SPARKIT: A Mind Map-Based MAS for Idea Generation Support

Masaki Ishizaka[1], Akihito Taya[2], and Yoshito Tobe[1]

[1] Grad. Sch. Aoyama Gakuin University, Sagamihara, Japan
{zacker,y.tobe}@rcl-aoyama.jp
[2] The University of Tokyo, Meguro, Japan
taya-a@iis.u-tokyo.ac.jp

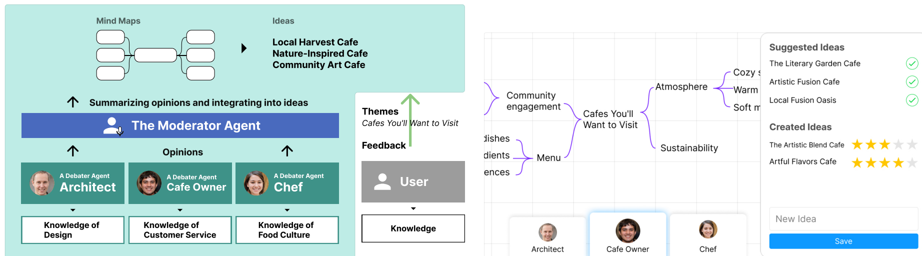
Abstract. Innovative solutions to societal challenges require the generation of creative ideas. Collaboration between humans and multi-agent systems (MAS) is a promising approach for idea generation, yet fostering creative discussions remains a challenge. This paper proposes Synergistic Platform for Advancing and Reinforcing Knowledge through Interactive Tools (SPARKIT) and SPARK-flow leveraging mind maps to facilitate idea generation within a MAS framework. SPARKIT supports idea generation with two types of large language model (LLM)-based agents: Debater Agents and the Moderator Agent. Debater Agents offer varied perspectives from their expertise, while the Moderator Agent structures discussions into a mind map to enhance user-agent collaboration. SPARK-flow is designed to stimulate creative idea generation by orchestrating discussions among these agents. A distinctive aspect of SPARKIT and SPARK-flow is that the agents facilitate the discussion and its structuring into a mind map, reducing the user's burden compared to existing methods. This paper compares different methods of discussion using language models and reveals that the impact of the discussion method on the creativity of ideas is consistent between humans and language models, with mind maps significantly enhancing creativity.

Keywords: Idea Generation · Mind Mapping · Human-Agent Interaction · Agent-based Discussion

1 Introduction

Addressing complex social challenges requires concrete and feasible solutions. However, especially in the early stages of idea generation, the lack of diverse perspectives due to insufficient information and knowledge boundaries poses a significant challenge.

To address these challenges, research on digital tools [3, 35, 4, 25] and agent-based systems [26, 13] that facilitate idea generation is advancing. These systems enhance idea generation by presenting users with a vast amount of information related to their input. Nonetheless, as most feedback from these systems is triggered by user actions, it is difficult for users without concrete ideas about the topic to utilize the system effectively.



(a) SPARKIT Architecture. **(b)** SPARKIT Interface.

Fig. 1: SPARKIT: A system for facilitating discussions through mind maps, involving Debater Agents, the Moderator Agent, and user interaction.

Idea generation requires organizing information. Mind maps are recognized as an effective tool for visually organizing thoughts and findings. [38, 30, 32]. Furthermore, combining concepts is crucial for generating creative ideas [23, 11, 10]. While mind maps help in understanding the relationships between concepts, existing tools [7, 12] primarily focus on their creation and do not sufficiently support the derivation and refinement of specific ideas from them.

This paper proposes Synergistic Platform for Advancing and Reinforcing Knowledge through Interactive Tools (SPARKIT), a system that supports idea generation through discussions involving multiple agents and users, facilitated by mind maps. As shown in Figure 1(a), when user inputs a theme based on a problem they intend to solve, SPARKIT utilizes two types of agents based on large language models (LLMs): Debater Agents and the Moderator Agent. These agents facilitate the generation of mind maps and the development of solution-oriented ideas relevant to the specified theme, thus enhancing the process of idea generation. Debater Agents are dynamically assigned expertise based on prompts, and they provide diverse perspectives to the user by presenting their opinions based on that knowledge. Users can also participate as Debater Agents, utilizing their knowledge. Furthermore, the Moderator Agent summarizes the discussion content into mind maps, combines them, and proposes concrete ideas, thereby alleviating the user's effort in idea generation. Additionally, SPARKIT has mechanisms to incorporate users' opinions and preferences through mind maps and ideas, steering discussions towards achieving ideas that conform to the users' objectives.

To obtain creative ideas, it is beneficial to discuss with multiple people who have different perspectives. However, too many opinions can cause the discussions to diverge and make it difficult to converge. In this study, we design SPARK-flow, a method within the SPARKIT that uses mind maps to balance divergent and convergent thinking in agent discussions. SPARK-flow draws inspiration from existing research on methods of discussion for creative idea generation [2, 34, 21, 18], integrating them into discussions among multi-agent systems. Section 2 describes the characteristics of creative ideas and existing research on methods of discussion. Section 3 explores how to apply these methods to multi-agent discussions, and Section 4 analyzes the impact of each method on the generated ideas. Our contributions are as follows:

1. We propose SPARKIT, a system that leverages multiple LLM-based agents with different expertise and utilizes mind mapping to support idea generation for users.
2. We design SPARK-flow, a method that extracts discussion topics and organizes them into a mind map, enabling creative idea generation through iterative discussions among agents.
3. We demonstrate the effectiveness of SPARK-flow by evaluating it based on five aspects of discussion methods and analyzing the generated mind maps to reveal the factors in creative idea generation.

2 Related Work

2.1 Mind Map-based Idea Generation

Mind maps are tools that place a core theme at the center, with related keywords branching out successively. This approach is beneficial for organizing thoughts in a structured manner. Mind maps were proposed by Tony Buzan [6] and are used in a wide range of fields, not limited to idea generation, but also in education [33], language acquisition [30], and more. Employing mind maps for brainstorming has been shown to enhance creativity [38] and increase the number of ideas compared to using bullet points [31]. Furthermore, deeper engagement in mind mapping tends to yield more unique ideas [24], supporting the theory that creativity thrives on integrating new elements into existing frameworks [11].

Tools for automatically creating mind maps have been proposed, such as methods that present technology and idea relationships using patent databases [25], aggregate crowd-sourced ideas into mind maps [7], and alternate mind map creation between humans and computers [12]. Although these tools ease mind map generation, they do not support the crucial steps of extracting and organizing ideas from the created mind maps. In contrast, SPARKIT not only generates mind maps but also offers support to utilize the information for the creation and refinement of specific ideas.

2.2 Enhancing Idea Generation with Digital Tools

Several systems have been proposed to support idea generation by presenting keywords [3] or images [35] related to the conversation topic. These methods stimulate users based on the conversation's content, thereby increasing the number of ideas. Nonetheless, the support becomes ineffective when the user stops speaking, as feedback is triggered by the user's utterances.

Cloud-based methods allowing numerous participants to share ideas in real-time can prevent discussion stagnation and have been shown to foster increasingly creative ideas over time [4]. However, these methods face challenges such as gathering participants and integrating diverse opinions.

The advent of LLM has introduced new possibilities for idea generation support. Tools that facilitate idea generation in a conversational format [13] and those that autonomously generate concepts [26] have been gaining attention. However, language models require user input to generate output, which means users must take the lead in the process. Additionally, the novelty and diversity of ideas generated by language models are often inferior to those obtained through crowdsourcing [26].

SPRKIT enables the generation of diverse ideas through discussions among multiple agents. Moreover, by structuring the content of discussions into mind maps, the system autonomously promotes the discussion, reducing the burden on users in generating ideas.

2.3 Activation of Idea Generation

Brainstorming is widely acknowledged as a method that fosters creativity. However, it has been observed that the efficiency of idea generation varies according to the structure of the brainstorming session. According to Al-Samarraie et al., group brainstorming, where participants interact simultaneously, is effective in producing a multitude of ideas. However, the number of ideas generated per person is less compared to the format where participants think of ideas individually [2]. Furthermore, it has been shown that for generating creative ideas, it is crucial to iteratively generate ideas within a short time [15], to engage in more associations [5], and to evaluate ideas when creating them iteratively [21].

The integration of concepts is a critical attribute of creative ideas. While directly combining different concepts has a limited impact on creativity, it has been demonstrated that adding new ideas based on already combined ideas enhances creativity [11, 23, 10]. Therefore, exposing users to ideas from diverse domains beforehand is considered effective [34]. SPARKIT leverages Debater Agents with diverse roles to provide users with multiple perspectives, facilitating rapid and iterative idea generation through the combination of these insights.

2.4 Evaluation of Idea

The creativity of ideas is evaluated using indicators such as novelty and usefulness [34, 29, 28]. According to Diedrich et al., novelty contributes more significantly to predicting creativity than usefulness, and in ideas with high novelty, usefulness contributes more substantially to creativity [14].

Kern et al. evaluated the responses to the Alternative Uses Test (AUT) using GPT-4 [1] in terms of novelty, feasibility, and value, with a notable positive correlation observed in the assessment of novelty compared to human evaluations [22]. Furthermore, it was demonstrated that the alignment between evaluations of GPT-4 and human evaluations strengthens as the agreement among human evaluators increases [27]. According to Hackl et al., multiple evaluations of macroeconomics test answers by GPT-4 resulted in an Intraclass Correlation Coefficient (ICC) of 0.999, indicating the reliability evaluation of GPT-4 [20]. These findings suggest that GPT-4 can consistently evaluate the creativity of ideas similarly to human evaluators. Therefore, this paper employs GPT-4 to assess the creativity of ideas generated by SPARKIT.

3 SPARKIT

SPARKIT is a system that supports idea generation through discussions among multiple agents, facilitated by mind maps. We implement SPARKIT as a web application. Users can edit the generated mind maps and provide feedback to the agents by evaluating the generated ideas, as shown in Figure 1(b).

To generate creative ideas, it is helpful to discuss with people from different perspectives. However, such discussion is not possible with only a single user. Moreover, too many opinions may cause the discussion to diverge, making it difficult to converge. To address this challenge, SPARKIT employs two types of agents: Debater Agents and the Moderator Agent, and conducts discussions using a method called SPARK-flow.

3.1 Debater Agents

Debater Agents are language models assigned roles through prompts and present opinions based on those roles. By establishing specific roles, language models can generate opinions based on the knowledge associated with those roles. This approach enables the production of more detailed responses than those possible without such role settings [37]. Additionally, by engaging multiple agents in discussions, even if the responses generated by the language model are initially incorrect, iterating through discussions can lead to improved and more accurate answers [16].

3.2 The Moderator Agent

Presenting the opinions generated by the Debater Agents directly to the user can lead to a proportional increase in text volume with the number of agents. This can make it difficult for users to grasp the content. To mitigate this, the Moderator Agent summarizes the discussion into a mind map, offering the following three advantages: (1) **Reduction of text volume by eliminating redundancies** This helps to control the increase in text volume associated with the rise in the number of agents, thereby reducing the cognitive load on the user. (2) **Visualization of the relationships between concepts** This representation facilitates the combination of concepts [33, 32, 30]. (3) **Clustering topics** By clustering topics, it becomes easier to identify topics of interest, which can facilitate the generation of ideas that meet the user's requirements. Additionally, supplying the mind map as a prompt to the language model enhances response performance [36]. By summarizing the discussion in a mind map, the Moderator Agent not only facilitates discussions among the agents but also aids in idea generation by organizing and presenting the discussion points to the user.

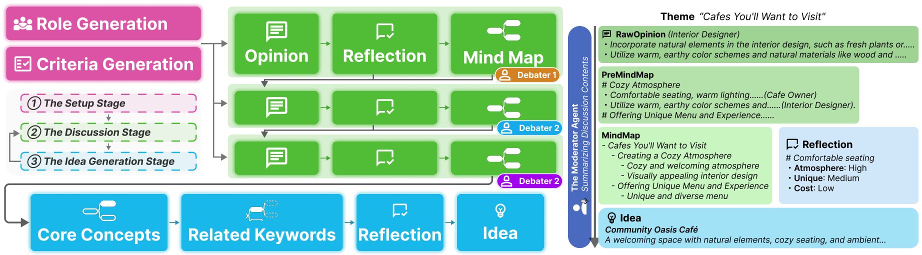


Fig. 2: SPARK-flow.

3.3 SPARK-flow

SPARK-flow is a mechanism within SPARKIT that orchestrates the flow of discussion among agents. SPARK-flow aims to: (1) generate creative ideas through the synergy of agent interactions and (2) create mind maps to present the discussion in an understandable format for users.

SPARK-flow consists of three main stages: the Setup Stage, the Discussion Stage, and the Idea Generation Stage. In the Setup Stage, the agents participating in the discussion are determined. In the Discussion Stage, Debater Agents engage in discussions on the theme. The Idea Generation Stage then focuses on generating ideas based on the content of these discussions. SPARK-flow repeats the Discussion Stage and Idea Generation Stage in cycles, each referred to as a *round*. Figure 2 illustrates the flow of the discussion and also shows the input-output relationships at each step, with inputs being conveyed to the agents using prompts.

The Setup Stage begins with the assignment of roles to Debater Agents $\mathcal{D} = \{p_{d_1}, p_{d_2}, \ldots, p_{d_{N_{\text{Roles}}}}\}$. The participants in the discussion have a significant impact on the ideas generated. If the role is not suitable for the theme, it can lead to forcibly connecting the theme and knowledge, resulting in the generation of ideas that lack relevance to the theme. To address this, we prompt the language model p to generate roles that are expected to yield creative ideas for the theme. In terms of the number of agents, it has been confirmed that increasing the number of agents beyond 3 to 4 leads to a deterioration in accuracy or a plateauing effect [9, 16]. Therefore, we set the number of agents $N_{\text{Roles}} = 3$.

While the amount of information generated increases in proportion to the number of agents, some of it does not contribute to creative ideas. It is essential to filter out irrelevant information to maintain a manageable volume of data for consideration. To address this, SPARK-flow produces N_{Criteria} evaluation criteria $\mathcal{C}$, used in later steps to assess items and determine their inclusion in mind maps or ideas. In subsequent steps, the generated content is evaluated based on these criteria, which are used as a basis for deciding whether to include them in the mind map or ideas.

The Discussion Stage begins once the Setup Stage is completed. Debater Agent p_{d_i} generates opinions O_i (*RawOpinion*) based on the assigned role and

Algorithm 1 The Discussion Stage in SPARK-flow

Require: $\mathit{theme}, N_{\text{Roles}}, N_{\text{Criteria}}$
Ensure: M
  $M \leftarrow$ None;
  $E \leftarrow$ None;
  $\mathcal{D} \leftarrow$ empty list;
  $\mathcal{C} \leftarrow$ empty list;
  for $i = 1$ to N_{Roles} **do**
    $d_i \sim p(\text{role}|\text{theme})$;
    Append p_{d_i} to $\mathcal{D}$;
  end for
  for $i = 1$ to N_{Criteria} **do**
    $c_i \sim p(\text{criterion}|\text{theme})$;
    Append c_i to $\mathcal{C}$;
  end for
  for all d_i in $\mathcal{D}$ do **do**
    $O_{d_i} \sim p_{d_i}(\text{o}|\text{theme}, M, E)$;
    $P_{d_i} \sim p_m(\text{g}|\text{theme}, M, O_{d_i})$;
    $E_{d_i} \sim p_m(\text{e}|\text{theme}, P_{d_i})$;
    $M_{d_i} \sim p_m(\text{m}|\text{theme}, M, P_{d_i}, E_{d_i})$;
    $M \leftarrow M_{d_i}$;
    $E \leftarrow E_{d_i}$;
  end for
  return M

the theme. Debater Agents are instructed to generate RawOpinion in bullet point format, as shown in Figure 2, to enable a wider range of independent perspectives than in conversational sentence format.

Directly incorporating the opinions of Debater Agents into a mind map can lead to a complex and bloated mind map because language models attempt to include all elements. To address this issue, the Moderator Agent p_m performs two steps before generating the mind map M_{d_i}(*MindMap*). First, the Moderator Agent generates a clustered representation of the opinions, called *PreMindMap* P_{d_i}, by extracting common elements from the opinions of the Debater Agents. The elements contained in the MindMap and PreMindMap are referred to as *nodes*. Although the formats of the MindMap and PreMindMap are similar, as shown in Figure 2, the text within the nodes of the PreMindMap uses the content of RawOpinion directly, while in the MindMap, these are summarized. This allows the Moderator Agent to decide on the content to be included in the MindMap by considering more information. Next, in the reflection step, the Moderator Agent p_m evaluates each node in P_{d_i} based on $\mathcal{C}$ to generate evaluations E_{d_i}, determining the priority for inclusion in the MindMap M_{d_i}. These two steps allow only the elements with a high potential for contributing to idea generation to remain in the mind map M_{d_i}.

After each Debater Agent generates an opinion, the Idea Generation Stage begins. The Moderator Agent selects nodes from the mind map to represent the core concept of an idea and chooses other nodes to combine with it, leveraging the benefits of concept combination using the mind map. Finally, the selected nodes are combined with the core concept based on the evaluation of the reflection step, resulting in the generation of new ideas.

By structuring the flow of discussion, SPARK-flow extracts elements from the discussion content that contribute to creative ideas and organizes them into ideas using a mind map.

3.4 Divergent and Convergent Thinking

According to Guilford, productive thinking can be categorized into two types: divergent and convergent thinking [19]. In divergent thinking, the ability to think in various directions is crucial, with traits such as fluency, flexibility, and originality being emphasized in creativity. On the other hand, convergent thinking is the process of consolidating thoughts into a unique answer. It is stated that divergent and convergent thinking do not occur entirely separately but often coexist.

SPARK-flow encourages divergent thinking by allowing Debater Agents to freely express opinions based on their role. The participation of multiple agents with diverse knowledge in the discussion enables thinking in more varied directions. Convergent thinking in SPARK-flow corresponds to the process of organizing these opinions into a mind map and generating ideas. To replicate the impact of convergent thinking on the subsequent process, the mind map generated during convergence is referenced by the Debater Agents when generating opinions. Additionally, this enables Debater Agents to reference the opinions of others, allowing them to supplement or develop the opinions of other agents.

Research has consistently shown that creative ideas are combinations of new elements or ideas with established technological components or methodologies [11, 10, 23]. In SPARK-flow, we apply this concept by summarizing the content of discussions into mind maps, which encourages the structural discovery of new relevancies and combinations. During idea generation, keywords are selected from the mind map, and related keywords are chosen to form new ideas. This approach enables the generation of new ideas unconstrained by existing concepts and the utilization of language model knowledge to make the ideas more concrete.

3.5 Reflection

The studies [21, 17] have consistently demonstrated the impact of evaluating ideas during the productive thinking phase. It has been found that a single evaluation of ideas can enhance the quality of subsequent ideas. To leverage the benefits of reflection, SPARK-flow implements reflection before mind mapping and idea generation, and references the results in later steps to enhance creativity. Additionally, SPARK-flow adopts multiple evaluation criteria. This approach encourages the generation of higher-quality ideas, even if the evaluation is low from some perspectives, by combining them with other elements.

Users can participate in the reflection step because all processes are conducted in natural language. By evaluating the generated ideas and opinions, users can clarify the ideas they are seeking, and the results affect the later process of SPARK-flow. This is expected to lead to the creation of ideas that meet the users' needs.

3.6 Settings of Experiments

Dataset In this experiment, SPARK-flow was executed with 30 to 50 themes generated by OpenAI's Chat Completions API (gpt-3.5-turbo). The cosine similarity between themes, calculated using embeddings from the Universal Sentence Encoder [8], averaged 0.56 with a standard deviation of 0.036, suggesting moderate diversity among the themes. In each step, specific prompts are given to the language model to execute SPARK-flow. The Debater Agents and the Moderator Agent were powered by gpt-3.5-turbo, each sequentially assuming three roles for opinion presentation. To eliminate the influence of roles and evaluation criteria, the comparative experiments utilized pre-generated common roles and three evaluation criteria ($N_{Criteria} = 3$). As a baseline method, we employ a technique that prompts gpt-3.5-turbo to generate creative ideas based on a given theme. In this paper, we refer to this approach as *Simple*. Simple has two variations: *Simple(w/ role)* and *Simple(w/o role)*, generating ideas based on a specific role and theme, and the theme alone, respectively.

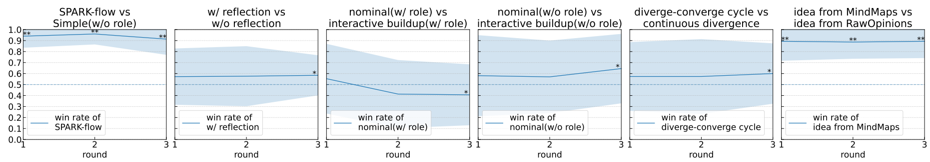
Idea Evaluation To compare the creativity of ideas generated by two different methods, ideas from each method were presented to OpenAI's Chat Completion API (gpt-4), which selected the more creative idea. By ensuring the source of the ideas was unidentifiable, we enabled an objective evaluation. It is known that explicitly asking for the rationale behind the evaluation improves the accuracy of the assessment [22]. Therefore, in this paper, GPT-4 not only selected ideas but also generated reasons for its choices. Approximately three ideas per theme and round are compared, with the win rate calculated for each theme and round. A one-sample t-test was conducted on win rates of all themes to test the null hypothesis that there is no difference in the quality of ideas generated by the two methods. If the null hypothesis is rejected and the win rate of one method exceeds 50%, that method is considered superior.

In addition to a comparison involving modifications to SPARK-flow, a common evaluation was conducted by comparing ideas from multiple methods against those generated by the baseline method, Simple(w/o role).

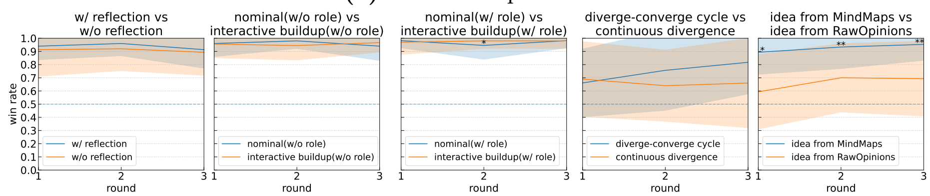
4 Experiments and Results

4.1 Comparative Evaluation of SPARK-flow and Simple Method in Idea Generation

This experiment aims to evaluate whether the ideas generated by SPARK-flow exhibit superior quality compared to those produced by a conventional method.



(a) Direct comparison.



(b) Baseline comparison.

Fig. 3: (a) Win rates for two methods over rounds, tested against a 50% baseline with p-values. (b) Win rate trajectories against Simple(w/o role) as baseline, with p-values assessing differences. The symbols * and ** indicate $p < 0.05$ and $p < 0.001$, respectively.

We compared the ideas generated by SPARK-flow with those generated by Simple(w/o role).

Figures 3 and Table 1 show the evolution of the win rate per round and the win rate, along with the percentage of themes with a win rate exceeding 50% in the final round, respectively. The win rate of SPARK-flow against Simple(w/o role) was 91.3% ($p < 0.01$), and since the win rate exceeded 50% in 96.7% of all themes, it was demonstrated that the ideas generated by SPARK-flow are more creative than those generated by Simple, regardless of the theme. Furthermore, when evaluating the impact of the presence or absence of roles on the creativity of ideas, the win rate of Simple(w/o role) against Simple(w/ role) was 55.1% (p=0.1), suggesting that the use of roles contributes to the generation of creative ideas.

4.2 Reflection Improves the Creativity of Ideas

In this experiment, we investigated the impact of reflection on the quality of ideas generated using a language model for ideation. The experiment compared the effects of the presence or absence of reflection and reference to the evaluation.

The results of this experiment suggest the potential for reflection to improve the quality of idea generation. The win rate when reflection was performed was statistically significant at 58.4% ($p < 0.05$), and given its low dependency on the theme, reflection is considered to play an important role in the idea generation process. In idea-generation systems utilizing language models, encouraging user reflection is key to leading to higher-quality ideas. These results indicate that SPARK-flow enables the generation of ideas through discussions among agents that cannot be obtained by a single agent.

Table 1: Win rates for Method A in direct comparisons against Method B across various themes

Method A	Method B	N_{Theme}	N_{Compare}	>50% Win Rate Theme Raio[%]
SPARK-flow	Simple(w/o role)	30	450	96.7
w/ reflection	w/o reflection	50	750	68.0
nominal(w/ role)	interactive buildup(w/ role)	50	450	32.0
nominal(w/o role)	interactive buildup(w/o role)	31	279	71.0
diverge-converge cycle	continuous divergence	50	450	66.0
idea from MindMaps	idea from RawOpinions	30	450	96.7

4.3 Comparing Nominal and Interactive Buildup Groups in Agent-Based Idea Generation

In the discussion methods for idea generation, Girotra et al. suggest that groups that discuss collectively from the beginning, termed "nominal groups," are less effective in generating superior ideas compared to groups that engage in individual thinking followed by group discussion, which they refer to as "interactive buildup" [18]. This experiment aims to verify whether this insight is also applicable to discussions involving language models. When Debater Agents generate opinions, in the nominal group scenario, they do not reference the opinions of others, whereas, in the interactive buildup scenario, they continuously form opinions while referencing the opinions of others. Furthermore, when generating ideas, all opinions are referenced and aggregated into the final idea. Additionally, the impact of roles was considered, and comparisons were made in the absence of roles.

With roles, the interactive buildup showed superior results, with a win rate of 59.3% ($p<0.05$), which is a trend opposite to that of [18]. In contrast, without roles, the nominal group showed a better win rate of 59.9% ($p<0.05$), consistent with the results of [18].

Assigning roles to language models can be assumed to bias their outputs. Therefore, in an interactive buildup with roles, opinions can be presented from different perspectives. Furthermore, the experiment by [18] demonstrated that interactive buildup facilitates the formation of ideas through the accumulation of others' opinions. Similarly, observations confirm that discussions utilizing language models also enable the supplementation of opinions presented by specific roles with insights from different viewpoints. On the other hand, when agents are not assigned roles, they tend to answer the given task more correctly, resulting in a lack of diversity in the discussion. Furthermore, discussions by language models differ from those by humans in that they do not experience stagnation.

Given this characteristic, it is expected that interactive buildup among agents with roles will provide the most beneficial impact by offering users diverse perspectives in idea generation.

4.4 Balancing Divergence and Convergence

In this experiment, we investigated how the balance between divergence and convergence in discussions among debating agents influences the quality of idea generation. In actual communication, discussions often oscillate between divergent modes, similar to brainstorming, and convergent modes, which involve detailed opinions and feedback. The effectiveness of these approaches in ideation using language models is still not fully understood. In standard interactions with existing language models, the user typically controls the direction of the discussion. However, there remain challenges in facilitating efficient and independent debates between agents. Our experiment specifically examined the role of agent collaboration in managing this balance for enhanced idea generation.

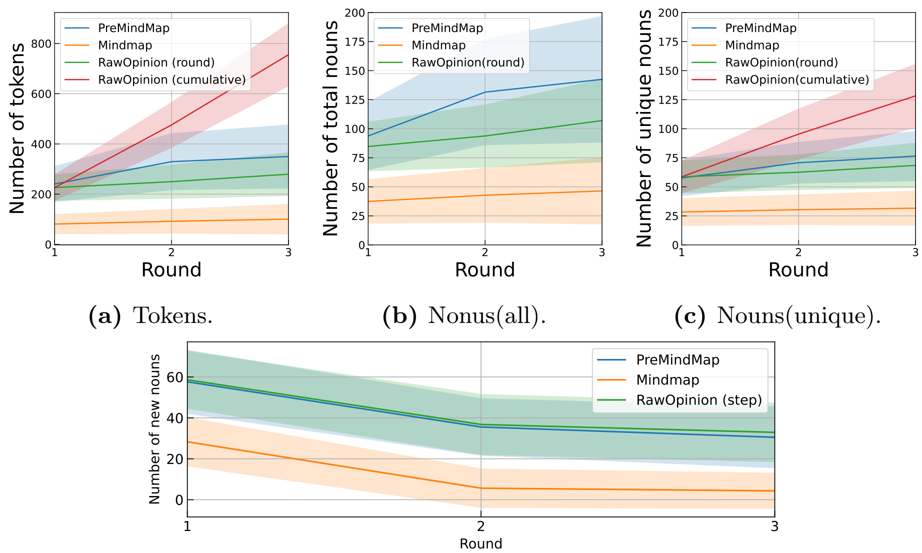
To compare the balance of divergence and convergence, we divided into two groups: (1) employing alternating divergence and convergence and (2) continuously diverging. The first group's agents would reflect and consolidate opinions in a mind map after each discussion, while the latter group's agents would immediately proceed to the next discussion without this consolidation step. To align the two groups, ideas were generated from opinions, not mind maps. We focused on the impact of divergence and convergence on discussion content, selecting ideation keywords exclusively from the compiled mind maps, and deliberately not generating related keywords using the language model's knowledge.

The results of the experiment revealed a notable trend: the group alternating between divergence and convergence showed an improvement in idea quality over time. This was evidenced by a statistically significant advantage in idea quality compared to a simpler approach without these dynamics (win rate: 72.3%, $p = 4.00 \times 10^{-6}$). On the contrary, the group engaging in continuous divergence demonstrated a decrease in idea quality over time, with a lower win rate of 66.3% ($p = 5.50 \times 10^{-2}$) compared to the Simple(w/o role).

The findings highlight the increasing significance of the convergence process as discussions progress. Furthermore, the large variance in win rates compared to Simple (w/o) for both methods indicates that generating ideas from opinions, rather than from a mind map, leads to unstable idea quality. This suggests the importance of organizing information from discussions into a mind map before generating ideas.

4.5 Enhancing Idea Generation through Mind Mapping

The comparative analysis conducted in this experiment focused on the use of mind maps for idea generation versus the method of directly referencing the opinions of Debater Agents. The aim was to assess how different methods of summarizing discussion content affect the quality of the generated ideas. Exceptionally, in the Idea Generation Stage, keyword generation related to the



(a) Tokens. **(b)** Nonus(all). **(c)** Nouns(unique).

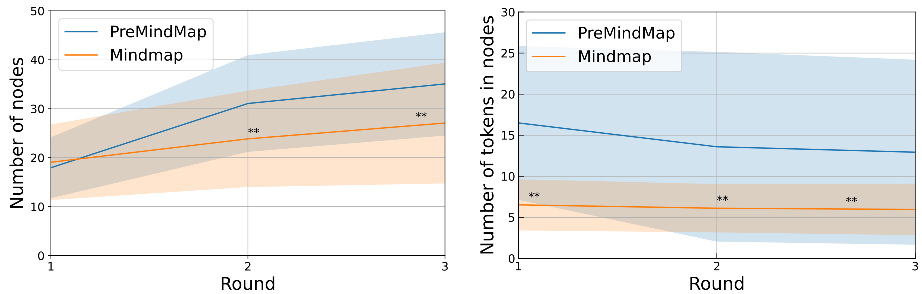
(d) New nouns.

Fig. 4: Sequential evolution of tokens and nouns in mind maps over three rounds. (a) Token counts per round, (b) Total noun counts, including repetitions, (c) Count of unique nouns, excluding repetitions, (d) New nouns not previously appearing. The 'RawOpinion (round)' indicates new elements per round, and 'Opinion (cumulative)' shows the aggregate up to each round.

discussion topics was omitted to maintain consistency with the original SPARK-flow, except for this alteration. When referencing opinions directly, all opinions presented by the Debater Agents up to that point were considered.

Idea generation using mind maps demonstrated a significantly higher win rate (89.3%, $p<0.05$) compared to the method of direct opinion reference. Moreover, the variance in win rates, when compared to Simple(w/o role), indicated a variance of 12.3% for ideas generated from mind maps, as opposed to 28.6% for those derived from direct opinion referencing. This disparity suggests that mind maps are more effective in consistently yielding creative ideas.

The enhancement in idea quality attributable to mind maps is linked to a reduction in information volume and the facilitation of concept association. Figure 4 shows the transition in the number of tokens and nouns in mind maps and opinions for each round. Direct referencing led to a per-round increase of approximately 200 tokens, while the increase for mind map-contained tokens was limited to about 30 tokens. This disparity underscores the efficiency of mind maps in organizing information, thereby reducing the volume of necessary information and stabilizing idea creativity. Figure 6 analyzes the hierarchical structure of nodes in the mind map and the relationships between nodes. As the hierarchy deepens, the similarity between nodes with the same parent increases, while the



(a) Nodes. **(b)** Tokens.

Fig. 5: (a) Trends in node counts across rounds, with t-test p-values for differences between Mindmap and PreMindMap. (b) Token counts per node, with significance tested similarly to (a).

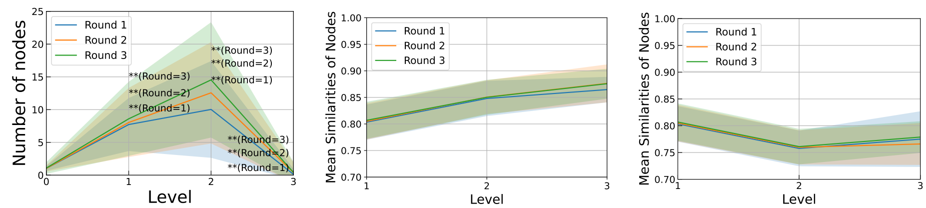
similarity between nodes with different parents decreases. This indicates that using a mind map makes it easier to combine related concepts with structural hints, facilitating the generation of new ideas. The ease of combining concepts considering their distance is believed to contribute to the originality of ideas.

4.6 Integrating Debater Agent Opinions into a Mind Map

This study analyzes how opinions generated by Debater Agents in SPARK-flow are organized into a mind map. Figure 4(a) shows an increasing trend in the number of tokens for RawOpinion, PreMindMap, and MindMap across successive rounds, indicating that the amount of information grows with each round. This trend can be attributed to the Debater Agents referencing the opinions of previous agents and generating opinions with an increased amount of information based on those references.

While PreMindMap and MindMap are structurally similar, there is a significant difference in their token counts. The token count for PreMindMap is approximately 300, whereas, for MindMap, it is about 100. This indicates that SPARK-flow references a large amount of information before summarizing it into the MindMap.

As shown in Figure 5, the comparison between the number of nodes in the MindMap and the PreMindMap indicates that the number of nodes in the MindMap is maintained from the PreMindMap. However, as demonstrated in Figure 5(b), the amount of text contained in each node is reduced. This organization of information contributes to the efficiency of presenting information to the user and the handling of necessary information during idea creation. Presenting the RawOpinion directly to the user would require referencing the output content of all Debater Agents. This leads to a tendency towards information overload, with an increase of approximately 300 tokens per round, as shown in



(a) Nodes by level. **(b)** Intra-parent node. **(c)** Inter-parent node.

Fig. 6: (a) Number of nodes per Level in mind maps, with p-values from t-tests comparing node counts of preceding Levels. The symbols * and ** indicate $p < 0.05$ and $p < 0.001$, respectively. (b) Similarity between nodes sharing the same parent, and (c) similarity between nodes with different parents, respectively.

Figure 4(a). In contrast, the MindMap limits the increase in token count to about 100, facilitating easier handling of information.

Figure 4(c) shows the round-by-round transition of the number of unique nouns, revealing that the number of unique nouns increases with each round for both methods. This suggests that the generation of opinions by multiple Debater Agents is increasing the number of concepts considered. Furthermore, when comparing the difference between the total number of nouns shown in 4(b) and the number of unique nouns, the magnitude of this difference is smaller for MindMap than for RawOpinion, indicating that MindMap significantly reduces the duplication of concepts contained in RawOpinion. Additionally, as shown in Figure 4(d), the number of newly introduced nouns is proportional between RawOpinion and MindMap. This explains why the token count in MindMap does not increase significantly, suggesting that the number of concepts presented by Debater Agents decreases and the information selected is narrowed down.

4.7 Structural Analysis of Mind Maps

Creative ideas are characterized by the integration of new concepts with existing combinations [11, 23, 10]. Additionally, as demonstrated in Section 4.5, utilizing mind maps facilitates the generation of creative ideas. This section investigates how the use of mind maps, through the analysis of their structural characteristics, contributes to the generation of creative ideas.

In this study, we introduce a hierarchical structure for analyzing mind maps, where the root node is arbitrarily designated as $Level = 0$. This categorization facilitates our examination of the distribution and similarity of nodes across different levels. Findings reveal a notable peak in the number of nodes at Level $=$ 2, as depicted in Figure 6(a), followed by a substantial decrease at Level $= 3$, where the presence of nodes is markedly sparse. While the number of nodes at Level $= 2$ increases with each round, there is no significant change in the number of nodes at Level $= 1$. Comparing the embedding vectors of the text in nodes at Level $= 1$ across rounds, the average cosine similarity between the most similar nodes is 0.99 with a standard deviation of 9.6×10^{-5}, indicating that the content of nodes at Level = 1 remains consistent across rounds. This suggests a strategy by the Moderator Agent during the creation of mind maps to reference mind maps from previous rounds and not alter the nodes at Level = 1, thereby maintaining the structure of the mind map. This approach likely contributes to the observation that the quality of ideas does not significantly vary from round to round, offering the advantage of providing users with organized information that does not undergo drastic changes, making it easier to comprehend.

Semantically interpreting the hierarchical structure of mind maps, Level = 1 represents broad perspectives related to the theme, while Level = 2 provides detailed opinions on those perspectives. The transition in the number of nodes across rounds reflects that in SPARK-flow generated mind maps, broad perspectives are formed in the initial rounds, followed by the addition of specific opinions in subsequent rounds. The analysis of the number of tokens in nodes presented in Figure 6 reveals that, regardless of the Level, there is no significant difference in the number of tokens contained in nodes, with an average of 5 tokens. This indicates that each node in the mind map consists of sentences of a certain length, rather than just words.

By comparing the similarity of embedding vectors of node texts, as shown in Figures 6(b) and 6(c), it is evident that nodes sharing the same parent node exhibit high similarity, while nodes with different parent nodes show low similarity. This suggests that combining nodes with the same parent node corresponds to combining closely related concepts, while combining nodes with different parent nodes involves merging distant concepts, indicating that the structure of mind maps is an effective tool for generating creative ideas.

5 Conclusion and Future Work

In conclusion, this study introduced the SPARKIT system and its algorithm, SPARK-flow, aimed at enhancing idea generation. The experimental results have shown that SPARKIT outperforms simpler methods in generating superior ideas, and the use of language models for idea generation aligns well with human-generated ideas in various aspects. The employment of mind maps was particularly highlighted for its ability to improve the quality of ideas. However, while the results are promising, there are concerns about the reproducibility of results due to differences in the LLMs used. This issue can be addressed by fixing the model or using open-source models.

For future work, we envision further enhancing the capabilities of language models through the application of reinforcement learning, as well as forming multiple agent groups to promote idea generation from diverse perspectives. This approach is expected to yield more creative and user-centric ideas.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 technical report. arXiv

preprint arXiv:2303.08774 (2023)
2. Al-Samarraie, H., Hurmuzan, S.: A review of brainstorming techniques in higher education. Thinking Skills and creativity **27**, 78–91 (2018)
3. Andolina, S., Klouche, K., Cabral, D., Ruotsalo, T., Jacucci, G.: Inspirationwall: supporting idea generation through automatic information exploration. In: Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition. pp. 103–106 (2015)
4. Andolina, S., Schneider, H., Chan, J., Klouche, K., Jacucci, G., Dow, S.: Crowdboard: augmenting in-person idea generation with real-time crowds. In: Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition. pp. 106–118 (2017)
5. Benedek, M., Neubauer, A.C.: Revisiting mednick's model on creativity-related differences in associative hierarchies. evidence for a common path to uncommon thought. The Journal of creative behavior **47**(4), 273–289 (2013)
6. Buzan, T., Buzan, B.: The mind map book. Pearson Education (2006)
7. Camburn, B., Arlitt, R., Anderson, D., Sanaei, R., Raviselam, S., Jensen, D., Wood, K.L.: Computer-aided mind map generation via crowdsourcing and machine learning. Research in Engineering Design **31**, 383–409 (2020)
8. Cer, D., Yang, Y., Kong, S.y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., Kurzweil, R.: Universal sentence encoder for English. In: Blanco, E., Lu, W. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 169–174. Association for Computational Linguistics, Brussels, Belgium (Nov 2018). https://doi.org/10.18653/v1/D18-2029, https://aclanthology.org/D18-2029
9. Chan, C.M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., Liu, Z.: Chateval: Towards better llm-based evaluators through multi-agent debate. arXiv preprint arXiv:2308.07201 (2023)
10. Chan, J., Dow, S.P., Schunn, C.D.: Do the best design ideas (really) come from conceptually distant sources. Engineering a Better Future: Interplay between Engineering, Social Sciences, and Innovation p. 111 (2018)
11. Chan, J., Schunn, C.D.: The importance of iteration in creative conceptual combination. Cognition **145**, 104–115 (2015)
12. Chen, T.J., Krishnamurthy, V.R.: Investigating a mixed-initiative workflow for digital mind-mapping. Journal of Mechanical Design **142**(10), 101404 (2020)
13. Di Fede, G., Rocchesso, D., Dow, S.P., Andolina, S.: The idea machine: LLM-based expansion, rewriting, combination, and suggestion of ideas. In: Proceedings of the 14th Conference on Creativity and Cognition. pp. 623–627 (2022)
14. Diedrich, J., Benedek, M., Jauk, E., Neubauer, A.C.: Are creative ideas novel and useful? Psychology of aesthetics, creativity, and the arts **9**(1), 35 (2015)
15. Dow, S.P., Heddleston, K., Klemmer, S.R.: The efficacy of prototyping under time constraints. In: Proceedings of the seventh ACM conference on Creativity and cognition. pp. 165–174 (2009)
16. Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., Mordatch, I.: Improving factuality and reasoning in language models through multiagent debate. arXiv preprint arXiv:2305.14325 (2023)
17. Georgiev, G.V., Georgiev, D.D.: Enhancing user creativity: Semantic measures for idea generation. Knowledge-Based Systems **151**, 1–15 (2018). https://doi.org/https://doi.org/10.1016/j.knosys.2018.03.016, https://www.sciencedirect.com/science/article/pii/S0950705118301394

18. Girotra, K., Terwiesch, C., Ulrich, K.T.: Idea generation and the quality of the best idea. Management science **56**(4), 591–605 (2010)
19. Guilford, J.P.: Creative abilities in the arts. Psychological review **64**(2), 110 (1957)
20. Hackl, V., Müller, A.E., Granitzer, M., Sailer, M.: Is GPT-4 a reliable rater? evaluating consistency in gpt-4 text ratings. arXiv preprint arXiv:2308.02575 (2023)
21. Hao, N., Ku, Y., Liu, M., Hu, Y., Bodner, M., Grabner, R.H., Fink, A.: Reflection enhances creativity: Beneficial effects of idea evaluation on idea generation. Brain and cognition **103**, 30–37 (2016)
22. Kern, F.B., Wu, C.T., Chao, Z.C.: Assessing novelty, feasibility, and value of creative ideas with an unsupervised approach using GPT-4 (2023)
23. Kim, D., Cerigo, D.B., Jeong, H., Youn, H.: Technological novelty profile and invention's future impact. EPJ Data Science **5**(1), 1–15 (2016)
24. Leeds, A.J., Kudrowitz, B., Kwon, J.: Mapping associations: exploring divergent thinking through mind mapping. International Journal of Design Creativity and Innovation **7**(1-2), 16–29 (2019)
25. Luo, J., Sarica, S., Wood, K.L.: Computer-aided design ideation using innogps. In: International design engineering technical conferences and computers and information in engineering conference. vol. 59186, p. V02AT03A011. American Society of Mechanical Engineers (2019)
26. Ma, K., Grandi, D., McComb, C., Goucher-Lambert, K.: Conceptual design generation using large language models. arXiv preprint arXiv:2306.01779 (2023)
27. Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. arXiv preprint arXiv:2305.18290 (2023)
28. Reiter-Palmon, R., Forthmann, B., Barbot, B.: Scoring divergent thinking tests: A review and systematic framework. Psychology of Aesthetics, Creativity, and the Arts **13**(2), 144 (2019)
29. Runco, M.A., Acar, S.: Divergent thinking as an indicator of creative potential. Creativity research journal **24**(1), 66–75 (2012)
30. Setiyawan, D.: Improving students' speaking skills in generating idea through new concept of mind mapping technique. In: International Conference on Educational Research and Innovation (ICERI 2019). pp. 227–231. Atlantis Press (2020)
31. Shih, P.C., Nguyen, D.H., Hirano, S.H., Redmiles, D.F., Hayes, G.R.: GroupMind: supporting idea generation through a collaborative mind-mapping tool. In: Proceedings of the 2009 ACM International Conference on Supporting Group Work. pp. 139–148 (2009)
32. Sun, M., Wang, M., Wegerif, R., Peng, J.: How do students generate ideas together in scientific creativity tasks through computer-based mind mapping? Computers & Education **176**, 104359 (2022)
33. Tao, M., Xie, R.: Mind map based computer network knowledge graph visualization research and application. In: International Symposium on Emerging Technologies for Education. pp. 3–12. Springer (2021)
34. Wahl, J., Hutter, K., FÜLLER, J.: How ai-supported searches through other perspectives affect ideation outcomes. International Journal of Innovation Management **26**(09), 2240028 (2022)
35. Wang, H.C., Cosley, D., Fussell, S.R.: Idea expander: supporting group brainstorming with conversationally triggered visual thinking stimuli. In: Proceedings of the 2010 ACM conference on Computer supported cooperative work. pp. 103–106 (2010)
36. Wen, Y., Wang, Z., Sun, J.: Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. arXiv preprint arXiv:2308.09729 (2023)

37. Xu, B., Yang, A., Lin, J., Wang, Q., Zhou, C., Zhang, Y., Mao, Z.: Expertprompting: Instructing large language models to be distinguished experts. arXiv preprint arXiv:2305.14688 (2023)
38. Zampetakis, L.A., Tsironis, L., Moustakis, V.: Creativity development in engineering education: The case of mind mapping. Journal of Management Development **26**(4), 370–380 (2007)