10), data science (3/10) and public health (1/10). All participants are fluent in two or more languages, one of which was English. See Table 1 for complete demographic information. Note that we did not restrict the specific languages spoken by participants, as multilingual speakers share many common traits, even when the languages they speak differ [3, 5] and our purpose is exploratory. All participants use text-based conversational agents, such as ChatGPT, on a regular basis, with eight of them using them multiple times a day, one using them once a day, and one using them for at least $4-6$ times a week. The participants reported using textual conversational agents for information retrieval, understanding concepts, translation between languages, writing assistance, and code generation.

We started each study with the information and consent process, followed by a questionnaire asking about participant's basic demographic information regarding language skills, usage of text-based conversational agent, and thoughts regarding multilingual conversational agents. Next, we asked users about their experiences and opinions on conversational agents as multilingual users, focusing on the challenges they faced, the strategies they have tried to overcome these challenges, and the technologies they envision. Each study lasted around 60 minutes in total and was done remotely via Zoom. Participants received $20 remuneration in the local currency. This study was approved by our institution's ethics board.

The study sessions were audio- and screen-recorded. The interviews were transcribed verbatim. The first author conducted thematic analysis on the transcripts to summarize the key patterns in participant challenges and strategies, and derive design considerations and dimensions. All authors discussed the themes until they reached agreement.

3.2 Design Considerations

We derived three design **c**onsiderations (**C1 - Predictability**, **C2 - Level of Detail**, **C3 - Traceability**) to address when determining possible aspects of multilingual language style that users want to manipulate.

C1: To what extent is the dynamics of the target multilingual style predictable? The multilingual style desired is somewhat structured, influenced by the context-dependent nature of participants' language usage. For example, conversations *"related to [their] culture easier to have in [their] native language"* -P10. However, unpredictability from both the agent and the human adds variability to the path to the target multilingual style. Participants report that

Table 2: Design Considerations (C#) and Design Dimensions (D#) derived from the formative study.

Consideration	Dimension	Values
C1. To what extent is the dynamics of the target multilingual style predictable?	**D1.** Timing	Side-by-Side, In-Situ
C2. On what levels of detail do users want to manipulate the multilingual style?	**D2.** Granularity	Phrase, Message, Conversation
C3. To what extent do users need to cross-reference different versions of multilingual style?	**D3.** Presentation	Pre-Formation, Post-Formation

the multilingual style used by the agent is often unexpected, leaving them uncertain about how they should interact with the agent to arrive at their intended outcome. P8 for one refrains from interweaving English and Chinese in their prompts, since *"the agent may respond in Chinese when [they] don't want that"*. While participants had some understanding of how to avoid unintended behavior, it remains largely unclear how to achieve the intended behaviour. The target multilingual style itself may also evolve ad-hoc, when participants realize the current language is no longer the most suitable. P9 explains that occasionally *"[he] starts talking to the agent in Portuguese until [he] gets to a point where [he] expresses [himself] better in English"*, prompting an abrupt switch to English. We see that the process of steering the conversational agent to use target multilingual style is not always clear and predictable.

C2: On what levels of detail do users want to manipulate the multilingual style? Participants exhibited varying preferences regarding the units of language they desired to manipulate when specifying multilingual style. Their needs ranged from broad control over the language of entire conversations to more nuanced adjustments of specific messages or phrases, influenced by language proficiency and habits. In some cases, context is important, such as when participants *"want to double check if [their] understanding is correct"* -P2, leading them to want to *"translate the entire conversation"* -P2. Other times, there is a specific issue, like when they *"only know a word in one language"* -P8, requiring just a small adjustment.

C3: To what extent do users need to cross-reference different versions of multilingual style? Different versions of multilingual style need to be available simultaneously for sense-making across multiple languages, whereas a single version at a time suffices when users only need the most effective one. For some participants, *"seeing different language versions might be useful because it allows [them] to perceive [the content] in different contexts"* -P1 and gain a more holistic understanding. Cross-referencing can also be helpful if participants *"gets something that doesn't make sense from the agent"* -P2. However, other times only one version is needed because ultimately *"[one] only needs the answer in one language"* -P5. For example, P3 is often content if the *"agent respond in same language [they] prompt in"*, and is more concerned about whether *"[the response] is correct"*. Overall, the need for different versions of multilingual style to be present simultaneously depends on participants' goals and the role of language in achieving them.

3.3 Design Dimensions

From the design considerations, we dissected multilingual style manipulation into three design **d**imensions (**D1 - Timing**, **D2 - Granularity**, **D3 - Presentation**). For each of the dimensions, we explain its focus and the associated design possibilities.

D1: Timing of Control. While the desired multilingual style follows certain patterns in a given scenario, uncertainty in achieving it arises due to the unpredictability of the agent's responses and the spontaneous nature of human preferences (**C1 - Predictability**). Therefore, when designing user control over multilingual style, it is important to consider when users can and prefer to steer it. User input can be provided at two possible time points: before (*Pre-Formation*) or after (*Post-Formation*) the agent's response. In the *Pre-Formation* approach, the multilingual style can be adjusted before the agent generates a response, which is useful when the target style and/or agent response style is predictable. In the *Post-Formation* approach, the multilingual style can be adjusted after the agent has generated a response, allowing users to refine it as needed. This is particularly helpful for ad-hoc needs.

D2: Granularity of Control. Depending on their current need, users want to manipulate different units of language when controlling multilingual style (**C2 - Level of Detail**). We thus need to consider the granularity of user manipulations on multilingual style. The natural design choices for this dimension are *Phrase-Level*, *Message-Level* and *Conversational-Level*. At the *Phrase-Level*, users can manipulate one or more words, for flexible and fine-grained control. This is helpful when there is only a small part of the language style they need to adjust. At the *Message-Level*, users manipulate entire messages at a time. This maintains more context while still allowing targeted control. At the *Conversational-Level*, the entire conversation is manipulated. Though there exists limited flexibility, the entire context is maintained, providing coherency.

D3: Presentation of Control. Users may prefer to access one or multiple versions of the agent's response in different multilingual styles simultaneously, depending on their goal (**C3 - Traceability**). When designing control for multilingual style, it is important to consider whether and how multiple versions should be presented to help users navigate and compare different possibilities. The design alternatives for this dimension are *Side-by-Side* and *In-Situ*. In the *Side-by-Side* case, multiple variants of the style can co-exist at a single point in time. This allows users to easily cross-reference and compare different versions. In the *In-Situ* case, only one language representation can exist at a single point in time. This maintains the conversational metaphor while still allowing users to steer the multilingual style.

4 MIRRIOS: MULTILINGUAL STYLE CONTROL

The design considerations and dimensions serve as guidelines for implementing users control over the multilingual style of conversational agents. We create a proof-of-concept probe that manifests all of our design dimensions to assess which design choices users favor the most, and how users perceive and use multilingual style.

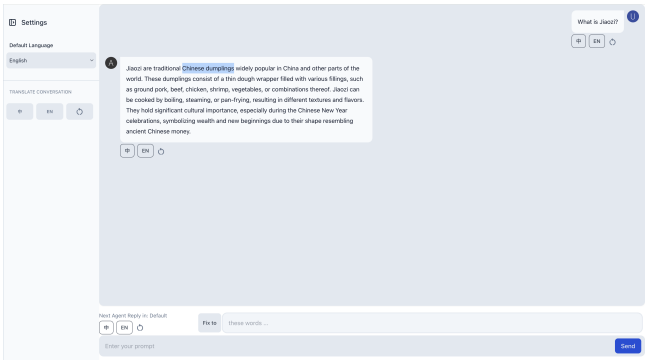

Figure 2: Overview of the design probe, Mirrios.

In particular, we developed a text-based conversational agent manifesting the design dimensions (Section 3.3) for multilingual style control (Figure 2). There are two variants of our probe, with different designs of the message interface, one for each value of the **Presentation** (**D3**) dimension; the details are in Section 4.1. Using our probe, users can freely manipulate the multilingual style of the agent responses to suit their needs. For simplicity, our probe is implemented for English and Chinese speakers as a proof-of-concept, but its features and design can be modified to support other language combinations.

4.1 User Interface Design

To materialize the **Presentation** dimension (**D3**), we designed two ways of displaying the user-initiated multilingual style manipulations (Figure 3). In the *Side-by-Side* design (Figure 3 Top), the new language representations appear beside the initial one. For message- or conversational-level operations, the entire new representation is rendered in a new message box below the original message. For phrase-level operations, the original representation is highlighted and the new representation appears as an annotation beside it. In the *In-Situ* design (Figure 3 Bottom), user-manipulated language representations replace the initial one. Depending on the granularity of operation, either entire messages or any sub-parts of it may be replaced by the new language representation.

4.2 Pre-Defining Multilingual Style

We developed a set of features (Figure 4A-C) that allow users to steer the multilingual style of the agent's response, before it's generated (**D1 - Timing**, *Pre-Formation*). Users can steer the agent's response towards a target multilingual style, pre-formation, at all **Granularity (D2)** levels. At the *Conversational-* (Figure 4A) and *Message-Level* (Figure 4B), users can instruct the agent to reply in a specific language for the rest of the conversation (*Conversational-Level*) or the next turn (*Message-Level*). At the *Phrase-Level* (Figure 4C), users can also choose the language in which certain phrases should appear in the next turn. The "Fix to" function takes a comma-separated list of phrases as input. The phrases can be in any language, i.e. both English and Chinese. Each phrase determines the language in which its corresponding semantic concept must always appear. For example, if the semantic concept "dumplings" is entered in English in the "Fix to" textbox, it will always appear in its English form in

Figure 3: Mirrios manifests the Presentation (D3) dimension, where language versions either co-exist (top; side-by-side) or replace each other (bottom; in-situ).

the following turn, overriding more global settings at *Message-* and *Conversational-Levels* if applicable. Conversely, if it were entered in Chinese, all instances of the semantic concept "dumplings" will appear in Chinese.

4.3 Adjusting Multilingual Style

We also defined a set of features (Figure 4D-F) that allow users to adjust the multilingual style of agent responses already generated (**D1 - Timing**, *Post-Formation*). Users can adjust the agent's response towards a target multilingual style, post-formation, at all **Granularity** (**D2**) levels. When the user decides to translate at the conversational-level, all existing messages are translated to the specified target language (*Conversation-Level*, 4D). If the user clicks on one of the buttons without selecting anything, the entire message is translated (*Message-Level*, 4E). Users can change a phrase (*Phrase-Level*, 4F) to a different language by selecting it and then clicking the corresponding target language button. 中 ("Zhong") is for translating to Chinese, EN for translating to English, and the reset button ("looping arrow") is for returning to the initial

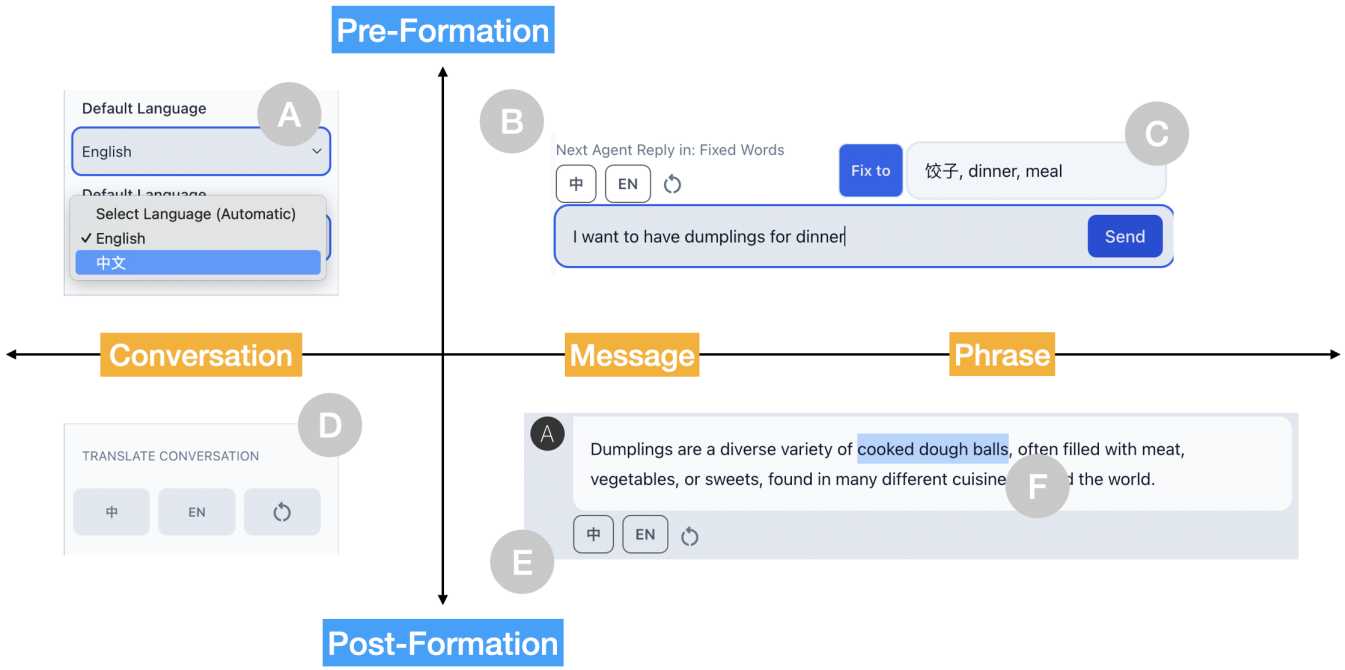

Figure 4: Mirrios manifests the Timing (D1) and Granularity (D2) dimensions by enabling the pre-definition (A - C) and adjustment (D - F) of multilingual style.

representation. Users can translate back and forth as many times as they want.

4.4 Implementation

We implemented the design probe as a web app. We used React for the front-end, and Node.js and Express.js for the backend. We leveraged OpenAI's GPT-4 model (`gpt-4-turbo-preview`) to generate the agent responses. The prompts to the GPT-4 model contained (1) the user's original message and (2) language style instructions based on the probe's state, i.e. user settings for the dimensions. We note that our goal is to provide a feasible manifestation of multilingual style control in conversational agents, and not an off-the-shelf conversational system. We do not claim our implementation is the best or even the only possibility, but simply that it is reasonable.

5 PROBE STUDY

To systematically explore our design considerations and dimensions, we conducted an user study where each participant experienced both variants of the probe, i.e. both the *Side-by-Side* and *In-Situ* interfaces, representing the **Presentation** (**D3**) dimension. The other dimensions, **Timing** (**D1**) and **Granularity** (**D2**), had the same implementation in both variants. Participants conducted one (1) task with each variant, with more details about the tasks in Section 5.1 and about the procedure overall in Section 5.2.

We examined user preferences and usage of different design choices (Section 5.3) and understand how they perceive multilingual style control (Section 5.4). These findings further provide us with implications on how to design multilingual style control, grounded in our design considerations and dimensions, which will be reported in Section 6.1.

5.1 Study Tasks

To encourage users to try out the different features, and to give them some starting points when interacting with the probe variants, we designed two scenarios to guide the users:

- **Travel Plans**: Suppose you are planning a month-long trip. You want to decide where to go and what to do. Possible locations include Beijing, Hainan, Shanghai, Ottawa, Montreal. Possible topics of interest include food, attractions, and events.
- **Culture Learning**: Suppose you want to learn more about the culture of different countries. Perhaps because you want to chat about it with friends. Example countries include China, Singapore, Canada, United States. Possible topics include literature, history, and art.

We choose the scenarios so that they did not require specific domain knowledge and could be done in any languages. Additionally, cultural-related topics were selected, since they are known to be language-sensitive and would serve to stimulate users to explore different multilingual styles. The scenarios served only as guidelines and we did not interfere with how users interacted with Mirrios as long as they remained meaningfully engaged. The combination and order of the probe variants and scenarios were balanced.

5.2 Participant and Procedure

We recruited 12 participants (7 Women, 5 Men, average age = 22.5 ± 1.24) through an university mailing list and snowball sampling. Five of participants were graduate students and seven were

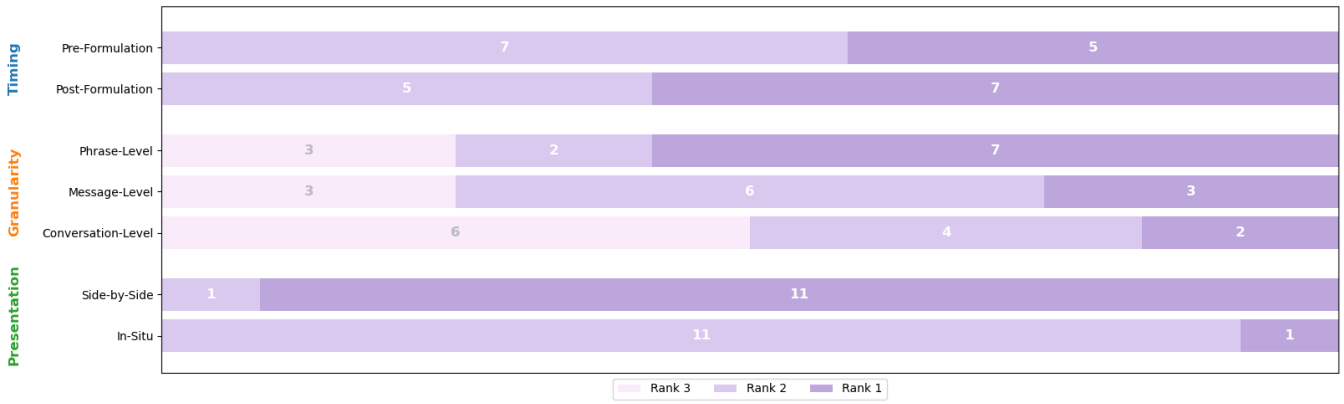

Figure 5: Distribution of participant rankings for the design dimensions, with Rank 1 being the best (lower = better).

undergraduates. Their field of study included computer science (5/12), mathematics (5/12), data science (1/12) and accounting (1/12). All participants were fluent in both English and Chinese. Additionally, participants use text-based conversational agents, such as ChatGPT and Ernie Bot, on a regular basis (at least $1-3$ times a week). Common usage of conversational agents by the participants include information retrieval, concept explanation, language translation, brainstorming, writing, and programming.

We began with the information and consent process, followed by a questionnaire asking about basic demographic information regarding language skills, conversational agent usage, and experience with multilingual conversational agents. Next, we gave the participants about 20 minutes to experience each of the probe variants. For each variant, we first explained how it works and then gave participants 5 minutes to freely play around with it, encouraging them to try all the available features to gain familiarity. Participants were informed that the probes were AI-powered and, as such, inherently imperfect. Next, we introduced the task scenario and offered participants 15 more minutes to engage with the probe, with the scenario in mind. Participants were encouraged to think-aloud [2] during the process, to help us can understand the reasoning behind their actions. At the end of the study, we gave participants a questionnaire asking them to rank and rate the values of the design dimensions. We wrapped up with a semi-structured interview, including questions on participants' thoughts on the design dimensions and values and the perceived utility of multilingual control as a whole. Each study session lasted around 70 - 80 minutes and were conducted in person. Participants were remunerated $20 in the local currency for their time. This study was approved by our institution's ethics board.

All study sessions were audio- and screen-recorded and then transcribed. The first author open-coded all interview transcripts, since we aimed to gain systematic and structured understanding [11] of how participants perceived having control over multilingual style and how they used and thought of the different dimensions and design choices. The initial codes were then discussed and iteratively refined until all authors reached an agreement. In the following sections, we present the key themes and patterns we extracted.

5.3 Findings: Exploring Design Dimensions

We explore the possible design choices across the design dimensions to investigate the desired manifestations of multilingual style control. The Likert-scale ratings are summarized in Figure 6, and the design choice rankings in each dimension are presented in Figure 5. We see that all design choices were perceived as neutral to positive, where the *Post-Formulation* (**D1 - Timing**), *Phrase-Level* (**D2 - Granularity**), and *Side-by-Side* (**D3 - Presentation**) values are the most preferred.

5.3.1 ***Timing (D1).*** Users generally preferred *Post-Formation* over *Pre-Formation* ($\mu_{\text{rank}} = 1.417 > 1.585$), because they *"want to see the message and then react"* -P2 and *"can only get confused after the model generated the message"* -P1. When users did prefer *Pre-Formation*, it was often because it *"kept the sense"* -P11 of being in a conversation. *Pre-Formation* was also found more convenient when participants could reasonably predict their own language needs. For example, some users liked the phrase-level pre-formation feature of "fix to," as it allowed them to set special words like *"academic terminology and place names"* -P6 to their preferred language, without needing to manually adjust the multilingual style each time the agent responded. In general, it seemed that the target multilingual style was usually not predictable. However, in the rarer cases where it could be accurately predicted, there was a notable benefit of convenience.

5.3.2 ***Granularity (D2).*** Users generally preferred *Phrase-Level* over *Message-Level* over *Conversational-Level* ($\mu_{\text{rank}} = 1.667 > 2.000 > 2.333$), though all three granularity were used in combination for different needs. *Phrase-Level* control was liked because users *"usually don't need the entire translation"* -P11 as they *"only don't know just one word"* -P8. Participants also liked phrase-level because it *"only changes one word or phrase, switching the [linguistic] sense less"* -P9 and making the conversation feel more coherent. Participants used *Message-Level* when *"phrase-level translation makes no sense and [they] want context"* -P2, or if they *"don't know the language well"* -P3. Participants tend to use *Conversational-Level* translations less, and think it's most useful for *"sending reports to others"* -P3 or making sense of the conversation as a whole. Generally, participants wanted the ability to manipulate the multilingual style at various granularities, rather than just one. Typically, user

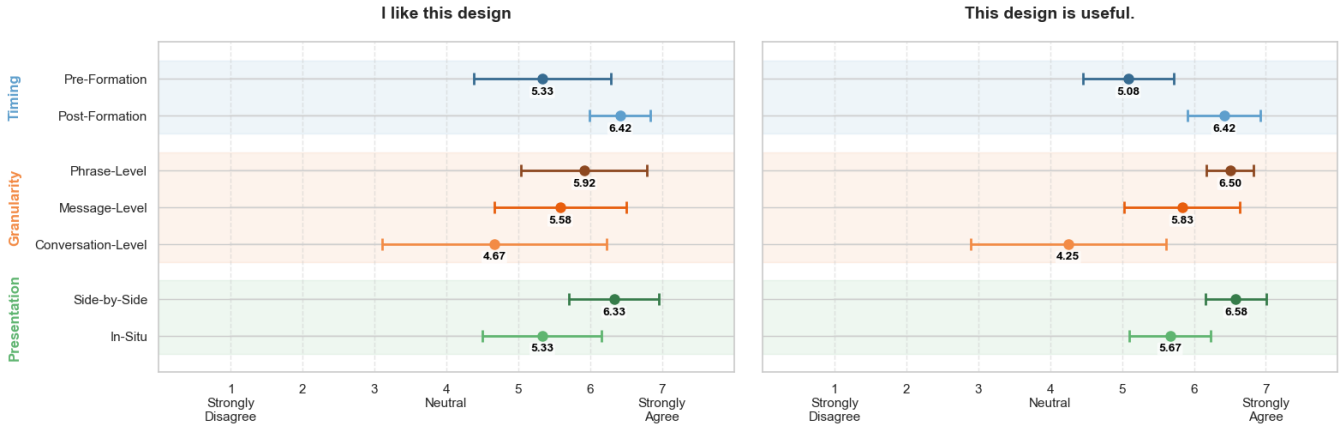

Figure 6: Distribution of Likert-scale ratings for the design dimension. Points represent the mean and error bars represent the 95% confidence interval from bootstrapping. The greater the rating, the better perceived the design choice.

manipulations were more fine-grained, and users sought additional context through coarser manipulations on demand.

5.3.3 ***Presentation (D3).*** Users generally preferred the *Side-by-Side* presentation over the *In-Situ* presentation ($\mu_{\text{rank}} = 1.083 > 1.917$). Participants preferred *Side-by-Side*, as they *"can see the context, what's wrong, and what was it before"* -P7 and could access *"more information at once"* -P2. However, participants thought *Side-by-Side* presentation was less suitable when they need to reference earlier messages, since *"the conversation became too long [...] and it was hard to locate and reference the translation"* -P4. On the other hand *In-Situ* was more ideal for maintaining the conversational metaphor, because it allows them to be *"more focusing on [themself]"* -P3 or because *"it feels more consistent"* -P9 and *"[they] don't need to switch in [their] head"* -P9 between different alternatives. Overall, while participants liked *Side-by-Side* for its cross-referencing support, there seems to exist a trade-off between traceability and naturalness and flow of conversation.

5.4 Findings: Utility of User Control

We examine participants' perception and usage of multilingual style control as a whole, in order to understand its utility. We found that the control offered the following benefits.

5.4.1 Reduce the need to constrain one's language habits. Participants reported that multilingual control reduced the need to adapt their language habits to ensure the conversational agent responded in the desired multilingual style. Notably, P2 set the agent's default reply language to Chinese but interacted with it in English. When asked, P2 explained that they *"preferred typing in English because it's more convenient"* while *"reading in Chinese is faster"* for them. P11 described another case where they want the agent to respond in a different language from the input: when using conversational agents for writing help, P11 wants to be able to *"provide the logic and outline in Chinese and then get the generated writing in English"*. P8 added that they have *"a set of words in English and a set of words in Chinese"*, and *"sometimes [they] don't know certain words"* in the other language. By setting the default reply language, they can freely switch between languages as it feels natural to them, without worrying about the output not aligning with their needs and preferences. P12 summarized that the *Pre-Formation* control enabled them to *"not [be] forced to use English when Chinese feels more comfortable"* and vice versa, *"reducing the mental load"* when interacting with the conversational agents.

5.4.2 Accommodates ad hoc and unexpected language needs. Participants found that multilingual control can accommodate their ad hoc and unexpected language needs. It is difficult to predict when multilingual users may need language support, as *"there is no standard answer"* -P2 for the gaps in their language proficiency. Participants confirmed that the presence or absence of language support in the conversational agent *"doesn't matter until [they] don't understand something"* -P2. *Post-Formation* control allows participants to address their language needs as they arise, where they can *"translate to know what it says"* -P2 for phrases that fall outside of *"[their] knowledge base"* -P3 in the agent's language. Multilingual control allowed participants used a on-demand and progressive approach to seek language support, where they *"first translate a word if I don't understand"* -P9 and then move up to translating whole messages *"if phrase-level translation makes no sense or if [they] want more context"* -P2. With controllability, participants' experiences are less reliant on the agent producing the perfect response, as they can easily recover from sub-optimal language choices by the agent.

5.4.3 Enables customized interaction with conversational agents. Participants believed that multilingual control allowed them to personalize and customize the language style of conversational agents to align with their preferences. Many expressed a preference for using specific languages for certain words or phrases, with some noting that they have *"different thinking ways in English versus Chinese"* -P9. *"Authenticity"* -P11 was a key consideration, as participants preferred proper nouns and specialized terminology to remain in their *"original language"* -P11 and wanted content to *"appear in the language [they] first learned it"* -P4. These key phrases *"[acted] as anchors"* -P10, and seeing them in the preferred language enhanced the flow and readability of the conversation. For example, P8 reports that they *"[fixed] the names of Western characters, movies*

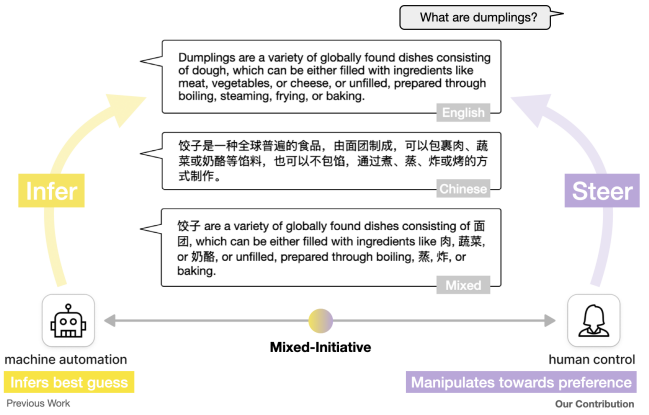

Figure 7: Envisioned mixed-initiative multilingual conversational agent. All three possible replies of the agent convey the same meaning, but are in different multilingual styles. Content in a desirable multilingual style is more digestible and preferred by users, but due to the ad-hoc and nuanced nature of multilingual preferences, agents cannot reliably infer the desired language style. In this research, we explored the concept of *multilingual style control*, to allow users to steer the agent towards the most desirable variant.

and books to English" while setting the rest to Chinese. Other participants fix terms to English *"when they notice the agent uses Chinese translations of English words"* -P1, for *"technical stuff like interview techniques"* -P10. Overall, participants found that multilingual control, especially on the *Pre-Formation Phrase-Level*, allowed them to fine-tune and tailor conversational agents to their liking.

6 DISCUSSION

We reflect on the design of the probe and the user studies, to propose design implications and directions for future research. We envision a future where user control and machine automation are meaningfully combined in mixed-initiative conversational agents, to allow arriving at the preferable multilingual style effectively and accurately (Figure 7).

6.1 Towards Mixed-Initiative Multilingual Conversational Agents

Based on our findings, we outline design implications for mixed-initiative [19] multilingual agents that combine automation and user control, and illustrate its utility with an envisioned usage scenario. Note that the usage scenario was designed with an "ideal" mixed-initiative multilingual agent in mind, grounded in our design implications, rather than being strictly limited to the study probe.

6.1.1 Design Implications. Our probe manifests all design dimensions and choices to help us systematically explore the design space. We reflect on the insights gained and the implications for future research and development of multilingual conversational agents.

Balance Agent Automation and Human Control. Allowing users to control multilingual style, rather than relying on existing approaches where the agent automatically determines the style, enables more natural and comfortable interactions. However, sometimes user also found that pure manual manipulation can be *"too much work [...] and if it's more automatic it would be more helpful"* -P2. Therefore, we advocate for a mixed-initiative approach [19], where agents proactively extrapolate and learn patterns based on user manipulations while maintaining user control. For example, our probe allows users to specify their preferred language for certain words and phrases using the `Fix To` function (*Pre-formulation*, *Phrase-Level*). A future mixed-initiative agent could infer preferred languages for related words, streamlining the process and reducing manual effort while still allowing users to edit or refine the list for optimal control. Ibn this contxt, user control can also be viewed as a teaching mechanism, where *Pre-formation* (**D1 - Timing**) reinforces correct usage, while *Post-formation* (**D1 - Timing**) highlights errors that need correction.

We note that implicit priming, as seen in existing work (e.g., [10]), and our manual steering approach each have their own advantages and limitations. Implicit priming is convenient when the user's intentions can be straightforwardly inferred from their language use – for example, when simply mimicking is sufficient. In contrast, manual steering becomes valuable when user goals and preferences are more latent or nuanced. For instance, in our probe study, we observed cases where users wanted to speak in one language but receive responses in another, or desired to adjust the language of specific words. These kinds of preferences are not easily inferred or handled automatically due to their dynamic and unpredictable nature. Overall, neither approach is perfect or universally optimal, and the trade-off between automation and agency is unavoidable [18]. We believe that an ideal mixed-initiative agent would navigate this trade-off effectively by allocating responsibility and initiative between user control and agent inference, depending on which is better suited for the task at hand.

Support Interactions across Different Levels of Granularity. While research in multilingual conversational agents have shifted from a monolithic language approach (i.e., either one language or another) to a more fine-grained lens of code-mixing [10, 12], in these work the multilingual style are determined automatically, which often do not align with user needs due to the nuanced and dynamic nature of language preferences. In our probe, we manifest the **Granularity** (**D2**) dimension and allow users to specify the multilingual style at *Phrase-*, *Message-*, and *Conversation-Levels*. As shown in our probe study, participants prefer starting with small adjustments – such as translating a technical term - to maintain the original context as much as possible, seeking additional context in another language only when necessary. Therefore, we must accommodate diverse user goals by enabling manipulation of multilingual style at varying levels of granularity. This helps users balance the trade-off between preserving the authenticity of the original message and meeting their own practical needs.

Facilitate Seamless Steering Without Disrupting User Flow. User control of multilingual style gives users more flexibility, but it can also interfere with their flow when interacting with conversational agents. In our probe, we manifested the **Presentation** (**D3**) dimension and found that while simultaneously displaying different language versions (*Side-by-Side*) helped users better reach their desired multilingual style, the added visual clutter could distract from their core goal or conversation. It is essential to recognize

that manipulating multilingual style is a means to enhance user comprehension and interaction with the agent's output, rather than an end in itself. User control should therefore be designed to be quick and intuitive, with the option to toggle it off when not needed. In a mixed-initiative multilingual conversational agent, the system should ideally learn as much as possible about user preferences, reducing the need for manual adjustments to only the edge cases.

Process Multilingual Styles in Meaningful Units. Although some work have considered language style in multilingual conversational agents, the focus has primarily been on the presence or absence of multilingual style considerations, with the style itself being determined randomly, such as by probabilistically determining the language of the next word [10]. In our studies, we found that while user preferences for multilingual style are somewhat ad hoc and not perfectly predictable, there are discernible patterns. For example, users preferred specialized terminology to appear in the language in which they first encountered it. To align with users and create more natural multilingual styles, conversational agents should learn these structures to use multilingualism in a more meaningful way, rather than relying solely on stochastic methods.

6.1.2 Envisioned Usage Scenario. Suppose Alice is a second-generation Chinese immigrant. Alice can read and understand basic Chinese but for her, writing and speaking is much easier in English since she moved to Canada at a young age. Alice will be traveling back to China in the summer and want to learn more about the street foods there. She wants to learn about the street foods in Chinese since it will be useful when she arrives in China, but type in English since it's easier for her. Alice sets the default reply language (**D1 - Timing**: *Pre-Formation*; **D2 - Granularity**: *Conversation-Level*) to Chinese so that the agent responds to her in Chinese even if she types in English. Alice can now ask about Chinese street foods in English, and read about them in Chinese, which perfectly fits her needs and habits. Occasionally, Alice comes across a word she doesn't understand in Chinese because she's immigrated for too long and forgot. When this happens, she selects the word to translate it (**D1 - Timing**: *Post-Formation*; **D2 - Granularity**: *Phrase-Level*) to English to understand. Alice wants to learn some Chinese, so she sets the view to *Side-by-Side* (**D3 - Presentation**) to be able to cross-reference English and Chinese versions. If translating a single word is insufficient, due to lacking context or too literal translations, Alice translates the entire message (**D1 - Timing**: *Post-Formation*; **D2 - Granularity**: *Message-Level*) to English for more context. After a few turns, Alice realizes that she is relying on English translations too much and perhaps defaulting agent responses to Chinese was not a good choice. Alice decides that only seeing food names in Chinese is good enough for her. She fixes a few such phrases she saw so far to Chinese (**D1 - Timing**: Pre-Formation, **D2 - Granularity**: Phrase-Level). By analyzing the words that were fixed to Chinese, the agent extrapolates that Alice wanted all cultural-specific terms in Chinese, and responds accordingly. Hence Alice doesn't always have to manually translate them or fix their language, which was convenient and allows her to focus on her information search. Ultimately, Alice successfully completes her research on street food in China, in a way that aligned with her language skills and preferences.

6.1.3 Technical Feasibility. Our design probe demonstrates how the interface of a conversational system can be leveraged to provide control and structure to the agent's language style. As we alluded, pure manual control can be tedious. An ideal agent should be mixed-initiative, balancing user agency and automation [18, 19]. The key technical challenge in achieving this lies in the division of roles and tasks between the user and the agent—in particular, determining when and what the agent should learn. Future work could explore both heuristic-driven strategies and more learning-based approaches to tackle this problem.

6.2 Multilingualism and the Gulfs of Execution and Evaluation

As the generative powers of large-language models increase, conversational agents are now able to respond regardless of the quality of the user input, this leads to a new problem of abstraction matching [27, 33]. Abstraction matching, or the act of having to overcome the gap between the nearly infinite natural language space and the much smaller space of well-formed prompts, widens the gulfs of evaluation and execution as originally defined in Norman's 7 stages of action [28]. We argue that multilingual users face an additional gap, which is the gap between the space of their natural, interweaved language usage, and the space of multilingual utterances that a conversational agent can meaningfully process and produce (Figure 1). As such, multilingual users not only struggle with content misalignment (*semantics*) but also misalignment in the multilingual presentation and style (*pragmatics*). Having to view agent responses in an undesirable language can be just as frustrating as encountering irrelevant content. By giving users the ability to specify the multilingual style of agents before generation, we can reduce the gulf of execution by minimizing randomness in style in agent responses. Additionally, allowing users to steer the multilingual style after the agent responds can reduce the gulf of evaluation, helping them better understand the agent's replies. Overall, user control of multilingual style can reduce misalignment between the user and the multilingual agent, enhancing their experience by reducing the need to constrain one's language habits, accommodating for ad hoc language needs, and facilitating customized interactions.

6.3 Limitations and Future Work

As an illustrative example, we recruited people who are fluent in English and Chinese, and did not consider other language pairs. Hence one limitation of our work is that due to our sampling population, findings from the probe study may not be fully generalizable to other multilingual populations speaking language pairs outside of English and Chinese. We do note that our focus is not on the preferred multilingual style itself, but rather the *user control* of multilingual style when interacting with conversational agents, which is arguably somewhat more language-agnostic. Existing work has shown that multilingual speakers do share some high-level behaviour patterns, such as switching to different languages in different contexts [4, 16]. Therefore, it is possible that key patterns we observed may be shared across different linguistic communities. Future work could validate whether this holds true and, if not, identify the key differences between linguistic groups in their perceptions and usage of multilingual style control in conversational agents.

The focus of our work is to derive and explore design considerations for user control of multilingual style in conversational agents. Due to the exploratory nature of our work, Mirrios naively manifests all the proposed design dimensions as a probe for understanding multilingual style control, and is neither intended nor suitable for use as an out-of-the-box tool. As a result, individual features may also not be optimized for direct usage. For example, the "Fix to" box clears after each turn rather than retaining entries, which can become tedious during actual use. While this design choice encourages users to engage more mindfully with the control features, it serves as a reminder of Mirrios's role as a research probe rather than a ready-to-use tool. Future work should apply the derived design implications to create a mixed-initiative multilingual conversational agent intended for end users, moving beyond a monolithic research probe. Since our goal was not to contribute a ready-to-use prototype, we did not compare conditions with and without user control, nor did we examine interactions between different design dimensions. These remain avenues for future work. Finally, we note that multilingual style is only one aspect of language style, and it would also be valuable to adapt our findings to other aspects of language style, such as formality [14], and to design user control in those contexts. In this work, we focused on multilingual style as an example since it makes a more practical difference,

7 CONCLUSION

Our research explores the concept of user control over multilingual style in conversational agents, addressing the limitations of current stochastic methods that fail to align with the nuanced and ad-hoc needs of multilingual users. Through a formative study, we identified design considerations and dimensions for multilingual style control. We implemented Mirrios, a prototypical conversational system with multilingual style control, and used it as a probe to understand the design possibilities and their broader utility. We found that user control of multilingual style can enhance personalization and flexibility, allowing users to navigate their multilingual needs more naturally. Based on our findings, we also propose design implications for user control to inspire the development of future mixed-initiative multilingual conversational agents.

ACKNOWLEDGMENTS

This work is supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) through the Discovery Grant #RGPIN-2020-03966 and the USRA program, and the University of Waterloo through the URA and URF programs. We thank all the participants for their time and our reviewers for their valuable insights. We acknowledge that much of our work takes place on the traditional territory of the Neutral, Anishinaabeg, and Haudenosaunee peoples. Our main campus is situated on the Haldimand Tract, the land granted to the Six Nations that includes six miles on each side of the Grand River.

REFERENCES

[1] [n. d.]. The CEFR Levels - Common European Framework of Reference for Languages (CEFR) - Www.Coe.Int. https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions.

[2] [n. d.]. Thinking Aloud: The #1 Usability Tool. https://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/.

[3] Christian Adjemian. 1976. On the Nature of Interlanguage Systems. *Language Learning* 26, 2 (1976), 297–320. https://doi.org/10.1111/j.1467-1770.1976.tb00279.x

[4] Christian Adjemian. 1976. On the Nature of Interlanguage Systems. *Language Learning* 26, 2 (1976), 297–320. https://doi.org/10.1111/j.1467-1770.1976.tb00279.x

[5] Mohammad Hamad Al-khresheh. 2015. A Review Study of Interlanguage Theory. *International Journal of Applied Linguistics and English Literature* 4, 3 (May 2015), 123–131. https://doi.org/10.7575/aiac.ijalel.v.4n.3p.123

[6] Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-Scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *Transactions of the Association for Computational Linguistics* 4 (2016), 463–476. https://doi.org/10.1162/tacl_a_00111

[7] Tal August, Kyle Lo, Noah A. Smith, and Katharina Reinecke. 2024. Know Your Audience: The Benefits and Pitfalls of Generating Plain Language Summaries beyond the "General" Audience. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–26. https://doi.org/10.1145/3613904.3642289

[8] Tal August, Katharina Reinecke, and Noah A. Smith. 2022. Generating Scientific Definitions with Controllable Complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 8298–8317. https://doi.org/10.18653/v1/2022.acl-long.569

[9] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2023. Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing. *ACM Trans. Comput.-Hum. Interact.* 30, 5 (Sept. 2023), 74:1–74:38. https://doi.org/10.1145/3589955

[10] Anshul Bawa, Pranav Khadpe, Pratik Joshi, Kalika Bali, and Monojit Choudhury. 2020. Do Multilingual Users Prefer Chat-bots That Code-mix? Let's Nudge and Find Out! *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 1–23. https://doi.org/10.1145/3392846

[11] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[12] Yunjae J. Choi, Minha Lee, and Sangsu Lee. 2023. Toward a Multilingual Conversational Agent: Challenges and Expectations of Code-mixing Multilingual Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3544548.3581445

[13] Peng Chu and Anita Komlodi. 2017. TranSearch: A Multilingual Search User Interface Accommodating User Interaction and Preference. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (<conf-loc>, <city>Denver</city>, <state>Colorado</state>, <country>USA</country>, </conf-loc>) *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 2466–2472. https://doi.org/10.1145/3027063.3053262

[14] Samuel Rhys Cox and Wei Tsang Ooi. 2022. Does Chatbot Language Formality Affect Users' Self-Disclosure?. In *Proceedings of the 4th Conference on Conversational User Interfaces (CUI '22)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3543829.3543831

[15] Danuta Gabryś-Barker. 2006. Language Activation in the Thinking Processes of a Multilingual Language User. *International Journal of Multilingualism* 3, 2 (July 2006), 105–124. https://doi.org/10.1080/14790710608668391

[16] Kieran Green. 2023. Identification of Commonalities across Different Languages. *Frontiers in Language Sciences* 2 (Nov. 2023). https://doi.org/10.3389/flang.2023.1172925

[17] Gualberto Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2017. Metrics for Modeling Code-Switching Across Corpora. In *Interspeech 2017*. ISCA, 67–71. https://doi.org/10.21437/Interspeech.2017-1429

[18] Jeffrey Heer. 2019. Agency plus Automation: Designing Artificial Intelligence into Interactive Systems. *Proceedings of the National Academy of Sciences* 116, 6 (Feb. 2019), 1844–1850. https://doi.org/10.1073/pnas.1807184115

[19] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems the CHI Is the Limit - CHI '99*. ACM Press, Pittsburgh, Pennsylvania, United States, 159–166. https://doi.org/10.1145/302979.303030

[20] Qian Hu and Zhao Pan. 2024. Is Cute AI More Forgivable? The Impact of Informal Language Styles and Relationship Norms of Conversational Agents on Service Recovery. *Electronic Commerce Research and Applications* 65 (May 2024), 101398. https://doi.org/10.1016/j.elerap.2024.101398

[21] Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. 2025. A Survey on Large Language Models with Multilingualism: Recent Advances and New Frontiers. https://doi.org/10.48550/arXiv.2405.10936

arXiv:2405.10936 [cs]

[22] Naveena Karusala, Aditya Vishwanath, Aditya Vashistha, Sunita Kumar, and Neha Kumar. 2018. "Only If You Use English You Will Get to More Things": Using Smartphones to Navigate Multilingualism. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174147

[23] Bo Young Kim, Qingyan Ma, and Lisa Diamond. 2024. "It's in My Language": A Case Study on Multilingual mHealth Application for Immigrant Populations With Limited English Proficiency. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3613905.3637125

[24] Eunhee Kim. [n. d.]. Reasons and Motivations for Code-Mixing and Code-Switching. ([n. d.]).

[25] Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. XLM-V: Overcoming the Vocabulary Bottleneck in Multilingual Masked Language Models. arXiv:2301.10472 (Oct. 2023). https://doi.org/10.48550/arXiv.2301.10472 arXiv:2301.10472 [cs].

[26] Chenjun Ling, Ben Steichen, and Alexander G. Choulos. 2018. A Comparative User Study of Interactive Multilingual Search Interfaces. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 211–220. https://doi.org/10.1145/3176349.3176383

[27] Michael Xieyang Liu, Advait Sarkar, Carina Negreanu, Benjamin Zorn, Jack Williams, Neil Toronto, and Andrew D Gordon. 2023. "What It Wants Me To Say": Bridging the Abstraction Gap Between End-User Programmers and Code-Generating Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–31.

[28] Don Norman. 2013. *The design of everyday things: Revised and expanded edition*. Basic books.

[29] Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2024. Multilingual Large Language Model: A Survey of Resources, Taxonomy and Frontiers. https://doi.org/10.48550/arXiv.2404.04925 arXiv:2404.04925 [cs]

[30] Valeria Resendez. 2020. A Very Formal Agent : How Culture , Mode of Dressing and Linguistic Style Influence the Perceptions toward an Embodied Conversational Agent? https://essay.utwente.nl/82242/.

[31] Ben Steichen, Chenjun Ling, and Silvia Figueira. 2023. Multilingual News Search—A Comparative User Study of Desktop and Mobile Interfaces. *International Journal of Human–Computer Interaction* 0, 0 (2023), 1–16. https://doi.org/10.1080/10447318.2023.2238978

[32] Jackson Stokes, Tal August, Robert A Marver, Alexei Czeskis, Franziska Roesner, Tadayoshi Kohno, and Katharina Reinecke. 2023. How Language Formality in Security and Privacy Interfaces Impacts Intended Compliance. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3544548.3581275

[33] Hari Subramonyam, Roy Pea, Christopher Pondoc, Maneesh Agrawala, and Colleen Seifert. 2024. Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt Based Interactions with LLMs. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3613904.3642754

[34] Samson Tan and Shafiq Joty. 2021. Code-Mixing on Sesame Street: Dawn of the Adversarial Polyglots. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 3596–3616. https://doi.org/10.18653/v1/2021.naacl-main.282

[35] Michelle M. E. van Pinxteren, Mark Pluymaekers, Jos Lemmink, and Anna Krispin. 2023. Effects of Communication Style on Relational Outcomes in Interactions between Customers and Embodied Conversational Agents. *Psychology & Marketing* 40, 5 (2023), 938–953. https://doi.org/10.1002/mar.21792

[36] Lindsey Vanderlyn, Gianna Weber, Michael Neumann, Dirk Väth, Sarina Meyer, and Ngoc Thang Vu. 2021. "It Seemed like an Annoying Woman": On the Perception and Ethical Considerations of Affective Language in Text-Based Conversational Agents. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, Arianna Bisazza and Omri Abend (Eds.). Association for Computational Linguistics, Online, 44–57. https://doi.org/10.18653/v1/2021.conll-1.4

[37] Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. PolyLM: An Open Source Polyglot Large Language Model. arXiv:2307.06018 (July 2023). https://doi.org/10.48550/arXiv.2307.06018 arXiv:2307.06018 [cs].

[38] Bryan B. Whaley and Wendy Samter. 2007. *Explaining Communication: Contemporary Theories and Exemplars*. Psychology Press.

[39] Zheng Xin Yong, Ruochen Zhang, Jessica Zosa Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Rowena Garcia, Thamar Solorio, and Alham Fikri Aji. 2023. Prompting Multilingual Large Language Models to Generate Code-Mixed Texts: The Case of South East Asian Languages. In *Sixth Workshop on Computational Approaches to Linguistic Code-Switching*.

[40] Fei Yuan, Shuai Yuan, Zhiyong Wu, and Lei Li. 2023. How Multilingual Is Multilingual LLM? https://doi.org/10.48550/arXiv.2311.09071 arXiv:2311.09071 [cs]

[41] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–21. https://doi.org/10.1145/3544548.3581388

[42] Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Indra Winata, and Alham Fikri Aji. 2023. Multilingual Large Language Models Are Not (Yet) Code-Switchers. https://doi.org/10.48550/arXiv.2305.14235 arXiv:2305.14235 [cs].

[43] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. WildChat: 1M ChatGPT Interaction Logs in the Wild. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=Bl8u7ZRlbM