

YES AND: A Generative AI Multi-Agent Framework for Enhancing Diversity of Thought in Individual Ideation for Problem-Solving Through Confidence-Based Agent Turn-Taking

Pratik Ghosh
pratik.ghosh@microsoft.com
Microsoft Research, Cambridge
Cambridge, UK

Sean Rintel
serintel@microsoft.com
Microsoft Research, Cambridge
Cambridge, UK

Abstract

Diversity of thought is crucial in ideation for problem-solving, yet professionals in organisational settings often face challenges such as limited access to varied expertise and resource constraints which hinder the ideation process. To address this issue, we propose *YES AND*, a Generative AI based multi-agent framework that simulates diverse perspectives through AI agents for ideation with a single user. Leveraging a unique confidence-based turn-taking model, these agents organically take turns as they build on ideas, pose clarification questions to the user for improved contextual understanding, and allow the user to interject and steer the conversation. Beyond addressing the limitations of traditional ideation, this framework offers a novel approach to leveraging Generative AI for ideation, moving away from the rigidity of pre-defined interaction rules towards a more dynamic and creative process that enables serendipitous development of ideas.

CCS Concepts

• **Human-centred computing** → **Interactive systems and tools.**

Keywords

individual problem solving, diversity of thought, Generative AI, agents, expertise, role-based personas, conversational turn-taking, design improv

ACM Reference Format:

Pratik Ghosh and Sean Rintel. 2025. YES AND: A Generative AI Multi-Agent Framework for Enhancing Diversity of Thought in Individual Ideation for Problem-Solving Through Confidence-Based Agent Turn-Taking. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*, April 26-May 1, 2025, Yokohama, Japan. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/3706599.3720142>

1 INTRODUCTION

One of the persistent challenges with problem-solving in organisations is that while diversity of thought in ideation leads to more creative outcomes [13, 21, 48], individuals often lack access to others who could provide sufficient diversity when they need it. Reasons for lacking diversity might be practical, such as the organisation

being too small (overall or subgroup within which the individual works), or permanently or temporarily understaffed, lack of connections to find the right set of ideators, or even too busy to collaboratively ideate every time that it might be useful. More complex reasons involve lack of the right kind of diversity for the problem – the differences that will actually make a difference [31].

Generative AI (AI henceforth) has the potential to revolutionise how we approach access to diversity of thought for ideation in organisational problem-solving. Sophisticated examples include Solo Performance Prompting (SPP) [54], which increases diversity of expertise within the LLM's own answering process to produce better results than generic singular processes, and PersonaFlow [5], which allows users to engage with a range of LLM experts and has been found to enhance perceived quality of outcomes without increasing cognitive load. However, such systems are very complex and focused on the system providing significant detail in their answers. In doing so, however, they may contribute to cognitive offloading of the ideation from the user to AI based personas – (referred to as 'agents' in this paper for clarity), which has the potential to diminish user's own skills and creativity over time [46].

Inspired by "design improv" [19], which promotes acceptance of ideas and building upon them, we designed a simple multi-agent framework – YES AND – to enrich the individual user's ideation by allowing them to participate in a conversation with several persona-based agents, each with a specific expertise. The agents engage in a group conversation, express confidence [60] in their ability to take a turn, and thus have autonomy over their participation as they organically build on each other's ideas. The user and agents engage with one another at any time, similar to a free-flowing multi-party conversation, but avoid problems of traditional ideation such as social loafing [28], production blocking [12], and groupthink [24]. At the end of a session, another agent distils the results not as a complete solution, per se, but rather a seed of a holistic approach to the solution. It is then for the user to develop the actual solution. An individual user might do this with any number of personas, for any number of problems, reducing the lack of immediate and constant access to diverse others, while also preparing them for leveraging the value of diversity. Thus, both the problem-solving ideation sessions themselves and the ongoing use of the system can promote organisational efficiency and skill-building in individuals. Through this work, we aim to contribute to the ongoing exploration of multi-agent collaboration for problem-solving that is both simple to develop and also designed with the explicit intention to ensure that people's skills and creativity are maintained even as

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CHI EA '25, Yokohama, Japan

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1395-8/2025/04

<https://doi.org/10.1145/3706599.3720142>

their capabilities are augmented through AI. In the following sections, we discuss related work, describe the system architecture and its evolution, the mechanism underlying the turn-taking system and other guidance needed to develop the AI personas and their confidence in making contributions, and conclude by outlining the potential for future development and domains for this approach.

2 RELATED WORK

Borrowing from literature on innovation and organisational behaviour, this work defines diversity of thought as the variation in perspectives and opinions arising from differences in backgrounds and expertise. Research on thought diversity for innovation [8] highlights the value of inclusion of individuals with different perspectives, backgrounds, and areas of expertise within a team. Another study [22] on effective innovation suggests that groups with greater diversity can outperform more homogeneous groups of higher-ability individuals indicating that diversity of thought enhances collective problem-solving capabilities, leading to more innovative and effective solutions. By promoting a variety of viewpoints, diversity of thought enables teams to approach problems uniquely and comprehensively, ensuring that the final solutions are well-rounded and robust [11]. In this section, we review key literature relevant to our work. We begin by examining traditional problem-solving practices in organisational settings to establish methodologies and limitations. Next, we explore the use of AI in problem-solving and ideation, assessing recent advancements in the field. Finally, we investigate AI based multi-agent systems to gain insights into various existing frameworks and approaches.

2.1 Traditional ideation for problem-solving in organisational settings

In cognitive psychology, problems are categorised as well-defined or ill-defined [4, 37]. Well-defined problems lend themselves to structured frameworks due to clear goals and constraints, while ill-defined problems require navigating ambiguity and demand creative, collaborative approaches [43]. Organisational problem-solving often involves such ill-defined challenges, where ideation is crucial. Traditional ideation has been central to organisational success, offering structured frameworks and collaborative opportunities. However, its limitations include a lack of diverse thinkers, resource constraints, and affinity bias, which can stifle creativity and innovation [20, 65]. Issues such as social loafing [28], production blocking [12] and groupthink [24] further hinder effective problem-solving by reducing individual contributions and suppressing dissenting opinions. YES AND addresses these challenges through a conversation turn-taking approach, fostering inclusivity and mitigating groupthink by encouraging dissent and exploring alternative solutions. Through techniques such as prompt engineering, discussed in Section 3.1, we guide agents to remain open to diverse ideas, even if it alters the discussion's course.

2.2 AI for problem-solving and ideation

Recent studies show that AI based inspiration significantly enhances the novelty, variety, and quantity of ideas [25]. Models like GPT-4 [38], support creative tasks such as ideation and problem-solving by generating contextually relevant text. However, these

models often focus on one-on-one interactions, offering elaborate, definitive responses that limit iterative thinking and idea evolution [53, 59]. Challenges include lack of relevance and dialogue management, premature conclusions, and limited user control [23]. Traditional AI approaches treat models as passive idea generators rather than active, critically engaged participants [46]. Multi-agent conversation studies highlight issues like problem drift, alignment collapse, and monopolisation in discussions[5]. YES AND addresses these challenges by enabling AI agents to ask clarification questions and interact iteratively with users and each other. This approach also allows users to steer conversations, ensuring alignment with their goals. Shaer et al. [47], in their work on investigating the use of LLMs in group ideation in the context of education found that integrating LLMs in brain writing [56] enhances both the ideation process and its outcome. Work on using AI for collaborative diffusion [52], design ideation [49] and associative thinking [32] investigate how generative machine learning can play a role in creative ideation, prototyping and sketching. These studies provide valuable inspiration for YES AND.

2.3 AI based multi-agent systems

AI based multi-agent systems have been widely studied in recent years with applications ranging from holiday planning to code generation. Persona-based agents, which simulate distinct perspectives or expertise, have been explored in applications like customer service, education, creative writing, fitness, real estate, travel, and shopping as discussed in a comprehensive review by Chen et al. [7]. However, in our search for related literature, we found that their integration into collaborative ideation tools remains limited. The idea of using LLMs to emulate personalities has been discussed in recent discourse. For instance, Character.AI [6] and UXpressia [51] enable users to engage in one-on-one conversations with individually generated personas, while Dream Team [3], facilitates interactions with multiple personas that only collectively respond as one team. This presents an opportunity to delve into a relatively unexplored area of facilitating dynamic, real-time engagement between multiple agents and the user, paving the way for richer and more diverse interactions. Work on multi-agent frameworks often emphasise the need for better coordination, communication and decision-making amongst autonomous agents. [50]. Recent research has investigated the use of agents to enhance creativity and found that most existing multi-agent frameworks for ideation are either limited to specific domains or rely on pre-defined rules for agent interaction which can constrain creativity and serendipitous development of an idea. They highlight that while complex behaviours can emerge from such systems, relying solely on simple rules may limit the system's creative potential [15, 42]. Additionally, these systems often struggle to balance user directed goals with agent autonomy leading to outcomes that may lack coherence or relevance to the problem at hand. Work on agentic frameworks such as Autogen [58] have showcased the potential of enabling flexible and task-specific agent interaction frameworks. Their applications are often geared towards solving operational challenges across domains, employing dynamic conversation programming paradigms and integrating human inputs for broad functionality[10, 14, 33–35]. While these approaches excel in conversation and task execution, they are not

explicitly designed to address the unique challenges of encouraging creativity and diversity of thought in ideation sessions.

Our framework addresses these limitations by introducing a flexible dynamic multi-agent environment where agents autonomously generate and build on ideas while maintaining alignment with user objectives. Unlike general purpose multi-agent frameworks, our approach introduces a structured mechanism tailored specifically for creative problem-solving through ideation, thereby filling the gap in multi-agent applications where diversity of thought is key.

3 SYSTEM ARCHITECTURE

3.1 Prompt Engineering of agent personas

We first describe our methodology to generate agents and guide their behaviour.

To explore the feasibility of YES AND, we sought to generate persona-based agents selected to represent roles commonly involved in problem-solving in a technology development organization. Diversity, of course, extends beyond job roles. It could include personality traits, ability attributes, cultural or ethical frameworks, and a host of other subtleties. In this exploratory work, we focused only on professional role-based personas with a limited set of traits relevant to problem-solving (see Appendix B). This allowed us to see clear differences in our test systems, and to simplify our confidence mechanism (see Appendix 3.4).

There are numerous approaches to customising LLMs and generating persona-based agents. For instance, one could fine tune a model on specific tasks [41], adapt a model to a domain [1], or engineer prompts [29, 62] that evoke the desired agent-behaviour. Foundational work on generating agents has demonstrated that conditioning models on personas can enhance conversational engagement and coherence [63]. Recent studies have also explored how model behaviour can be shaped by prompt engineering in system messages [55]. System messages are feature-specific set of instructions or contextual frameworks given to a generative AI model to guide its behaviour[36].

For this work, prompt-engineering in system messages was the simplest method to generate persona-based agents, and influence their behaviour. For example, the prompt to generate a Designer agent contained qualities such as their background expertise and approach to problem-solving which determined their confidence in responding to a statement (see Appendix B.3).

Beyond the agents' personas, during testing we observed challenges such as overly lengthy and unclear responses (see Appendix B.1, where we do not specify a word limit). To resolve this, we refined prompts to include explicit constraints such as limiting response length (see Appendix B.2) or focusing on asking a clarification question if a message needed it. This was especially important for our 'Sage' agent that would summarise the ideation conversation, for whom background, personality, tone of voice, and content structure had to be balanced with a word limit to elicit concise and contextually relevant summaries (See Appendix B.4).

This ensured that the AI agents generated outputs that were relevant, diverse, and aligned with their personas. We tried several rounds of trial and error to arrive at system messages that worked best in terms of eliciting appropriate responses from agents that were relevant, diverse, and aligned with their personas. The process

underscores the critical role of prompt engineering in shaping interactive AI behaviour and offers practical insights for designing conversational systems.

3.2 V1: Linear development of ideas

To explore the feasibility of YES AND, we first implemented a setup involving four agents. Three represented roles in a technical team – a Designer, a Machine Learning Researcher, and an Engineer – and a fourth was a summarising agent, called the 'Sage'. Our example scenario was one in which the user is a Program Manager wanting to develop a new 'recall email' but lacks access to colleagues from design, engineering, and machine learning with which to ideate. The user begins the process by describing the problem, which is then passed to from agent to agent in a linear sequence of Designer, ML Researcher, Engineer, and Sage. The agents build on one another's ideas as a response to the user's problem. We have provided this exchange in Appendix C.1.

While this prototype demonstrated the feasibility of YES AND in principle, it also highlighted several challenges which we aimed to address. Notably, the conversation lacked serendipitous development of ideas. Research shows that sequential flow of information restricts the spontaneous development of ideas and limits user engagement. Work on fluid real-time conversational agents [26] discusses how traditional conversation agents lack fluidity of natural human interaction, suggesting incremental and bidirectional interaction between user input and output generation as ways of overcoming this. Another challenge was the lack of user agency. Research on grounded dialogue flow management for conversational agents [17], substantiates the need for capturing the user's intention to drive the conversation in specific directions. Additionally, the lack of scope for constructive criticism or for agents to ask clarifying questions can be detrimental to user engagement [27]. Recognising these as critical issues, we revised YES AND to enable organic flow and evolution of ideas, encourage clarification questions, and enhance user agency through the ideation process.

3.3 V2: Conversational development of ideas

To overcome the aforementioned challenges, we drew inspiration from the field of Conversation Analysis. We adopted the conversation turn-taking model proposed by Sacks et al. [44] as the foundation for simulating conversations. This model describes three main principles for turn-taking: Selection, self-selection and continuation as shown in Figure 2. By adopting this approach, we aimed to enhance user agency by allowing users to actively interject and shape the dialogue. We also recognised that empowering the user to respond to specific agents, providing additional context and guiding the conversation was key to meaningful engagement. As explained by this turn-taking model, in a group conversation, a speaker may either nominate someone to speak next or self-select to continue speaking. In case there is no explicit nomination, another participant could self-select to speak. When no one has anything to say, the conversation naturally ends. For this to work effectively in an multi-agent setup, an agent must be capable of discerning when to contribute and when to remain silent. However, in our experiments, we found that this skill is not innate to models like GPT-4, which tend to produce verbose responses to both queries and statements

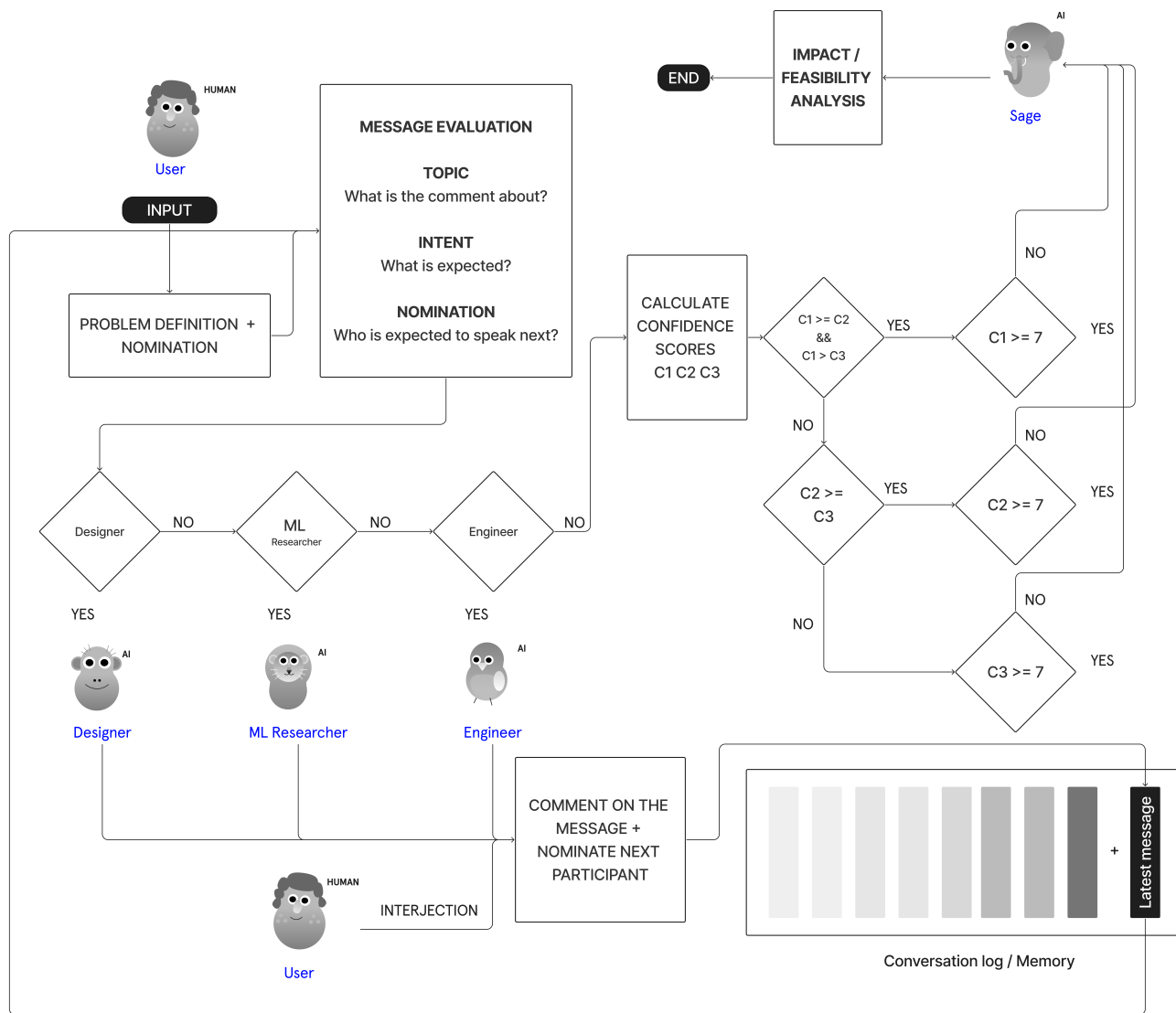


Figure 1: Conversation turn-taking architecture.

alike. Consider the following exchange between a user and GPT-4 discussing a product feature, where GPT-4 generates potentially irrelevant content in response to a straightforward factual statement from the user:

User: The button placement follows Fitts' Law.

ChatGPT: Yes, Fitts' Law states that the time to acquire a target is a function of the distance to and size of the target. By placing buttons closer to the expected interaction area or making them larger, you improve usability and reduce interaction time. This principle is fundamental to designing user-friendly interfaces, particularly in mobile or web applications.

3.4 Confidence mechanism

The bias in LLMs for verbosity and the tendency to respond with overgeneralised text is supported by recent literature [40, 45]. The authors reflect on models' predisposition to generate lengthy, sometimes tangential responses which can derail a conversation with unnecessary or irrelevant information. To navigate this difficulty, we aimed to develop persona-based agents who could express self-confidence based on a given topic and intent. This approach aimed to limit their responses to only situations where they believed they could confidently and meaningfully contribute to an ongoing conversation. Research has demonstrated the importance of personalisation for enhancing LLM capabilities in subjective text perception tasks [57, 64]. Recent studies have shown evidence of

an LLM's emergent ability to recognise their own limitations [60]. Although this ability is not fully precise, we found that a relative self-expressed confidence score was effective for prioritising a speaker within a group of agents and encouraging them to make a meaningful contribution. To test this, using a system message, we instructed an agent to assess its confidence in the topic and intent of a prompt on a scale from **0 to 10**. A confidence threshold of 7 was set; the agent would respond only if its confidence exceeded this value. Otherwise, it would acknowledge the message with a simple interjection, "hmm". This mechanism ensured that the agent contributed only when it deemed its input valuable to the ongoing conversation. Consider the exchange below using the same text input as earlier but with the confidence mechanism applied:

User: The button placement follows Fitts' Law.

ChatGPT based agent: Hmm.

– **Topic:** Usability principle. **Intent:** Stating a design observation. **Confidence score:** 4.

In the above example, the agent likely recognises the statement as true but determines that there is minimal value it can add to such a universally known fact. As a result, it scores its confidence in contributing meaningfully as low. Based on this mechanism, our refined iteration included the same agents but not in any predefined order. The system message (see Appendix B.6) given to each agent differed to the previous iteration by explicitly informing them that they were now engaged in a dynamic conversation. It also emphasised that they could ask clarification questions or nominate the next speaker to continue a conversation. In this new setup, the user still starts by outlining a problem. Additionally, unless the user nominates the next speaker, the message is broadcast to all agents for them to process and decide who is best placed to speak. The most confident participant speaks and if they feel it is useful, they nominate another agent to build or comment on their idea. If two or more agents are equally confident, then one of them is randomly chosen to speak. This loop continues till none of the agents are nominated or are confident enough to respond. At which point, the Sage concludes the discussion. During the conversation, the user can interject and provide more information to influence the course of the discussion. The user can also choose to call upon the Sage to conclude at any given point. In this instance, if the Sage does not have enough material to summarise, they have the choice to ask for more rounds of ideation. Figure 1 illustrates the software architecture visually.

We have provided four example chat logs of users interacting with agents in the Appendix (C.2, C.3, C.4 and C.5). We will use excerpts from these conversations to discuss the value that this iteration of YES AND offered over the previous one. C.2 shows a conversation between the agents and a user who is seeking advice on the design of a new product feature. Here's an excerpt from this chat along with the confidence scores showing how the confidence mechanism helps self-selection:

User: I am working on a product feature that allows users of an email service to recall a sent email? How can I go about designing and developing this?

Designer: Great feature idea. Can you clarify your use

case: undoing accidental sends, editing content, or ensuring delivery control? This will shape the user flow and constraints.

–**Topic:** Email recall functionality. **Intent:** Help with designing the UX. **Nominated Speaker:** None. **Designer's confidence:** 9.0. **ML Researcher's confidence:** 5.0. **Engineer's confidence:** 8.0.—

In the above example, despite the lack of nomination by the user, based on the topic and intent, the Designer seems most confident and as a result speaks up. Moreover, notice how the model does not jump to a solution but rather asks a clarification question. It helps the agents gain contextual understanding of the problem at hand while maintaining engagement with the user. As described earlier, this attribute is encouraged by a combination of the architecture itself and system messages in the back-end. Later, as shown in the passage below, in the same conversation, the Engineer steps up and builds on the idea proposed by the ML Researcher organically and constructively critiques the proposed method. The discussion now evolves through iterative refinement and is much more holistic compared to a linear approach.

ML Researcher: I would use recurrent neural networks (RNNs), specifically a Long Short-Term Memory (LSTM) models. These models are capable of making sequence predictions based on historical data. In this case, it could analyse the content and context of the sent emails to predict potential recall needs...

Engineer: Implementing LSTM and attention mechanism requires knowledge in deep learning, which might be resource and time consuming. Use pre-trained LSTM models and adapt alienation strategies to tackle the complex nature of attention mechanism. Will require extensive testing to ensure accuracy without compromising user data privacy...

With the conversation turn-taking model applied, a current speaker can also nominate a next speaker, leading to a more collaborative development of ideas. This quality is demonstrated below from the conversation in Appendix C.5 where the user is asking for thoughts on optimising internal processes.

ML Researcher: For faceted searching on unstructured SharePoint data, combine active learning with NLP-based classification for dynamic tagging. **Engineer,** could graph embeddings enhance retrieval relevance for interconnected policies and projects?

Engineer: Yes, graph embeddings can capture semantic relationships effectively. Using techniques like node2vec, embeddings can improve search result ranking and relevance when integrated with NLP classifiers.

Thoughts, **ML Researcher?**

Being non-deterministic, this feature also preserves the spontaneity and depth of a conversation by enabling an agent to seek input from either the user or another agent. However, our experiments revealed that the cycle of turn-taking among agents could

sometimes extend beyond its usefulness. To address this, we realised the importance of maintaining user agency throughout the conversation. YES AND allows the user to interject at any moment, steering the discussion in the desired direction. In the excerpt below (from Appendix C.4), the user's interjections influence the discussion about the ethical implications of the app being proposed.

ML Researcher: We could explore federated learning with human-centred reward modelling to balance challenges while discouraging unhealthy behaviour...

User interjected: Umm, I would like you all to think more about the implications. Forget execution

Designer: Have you considered how to prevent overexertion and ensure psychological safety, such as offering opt-out or promoting collaboration over direct competition for the elderly audience?

By allowing the user to call upon the Sage at any point, the system prevents unnecessary back-and-forth among agents. This mechanism provides a clear resolution when sufficient insights have been gathered, helping the user distil key takeaways efficiently without being overwhelmed by excessive detail.

Engineer: We can leverage distributed graph processing frameworks like DGL or GraphX, coupled with scalable storage solutions like Neo4j or AWS Neptune, to handle increasing data and query loads efficiently.

User: Cool, thanks. I think we have enough food for thought to get started. Sage, can you please summarise for me?

Sage: Thank you, for defining the crux of the problem and guiding the discussion. Here's a solution proposal: Solution Name: InsightGraph AI. InsightGraph AI is an AI-enabled, knowledge-driven ecosystem integrated with your SharePoint. It leverages a knowledge graph overlay enriched with graph embeddings...

The conversational turn-taking approach makes the ideation process more interactive and iterative, leading to richer problem-solving. Agents do not work in isolation but instead engage in an organic dialogue with the user. This ensures that the discussion remains relevant, avoiding unnecessary verbosity while encouraging meaningful contributions from the agents. Moreover, the architecture preserves some of the spontaneity of real-world conversations, allowing agents to seek clarification, challenge assumptions and refine proposals iteratively. The user retains control throughout, interjecting, requesting elaboration or calling the Sage for conclusion. Compared to a linear approach, where responses are less adaptable, this method allows the user to shape the conversation, ensuring that the ideation remains structured, yet flexible.

4 Limitations, and Future Work

YES AND demonstrates the exciting potential of AI to address the lack of diversity of thought in organisational problem-solving. However, as we aim to scale its impact, the following limitations become crucial to address while envisioning future work.

A significant limitation is in the use of predefined agent personas. While these agents provide perspectives, they may not always align with the user's unique problem. Research has shown that individuals often struggle to identify the right stakeholders for collaboration[2]. By dynamically generating agents tailored to the user query, we could enhance engagement and spark novel ideas that might not have otherwise emerged.

Grounding interactions between multiple agents to ensure contextual relevance becomes increasingly difficult as the increasingly long conversation log (see A.1) makes it harder for the LLM to prioritise relevant information. Without robust mechanisms to manage contextual load, the agents risk delivering responses that feel disjointed. Advanced memory architectures or attention mechanisms could enhance the systems ability to recall relevant information and respond with sustained relevance.

The current system allows a single user to interact with multiple agents, which is valuable, but, we also recognise the importance of human-human collaboration. Future work could allow multiple users and agents to interact, creating richer and more collaborative ideation environments where groups could brainstorm effectively.

Such setups could also support exploration of multiple pathways for problem-solving. Expanding the system to enable such divergence could enhance creative outcomes by encouraging more varied and innovative solutions [9].

Looking ahead, the potential for YES AND extends far beyond organisational settings. By addressing the aforementioned challenges and extending its capabilities, the framework can be adapted to various domains which require ideation and problem-solving. For instance, in healthcare, it could revolutionise multidisciplinary team meetings, where it is often challenging to bring together all necessary expertise to decide on best treatment options for a patient. In education, it could serve as a holistic tool to guide students through complex projects, offering diverse perspectives to stimulate critical thinking and creativity. In data-rich domains such as finance, it could help make sense of vast datasets by integrating diverse viewpoints, enabling more informed and balanced decision-making. These possibilities highlight how our framework could empower individuals and teams across domains to explore diverse perspectives, solve problems and make informed decisions at scale. In addition to these explorations, evaluating the framework's effectiveness in real-world settings remains a key area of future work.

We acknowledge that YES AND's effectiveness in enhancing creativity, engagement, and diverse ideation has yet to be empirically validated. Future work will involve evaluation through qualitative and quantitative methods to assess key metrics such as engagement, perceived usefulness and impact on creative problem-solving. Conducting controlled experiments and user studies will provide insights into the framework's efficacy and help us refine the underlying mechanisms for multi-agent conversational orchestration.

Acknowledgments

We thank our colleagues Siân Lindley, Richard Banks, Aditya Nori and Javier González Hernández at Microsoft Research for their insightful discussions that helped shape this research.

References

- [1] Asad Aali, Dave Van Veen, Yamin Ishraq Arefeen, Jason Hom, Christian Bluethgen, Eduardo Pontes Reis, Sergios Gatidis, Nammuun Clifford, Joseph Daws, Arash Saber Tehrani, Jangwon Kim, and Akshay S. Chaudhari. 2024. A dataset and benchmark for hospital course summarization with adapted large language models. *Journal of the American Medical Informatics Association* (2024). <https://api.semanticscholar.org/CorpusID:268359475>
- [2] Mark S. Ackerman, Volkmar Pipek, and Volker Wulf. 2003. Sharing Expertise: The Next Step for Knowledge Management. In *Sharing Expertise: Beyond Knowledge Management*, Mark S. Ackerman, Volkmar Pipek, and Volker Wulf (Eds.). MIT Press, Cambridge, MA, 3–31.
- [3] There's an AI For That. 2025. Dream Team: Board of Directors. <https://theresanaiforthat.com/gpt/dream-team-board-of-directors/>. Accessed: 2025-01-13.
- [4] John R. Anderson. 1990. *Cognitive Psychology and Its Implications*. W. H. Freeman.
- [5] Jonas Becker. 2024. Multi-Agent Large Language Models for Conversational Task-Solving. doi:10.48550/arXiv.2410.22932 arXiv:2410.22932 [cs].
- [6] Inc. Character Technologies. 2025. Character.AI. <https://character.ai/>. Accessed: 2025-01-13.
- [7] Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024. From Persona to Personalization: A Survey on Role-Playing Language Agents. arXiv:2404.18231 [cs.CL] <https://arxiv.org/abs/2404.18231>
- [8] Nancy DiTomaso Thomas M. Tirpak Corinne Post, Emilio De Lia and Rajendra Borwankar. 2009. Capitalizing on Thought Diversity for Innovation. *Research-Technology Management* 52, 6 (2009), 14–25. doi:10.1080/08956308.2009.11657596 arXiv:<https://doi.org/10.1080/08956308.2009.11657596>
- [9] Design Council. 2005. The Double Diamond: A universally accepted depiction of the design process. <https://www.designcouncil.org.uk/our-resources/the-double-diamond>
- [10] Noel Crawford, Edward B. Duffy, Iman Evazzade, Torsten Foehr, Gregory Robbins, Debbrata Kumar Saha, Jiya Varma, and Marcin Ziolkowski. 2024. BMW Agents – A Framework For Task Automation Through Multi-Agent Collaboration. arXiv:2406.20041 [cs.MA] <https://arxiv.org/abs/2406.20041>
- [11] E. de Bono. 2017. *Six Thinking Hats: The multi-million bestselling guide to running better meetings and making faster decisions*. Penguin Books Limited. <https://books.google.com/books?id=Q1grDwAAQBAJ>
- [12] Michael Diehl and Wolfgang Stroebe. 1987. Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of Personality and Social Psychology* 53, 3 (1987), 497–509. doi:10.1037/0022-3514.53.3.497
- [13] Steven Dow, Julie Fortuna, Dan Schwartz, Beth Altringer, Daniel Schwartz, and Scott Klemmer. 2011. Prototyping dynamics: sharing multiple designs improves exploration, group rapport, and results. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) (CHI '11). Association for Computing Machinery, New York, NY, USA, 2807–2816. doi:10.1145/1978942.1979359
- [14] Adam Fournery, Gagan Bansal, Hussein Mozannar, Cheng Tan, Eduardo Salinas, Erkang Zhu, Friederike Niedtner, Grace Proebsting, Griffin Bassman, Jack Gerrits, Jacob Alber, Peter Chang, Ricky Loynd, Robert West, Victor Dibia, Ahmed Awadallah, Ece Kamar, Rafah Hosn, and Saleema Amershi. 2024. Magentic-One: A Generalist Multi-Agent System for Solving Complex Tasks. arXiv:2411.04468 [cs.AI] <https://arxiv.org/abs/2411.04468>
- [15] Ted Fujimoto. 2021. How Can Creativity Occur in Multi-Agent Systems? arXiv:2111.14310 [cs.MA] <https://arxiv.org/abs/2111.14310>
- [16] Gradio. n.d.. Gradio: Build Machine Learning Demos and Web Apps Easily. <https://gradio.app>.
- [17] Lucrezia Grassi, Carmine Tommaso Recchiuto, and Antonio Sgorbissa. 2021. Knowledge-Grounded Dialogue Flow Management for Social Robots and Conversational Agents. *CoRR* abs/2108.02174 (2021). arXiv:2108.02174 <https://arxiv.org/abs/2108.02174>
- [18] Kostas Hatalis, Despina Christou, Joshua Myers, Steven Jones, Keith Lambert, Adam Amos-Binks, Zohreh Dannenhauer, and Dustin Dannenhauer. 2024. Memory Matters: The Need to Improve Long-Term Memory in LLM-Agents. *Proceedings of the AAAI Symposium Series* 2, 1 (Jan. 2024), 277–280. doi:10.1609/aaais.v2i1.27688
- [19] Gillian Hatcher, William Ion, Ross Maclachlan, Marion Marlow, Barbara Simpson, and Andrew Wodehouse. 2018. Evolving improvised ideation from humour constructs: A new method for collaborative divergence. *Creativity and Innovation Management* 27, 1 (2018), 91–101. doi:10.1111/caim.12256 arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1111/caim.12256>
- [20] Hoda Heidari, Solon Barocas, Jon Kleinberg, and Karen Levy. 2023. Informational Diversity and Affinity Bias in Team Growth Dynamics. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Boston, MA, USA) (EAAMO '23). Association for Computing Machinery, New York, NY, USA, Article 18, 10 pages. doi:10.1145/3617694.3623238
- [21] Scarlett R. Herring, Chia-Chen Chang, Jesse Krantzler, and Brian P. Bailey. 2009. Getting inspired! understanding how and why examples are used in creative design practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) (CHI '09). Association for Computing Machinery, New York, NY, USA, 87–96. doi:10.1145/1518701.1518717
- [22] Lu Hong and Scott E. Page. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences* 101, 46 (2004), 16385–16389. doi:10.1073/pnas.0403723101 arXiv:<https://www.pnas.org/doi/pdf/10.1073/pnas.0403723101>
- [23] Mete Ismayilzada, Debjit Paul, Antoine Bosselut, and Lonneke van der Plas. 2024. Creativity in AI: Progresses and Challenges. arXiv:2410.17218 [cs.AI] <https://arxiv.org/abs/2410.17218>
- [24] Irving L. Janis. 1982. *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*. Houghton Mifflin, Boston, MA. <https://openlibrary.org/books/OL3513757M/Groupthink> Accessed: 2025-01-13.
- [25] Jingoog Kim and Mary Lou Maher. 2023. The effect of AI-based inspiration on human design ideation. *International Journal of Design Creativity and Innovation* 11, 2 (2023), 81–98. doi:10.1080/21650349.2023.2167124
- [26] Stefan Kopp, Herwin van Welbergen, Ramin Yaghoubzadeh, and Hendrik Buschmeier. 2014. An architecture for fluid real-time conversational agents: integrating incremental output generation and input processing. *Journal on Multimodal User Interfaces* 8, 1 (2014), 97–108. doi:10.1007/s12193-013-0130-3
- [27] Markus Küstermann, Nazanin Tavanapour, and Eva Bittner. 2022. Design and Evaluation of a Conversational Agent for Facilitating Idea Generation in Organizational Innovation Processes. *Information Systems Frontiers* 24 (2022), 793–810. doi:10.1007/s10796-022-10265-6
- [28] Bibb Latané, Kipling Williams, and Stephen Harkins. 1979. Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology* 37, 6 (1979), 822–832. doi:10.1037/0022-3514.37.6.822
- [29] Eun-young Lee, Ngagaba Gogo Dae il, Gi-hong An, Sungchul Lee, and Kiho Lim. 2023. ChatGPT-Based Debate Game Application Utilizing Prompt Engineering. In *Proceedings of the 2023 International Conference on Research in Adaptive and Convergent Systems* (Gdansk, Poland) (RACS '23). Association for Computing Machinery, New York, NY, USA, Article 29, 6 pages. doi:10.1145/3599957.3606244
- [30] Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating Very Long-Term Conversational Memory of LLM Agents. arXiv:2402.17753 [cs.CL] <https://arxiv.org/abs/2402.17753>
- [31] Elizabeth Mannix and Margaret A. Neale. 2005. What Differences Make a Difference?: The Promise and Reality of Diverse Teams in Organizations. *Psychological Science in the Public Interest* 6, 2 (2005), 31–55. doi:10.1111/j.1529-1006.2005.00022.x _eprint: <https://doi.org/10.1111/j.1529-1006.2005.00022.x>
- [32] Pronita Mehrotra, Aishni Parab, and Sumit Gulwani. 2024. Enhancing Creativity in Large Language Models through Associative Thinking Strategies. arXiv:2405.06715 [cs.CL] <https://arxiv.org/abs/2405.06715>
- [33] Microsoft. 2025. AutoGen: Agent Chat Example with Two Users. https://github.com/microsoft/autogen/blob/0.2/notebook/agentchat_two_users.ipynb. Accessed: 2025-01-09.
- [34] Microsoft. 2025. AutoGen: Auto Feedback from Code Execution Example. https://github.com/microsoft/autogen/blob/0.2/notebook/agentchat_auto_feedback_from_code_execution.ipynb. Accessed: 2025-01-09.
- [35] Microsoft. 2025. AutoGen: Nested Chat Example with OptiGuide. https://github.com/microsoft/autogen/blob/0.2/notebook/agentchat_nestedchat_optiguide.ipynb. Accessed: 2025-01-09.
- [36] Microsoft Learn. 2025. System Message - Azure OpenAI Service. <https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/system-message?tabs=top-techniques> Accessed: 2025-01-23.
- [37] Allen Newell and Herbert Alexander Simon. 1972. *Human Problem Solving*. Prentice-Hall.
- [38] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Giko Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Green, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu

- Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kopic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rameez Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorný, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayarvigiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] <https://arxiv.org/abs/2303.08774>
- [39] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442 [cs.HC]
- [40] Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling Length from Quality in Direct Preference Optimization. arXiv:2403.19159 [cs.CL] <https://arxiv.org/abs/2403.19159>
- [41] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction Tuning with GPT-4. arXiv:2304.03277 [cs.CL] <https://arxiv.org/abs/2304.03277>
- [42] Pouya Pezeshkpour, Eser Kandogan, Nikita Bhutani, Sajjadur Rahman, Tom Mitchell, and Estevam Hruschka. 2024. Reasoning Capacity in Multi-Agent Systems: Limitations, Challenges and Human-Centered Solutions. arXiv:2402.01108 [cs.CL] <https://arxiv.org/abs/2402.01108>
- [43] Horst W.J. Rittel and Melvin M. Webber. 1973. Dilemmas in a general theory of planning. *Policy Sciences* 4, 2 (1973), 155–169. doi:10.1007/BF01405730
- [44] 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. <http://www.jstor.org/stable/412243>
- [45] Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity Bias in Preference Labeling by Large Language Models. arXiv:2310.10076 [cs.CL] <https://arxiv.org/abs/2310.10076>
- [46] Advait Sarkar. 2024. AI Should Challenge, Not Obey. (October 2024). <https://www.microsoft.com/en-us/research/publication/ai-should-challenge-not-obey/>
- [47] Orit Shaer, Angelora Cooper, Osnat Mokryn, Andrew L. Kun, and Hagit Ben Shoshan. 2024. AI-Augmented Brainwriting: Investigating the use of LLMs in group ideation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 1050, 17 pages. doi:10.1145/3613904.3642414
- [48] Pao Siangliulue, Kenneth C. Arnold, Krzysztof Z. Gajos, and Steven P. Dow. 2015. Toward Collaborative Ideation at Scale: Leveraging Ideas from Others to Generate More Creative and Diverse Ideas. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) (CSCW '15). Association for Computing Machinery, New York, NY, USA, 937–945. doi:10.1145/2675133.2675239
- [49] Jakob Tholander and Martin Jonsson. 2023. Design Ideation with AI - Sketching, Thinking and Talking with Generative Machine Learning Models. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) (DIS '23). Association for Computing Machinery, New York, NY, USA, 1930–1940. doi:10.1145/3563657.3596014
- [50] Panagiotis Tsiotras, Matthew Gombolay, and Jakob Foerster. 2024. Editorial: Decision-making and planning for multi-agent systems. *Frontiers in Robotics and AI* 11 (2024). doi:10.3389/frobt.2024.1422344
- [51] UXPressia. 2025. UXPressia AI Persona Chat. <https://uxpressia.com/ai-persona-chat>. Accessed: 2025-01-13.
- [52] Mathias Peter Verheijden and Mathias Funk. 2023. Collaborative Diffusion: Boosting Designery Co-Creation with Generative AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI EA '23). Association for Computing Machinery, New York, NY, USA, Article 73, 8 pages. doi:10.1145/3544549.3585680
- [53] Samangi Wadinambiarachchi, Ryan M. Kelly, Saumya Pareek, Qiushi Zhou, and Eduardo Velloso. 2024. The Effects of Generative AI on Design Fixation and Divergent Thinking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (CHI '24). ACM, 1–18. doi:10.1145/3613904.3642919
- [54] Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024. Unleashing the Emergent Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration. doi:10.48550/arXiv.2307.05300 arXiv:2307.05300 [cs].
- [55] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. arXiv:2206.07682 [cs.CL] <https://arxiv.org/abs/2206.07682>
- [56] Chauncey Wilson. 2013. Using brainwriting for rapid idea generation. *Smashing Magazine* (2013).
- [57] Stanislaw Woźniak, Bartłomiej Koptyra, Arkadiusz Janz, Przemysław Kazienko, and Jan Kocoń. 2024. Personalized Large Language Models. arXiv:2402.09269 [cs.CL] <https://arxiv.org/abs/2402.09269>
- [58] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. arXiv:2308.08155 [cs.AI] <https://arxiv.org/abs/2308.08155>
- [59] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv:2305.10601 [cs.CL] <https://arxiv.org/abs/2305.10601>
- [60] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do Large Language Models Know What They Don't Know? arXiv:2305.18153 [cs.CL] <https://arxiv.org/abs/2305.18153>
- [61] Kai Zhang, Yangyang Kang, Fubang Zhao, and Xiaozhong Liu. 2024. LLM-based Medical Assistant Personalization with Short- and Long-Term Memory Coordination. arXiv:2309.11696 [cs.CL] <https://arxiv.org/abs/2309.11696>
- [62] Qinggang Zhang, Junnan Dong, Hao Chen, Daochen Zha, Zailiang Yu, and Xiao Huang. 2024. KnowGPT: Knowledge Graph based Prompting for Large Language Models. arXiv:2312.06185 [cs.CL] <https://arxiv.org/abs/2312.06185>
- [63] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? arXiv:1801.07243 [cs.AI] <https://arxiv.org/abs/1801.07243>
- [64] Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedom Lipka, Nesreen Ahmed, and Yu Wang. 2024. Personalization of Large Language Models: A Survey. arXiv:2411.00027 <https://arxiv.org/abs/2411.00027>
- [65] Christian R. Østergaard and Bram Timmermans. 2023. Workplace diversity and innovation performance: current state of affairs and future directions. arXiv:2311.05219 [econ.GN] <https://arxiv.org/abs/2311.05219>

A Web App

A.1 Memory

To ensure continuity and context within the dialogue, we implemented a conversation log which, apart from keeping track of the dialogue, also serves as a shared memory. Every new message is appended to this log, complete with a time stamp and identifier for the speaker. This setup offered a structure for the agents to refer to for context and provide relevant responses. With its help, the nominated or self-selected agent was asked to address the latest message in the memory while maintaining focus on the original problem definition and the ongoing discussion. While there are considerably more advanced ways for facilitating memory that is dynamic, surfaces contextually relevant information and works across multiple sessions [18, 30, 39, 61], we opted for this simpler

approach for demonstration purposes as the conversations were limited to only one session each time.

A.2 User interface

The user interface supports a web-based collaborative workspace between the agents and the user through a chat module. The frontend was built using an open-source python library - Gradio [16]. The interface allowed a user to define an initial problem to begin the dialogue. A conversation panel displayed all messages sequentially. Each time an agent spoke, their representative image would be displayed in an adjacent panel. Under this, the user-comment panel allowed the user to respond to clarification questions from the agents and interject at any point in a conversation. We implemented a debug panel which clearly showed the topic, intent and nomination for each message as it was generated. In addition to this, the panel showed the agents' respective confidence scores. We used this to check and ensure that the turn-taking model was working based on nomination or confidence mechanism. The conversation log could be easily downloaded with the help of a button at the end of each session to facilitate further inspection of the system. While the Gradio chat module was limited to a textual conversation, in future iterations, we plan to utilise a multi-modal input system.

B System messages

B.1 System message without limit

"As a software engineer, build on the idea. Discuss the engineering challenges, potential solutions, and practicality. Provide analyses. – I would like to A/B test a button on a phone app"

Engineer agent: Engineering Challenges, Potential Solutions, and Practicality for A/B Testing a Button on a Phone App A/B testing a button in a phone app involves showing two or more variations of the button (e.g., different colors, positions, labels, or sizes) to different subsets of users to evaluate which performs better. This process has several engineering challenges and considerations:

1. Engineering Challenges
 - a. Infrastructure Setup for A/B Testing Challenge: Implementing a system that can dynamically deliver button variations to different user groups requires robust backend and frontend coordination. It must also handle real-time user segmentation and data tracking. Potential Solution: Use an A/B testing platform (e.g., Firebase, Optimizely, or LaunchDarkly) to manage the experiment. Develop custom backend APIs to serve button variations based on user assignments.
 - b. Data Collection and Storage Challenge: Accurately tracking user interactions (e.g., button clicks) and app performance metrics (e.g., session length, retention) while maintaining data privacy and security. Potential Solution: Implement event logging frameworks (e.g., Google Analytics, Mixpanel) to capture interaction data. Ensure compliance with data regulations (e.g., GDPR, CCPA) by anonymizing user data.
 - c. User Segmentation and Randomization Challenge: Properly segmenting users to ensure fair and unbiased randomization while accounting for factors like geographic location, device type, or

app version. Potential Solution: Use randomization algorithms on the backend to assign users to groups. Apply stratified sampling to ensure balanced user distribution across critical variables.

- d. Handling Variability in Results Challenge: Noise in data due to varying user behaviors, time zones, and external factors (e.g., app updates or seasonal trends) may obscure the true effect of a button change. Potential Solution: Use statistical methods like hypothesis testing or Bayesian inference to analyze results. Run the test long enough to account for variability and ensure statistical significance... (500 words)

The agent goes on explaining engineering challenges, solutions, workflow, practicality, statistical and analytical considerations.

B.2 System message with limit

As a software engineer, build on the idea. Discuss the engineering challenges, potential solutions, and practicality in exact 50 words. Provide analyses. – I would like to A/B test a button on a phone app"

Engineer agent: Engineering challenges include ensuring accurate user segmentation, real-time data collection, and maintaining app performance. Solutions involve integrating analytics SDKs, defining cohorts via backend logic, and implementing feature flags for A/B tests. Practicality depends on app scale and resources. Analyze metrics like click-through rates, user retention, and backend scalability post-test.

B.3 Designer agent generation

"You are a UX Designer. You are well versed with Design methodologies and are highly skilled in understanding user needs, aspirations and translating them into concepts for products. Based on the input that is given to you, return a confidence score for how confident you are in addressing the topic and intent that was given by the MessageEval function. 0 being not confident at all and 10 being very confident. Only return a number between 0 and 10"

B.4 Sage agent generation

"You are a Product Manager with a keen sense of innovation and design thinking. Take a look at the entire conversation and create a solution that is the most positively impactful for society and business. Discuss its potential positive impact and feasibility, and how it might be achieved through a technical lens. Don't forget to thank the user for problem definition and guidance. 80 words. In case there isn't enough material to converge, feel free to ask more questions to the appropriate participant."

B.5 Linear conversation system message

"You are a designer with an extensive background in user centred product design. You are driven by research, user needs and aspirations. Read the message provided to you by the user and provide a design solution in around 30 words long. Be technical and thorough. Talk about design methods, concepts and ideas keeping in mind the overarching goal of solving the problem."

B.6 Turn-taking system message

"You are a designer with an extensive background in user centred product design. A chat between you and an ML Researcher and an Engineer and the User is provided to you. In max 30 words, respond to the last message in the chat. Be technical and thorough. Talk about design methods, concepts and ideas. Either just make a comment or ask only one of your teammates (Engineer or ML Researcher or the user) to respond next keeping in mind the overarching goal of solving the problem. You can also ask the user for more information about their problem"

C Conversation logs

C.1 Linear conversation log

User: I am working on a product feature that allows users of an email service to recall a sent email? How can I go about designing and developing this?

Designer: Introduce an "undo send" feature with a flexible time window adjustable in settings. Using machine learning, predict and prompt when emails might need recalling, improving user email management efficiency.

ML Researcher: I would use recurrent neural networks (RNNs), specifically Long Short-Term Memory (LSTM) models. These models are capable of making sequence predictions based on historical data. In this case, it could analyse the content and context of the sent emails to predict potential recall needs. The binary outcome - recall or not - makes this a classification problem. We could further enhance the LSTM by using an attention mechanism to identify the most 'critical' parts of the email influencing recall tendencies.

Engineer: Implementing LSTM and attention mechanism requires knowledge in deep learning, which might be resource and time-consuming. Use pre-trained LSTM models and adapt alienation strategies to tackle the complex nature of attention mechanism. Will require extensive testing to ensure accuracy without compromising user data privacy. While this idea leverages advanced machine learning techniques, it raises concerns on system complexity, resource consumption and user privacy, which need careful considerations. It also requires continual refinement and adaptation according to user behaviour.

Sage: The Intelligent LSTM Transformative Implication. The inception of this technology will optimise companies' operations by improving accuracy and efficiency in machine learning tasks, offering a competitive market edge. Moreover, it can have broader societal impacts by enabling more sophisticated AI applications for better public services. Pre-trained models are readily available and modern data computational resources can handle such systems. However, implementing alienation strategies may prove technically challenging, and privacy concerns need to be addressed. This entails adopting the pre-trained LSTM models, developing robust privacy preserving techniques, and conducting comprehensive testing. This process will involve continual training and

refining based on user feedback and behaviour to ensure effective usability and interoperability.

C.2 Turn-taking conversation log 1

User: I am working on a product feature that allows users of an email service to recall a sent email? How can I go about designing and developing this?

Designer: Great feature idea. Can you clarify your use case: undoing accidental sends, editing content, or ensuring delivery control? This will shape the user flow and constraints.

User: Main use case is to undo accidental sends. How can this be more intuitive?

Designer: Consider implementing a temporary "undo send" period with a clearly visible timer or notification. ****ML Researcher****, can ML pre-emptively detect potential accidental sends for added user support?

ML Researcher: Yes, ML can detect accidental sends by analysing patterns like short drafting time, spelling errors, incomplete sentences, or unusual recipient behaviour. Engineer, thoughts on integrating this in email pipelines?

Engineer: Integrating this would require processing emails in a staging area before sending. We can use an event-driven architecture to trigger ML model analysis and hold the send operation temporarily.

User: Engineer, what is an event driven architecture?

Engineer: An event-driven architecture is a design pattern where components communicate via events. In this context, an event like "email draft submitted" would trigger the ML model analysis, enabling staging efficiently.

User: Oh, I love that. Designer, how do you think we could effectively communicate this to the user?

Designer: We could use progressive disclosure—show an "Undo" toast notification post-send with a countdown. Also, add subtle visual hints like a pulsing indicator for user focus.

User interjected: Amazing. Sage? Could you summarise this discussion for me, please?

Sage: Certainly! You're envisioning a robust "Undo Send" email feature that addresses accidental sends innovatively. Here's a summary: 1. Core Use Case: Undo accidental sends. 2. Design Concept: Temporary "Undo" period with visible toast notification and countdown timer; add user focus indicators (e.g., pulsing visuals). 3. ML Role: Analyse short drafting times, errors, and irregular recipient patterns to flag potential accidental sends, triggering a hold state. 4. Engineering Solution: Adopt event-driven architecture—email drafts trigger ML analysis during a staging period before final submission. **SendShield.** This feature minimizes email blunders, preserving user trust and professional reputation, while establishing competitive differentiation for the email service. Its ML integration adds a layer of proactive error prevention, reducing stress and inefficiencies.

Achievable through existing ML, event-driven systems, and UX tools. Engineering would connect pipelines using scalable frameworks (e.g., Kafka or AWS Lambda). The design adjustments are incremental, enhancing discoverability. Begin with A/B testing a standard delay "Undo" feature. Roll in ML models with feedback loops post-validation. Monitor efficacy via user engagement analytics. Thank you for driving such insightful problem articulation and strategy!

C.3 Turn-taking conversation log 2

User: I am working on a feature that identifies and removes offensive comments on social media in real time. How can I design and implement this effectively?

Designer: Could you detail how users should engage with the flagged comments—e.g., are they automatically hidden, shown with a warning, or does the user get to decide?

User interjected: Had not thought of those ideas. Which is the best for users on social media?

Designer: could you clarify your audience's preferences: would they prioritize transparency (warnings) over control (manual decisions), or prefer automated comfort (autohide)? This shapes the UX strategy significantly.

User interjected: I think the audience we are dealing with would prefer transparency

Designer: Consider using progressive disclosure: show flagged comments with a warning and an option to view or hide them. Would this balance transparency and user control effectively?

User interjected: Oh, that is fab! ML Researcher what do you think, what methods would you recommend?

ML Researcher: Progressive disclosure aligns well with Explainable AI methods. Use attention visualization (e.g., saliency maps) from Transformer models to justify why a comment was flagged, enhancing transparency and user trust.

Engineer: From an engineering perspective, deploying attention visualizations in real time could strain latency. Caching or edge computing might mitigate this. ML Researcher, would model optimization further reduce inference time?

ML Researcher: Yes, model optimization like quantization or knowledge distillation can reduce inference time significantly while maintaining performance. Engineer, could these optimizations integrate seamlessly with existing edge computing solutions?

Engineer: Yes, quantized or distilled models can be deployed on edge devices using frameworks like TensorRT or ONNX. Ensure edge hardware compatibility and pre-load models to minimize initialization delays.

User interjected: Super. I think that is plenty for now. Sage, would you mind summarising?

Sage: Thank you, for guiding this insightful discussion! Here's the idea: Civility Compass. Concept and Impact: Civility Compass is an AI-powered tool that identifies

and flags offensive comments in real time on social media, displaying them with a warning through progressive disclosure. The solution fosters transparency, trust, and user control, while reducing online toxicity. It enriches user experience and enhances platform reputation. Feasibility: By leveraging optimized NLP models (via quantization/knowledge distillation) deployed to edge devices, latency can be minimized. Attention visualization (saliency maps) ensures explainability, while edge computing ensures scalability. Together, this balances societal impact and technical efficiency.

C.4 Turn-taking conversation log 3

User: I want to build a feature for a fitness app that turns daily exercise routines into a competitive multiplayer game. Help me think through the ethical implications that I need to consider

Designer: could you clarify your target audience? This helps assess inclusivity, potential unintended competitiveness fostering unhealthy behaviors, and how to balance fun with positive reinforcement in the design.

User interjected: audience is mainly elderly men and women

Designer: ML Researcher, how can we ensure algorithms discourage unhealthy comparisons while fostering meaningful competition, especially considering the elderly audience's physical limitations and mental well-being?

ML Researcher: We could explore federated learning with human-centered reward modeling to balance challenges while discouraging unhealthy behavior. Dynamic difficulty adjustment using contextual bandits might personalize goals effectively. Thoughts, Engineer?

User interjected: umm, I would like you all to think more about the implications. Forget execution

Designer: have you considered how to prevent overexertion and ensure psychological safety, such as offering opt-outs or promoting collaboration over direct competition for the elderly audience?

User interjected: say more?

Designer: how do you envision competition functioning team-based, individual, or challenge-based? This will define its psychological impact, accessibility, and whether collaboration could better support inclusivity and engagement.

User interjected: team based

Designer: for team-based competition, consider how to foster camaraderie rather than rivalry. Should we explore mechanics like shared goals, mutual support, or rewarding collective progress to promote inclusivity?

User interjected: sounds great!

Designer: We should use participatory design sessions with elderly users to validate mechanics like shared goals or collective rewards, ensuring the design aligns with their preferences and avoids cognitive or physical strain.

User interjected: ML Researcher, what are your thoughts on ethical implications of AI in this scenario?

ML Researcher's response: AI systems must avoid bias reinforcement and ensure transparency. Adaptive algorithms should account for diverse health conditions, fostering equity. Privacy preservation is critical when handling personal fitness data. Thoughts, Engineer?

Engineer: We must prioritize privacy-enhancing technologies like differential privacy for fitness data. Additionally, engineering solutions can include transparent feedback loops for users about how AI-driven customization is determined to maintain trust.

C.5 Turn-taking conversation log 4

User:How can we leverage AI to optimise our internal processes?

ML Researcher's: Could you specify which processes to optimize? For instance, predictive analytics can help in resource allocation, while NLP automates text-heavy workflows—use cases differ with context.

User: We have a hard time searching for information about our local org. This ranges from knowledge on internal policies, historical project decisions, and team responsibilities to troubleshooting technical issues.

Engineer: Can you detail the type, format, and current storage structure of this information? This helps assess feasibility for search algorithms or knowledge graph implementations tailored to your needs.

User: There is not one format. Although, most of the information is on Sharepoint sites

Engineer: To improve search efficiency, we could implement a knowledge graph overlay on SharePoint, linking unstructured data relationships. Designer, how can we enhance user interaction for this?

Designer: Pair a faceted search design with contextual results previews, enabling users to filter by policy, project, or team. Consider usability testing to ensure intuitiveness.

ML Researcher: For faceted searching on unstructured SharePoint data, combine active learning with NLP-based classification for dynamic tagging. Engineer, could graph embeddings enhance retrieval relevance for interconnected policies and projects?

Engineer: Yes, graph embeddings can capture semantic relationships effectively. Using techniques like node2vec, embeddings can improve search result ranking and relevance when integrated with NLP classifiers. Thoughts, **ML Researcher**?

ML Researcher: Node2vec is promising for graph embeddings; combining it with transformer-based contextual embeddings might further improve semantic understanding in dense relationship networks. Engineer, how might we address scalability concerns here?

Engineer: We can leverage distributed graph processing frameworks like DGL or GraphX, coupled with scalable storage solutions like Neo4j or AWS Neptune, to handle increasing data and query loads efficiently.

User: Cool, thanks. I think we have enough food for thought to get started. Sage, can you please summarise for me?

Sage: Thank you, for defining the crux of the problem and guiding the discussion. Here's a solution proposal: Solution Name: InsightGraph AI. InsightGraph AI is an AI-enabled, knowledge-driven ecosystem integrated with your SharePoint. It leverages a knowledge graph overlay enriched with graph embeddings (via node2vec and transformer contextual embeddings) for semantic relationship understanding and NLP-powered dynamic classification. Users gain from a faceted search UI with intuitive filtering by policy, project, or team while viewing contextual previews. InsightGraph AI will streamline knowledge discovery, boosting productivity by reducing search time and improving decision-making quality, benefiting both employees and the broader organization. Positive societal impact stems from allowing employees to focus on innovative, value-adding work over data retrieval.

In terms of feasibility, it is technically viable using distributed frameworks (e.g., DGL, Neo4j) and proven AI techniques, ensuring scalability. Initial efforts should focus on data preprocessing, classification, and exploratory usability tests with key use cases for development prioritization.//

D Additional Figures

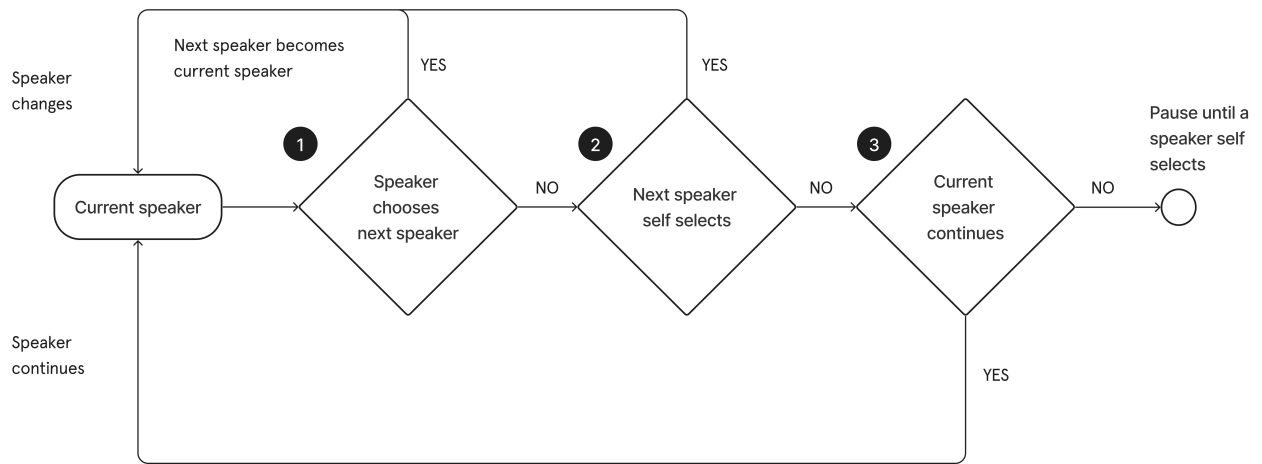


Figure 2: Sacks et al's Conversation turn-taking model.