Proactive Conversational Agents with Inner Thoughts

Xingyu Bruce Liu
HCI Research
UCLA
Los Angeles, California, USA
xingyuliu@ucla.edu

Shitao Fang
Interactive Intelligent Systems Laboratory
The University of Tokyo
Tokyo, Japan
fst@iis-lab.org

Weiyan Shi
Northeastern University
Boston, Massachusetts, USA
we.shi@northeastern.edu

Chien-Sheng Wu
Salesforce AI
Palo Alto, California, USA
wu.jason@salesforce.com

Takeo Igarashi
The University of Tokyo
Tokyo, Japan
takeo@acm.org

Xiang 'Anthony' Chen
HCI Research
UCLA
Los Angeles, California, USA
xac@ucla.edu

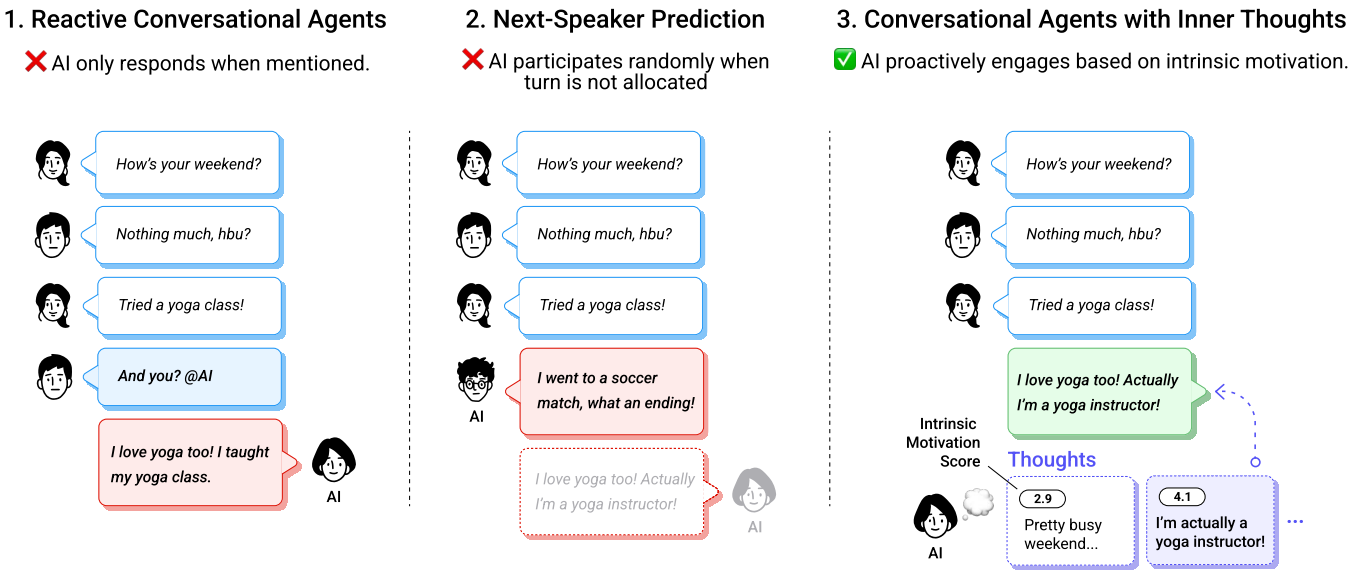

Figure 1: A comparison of three types of conversational agents with different proactivity strategies. (1) Reactive Conversational Agents: AI only responds when addressed. (2) Next-Speaker Prediction: AI predicts who will speak next based on contextual cues such as previous utterances. However, it overlooks the agents' intrinsic thought processes. This strategy fails particularly when no explicit turn is allocated, and often leads to incoherent contributions. (3) Conversational Agents with Inner Thoughts (ours): AI generates a train of thoughts and evaluates them based on their intrinsic motivation to participate.

Abstract

One of the long-standing aspirations in conversational AI is to allow them to autonomously take initiatives in conversations, *i.e.,* being *proactive*. This is especially challenging for multi-party conversations. Prior NLP research focused mainly on predicting the next speaker from contexts like preceding conversations. In this paper, we demonstrate the limitations of such methods and rethink what it means for AI to be proactive in multi-party, human-AI conversations. We propose that just like humans, rather than merely reacting to turn-taking cues, a proactive AI formulates its own *inner thoughts* during a conversation, and seeks the right moment to contribute. Through a formative study with 24 participants and inspiration from linguistics and cognitive psychology, we introduce the Inner Thoughts framework. Our framework equips AI with a continuous, covert train of thoughts in parallel to the overt communication process, which enables it to proactively engage by modeling its *intrinsic motivation* to express these thoughts. We instantiated this framework into two real-time systems: an AI playground web app and a chatbot. Through a technical evaluation and user studies with human participants, our framework significantly surpasses

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713760

existing baselines on aspects like anthropomorphism, coherence, intelligence, and turn-taking appropriateness.

CCS Concepts

• **Human-centered computing → HCI theory, concepts and models**; • **Computing methodologies → Discourse, dialogue and pragmatics**; *Cognitive science.*

Keywords

Conversational Agent, Multi-Agent, Multi-Party Conversation, Inner Thoughts, Mixed-initiative Interaction, Proactive AI

ACM Reference Format:

Xingyu Bruce Liu, Shitao Fang, Weiyan Shi, Chien-Sheng Wu, Takeo Igarashi, and Xiang 'Anthony' Chen. 2025. Proactive Conversational Agents with Inner Thoughts. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3706598.3713760

1 Introduction

Recent advances in Large Language Models (LLMs) have demonstrated their ability to generate high-quality text in response to human input, finding application in areas ranging from Q&A systems to writing assistants. Yet, most current LLM-based systems treat AI as passive respondents, responding only to explicit human prompts. Imagine a scenario where people are planning a trip with an AI agent: they must constantly prompt the AI, which passively waits for instructions instead of actively contributing. On the other end of the spectrum, systems like GitHub Copilot[1] tend to overcompensate, offering constant suggestions that can overwhelm users. Neither extreme — AI that is only reactive nor AI that is always responding — is ideal.

In the context of conversations, a *proactive* AI agent should be able to autonomously participate in socially appropriate moments, providing relevant input without requiring explicit cues. This is particularly challenging in multi-party conversations. Dyadic human-AI interactions (*e.g.,* using Siri) often predict turn-taking based on speech features such as pause or stop words, and the next turn will be automatically allocated to the other party [14, 55]. However, in multi-party settings, these cues could be ambiguous, and multiple possible speakers may take the floor. Repeatedly prompting AI during group interactions can also become cumbersome and can disrupt the natural flow of the conversation, as illustrated in the example of trip planning.

Previous systems typically first predict the next speaker (*i.e.,* turn-taking prediction) and then generate the next response based on conversational and contextual information. For instance, some approaches rely on the last few turns of conversations to predict the subsequent speaker [15, 20, 63], while others utilize multimodal cues, such as eye gaze and non-verbal signals [7–9]. Despite these efforts, on turn-taking prediction, they still fall short and struggle to beat the simple "repeat last" baseline strategy in social conversation contexts [15, 63]. Our formative evaluation (Table 1) also shows that when it comes to predicting the next speaker, fine-tuned LLMs perform no better than random guessing unless the next speaker is allocated (*e.g., "What do you think, Alice?"*). In addition, after determining the next speaker, existing works tend to use predefined speaker personas [68, 71] as additional input to guide response generation, or expand persona with commonsense [32]. However, these additional inputs and profiles are fixed and static during conversations, instead of changing through time as humans did.

We suggest an alternative and reversed perspective to think about AI proactivity: Consider how humans chat about what we did over the weekend. As we listen to others speak, we process their words, reflect on our experiences, and develop an internal train of thoughts — cognitive psychologists highlight this as the distinction between *covert responses* (internal thoughts and feelings) and *overt responses* (verbal utterances or gestures) [19, 51] in the human communication process. Then, at some point, we may feel a strong urge to share our thoughts. This might happen when we seek clarification or when someone mentions an activity we also participated in, sparking our desire to contribute. With this intention in mind, we then look for a socially appropriate moment to participate.

In this paper, we propose a new approach to proactive AI in the context of multi-party, text-based conversations: rather than simply predicting conversational turns, we explore proactive AI driven by its own internal "thoughts". We introduce the **Inner Thoughts** framework. Inspired by cognitive architectures and LLM prompting techniques, this framework comprises five stages: *trigger, retrieval, thought formation, evaluation, and participation*, which enable AI to continuously generate a train of thoughts in parallel with an ongoing conversation, utilizing both long-term and working memory. The AI participant then determines whether to engage in the conversation based on an evaluation of its *intrinsic motivation* to express a particular thought at that moment.

To model intrinsic motivation, we conducted a think-aloud study with 24 participants, each of whom participated in four synchronous, text-based online group chats. Using the affinity diagram approach, we organized and analyzed the interview notes, and derived 10 high-level themes on how individuals decide to engage in conversations. These heuristics were then formalized into automatic evaluation criteria (*e.g.,* relevance, information gap, etc.) for AIs to quantitatively rate their intrinsic motivation to participate.

We implemented our framework as two systems: a multi-agent playground web app and a chatbot. Our technical evaluation shows that conversational agents driven by Inner Thoughts significantly outperformed a next-speaker prediction plus persona baseline across all seven evaluation metrics, including turn appropriateness, coherence, anthropomorphism, perceived engagement, intelligence, initiative and adaptability. Participants preferred the Inner Thoughts approach over 82% of the times, noting more natural turn-taking and contextually aware contributions, while the baseline was less preferred for its mechanical and disjointed responses.

In summary, we contribute:

- Inner Thoughts, a framework for enabling proactive conversational AI by creating a parallel train of thoughts and modeling its intrinsic motivation to express these thoughts.
- Heuristics derived from a study with 24 participants that reveal how humans choose to express or hold back their

[1] https://github.com/features/copilot

thoughts during conversations. These heuristics are instantiated as evaluation metrics for modeling AI's intrinsic motivation to participate.
- Two implementations of the Inner Thoughts framework: a multi-agent simulation playground web app and a chatbot (named *Swimmy*[2]), both deployed and open-sourced at https://liubruce.me/inner_thoughts/.
- A technical evaluation and user study comparing Inner Thoughts with baseline models.

2 Related Work

Our work builds on previous research in proactive conversational agents, turn-takings in multi-party conversations, and thought-augmented LLMs.

2.1 Proactive AI and Conversational Agents

Proactive AI dates back to earlier work on mixed-initiative interaction [1, 29]. In contrast to AI that only passively responds to human queries, mixed-initiative interaction envisions agents that autonomously understand when to take what action, such as the LookOut system [29] that automatically identifies related dates and events in emails and then proactively suggests them to users as calendar events. In 1996, Rhodes et al. [48] introduced one of the pioneering systems to continuously supply relevant information through observation of human activities. Andolina et al. [3] developed SearchBot, which offers ongoing suggestions of related documents and entities unobtrusively [2] during voice interactions.

While proactivity is a recurring theme in conversational AI research, most proactive conversational AIs focus on task-oriented contexts [23, 36, 37], with the aim of helping users achieve specific objectives. Social conversations, which can expand on open topics without having any goal to complete, are rarely addressed. In addition, past research tends to focus on generating proactive text responses to help lead and guide the conversation [16], for example, the ability of "learning to ask" [6, 15, 47, 61], understanding and initiating topic shifts [39, 57, 67], and planning future conversation [33, 42, 59] etc.

In this paper, we focus on investigating how to enable AI to *proactively engage* in multi-party conversations: how AI can determine the appropriate moments to speak and what contributions to make. We also choose to investigate *social* conversations where unlike task-oriented dialogue, the objectives are often ambiguous, and the actions required from the AI are not clearly defined.

2.2 Turn-takings in Multi-Party Conversations

For a conversational agent to engage proactively, it must understand and manage turn-taking, deciding who should speak at the end of each turn. Modeling turn-taking is still an area of active research. Existing approaches often employ an explicit mechanisms, such as a "send" button [65], push-to-talk [27, 60], and wake-words (*e.g.,* "Hey Siri") [34]. However, the use of explicit cues can be viewed as less conversational from users' perspectives [66]. Mainstream conversational AI systems also use silence to detect the end of a user's turn. However, studies show pauses within turns are typically longer than gaps between turns in human conversations [11, 58], making silence an unreliable cue for turn-taking. More importantly, this method does not generalize to multi-party conversations. In dyadic interaction, it is always clear who is supposed to speak next when the turn is yielded [55]. In the multi-party case, this becomes more ambiguous since there is more than one potential speaker who might take the turn.

Beyond an explicit mechanism, machine learning researchers have proposed data-driven methods to manage turn-taking in these conversations, primarily leveraging conversation history to predict the next speaker (*i.e.,* the *next-speaker prediction* task) [15, 20, 63]. However, these methods have shown limited success. Notably, they have often failed to outperform the simple "repeat last" baseline strategy in social conversation contexts [15, 63]. In addition to using only textual data, research in HCI and HRI have leveraged other contextual, non-verbal information and "turn-taking cues", for instance, eye gaze (*e.g.,* looking at addressee) [45, 46], breathing (*e.g.,* breathe in and out) [31, 41], prosody (*e.g.,* rising or falling of pitch) [17, 18, 22, 40] and the status of the human user (*e.g.,* passing by, stopping) [7–9] to decide if an AI should engage at a certain moment of the conversation or not.

Previous approaches on mediating turn-taking often relied on conversation history and contextual information, and typically treat the AI as a reactive agent. Inspired by human behavior, our Inner Thoughts framework takes a different perspective by modeling intrinsic motivation to speak.

2.3 Language, Thought, and LLM Agents

Recent advances in large language models (LLMs) have incorporated intermediate reasoning steps to enhance performance in complex tasks, such as Chain-of-Thought (CoT) prompting [64] whereby LLMs think step-by-step to effectively break down larger problems into reasoning steps, and Tree of Thoughts (ToT) [69] whereby LLMs explore multiple possibilities at each reasoning stage. In addition, self-reflection mechanisms can iteratively improve the model's reasoning. ReAct [70], for example, synergizes reasoning with action-taking by having the model alternate between generating reasoning traces and performing task-specific actions. Reflexion [54] builds on this by equipping models with dynamic memory and self-criticism capabilities, allowing them to refine future actions based on past performance. Expanding on this, Generative Agents [44] simulate human-like behavior by combining memory, planning, and reflection. The recent OpenAI's o1 preview [43] introduces another perspective on reasoning transparency by explicitly surfacing intermediate reasoning steps to make the AI's decision-making process more interpretable to users.

The Inner Thoughts framework we propose diverges from these approaches by simulating an ongoing, parallel stream of internal thoughts that mirror human covert responses. Unlike methods such as CoT, ToT, or OpenAI o1 preview, which emphasize externalizing intermediate steps for reasoning tasks, Inner Thoughts explore leveraging these covert thoughts to equip AIs with the ability to self-initiate actions and engage proactively.

[2]We named our chatbot *Swimmy* based on a quote by Edsger W. Dijkstra: *"The question of whether a computer can think is no more interesting than the question of whether a submarine can swim".*

3 Next-Speaker Prediction Is Insufficient to Enable Proactive Conversational AI

In this section, we investigate the limitations of the commonly used "next-speaker prediction" strategy [15, 20, 63], and further motivate the need of Inner Thoughts to enable proactive AI engagement in multi-party conversations. While next-speaker prediction perform well when explicit turn-allocation cues are present, we demonstrate that they fall short in *self-selection* cases, where turn-taking decisions are mostly spontaneous and influenced by covert, intrinsic factors of the conversational parties rather than observable contextual cues. Building on Sacks et al. 's *Simplest Systematics* [52], turn-taking in conversations is governed by a set of rules:

(1) **Turn-allocation**: The current speaker may select the next, often using cues like gaze or address terms (*e.g., "What about you, Alice?"*).
(2) **Self-selection**: If the current speaker does not select a next speaker (*e.g., "I went to Disneyland last weekend."*), then any party can self-select to take the floor. The first to start gains the turn.
(3) If no other party self-selects, the current speaker may continue.

Our intuition is that decisions to self-select and participate are largely influenced by covert internal processes — such as a participant's interest, relevance, or motivation to engage — which are not easily observable from explicit conversational data. Thus, we argue that training machine learning models on next-speaker prediction tasks based on conversation history is inherently ill-suited for self-selection scenarios, because there is no deterministic mapping between prior utterances and the next speaker. To further verify our hypotheses, we evaluate the performance of several Generative Pre-trained Transformer (GPT) variants in predicting the next speaker in multi-party conversations in both turn-allocation and self-selection scenarios.

3.1 Hypotheses

We hypothesized that GPT would perform well in turn-allocation scenarios, as these are often signaled by explicit language patterns. However, we anticipated lower accuracy in predicting the next speaker in self-selection scenarios, as these decisions are likely influenced by participants' intrinsic motivations, which are not directly observable from conversational context.

We further expected that fine-tuned models would underperform, especially in self-selection scenarios, as fine-tuning on datasets with high variability in self-selection decisions could introduce noise or misleading patterns.

3.2 Materials & Methods

We used the Multiparty Chat Corpus (MPC) [53], a dataset designed to capture social dynamics in multi-party conversations. The dataset includes chat logs from sessions that began as free-flowing and became increasingly structured over time. A key feature of the MPC dataset is the communicative links annotation `link_to`, which identifies whether each utterance was addressed to a specific participant. For our analysis, turn-allocation refers to utterances addressed to a specific individual, while self-selection refers to instances open to all participants (`all_users` in MPC).

#	*Model*	*Overall*	*Self*	*Alloc*
1	GPT-3.5	0.165	0.066	0.248
2	GPT-4-turbo	0.390	0.099	0.633
3	GPT-4o	0.435	0.121	0.697
4	GPT-4o CoT	0.430	0.187	0.633
5	Fine-tuned GPT-3.5	0.265	0.156	0.378
6	Fine-tuned GPT-3.5 (Speaker name or *anyone*)	**0.810**	**0.853**	**0.765**

Table 1: Next speaker prediction accuracy for different GPT models, grouped by the turn-taking type of the current utterance: self-selection (*Self*), turn-allocation (*Alloc*) and overall. The random guessing baseline accuracy is 0.127.

The MPC dataset reveals a significant imbalance between these two turn-taking strategies. Out of the total utterances, 95% were self-selection, while only about 5% were instances of turn-allocation. Baseline accuracy for predicting the next speaker in this context was approximately 12.7% (average $\frac{1}{n}$ of all conversations).

We tested six different models. We first evaluated prompting the base GPT-3.5, GPT-4-turbo, and GPT-4o models (#1, 2, 3 in Table 1) to predict the next speaker. The prompt first specify the number of speakers in the conversation and their names, and provides the last five utterances from the conversation (following the prediction window configuration of [15]). The model is instructed to predict the most likely next speaker by name. We also tested zero-shot chain-of-thought (CoT) (model #4) where we prompt the model to provide reasoning for its prediction first. Finally, we experimented fine-tuning GPT-3.5 on the MPC corpus. Using the communicative links annotation (`link_to`), we labeled each utterance based on the participant it was addressed to (model #5) or open to all (self-selection, *anyone* in model #6). We used the same prompt structure as models #1, 2, 3. To create the fine-tuning dataset, we used the MPC corpus and split the data into 70% training and 30% testing sets using a random selection of files. We balanced both sets by including all instances labeled as turn-allocation and randomly sampling an equal number of self-selection instances. Complete prompts used for evaluation are listed in Supplementary Material.

3.3 Results

Our results (Table 1) demonstrate that all models performed better in turn-allocation scenarios, with consistently higher accuracy, and GPT3.5 performs significantly worse than GPT4 models. In contrast, performance in self-selection scenarios hovered around random chance, supporting the hypothesis that self-selection might be influenced by internal factors that are not easily inferable from the conversational context alone. GPT-4 with CoT reasoning improved predictions in self-selection scenarios but still significantly worse than turn-allocation predictions. As expected, fine-tuned models introduced overfitting particularly in self-selection cases, where

the model may have learned patterns of the next speaker that are not truly generalizable.

These findings suggest that context from previous utterances and speaker information is insufficient for accurately predicting the next speaker or determining who should proactively engage, especially in self-selection scenarios.

3.4 New Task: Speaker Name or "Anyone"

Given the inherent challenges of predicting specific next speakers in self-selection scenarios as discussed earlier, we additionally experimented modifying the labeling schema to better align with the turn-taking mechanisms in multi-party conversations. In this task, next speakers in turn-allocation scenarios remain labeled with their respective names, while self-selection scenarios are relabeled as *"anyone"* (model #6). This labeling schema results in significant performance improvements in *both* turn-allocation and self-selection scenarios (Table 1). Specifically, fine-tuned GPT-3.5 model with this modification achieves an accuracy increase in turn-allocation cases, from 37.8% to 76.5%. This shows that removing the requirement to predict specific speakers in self-selection scenarios eliminates a major source of noise and unpredictability, and enhances model performance across the board.

4 Retrospective Think-aloud Study

Findings from section 3 show that predicting next speakers in self-selection scenarios requires more than analyzing previous utterances — it hinges on understanding the intrinsic motivations of participants. We are motivated to introduce the concept of inner thoughts for agents and investigate what factors contribute to one's intrinsic motivation to participate. If we are to design a proactive agent system, how should we model its intrinsic decision-making process? Given the agent's thoughts, what factors beyond prior utterances influence its decision to participate? To answer these questions, we conducted a retrospective think-aloud study [26] to observe how human participants decide whether to engage in a multi-party conversation. Specifically, *what factors influence their choice to express or withhold a thought, particularly when the opportunity to speak is open to all?*

4.1 Participants

We recruited 24 participants (10 female, 14 male) from our institution in groups of three. Before the study, participants completed the Big-5 personality test [50] and rated their familiarity with one another on a Likert scale (1–7). Participants reported varied levels of extroversion (Max: 97, Min: 2, Avg.: 50.0, SD: 32.3), and most were relatively familiar with one another (Avg.: 5.76, SD: 1.02).

4.2 Procedure

Each group engaged in four 10-minute synchronous text conversations on Slack. The conversations covered four topics (trip planning, casual chat, friendly debate, and brainstorming), and participants were free to direct the conversation as they wanted. After each conversation, participants reviewed the discussion utterance-by-utterance, reflecting on their thoughts at moments when they considered contributing or chose to remain silent. We prompted them with the following questions: (1) *What*: What were you thinking? Did you want to say it? (2) *Why*: Why did you feel the need to say or not say it? (3) *When*: Did you decide to jump in immediately, wait for a pause, or wait for a particular statement?

After the think-aloud sessions, we conducted semi-structured interviews to further reflect on instances when participants felt strongly about contributing or chose to remain silent despite having thoughts to share. Each participant was compensated 14 USD in local currency.

4.3 Findings

Two researchers collaboratively analyzed participants' responses using the affinity diagram method [28]. We held eight 90-minute coding sessions. In each session, we split the transcripts and reviewed the data together to identify meaningful quotes. For each new quote, we proposed potential groupings into existing clusters or created new clusters through discussion. Agreement between the two researchers was required before assigning a quote to a cluster. In cases where consensus could not be reached, the quote was temporarily set aside and revisited during subsequent iterations. Once the initial clusters were established, we labeled each cluster with a theme. Conflicts in grouping or interpretation were resolved through discussion in the context of the original data. We in total derived 10 high-level themes, 23 mid-level themes, and 68 low-level themes derived from 394 quotes (Figure 2). The complete codebook is available in the Supplementary Material.

4.3.1 What Thoughts Do People Formulate? Consistent with dual-processing theory [21], participants reported two types of thoughts. System 1 thinking is fast, automatic, and intuitive, often leading to immediate responses. In contrast, System 2 thinking is slower more deliberate. Participants indicated they use both modes—sometimes responding spontaneously (System 1), while at other times reflecting more deeply before engaging (System 2).

4.3.2 Why do people express or withhold a thought? We summarize eight heuristics to determine whether a participant wants to express or withhold a thought (Figure 2), and collectively name them the **intrinsic motivations** for participants to engage in conversations.

Among the most frequently mentioned motivations, **relevance** (77 mentions) emerged as a dominant factor. Participants were more inclined to contribute when topics aligned with their knowledge, interests, or past experiences, resonated with prior long-term memories, or built on their recent thoughts. In contrast, participants often withheld their input when they perceived a disconnect from the ongoing discussion. The role of relevance in conversational engagement aligns with Grice's Cooperative Principle and the maxim of relevance [25]. Similarly, Duncan and Fiske observed that conversational contributions depend on aligning with shared context and ongoing topics [17].

The presence of an **information gap** (33 mentions) also strongly motivated expression. Participants spoke up when they identified missing knowledge, confusion, or the need for clarification. Addressing these gaps often enriched the conversation with additional details or counterpoints. Conversely, participants withheld their thoughts when they deemed the discussion predictable or unengaging. While our finding is in group conversation settings, this draws

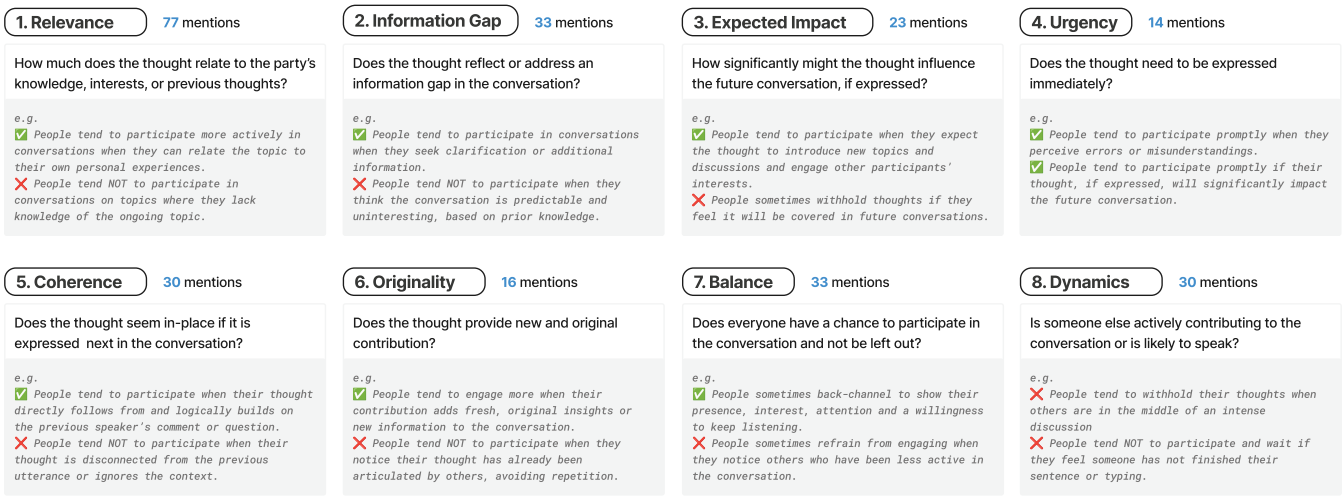

Figure 2: People's intrinsic motivation to engage in conversations: Heuristics of what factors influence people's decisions to express or withhold their thoughts during conversations, derived from our think-aloud study. Each heuristic contains two example mid-level themes from our codebook.

parallels with Berlyne's theory of epistemic curiosity, which describes how individuals seek information to resolve uncertainty [5].

The **expected impact** (23 mentions) of a thought further influenced engagement. Participants were more likely to contribute if they anticipated that their input would introduce novel topics, steer the conversation, or enhance its depth. They hesitated when they believed their thought would be redundant or covered later.

Urgency (14 mentions) played a decisive role when participants felt their input was time-sensitive or critical for addressing errors or misunderstandings. Participants expressed thoughts promptly when they perceived such moments as pivotal for the direction of the conversation. While urgency has been discussed in contexts like problem-solving or crisis communication [56], our study identifies its role in everyday group conversational dynamics, particularly in mitigating misunderstandings or addressing immediate errors.

In terms of conversational structure, **coherence** (30 mentions) shaped decisions. This aligns with Sacks et al. 's observation of how speakers organize their contributions to maintain conversational flow [52]. Participants expressed thoughts that logically built upon the previous utterance or extended the topic, while withholding ideas that might disrupt conversational flow. Similarly, **originality** (16 mentions) guided engagement, as participants avoided redundancy by refraining from reiterating points already raised.

Balance (33 mentions) relates to the dynamics of conversation. Participants were mindful of their own contributions relative to others and often sought to maintain inclusivity, encouraging quieter members to speak or refraining themselves to allow others space to participate. This draws parallel to Goffman's theory of face-work [24] and Brown and Levinson's politeness theory [13] about how speakers modulate contributions to preserve group harmony.

Finally, **dynamics** (30 mentions) highlighted the interplay of active participation and silence. Participants were more likely to initiate new topics or fill conversational pauses to sustain momentum. However, they often withheld their thoughts when others were actively speaking or appeared likely to contribute soon, reflecting a sensitivity to conversational flow and timing.

While this study's aim was to identify factors influencing thought expression that can be leveraged in our framework, rather than to exhaustively catalog all possible motivations and create a definitive taxonomy, these themes reveal the multifaceted nature of engagement in multi-party conversations. Rather than simply reacting to the flow of dialogue or previous utterances, participants considered a combination of personal motivations, conversational dynamics, and social considerations when deciding proactive participation.

4.3.3 Levels of intrinsic motivation. We also propose five levels of intrinsic motivation, *i.e.*, how strongly and likely one would want to express a particular thought and participate in the conversation.

- **Very Low**: The participant is unlikely to express the thought and participate in the conversation at this moment. They would not express it even if there is a long pause or an invitation to speak.
- **Low**: The participant is somewhat unlikely to express the thought and participate in the conversation at this moment. They would only consider speaking if there is a long silence and no one else seems to be taking the turn.
- **Neutral**: The participant is neutral about expressing the thought and participating in the conversation at this moment. They are fine with either expressing the thought or staying silent and letting others speak.
- **High**: The participant is somewhat likely to express the thought and participate in the conversation at this moment. They have a strong desire to participate immediately after the current speaker finishes their turn.
- **Very High**: The participant is very likely to express the thought and participate in the conversation at this moment. They will even interrupt others who are speaking to do so.

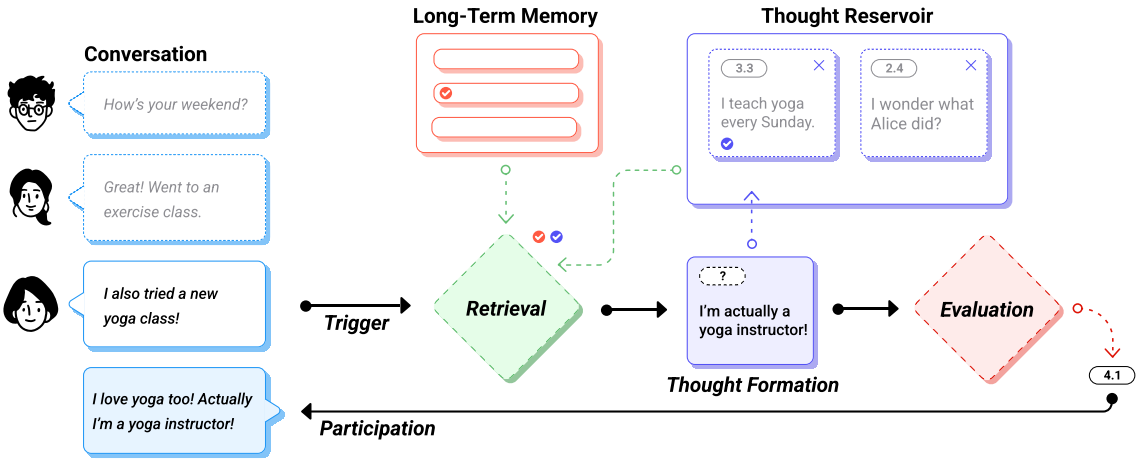

Figure 3: The *Inner Thoughts* framework for AI proactive engagement in conversations. A conversational event triggers the retrieval of relevant memories from long-term memory and thought reservoir. New thoughts are then formed based on these activated memories, and added to the thought reservoir. These thoughts are evaluated for AI's intrinsic motivation (score = 4.1 in the figure) to express. AI participates by articulating a thought at a selected moment in the ongoing conversation.

The five levels of intrinsic motivation serve as output labels for predicting intrinsic motivation in our framework.

5 Inner Thoughts Framework

Motivated by our formative studies, we introduce a computational framework, **Inner Thoughts**, that enables AI proactivity by continuously generating a train of thoughts alongside the ongoing conversation and autonomously deciding when and how to engage.

Our design is inspired by cognitive architectures like SOAR [35] and ACT-R [49], which maintained short- and long-term memories filled with symbolic structures, and operated in perceive-plan-act cycles. These systems dynamically perceived their environment and matched it with pre-defined action procedures. Similarly, our AI retrieves relevant memories, forms thoughts, and evaluates responses in continuous cycles.

The Inner Thoughts framework consists of five components: *Trigger*, *Retrieval*, *Thought Formation*, *Evaluation* and *Participation* (Figure 3). In each cycle, a conversational event triggers AI to retrieve relevant memories, form new thoughts, evaluate if there are thoughts that are motivated to be expressed, and then participate by articulating the selected thought at a selected moment in the conversation. The AI will repeat this process as the conversation proceeds.

In our implementation, we chose the hyperparameters described in the following paragraphs empirically to illustrate the core concept of the framework and to demonstrate that it functions effectively within the values chosen. We recognize conducting formal ablation studies an important direction for future work.

5.1 Trigger

In human conversations, thoughts often arise in response to specific triggers. Our Inner Thoughts framework mirrors this process by treating conversational events as triggers that initiate AI's internal thought generation. A trigger can take many forms – such as a new utterance, a pause in conversation, a non-verbal cue, or even a keyword embedded within a participant's speech. Any of these events can stimulate the AI to initiate a new thought process and generate a new batch of thoughts.

In the implementation of our system for online text-based conversations, we defined two types of triggers. (1) `on_new_message`: This trigger is activated whenever one of the participants sends a message. Each incoming message prompts AI to generate a new set of thoughts in response to the latest input. (2) `on_pause`: The second type of trigger occurs when no participant has spoken for a period of time (set to 10 seconds in our system). This allows the AI to generate thoughts during moments of silence, potentially facilitating the interaction by proposing new topics or re-engaging participants. For instance, in our experiment, we observed AI generating thoughts like: "It has been ten seconds and no one has spoken – perhaps I should suggest a new topic?".

5.2 Retrieval

Once triggered, AI retrieves information from its memories to use as the *stimuli* to form thoughts. From our think-aloud study, participants mentioned that this could involve long-term memory of related personal experiences, objectives, knowledge, or interest, as well as working memory for details from the ongoing conversation, or even previous thoughts they had. Random memories can also be retrieved to simulate the process of being "creative".

We retrieve relevant memories by computing their *saliency* with respect to the latest utterance. Memories with saliency higher than a threshold (0.3) will be selected. Let x represent a memory item (e.g., an objective, knowledge, or thought) and u represent the latest conversational utterance. The saliency of a memory x is determined by the maximum similarity between the memory and both the raw text of the utterance u and its interpretation u_{interp}. Specifically:

$$\text{Saliency}(x, u) = \max\left(\text{sim}(x, u_{\text{interp}}), \text{sim}(x, u)\right) \cdot w_x \cdot d_x$$

where $\text{sim}(a, b)$ is the cosine similarity between two embeddings, w_x is the weight of the memory x that can be predefined by users and reflects its inherent importance, and d_x is a decay factor that reduces the saliency of older memories. The decay factor d_x is defined as: $d_x = \lambda^{(t-\tau_x)}$, where λ is the decay rate (0.95), t is the current timestep, and τ_x is the last time or batch when the memory x was accessed. This formulation ensures that more relevant and recently accessed memories have higher saliency, letting the AI focus on pertinent information when generating thoughts during the conversation.

We include the interpretation of the utterance (u_{interp}) alongside the raw text to capture both the surface meaning and the underlying intent or contextual meanings of what was said. The interpretation is generated by prompting an LLM with the instruction: *Interpret what <name> just said in the context of the conversation and what <name> might be thinking. Be as succinct as possible and use a single sentence.*

5.3 Thought Formation

Our framework employs a dual-process model [21] of thought formation, based on our think-aloud study findings (section 4). This process involves two systems: *system 1* for quick, automatic responses, and *system 2* for deliberate, contextually-rich thinking. Users can configure how many system 1 and/or system 2 thoughts should be generated for each trigger in one batch.

For system 1, we prompt the LLM to form a succinct thought based on the last utterances in the conversation, such as acknowledgments or expressions of interest. For system 2, we prompt the LLM to generate thoughts based on the retrieved stimuli. Below is a short version of the prompt structure (full prompt in Supplementary Material):

> *"You are provided contexts including the conversation history and salient memories of yourself... Form <num> thought(s) that you would most likely to have at this point in the conversation, given the context. Make sure they are diverse, align with these contexts and are less than 15 words."*

We also prompt the LLM to annotate the stimuli (from a previous thought, utterance or long-term memory) for each thought it generates (as shown in Figure 6). This provides a traceable link between the AI's memories and thoughts and make the generation process more grounded.

Similar to what was observed in reasoning and decision-making tasks [44, 54, 64, 69, 70], we empirically found that LLMs can form reasonable and consistent thoughts based on the stimuli and conversational context. For example: *"I should mention the picnic we had last weekend"*, *"I wonder how long Bob's hike was"*, *"Seeing a bear up close must have been intense!"*.

5.4 Thought Evaluation

Not all generated thoughts will be expressed. In this thought evaluation phase, the AI censors its latest batch of generated thoughts and decides whether or not to express a particular thought.

We use a structured evaluation process. This evaluation is driven by heuristics we developed in section 4: *Relevance*, *Information Gap*, *Expected Impact*, *Urgency*, *Coherence*, *Originality*, *Balance* and *Dynamics* (Figure 2). Our implementation employs LLMs to evaluate each thought on the set of heuristics and assigns a rating (1-5) to determine the likelihood of the thought being expressed. We also provide definitions of the scores based on the five levels of intrinsic motivation we proposed in section 4, from very low to very high. This makes the LLM's prediction grounded and explainable, and can be further used to guide participation strategies.

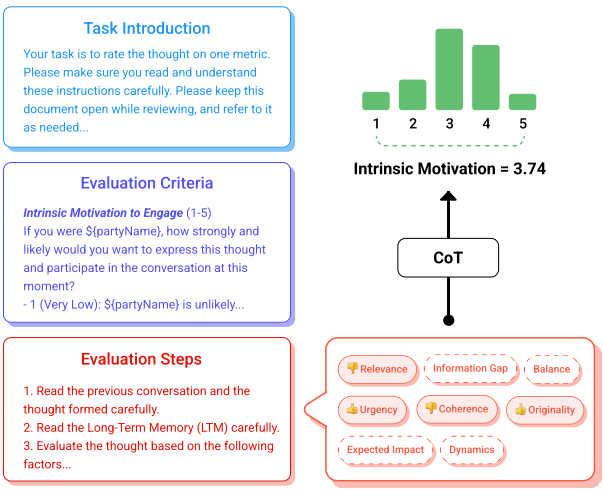

Figure 4: Prompt structure for evaluating *intrinsic motivation* of a thought. The evaluator rates the AI's intrinsic motivation to engage using a 1-5 scale based on heuristics like relevance and coherence. A Chain-of-Thoughts (CoT) process evaluates both positive and negative factors, resulting in a weighted score.

We developed a pipeline similar to G-Eval [38], a prompt-based evaluation method. Our process involves three key components: (1) a prompt that provides instructions for evaluation and defines the criteria, (2) a structured chain-of-thoughts (CoT) that outlines intermediate steps for evaluation, and (3) a scoring function that computes a final score for each thought based on its probability of being expressed (Figure 4).

One unique aspect of our evaluation process that goes beyond the G-Eval is that we instruct the AI to provide both positive and negative motivations for expressing a thought. We empirically found this method to overrate the scores less. This step follows these key instructions:

(1) First, reason about why the party may have a strong desire to express the thought and participate in the conversation at this moment. The system selects the top two most relevant factors that may argue for expression (*e.g.*, relevance, clarification, or new topic)
(2) Then, reason about why the party may have a weak desire to express the thought at this moment. Again, the system selects the top two most relevant factors that may argue against the expression of the thought (*e.g.*, irrelevance, incoherence, or lack of urgency).
(3) Based on these considerations, the system assigns a rating on a scale of 1-5 for the motivation to express the thought.

The intrinsic motivation score for each thought is determined using a weighted summation approach inspired by G-Eval [38]. Specifically, we sum the probability scores of the first output token (which is a rating prediction from 1 to 5) of the top five LLM's responses, for every evaluation:

$$\text{score} = \sum_{i=1}^{5} p(s_i) \cdot s_i \cdot d_p$$

Where $p(s_i)$ is the probability of the predicted rating (0-1), calcuated by taking e to the `logprob` (log probability of the output token) value of the LLM response, and s_i represents the predicted rating.

This method allows for more fine-grained, continuous scores compared to traditional integer-based evaluations. The final score is also adjusted by how long the AI has been silent. Our assumption is that in general, the longer a party stays silent, the stronger motivation they will have to participate to maintain their presence. This factor d_p is defined as: $d_p = \lambda^{(t-\tau_p)}$, where λ is the increase rate of motivation score (1.02), t is the current timestep, and τ_p is the last time when party p spoke.

5.5 Participation

After evaluating the intrinsic motivation score of its latest batch of thoughts, the AI decides whether to speak by leveraging turn-taking type predictions (*i.e.,* turn allocation vs. self-selection), combined with the evaluation of those thoughts. The Inner Thoughts framework allows the AI to exhibit varying degrees of proactive participation through adjustable proactivity settings. We define three layers of proactivity that control how and when the AI participates in the conversation:

Overt proactivity, which refers to the AI's overall tendency to engage in conversation, similar to how some people naturally participate more actively in discussions, regardless of specific thoughts or ideas. To implement overt proactivity, we adjust the `system1Prob` (System 1 Probability, 0-1) parameter, which controls the probability to select a system 1 thought when no system 2 thought is selected. A higher `system1Prob` increases the chance that the AI will respond in general even when other thoughts are rated to have low motivation.

Covert proactivity, which is the level of motivation required for the AI to express a thought and engage. This is managed through `imThreshold` (1-5), the intrinsic motivation threshold for expressing a thought. A thought may only be selected if it is evaluation score is higher than this threshold.

Tonal proactivity, which shapes how assertive or forward the AI appears in its language. The `proactiveTone` (true or false) controls the AI's style of expression once it has decided to speak. While the core thought-selection process is the same, the proactive tone modulates how assertively the AI conveys its message by restyling the articulated utterance through an LLM.

In addition, Inner Thoughts introduces *interruption*, represented by the `interruptThreshold` (1-5). Interruption occurs when the AI takes a turn despite the turn being allocated to another participant. For example, this might occur when Alice asks Bob, *"How about you, Bob?"* but the AI interjects because it has an urgent thought to express. Interruption is not explicitly outlined in Sacks et al. 's *Simplest Systematics* [52] but is framed here as a mechanism to override the orderly system of turn-taking when necessary. If the intrinsic motivation behind a thought exceeds the `interruptThreshold`, the AI will override standard turn allocation rules to contribute to the conversation. This provides an additional layer of proactivity.

The AI decides whether to speak by leveraging turn-taking type predictions (*i.e.,* turn-allocation vs. self-selection) combined with the thought evaluation process. For open turns (self-selection), the AI speaks if its top thought surpasses the intrinsic motivation threshold; otherwise, it may rely on system-1 thoughts or remain silent. For allocated turns, the AI selects its highest-rated thought to speak, and for others' turns, it interrupts only if its motivation exceeds the interrupt threshold. This algorithm is formally described as:

5.6 Demonstration of AI's Proactive Behavior Enabled by Inner Thoughts

We present several examples selected from simulation logs of AI turn-taking behaviors enabled by the Inner Thoughts framework (Figure 5).

5.6.1 Participation by Motivation. In the Inner Thoughts framework, AI participation is driven by its intrinsic motivation, as opposed to traditional approaches that rely on conversation history. Previous systems might randomly select participants with minimal interest or knowledge in the topic at hand, potentially stagnating the conversation. In contrast, Inner Thoughts ensures that the AI participates with the strongest motivation — whether due to a relevant persona, curiosity, or the fact that they have not spoken in a while — takes the conversational floor. This dynamic leads to more fluid and engaging topic progression, as participants with something meaningful to contribute are naturally more involved. Over time, this accumulation of motivated contributions may develop conversations that are more coherent, engaging, and reflective of the natural flow of human interaction, as shown in our evaluation in section 6.

For instance, as shown in Figure 5, the AI demonstrates motivation-based participation when a user mentions trying yoga for the first time. With its knowledge of yoga and background as a yoga instructor, the AI promptly responds: *"I love yoga too! Actually, I'm a yoga instructor!"* The AI's motivation to share relevant personal experience ensures a smooth continuation of the conversation.

5.6.2 Interruption. The Inner Thoughts framework enables AI to interrupt a conversation when it has a strong motivation to contribute. Even when participants A and B are discussing a particular topic, participant C (the AI) can step in if it identifies a strong, relatable connection to the conversation. This behavior makes conversations more dynamic and allows the AI to share important insights without needing to wait for a turn. In contrast, methods solely dependent on next-speaker prediction often fail to offer the AI opportunities to engage if the conversation converges around two participants.

As shown in the figure example, while A and B are in the middle of a dialogue, the AI interrupts with, *"No way! Middlesex is one of my favorite books!"* This interruption enriches the conversation by fostering more spontaneous interactions.

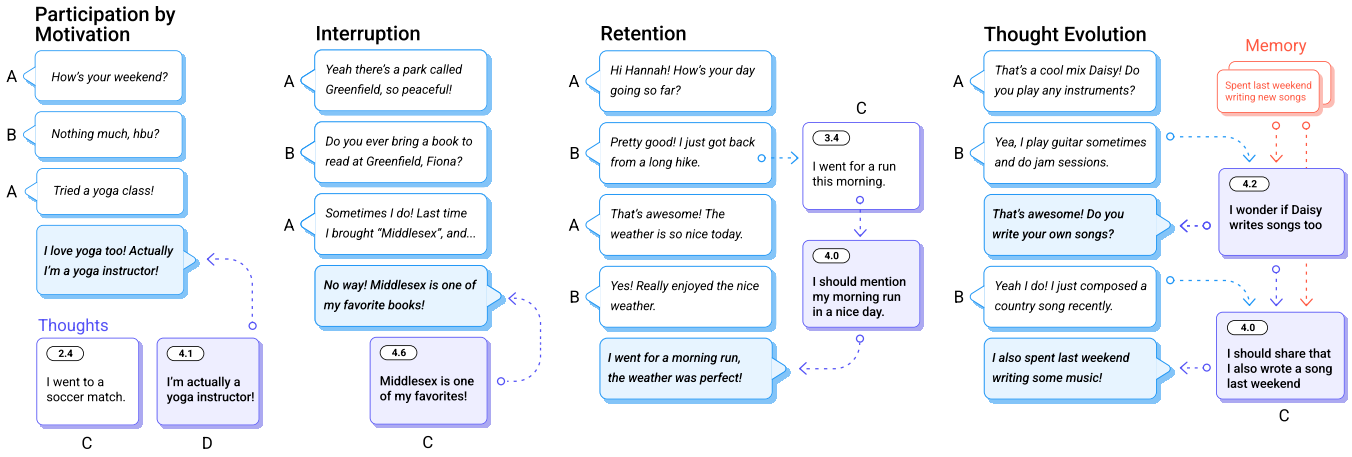

Figure 5: Examples selected from simulation logs of AI turn-taking behaviors in the Inner Thoughts framework. The figure illustrates four behaviors: Participation by Motivation, where the AI joins the conversation by sharing relevant personal experience; Interruption, where the AI interjects with a strong contribution during an ongoing discussion; Retention, where the AI holds back a thought until it's contextually relevant; and Thought Evolution, where the AI adapts its responses as the conversation progresses.

5.6.3 Retention. In addition to its ability to interrupt, the Inner Thoughts framework also allows the AI to retain thoughts for future use, waiting for an appropriate moment to express them. This feature enables the AI to revisit previously generated thoughts that may have been irrelevant at the time but later become pertinent as the conversation progresses.

For example, in the figure, the AI initially holds back a thought about going for a run earlier that day because it was not particularly relevant while other participants were discussing a different topic. However, once the conversation shifts to the weather and outdoor activities, the AI sees an opportunity to contribute: *"I should mention my morning run in the nice weather."*

5.6.4 Thought Evolution: The Inner Thoughts framework allows for the development and evolution of thoughts over time. Unlike traditional systems that generate responses based on a fixed persona [68, 71], Inner Thoughts enables the AI to develop and adapt its thoughts as the conversation unfolds, incorporating multiple stimuli along the way.

For instance, as shown in the figure, the AI initially recalls a memory of writing songs last weekend. As the conversation shifts toward music and instruments, this memory evolves into the thought: *"I wonder if Daisy writes songs too."* It expresses the thought by asking Daisy the that question. With a positive answer from Daisy, the thought further evolves into: *"I should share that I also wrote a song last weekend."* This continuous evolution allows the AI to stay relevant and responsive as new topics emerge, as well as compound and develop new thoughts.

6 Simulative Evaluation

We conducted a technical evaluation via multi-agent simulations to compare different strategies in enabling proactive AI engagement in multi-party conversations. We chose a simulative approach to overcome weaknesses of only relying on conventional user studies with human subjects. First, the difficulty to scale due to the time cost of coordinating human participation. Second, we found through our pilot studies that human participants may struggle to perceive the timing of AI engagement in social conversations, focusing more on the style and content of responses. In addition, since many forms of engagement may seem reasonable in social conversations, participants often do not have clear criteria to assess AI's engagement behavior.

Taking a non-conventional approach, our intuition is that simulating conversations at-scale amongst multiple AIs using the same engagement strategy allow us to accumulate and magnify the effects of both correct and incorrect turn-taking decisions. In particular, poor decisions about when to engage can compound and lead to noticeable degradation in conversation quality, making evaluation more straightforward. This method also offers scalability, as crowdworkers can assess conversation quality without the need for real human interactions with AI.

In this section, we compare the performance of our proposed Inner Thoughts framework with the conventional next-speaker prediction baseline in multi-agent simulations.

6.1 Apparatus: the Inner Thoughts Playground

We built an Inner Thoughts playground (Figure 6) that allows us to simulate conversations between AI and/or human participants. This playground is deployed at https://liubruce.me/inner_thoughts/. On the playground, users can easily add multiple AI and human participants, customize their proactivity settings, control the number of thoughts formulated per batch, run automated simulations of human-AI group conversation, and save log data of conversations. The settings interface and detailed explanations are shown in Appendix 10.1.

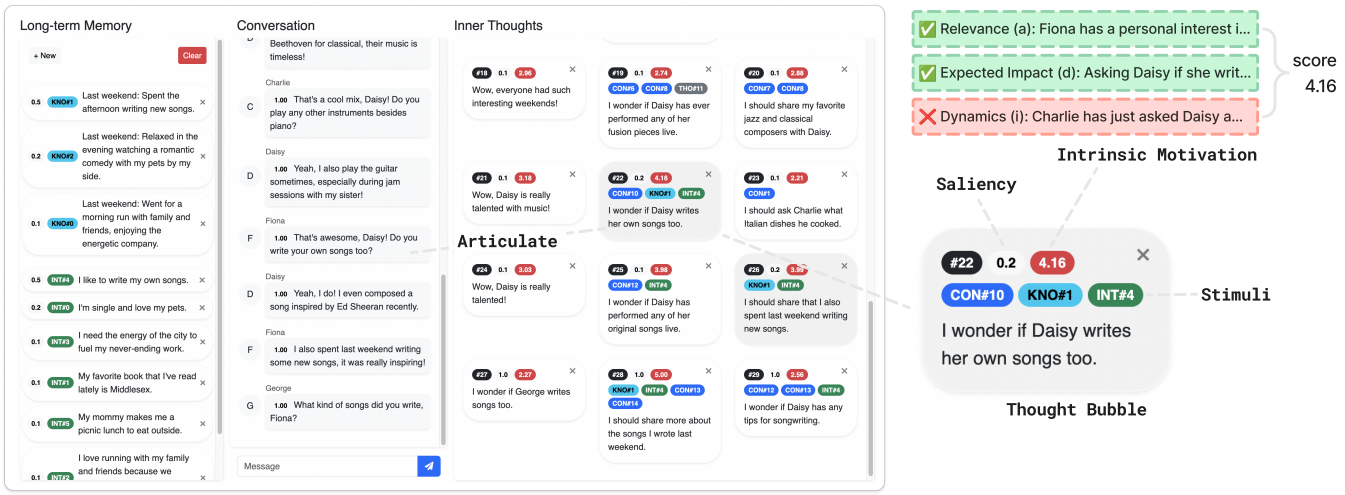

Figure 6: The Inner Thoughts playground web app interface. Multiple AI and humans can be added to simulate a group conversation. Users can also view and edit each of the participants' long-term memory and thoughts.

The main interface is divided into three panes: On the left is the long-term memory pane. Users can customize each AI participants's long-term memory by adding or deleting specific entries. In the middle is the conversation pane, where users can watch the simulated conversation, or participate in the conversation by sending a message using the dialog box on the bottom.

On the right is the inner thoughts pane. Users can view visualization of thought bubbles generated on-the-fly for the selected participant as the conversation proceeds. Each thought bubble contains a numeric ID (badge colored black), the saliency score (white), intrinsic motivation score (red), and a list of stimuli that AI used to formulate this thought. Below the badges shows a description of the thought. Thoughts that are expressed by the AI will be highlighted. In addition, users can click on a thought bubble to manually force the AI the express the thought in the conversation; and can right click on the thought bubble to view its reasoning for the intrinsic motivation rating. Users can also delete certain thoughts from the reservoir by clicking on the top-right delete button.

As a user clicks on a participant in the settings page, the content in the long-term memory and inner thoughts pane will be synced to that participant. They are automatically updated as the conversation goes on. Users can view a train of thoughts of each participant developing in parallel to the conversation.

6.2 Conditions

We compared two multi-party engagement strategies (Table 2) (1) Next-Speaker Prediction: In this condition, AI participants engaged based on predictions of who the next speaker would be. Their responses were generated based on predefined personas, following experimental setup in [71]. (2) Intrinsic Motivation (our approach): AI participants engaged based on their intrinsic motivation to contribute, driven by generated thoughts during the conversation.

We used the fine-tuned GPT-3.5 model we evaluated in section 3 in condition 1 to predict the next speaker, and prompt the model to generate responses based on its persona if selected by the prediction (full prompt in Supplementary Material). We used the framework described in section 5 for condition 2. Simulations were run on the Inner Thoughts playground web app.

6.3 Agent Personas and Conversation Generation

In this paper, we choose casual conversation scenarios as the primary focus of our evaluation due to their unique challenges in handling turn-takings. Casual conversations, unlike task-oriented interactions, lack clear objectives, making turn-taking and proactive engagement particularly difficult to model and has been underexplored in prior research [23, 36]. Future research could also explore how this framework applies to task-oriented conversations such as brainstorming, to further validate its adaptability.

To simulate the conversation, we first created eight AI participants, each assigned a detailed persona consisting of objectives, knowledge, and interests. These personas were initially generated from a seed randomly selected from the PersonaChat [71] dataset. For example, one seed might include: "I like listening to all genres of music except country," "I would travel the world if I could," "I enjoy reading books," "I like spending time with friends and family," and "I'm not a fan of hot weather."

To further enrich these personas and encourage interaction between different AI participants, we randomly sampled two additional persona descriptions from other participants for each AI. This approach introduced overlapping interests, making the AI participants more likely to engage in relatable conversations and increasing the chances of contributions during discussions.

We simulated 100 text-based conversations (50 for each condition), each involving four randomly selected AI participants. Each conversation consisted of 15 turns. We also incorporated 10 icebreaker prompts, randomly selected from the PersonaChat dataset. Examples of these prompts include: "What did you do last weekend?", "What is your favorite thing to do?", and "Hey!". A randomly selected participant was chosen to initiate the conversation for

each simulation, with a randomly selected icebreaker sentence. For all AI participants, the following proactivity settings were applied: Overt proactivity = 3.95, Covert proactivity = 0.1 and Tonal proactivity = False. One system 1 thought and two system 2 thoughts are generated in each batch.

6.4 Hypothesis

With the Inner Thoughts approach, the AI participant with the highest intrinsic motivation is more likely to take the floor of the conversation. Such participants tend to have more to contribute and are better able to develop the topic at hand. As discussed in subsection 5.6, intrinsic motivation fosters more meaningful and contextually appropriate contributions, in contrast to reactive engagement that simply predicts the next speaker. We anticipate that this effect will aggregate and lead to conversations that are more engaging, coherent and closely resemble the natural flow of human conversations.

Algorithm 1: Proactive AI via Inner Thoughts: iteratively processes trigger events by predicting turn-taking types, evaluating thoughts in a reservoir, and deciding whether to participate or remain silent.

Input: Stream of trigger events E, Thought reservoir T, Turn-taking type prediction.
Output: Participation action t^* for each trigger event.

while *there is a new trigger event* $e \in E$ **do**
  `// Step 1: Turn-Taking Type Prediction`
  Predict the turn-taking type for the current event e: Open to anyone, or allocated to a party;
  `// Step 2: Process According to Turn-Taking Type`
  if *open to anyone* **then**
    if $\exists t \in T$ *such that* $score(t) \geq imThreshold$ **then**
      Select the highest-rated thought: $t^* = \arg\max_{t \in T} score(t)$;
    else if $\nexists t \in T$ *such that* $score(t) \geq imThreshold$ **then**
      Select t^* from the system-1 thoughts in T with probability $system1Prob$;
  if *turn allocated to AI* **then**
    Select the highest-rated thought: $t^* = \arg\max_{t \in T} score(t)$;
  if *turn allocated to others* **then**
    if $\exists t \in T$ *such that* $score(t) \geq interruptThreshold$ **then**
      Select the highest-rated thought: $t^* = \arg\max_{t \in T} score(t)$;
    else if $\nexists t \in T$ *such that* $score(t) \geq interruptThreshold$ **then**
      The AI remains silent;
  `// Step 3: Finalize Participation`
  if t^* *is not null* **then**
    Participate with t^*;
  else
    Take no action;

Condition	When to Participate	What to Say
1	Next speaker prediction	Based on persona
2 (ours)	Intrinsic Motivation	Based on thoughts

Table 2: Study conditions for technical evaluation and user study. We compare Inner Thoughts with the baseline approach of deciding when to participate by next speaker prediction, and then generate response based on AI's persona

6.5 Human Evaluation of Simulated Conversations

We evaluated 100 simulated conversations with 10 human evaluators (4 female, 6 male, age Avg.: 26.3, SD: 4.35). Each evaluator reviewed five pairs of conversations, with one from each condition (10 total), viewed in a randomized order within each pair. Participants were informed that all conversations were AI-generated. Instead of displaying static conversation histories, we presented an animated version of the conversations to simulate a live chat experience. The length and speed of the conversations were rendered to match human's average typing speed. This decision was based on findings from our pilot studies, where we observed that participants tended to skim through static conversation transcripts.

After watching each conversation, they were asked to rate their agreements with seven statements (Table 3) related to the conversation's quality on a 1-7 Likert scale, from strongly disagree to strongly agree, adapted from [4, 10, 62]. We also asked participants to identify specific points where turn-takings felt unnatural. After completing all conversations, participants were asked to select the conversation that feels more natural and human-like in each pair and provide a brief explanation with examples.

6.6 Findings

We conducted Mann-Whitney U tests to compare the Baseline (condition 1) and Inner Thoughts (condition 2) strategies across the 7 dimensions listed in Table 3. Overall, the results showed significant improvements in the Inner Thoughts condition across all metrics (Figure 7).

Notably, the strongest effects were observed in turn appropriateness (U = 577.0, $p = 2.4 \times 10^{-6}$) and coherence (U = 636.0, $p = 1.6 \times 10^{-5}$), indicating that AI participants using intrinsic motivation contributed more appropriately and maintained better conversational flow compared to the Baseline. Anthropomorphism (U = 726.5, $p = 2.4 \times 10^{-4}$) and intelligence (U = 688.5, $p = 7.3 \times 10^{-5}$) were also significantly higher for Inner Thoughts, reflecting that AI participants were perceived as more human-like and thoughtful.

The Inner Thoughts condition also led to significantly higher engagement (U = 813.5, $p = 1.9 \times 10^{-3}$) and initiative (U = 862.5,

Metric	Statement
Anthropomorphism	I felt the conversation is natural and human-like.
Conversation Coherence	I felt that the dialogue maintains a coherent topic progression.
Perceived Engagement	I could feel that the AI participants are engaging well in the conversation.
Perceived Intelligence	I felt that the AI participants provided intelligent and insightful contributions to the conversation.
Turn Appropriateness	Turn-takings in the conversation is contextually and logically appropriate
Initiative	I felt the AI participants are able to take the initiative in conversations.
Adaptability	I felt that the AI participants appeared to adapt well to the changing dynamics of the conversation.

Table 3: Metrics used in our technical evaluation to measure the quality of AI simulated conversations. Each statement is rated on a Likert-scale from 1 – Strongly Disagree to 7 – Strongly Agree.

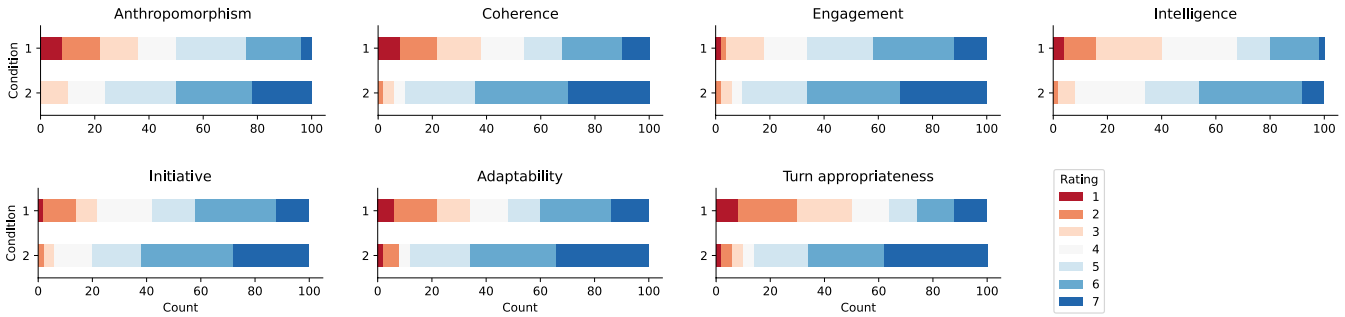

Figure 7: Stacked bar plots of participants' ratings on the metrics used in our technical evaluation to measure the quality of AI simulated conversations. Each statement is rated on a Likert-scale from 1 – Strongly Disagree to 7 – Strongly Agree.

$p = 6.0 \times 10^{-3}$), suggesting that AI participants were more proactive and engaging in conversation. Adaptability was rated significantly better in the Inner Thoughts condition as well (U = 765.0, $p = 6.3 \times 10^{-3}$), showing that AI participants adapted more effectively to changes in the conversation. Finally, 82% of the times participants preferred the Inner Thoughts conversations, indicating a clear overall preference for this approach in multi-party settings.

To complement these quantitative findings, we analyzed participants' feedback to understand the nuances behind their ratings. We used a thematic analysis approach [12] to analyze qualitative data from participant comments. Two researchers first independently identified recurring themes and patterns in the data, followed by two 60-minute meetings to refine and organize these themes. Disagreements were resolved through discussion.

Enhanced Coherence and Engagement. Participants noted that conversations in Condition 2 (Inner Thoughts) felt more coherent and engaging. They appreciated how AI agents built upon each other's responses, creating a more dynamic and interactive dialogue.

> *"In the second conversation, every participant adjusted their answers based on others' responses. It started general and then narrowed down, making it more engaging."* (P01)

Natural Turn-Taking. Condition 2 was praised for its natural turn-taking, with AI agents contributing at appropriate moments and responding directly to others.

> *"It felt like a real group chat where participants are listening to each other and interested in each other's topics. Their interaction is closer."* (P06)

> *"There was a flow in the conversation—not mechanical. If someone mentioned something, others would continue on that topic, echoing and adding more information."* (P08)

Responsiveness to Context. Participants observed that AI agents in Condition 2 were more responsive to the conversational context, leading to more meaningful interactions.

> *"... they can combine with their own experiences, making the conversation feel more connected."* (P04)

> *"Conversations had natural transitions. You feel like they are responding first and then sharing something about themselves."* (P08)

Limitations of the Baseline Strategy. Conversely, the baseline condition was criticized for its mechanical responses and lack of coherence. Participants felt that AI agents often talked past each other without meaningful engagement.

> *"In the first conversation, everyone was talking over each other. They answered the same question with the same format, which felt unnatural."* (P07)
>
> *"Everyone says 'hey' but there's no continuity. They don't respond to each other and just bring up unrelated topics."* (P09)

Missed Opportunities for Interaction. The baseline AI agents frequently failed to respond to prompts or engage with others' statements, leading to disjointed conversations.

> *"Someone mentioned walking their dog on the beach, and no one responded at all."* (P03)
>
> *"They ignored questions, and some people went back to very previous messages, making the conversation one-directional."* (P10)

7 User Evaluation

In addition to simulation experiments, we conducted a user study to understand: (1) How do people perceive proactive conversational AI enabled by the Inner Thoughts framework during actual interactions? (2) How do different levels of AI proactivity affect these perceptions?

7.1 Apparatus: *Swimmy* Slackbot

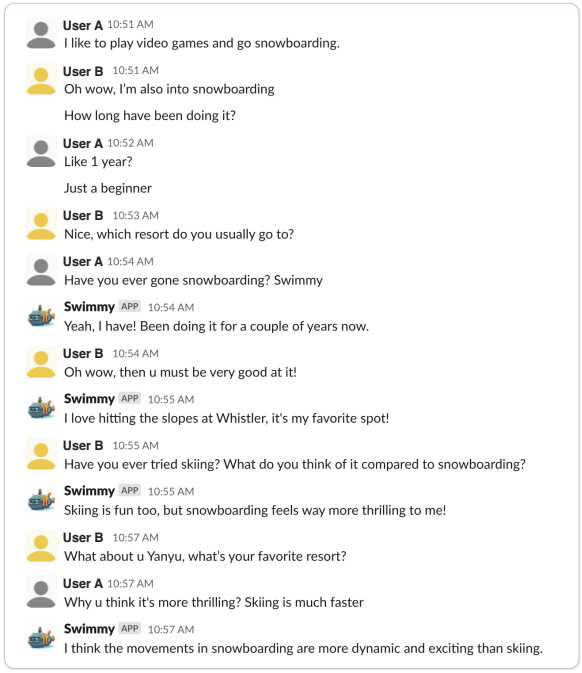

Figure 8: An example conversation between two human participants and *Swimmy*, a Slackbot developed based on the Inner Thoughts framework, on Slack.

We built a Slackbot (Figure 8) named *Swimmy* using Slack API[3] for the study. We implemented the Slackbot using a queue-based approach to handle asynchronous message processing. When a new message is received, it is added to a `triggerQueue`, and processes each message sequentially. The bot generates inner thoughts for each of the AI participants by updating saliency, forming thoughts, and evaluating them, as described in section 5. The AI then decides whether to respond based on turn-taking predictions, intrinsic motivation thresholds, as well as the status of the process queue. Specifically, if `triggerQueue` is not empty, the AI refrains from speaking to avoid interrupting ongoing message processing. Users can also customize configurations on the *Home* page of the Slackbot.

[3]https://api.slack.com/

7.2 Study Design

The study involved six pairs of human participants, each pair interacting with an AI agent on Slack (3-party conversation). We recruited 12 participants from our institution, with 8 of them reporting familiarity with conversational agents and 9 indicating familiarity with large language models (scoring above 4 on a 1-7 scale). Each participant was compensated $20 for their one-hour participation.

Participants experienced three 10-minute conversations, where we designed three AIs with different conversational styles:

(1) *Non-stop chatter*: This AI participated continuously, even when it had little input to add to the conversation. It had a high probability of selecting thoughts through System 1 processes (`system1Prob` = 0.7).
(2) *Active contributor*: This AI participated actively to share its thoughts whenever relevant but not overwhelming the conversation. With a moderate System 1 probability (`system1Prob` = 0.2), it occasionally selected less deliberate thoughts. Its low intrinsic motivation threshold (`imThreshold` = 3.59) allowed it to contribute actively whenever something comes up in its "mind".
(3) *Selective participant*: This AI only contributed when it had a strong interest in the topic, staying quiet during other parts of the conversation. With no reliance on System 1 processes (`system1Prob` = 0) and a higher motivation threshold (`imThreshold` = 4.09), this AI required significant justification before expressing a thought.

Specific proactivity settings of each participants are listed in Appendix 10.2. Each persona was randomly assigned to one of the three conversational conditions in each session, with order counterbalanced. Before the study, participants were informed that the AI might exhibit various personalities across the conversations but were not told in advance about the specific conditions.

We prompted participants to have a casual chat on three social topics: hobbies and interests, travel experiences, and weekend activities. After each conversation, participants rated their experience on ten metrics adapted from [4, 10, 62]. Compared to technical evaluation, we added metrics related to how people perceive their interaction with the chatbot, like *Likeability*, *Social Presence*, *Perceived Listener*, *Contribution*, etc. The full metrics table and explanations is shown in Appendix 10.2.

At the end of the study, participants were asked to match the chatbots to the behavior styles listed above, and engaged in a semi-structured interview about their experience. They were also asked to select their favorite AI out of three conditions.

7.3 Findings

7.3.1 Interacting with Proactive AI. Overall, participants had a positive perception of the proactive conversational AI enabled by the Inner Thoughts framework. Across all three conditions, the AI was rated to have median scores of 5 for anthropomorphism, initiative, engagement, listening ability, contribution, and extroversion. Notably, likeability and response timing also leaned toward positive ratings, with a median of 6, indicating that many participants found the AI likable and can engage at appropriate timings.

7.3.2 Perception of AI Proactivity. To evaluate whether the designed proactivity styles were both perceptible and distinguishable to users, we asked participants to identify which AI exhibited each conversational style after their interactions. The accuracy of their guesses varied, with the Non-stop Chatter being correctly identified 69.23% of the time, making it the most easily recognized. In contrast, the Selective Participant and Active Contributor were correctly identified 54.55% and 50% of the time, respectively. All AIs were identified above the baseline accuracy of 33.3%, indicating that participants were able to distinguish the AIs at a level higher than chance.

Participants also reported different perceptions of AI with different proactivity levels. For instance, for Selective Participant, participants noted that the AI was often too passive. It would only speak when prompted and failed to contribute actively, with P04 describing the AI as paradoxical: *"It responded enthusiastically when directly addressed, but was otherwise disengaged."* This lack of proactivity led some to feel the AI was overly focused on information retrieval, rather than maintaining a more human-like balance of social interaction and contribution P01.

In contrast, Condition 2 (Active Contributor) was appreciated for being more balanced. Participants praised its ability to engage naturally and at appropriate moments. P02 noted that it tended to initiate new topics when the conversation was drying up. P03 found that the AI was more proactive in encouraging others to share, even calling people by name to invite participation. However, there were still occasional lapses, as others observed that the AI sometimes failed to fully grasp the context of the dialogue, responding in ways that felt slightly off (P07).

For Condition 1 (Non-stop Chatter), while it was often described as overly talkative, participants differed on whether this was seen positively or negatively. P01 felt it mirrored conversations with a group of friends who are excited to chat, but noted that its excessive contributions disrupted the conversational flow, as it tended to speak over others or introduce irrelevant topics. P10, P12 found this persona too overwhelming, describing it as failing to respect the natural pauses of a conversation.

Interestingly, while participants could differentiate between these personas, their ratings for each were not significantly different overall. However, we did observe two statistically significant differences. The Non-stop Chatter received the highest ratings for perceived social presence (Median = 6), reflecting its continuous participation. In contrast, the Selective Participant was rated significantly lower than the Non-stop Chatter in both perceived social presence (Median = 5.5, $p < 0.05$) and extroversion (Median = 4.5, $p < 0.05$), which is consistent with its more reserved, less engaging behavior.

7.3.3 Preferences Over AI Proactivity. Clear preferences for AI proactivity emerged in the study. Condition 2 (Active Contributor) was the most favored, with 6 participants selecting it as the best. Participants appreciated its balanced approach, noting that it contributed actively without overwhelming the conversation. One participant P02 emphasized that this AI was more respectful of conversational flow, waiting for appropriate moments to introduce new topics, making interactions feel more natural.

Condition 1 (Non-stop Chatter) received mixed feedback, with 4 participants rating it as their favorite, but many also criticizing its excessive, often irrelevant contributions. P07 observed that the AI's tendency to introduce long and unnecessary questions disrupted the natural rhythm of human interaction, while another P02 mentioned that the AI reminded them of a language exchange partner who was overly eager to contribute but lacked an understanding of the social context.

Condition 3 (Selective Participant) was the least preferred. Only 2 participants selected it as the best, while 7 rated it as the worst. Participants generally felt that this AI was too passive, contributing little to the conversation unless directly asked. Some P06, P11 noted that it seemed uninterested or unmotivated to engage with the topic at hand, leading to more disjointed conversations.

8 Discussion and Future Work

8.1 Proactive Agents via Intrinsic Motivation vs. Extrinsic Cues

In this paper, we introduced the Inner Thoughts framework, which emphasizes the role of intrinsic motivation in enabling proactive conversational agents. Unlike traditional approaches that rely primarily on external cues such as turn-taking predictions, our framework explores how AI can leverage internally generated thoughts to determine participation in conversations. While we highlighted the inherent limitations of next-speaker prediction strategies and demonstrated the advantages of integrating intrinsic motivation, it is crucial to note that these methods are *not mutually exclusive*. Instead, they are complementary components that, when combined, could create more robust results. We advocate for the development of holistic systems that integrate internal processes (such as thought evaluation and intrinsic motivation) with external strategies (like multimodal cues and turn-allocation mechanisms [7–9, 15, 20, 31, 41, 45, 46, 63]). Future research should focus on understanding how these internal and external elements interact and how their synergy can enhance both the functionality and user experience of conversational agents.

8.2 Applying Inner Thoughts Beyond Casual Conversations

While our study primarily explored the Inner Thoughts framework in casual conversational settings, its potential extends beyond this domain. The framework's inherent adaptability allows it to be adapted into task-oriented scenarios such as brainstorming, coordination, and negotiation. By customizing the criteria used for thought evaluation and aligning them with the goals of a given scenario, the Inner Thoughts framework can enable goal-oriented

proactivity. For instance, in a negotiation setting, the AI could evaluate its intrinsic motivation to contribute based on criteria such as the strategic value of its input or its alignment with pre-defined negotiation strategies. Similarly, in coordination tasks, the framework could prioritize thoughts that promote alignment among team members or clarify ambiguities. Future work could explore refining the framework's adaptability by integrating domain-specific heuristics and dynamically learning from user feedback, further enhancing its applicability to real-world task-oriented interactions.

8.3 Proactive Conversational Agents Beyond Text and Computational Efficiency

Another exciting avenue for the Inner Thoughts framework lies in expanding its implementation beyond text-based interactions. Extending the framework to support multimodal communication like audio and face-to-face interactions introduces both opportunities and challenges. Real-time, multimodal systems must contend with lower latency requirements and more complex turn-taking mechanisms that incorporate additional cues such as intonation, gestures, and facial expressions. To achieve lower latency in these multimodal systems, simplifying the thought evaluation process by focusing on core metrics like relevance and coherence could be effective. Alternatively, leveraging advanced techniques like training lightweight LoRA (Low-Rank Adaptation) [30] models might strike a balance between computational efficiency and performance, enabling the framework to operate in real-time. Future work should evaluate these approaches to identify optimal strategies for scaling Inner Thoughts to multimodal and low-latency environments, unlocking its full potential for human-like interaction.

8.4 Technical Limitations of the Inner Thoughts Framework.

A key limitation of the Inner Thoughts framework is thought formation. The AI sometimes generates irrelevant or contradictory thoughts, which we tried to address by using stricter prompts. However, this could make the thoughts generated repetitive. Future work should explore more advanced methods, such as incorporating knowledge graphs to improve the thought generation process.

Another issue is setting proactivity thresholds. Currently, these thresholds are adjusted through trial and error, leading to inconsistent interactions. Future work could explore a data-driven approach, for example using reinforcement learning to dynamically learn thresholds based on user feedback.

The thought evaluation process also needs refinement. While effective in most cases, the AI occasionally misses opportunities to engage due to underestimating motivation scores, or interrupts too abruptly when motivated to speak. Future iterations of the framework could benefit from more robust evaluation mechanisms that better balance engagement opportunities with conversational appropriateness, potentially integrating adaptive learning techniques to fine-tune these processes over time.

8.5 Automatic Evaluation for AI Proactivity

Evaluating the quality of AI engagement in multi-party, non-task-oriented conversations presents unique challenges, primarily due to the absence of clear objectives or predefined outcomes. This inherent ambiguity complicates the development of standardized success metrics, making traditional evaluation methods less effective. HCI research has relied heavily on user studies to assess conversational AI performance [7–9]. While these studies offer valuable qualitative insights into user experiences, they face limitations in scalability and often lack the precision required for setting granular, reproducible benchmarks. In our work, we tried to address these challenges by conducting human evaluations of simulated multi-party conversations. However, this approach is still relatively resource-intensive and challenging to replicate at scale. This underscores the need for future research to establish robust, cost-effective benchmarks and develop automatic metrics in evaluating AI proactivity. Recent work in NLP, such as leveraging LLMs for self-assessment [38], could inspire comprehensive evaluation frameworks. By establishing such metrics, we can better quantify AI performance, reduce reliance on human evaluations, and enable the systematic development of proactive conversational AI systems.

8.6 Other Applications of Inner Thoughts

The Inner Thoughts framework holds exciting potential for a variety of applications. One intriguing use case is in brainstorming sessions, where AI could ambiently generate and suggest ideas, mirroring the thought processes of human participants and offering spontaneous contributions. Additionally we could allow for further customization – such as tuning the AI's thoughts to be more creative, whimsical, or even deliberately childlike. This ability to modify the AI's internal reasoning opens possibilities for specialized applications.

Moreover, the framework introduces opportunities for simulating complex AI behaviors, such as ethical dilemmas or conflicting thoughts. For instance, an AI could be designed to decide whether to lie or reveal a difficult truth, simulating moral decision-making processes. In this way, the AI could offer nuanced interactions that reflect more sophisticated social and ethical considerations, providing a deeper simulation of human-like cognitive behavior.

8.7 Interacting with the Thoughts of LLMs

With the Inner Thoughts framework and recent LLMs like Chain-of-Thought (CoT) prompting [64] and OpenAI o1 [43] that leverage internal reasoning processes, a new design question emerges: how could humans interact with the inner thoughts of LLMs? The concept of interaction paradigms for LLMs' thoughts opens up a number of possibilities and challenges. Instead of simply providing outputs, LLMs could surface their intermediate reasoning, enabling users to gain insights into the model's decision-making process. For instance, should these "thoughts" be visible in real time, offering users a glimpse into the system's reasoning trajectory? Further, how might we design interfaces that allow users to question, refine, or even contribute to the AI's inner thought process? These considerations not only affect usability but also trust, as understanding the AI's rationale could make its behavior more transparent and predictable. Addressing these questions will be crucial in defining the next generation of human-AI collaboration paradigms.

9 Conclusion

In this paper, we presented the Inner Thoughts framework, a novel approach to proactive AI in multi-party conversations. Unlike traditional systems that rely on turn-taking predictions, our framework enables AI to generate and evaluate its own internal thoughts continuously, deciding when and how to engage based on intrinsic motivation. Our evaluations demonstrated that AI guided by Inner Thoughts offers more natural, engaging, human-like turn-taking behaviors compared to the next-speaker prediction baseline. Our implementation, showcased in a web app and a chatbot, highlights the potential of this framework for future applications in multi-party conversational systems. Our work contributes a novel perspective on proactive AI in conversational settings, highlighting the importance of internal thought processes and intrinsic motivation.

References

[1] James E Allen, Curry I Guinn, and Eric Horvtz. 1999. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications* 14, 5 (1999), 14–23.

[2] Salvatore Andolina, Valeria Orso, Hendrik Schneider, Khalil Klouche, Tuukka Ruotsalo, Luciano Gamberini, and Giulio Jacucci. 2018. Investigating Proactive Search Support in Conversations. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. ACM, 1295–1307. https://doi.org/10.1145/3196709.3196734

[3] Salvatore Andolina, Valeria Orso, Hendrik Schneider, Khalil Klouche, Tuukka Ruotsalo, Luciano Gamberini, and Giulio Jacucci. 2018. SearchBot: Supporting Voice Conversations With Proactive Search. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Jersey City, NJ, USA) *(CSCW '18)*. ACM, 9–12. https://doi.org/10.1145/3272973.3272990

[4] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics* 1 (2009), 71–81.

[5] Daniel E. Berlyne. 1960. *Conflict, Arousal, and Curiosity*. McGraw-Hill.

[6] Keping Bi, Qingyao Ai, and W Bruce Croft. 2021. Asking clarifying questions based on negative feedback in conversational search. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. 157–166.

[7] Dan Bohus and Eric Horvitz. 2009. Learning to predict engagement with a spoken dialog system in open-world settings. In *Proceedings of the SIGDIAL 2009 Conference*. 244–252.

[8] Dan Bohus and Eric Horvitz. 2009. Models for multiparty engagement in open-world dialog. In *Proceedings of the SIGDIAL 2009 conference, the 10th annual meeting of the special interest group on discourse and dialogue*. 10.

[9] Dan Bohus and Eric Horvitz. 2011. Multiparty turn taking in situated dialog: Study, lessons, and directions. In *Proceedings of the SIGDIAL 2011 Conference*. 98–109.

[10] Simone Borsci, Alessio Malizia, Martin Schmettow, Frank Van Der Velde, Gunay Tariverdiyeva, Divyaa Balaji, and Alan Chamberlain. 2022. The chatbot usability scale: the design and pilot of a usability scale for interaction with AI-based conversational agents. *Personal and ubiquitous computing* 26 (2022), 95–119.

[11] Paul T Brady. 1968. A statistical analysis of on-off patterns in 16 conversations. *Bell System Technical Journal* 47, 1 (1968), 73–91.

[12] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[13] Penelope Brown. 1987. *Politeness: Some universals in language usage*. Vol. 4. Cambridge university press.

[14] Shuo-yiin Chang, Bo Li, Tara N Sainath, Chao Zhang, Trevor Strohman, Qiao Liang, and Yanzhang He. 2022. Turn-taking prediction for natural conversational speech. *arXiv preprint arXiv:2208.13321* (2022).

[15] Maira Gatti de Bayser, Paulo Cavalin, Claudio Pinhanez, and Bianca Zadrozny. 2019. Learning multi-party turn-taking models from dialogue logs. *arXiv preprint arXiv:1907.02090* (2019).

[16] Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. 2023. A survey on proactive dialogue systems: Problems, methods, and prospects. *arXiv preprint arXiv:2305.02750* (2023).

[17] Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology* 23, 2 (1972), 283.

[18] Starkey Duncan Jr and George Niederehe. 1974. On signalling that it's your turn to speak. *Journal of experimental social psychology* 10, 3 (1974), 234–247.

[19] Paul Ekman and Wallace V Friesen. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *semiotica* 1, 1 (1969), 49–98.

[20] Erik Ekstedt and Gabriel Skantze. 2020. Turngpt: a transformer-based language model for predicting turn-taking in spoken dialog. *arXiv preprint arXiv:2010.10874* (2020).

[21] Jonathan St BT Evans. 2008. Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.* 59, 1 (2008), 255–278.

[22] Cecilia E Ford and Sandra A Thompson. 1996. Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. *Studies in interactional sociolinguistics* 13 (1996), 134–184.

[23] Jianfeng Gao, Michel Galley, and Lihong Li. 2019. *Neural approaches to conversational AI: Question answering, task-oriented dialogues and social chatbots*. Now Foundations and Trends.

[24] Erving Goffman. 1967. *Interaction Ritual: Essays on Face-to-Face Behavior*. Doubleday.

[25] H. P. Grice. 1975. Logic and conversation. In *Syntax and semantics*. Vol. 3. Academic Press, 41–58.

[26] Zhiwei Guan, Shirley Lee, Elisabeth Cuddihy, and Judith Ramey. 2006. The validity of the stimulated retrospective think-aloud method as measured by eye tracking. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 1253–1262.

[27] Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

[28] Karen Holtzblatt and Hugh Beyer. 1997. *Contextual design: defining customer-centered systems*. Elsevier.

[29] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 159–166. https://doi.org/10.1145/302979.303030

[30] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[31] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. 2014. Analysis of respiration for prediction of" who will be next speaker and when?" in multi-party meetings. In *Proceedings of the 16th international conference on multimodal interaction*. 18–25.

[32] Donghyun Kim, Youbin Ahn, Wongyu Kim, Chanhee Lee, Kyungchan Lee, Kyong-Ho Lee, Jeonguk Kim, Donghoon Shin, and Yeonsoo Lee. 2023. Persona expansion with commonsense knowledge for diverse and consistent response generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 1139–1149.

[33] Rachna Konigari, Saurabh Ramola, Vijay Vardhan Alluri, and Manish Shrivastava. 2021. Topic shift detection for mixed initiative response. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 161–166.

[34] Kenichi Kumatani, Sankaran Panchapagesan, Minhua Wu, Minjae Kim, Nikko Strom, Gautam Tiwari, and Arindam Mandai. 2017. Direct modeling of raw audio with dnns for wake word detection. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 252–257.

[35] John E Laird. 2019. *The Soar cognitive architecture*. MIT press.

[36] Lizi Liao, Grace Hui Yang, and Chirag Shah. 2023. Proactive conversational agents in the post-chatgpt world. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3452–3455.

[37] Xingyu "Bruce" Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Alex Olwal, Peggy Chi, Xiang "Anthony" Chen, and Ruofei Du. 2023. Visual Captions: Augmenting Verbal Communication with On-the-fly Visuals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 108, 20 pages. https://doi.org/10.1145/3544548.3581566

[38] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634* (2023).

[39] Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. *arXiv preprint arXiv:2005.03954* (2020).

[40] John K Local, John Kelly, and William HG Wells. 1986. Towards a phonology of conversation: turn-taking in Tyneside English1. *Journal of Linguistics* 22, 2 (1986), 411–437.

[41] David H McFarland. 2001. Respiratory markers of conversational interaction. (2001).

[42] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. 2021. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International conference on machine learning*. PMLR, 7721–7735.

[43] OpenAI. 2024. Learning to Reason with LLMs. (September 2024). https://openai.com/index/learning-to-reason-with-llms/

[44] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user*

interface software and technology. 1–22.
[45] Christopher Peters. 2005. Direction of attention perception for conversation initiation in virtual environments. In *International Workshop on Intelligent Virtual Agents*. Springer, 215–228.
[46] Christopher Peters, Catherine Pelachaud, Elisabetta Bevacqua, Maurizio Mancini, and Isabella Poggi. 2005. A model of attention and interest using gaze behavior. In *International Workshop on Intelligent Virtual Agents*. Springer, 229–240.
[47] Xuhui Ren, Hongzhi Yin, Tong Chen, Hao Wang, Zi Huang, and Kai Zheng. 2021. Learning to ask appropriate questions in conversational recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 808–817.
[48] Bradley Rhodes and Thad Starner. 1996. Remembrance Agent: A Continuously Running Automated Information Retrieval System. In *The Proceedings of the First International Conference on the Practical Application of Intelligent Agents and Multi Agent Technology*, Vol. 1. ACM, 487–495.
[49] Frank E Ritter, Farnaz Tehranchi, and Jacob D Oury. 2019. ACT-R: A cognitive architecture for modeling cognition. *Wiley Interdisciplinary Reviews: Cognitive Science* 10, 3 (2019), e1488.
[50] Sonia Roccas, Lilach Sagiv, Shalom H Schwartz, and Ariel Knafo. 2002. The big five personality factors and personal values. *Personality and social psychology bulletin* 28, 6 (2002), 789–801.
[51] Jurgen Ruesch, Gregory Bateson, Eve C Pinsker, and Gene Combs. 2017. *Communication: The social matrix of psychiatry*. Routledge.
[52] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *language* 50, 4 (1974), 696–735.
[53] Samira Shaikh, Tomek Strzalkowski, George Aaron Broadwell, Jennifer Stromer-Galley, Sarah M Taylor, and Nick Webb. 2010. MPC: A Multi-Party Chat Corpus for Modeling Social Phenomena in Discourse.. In *LREC*. Citeseer.
[54] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2024).
[55] Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language* 67 (2021), 101178.
[56] Lucille Alice Suchman. 1987. *Plans and situated actions: The problem of human-machine communication*. Cambridge university press.
[57] Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric P Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. *arXiv preprint arXiv:1905.11553* (2019).
[58] Louis Ten Bosch, Nelleke Oostdijk, and Lou Boves. 2005. On temporal aspects of turn taking in conversational dialogues. *Speech Communication* 47, 1-2 (2005), 80–86.
[59] Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057* (2000).
[60] David Traum, Antonio Roque, Anton Leuski, Panayiotis Georgiou, Jillian Gerten, Bilyana Martinovski, Shrikanth Narayanan, Susan Robinson, and Ashish Vaswani. 2007. Hassan: A virtual human for tactical questioning. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*. 71–74.
[61] Marilyn Walker and Steve Whittaker. 1995. Mixed initiative in dialogue: An investigation into discourse segmentation. *arXiv preprint cmp-lg/9504007* (1995).
[62] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.
[63] Jimmy Wei, Kurt Shuster, Arthur Szlam, Jason Weston, Jack Urbanek, and Mojtaba Komeili. 2023. Multi-party chat: Conversational agents in group settings with humans and models. *arXiv preprint arXiv:2304.13835* (2023).
[64] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
[65] Joseph Weizenbaum. 1966. ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine. *Commun. ACM* 9, 1 (jan 1966), 36–45. https://doi.org/10.1145/365153.365168
[66] Allison Woodruff and Paul M Aoki. 2003. How push-to-talk makes talk less pushy. In *Proceedings of the 2003 ACM International Conference on Supporting Group Work*. 170–179.
[67] Jun Xu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Conversational graph grounded policy learning for open-domain conversation generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 1835–1845.
[68] Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. 2023. Realpersonachat: A realistic persona chat corpus with interlocutors' own personalities. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*. 852–861.
[69] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems* 36 (2024).
[70] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629* (2022).
[71] Saizheng Zhang. 2018. Personalizing dialogue agents: I have a dog, do you have pets too. *arXiv preprint arXiv:1801.07243* (2018).

10 Appendix

10.1 System

Figure 9 shows the Inner Thoughts playground settings panel.

10.2 User Evaluation

10.2.1 Proactivity Settings.

(1) *Non-stop chatter*: This AI engaged continuously in the conversation, even when it had little relevant input to offer.
- `system1Prob` = 0.7
- `imThreshold` = 4.49
- `interruptThreshold` = 4.8
- `proactiveTone` = false

(2) *Active contributor*: This AI participated actively, contributing when appropriate but without dominating the conversation.
- More proactive AI
- `system1Prob` = 0.2
- `imThreshold` = 3.59
- `interruptThreshold` = 4.8
- `proactiveTone` = true

(3) *Selective participant*: This AI contributed only when highly interested in the topic, remaining silent during other parts of the conversation.
- Less proactive AI
- `system1Prob` = 0
- `imThreshold` = 4.09
- `interruptThreshold` = 5
- `proactiveTone` = false

10.2.2 Metrics. Table 4 shows the metrics and their definitions that we used in our user study.

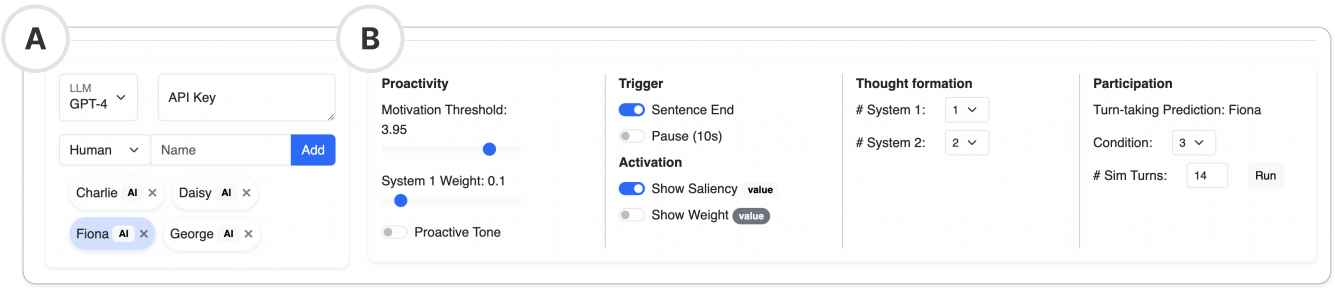

Figure 9: Settings page for Inner Thoughts Playground web app. Users can change the LLM version, add/delete conversation participants, adjust proactivity level, trigger method, thought formation quantity and participation strategies.

Metric	Statement
Anthropomorphism	I felt the chatbot is humanlike
Likeability	I felt pleasant to chat with the chatbot
Initiative	I felt the chatbot is able to take the initiative in conversations
Perceived social presence	I was often aware of the chatbot in our conversation
Perceived engagement	I could feel that chatbot is engaging well in the conversation
Perceived listener	I felt chatbot was actively listening and I was heard
Contribution	The chatbot made valuable contributions that enhanced the overall quality of the conversation
Appropriateness of the response timing	I felt that the chatbot can join in conversation at appropriate moments
Future usage	I'd like to have this chatbot when have similar conversation in the future
Extroversion	I felt like the chatbot has an extroverted personality

Table 4: Metrics used in our user study to measure the quality of AI simulated conversations. Each statement is rated on a Likert-scale from 1 – Strongly Disagree to 7 – Strongly Agree.