Creativity in LLM-based Multi-Agent Systems: A Survey

Yi-Cheng Lin* **Kang-Chieh Chen*** **Zhe-Yan Li*** **Tzu-Heng Wu***
Tzu-Hsuan Wu* **Kuan-Yu Chen*** **Hung-yi Lee** **Yun-Nung Chen**
National Taiwan University, Taipei, Taiwan
{f12942075, r13944050}@ntu.edu.tw y.v.chen@ieee.org

Abstract

Large language model (LLM)-driven multi-agent systems (MAS) are transforming how humans and AIs collaboratively generate ideas and artifacts. While existing surveys provide comprehensive overviews of MAS infrastructures, they largely overlook the dimension of *creativity*, including how novel outputs are generated and evaluated, how creativity informs agent personas, and how creative workflows are coordinated. This is the first survey dedicated to creativity in MAS. We focus on text and image generation tasks, and present: (1) a taxonomy of agent proactivity and persona design; (2) an overview of generation techniques, including divergent exploration, iterative refinement, and collaborative synthesis, as well as relevant datasets and evaluation metrics; and (3) a discussion of key challenges, such as inconsistent evaluation standards, insufficient bias mitigation, coordination conflicts, and the lack of unified benchmarks. This survey offers a structured framework and roadmap for advancing the development, evaluation, and standardization of creative MAS.[1]

1 Introduction

Advances in LLMs and deep learning have fueled rapid growth in MAS research (Guo et al., 2024a; Tran et al., 2025). Single-agent pipelines, such as one-shot or simple iterative LLM prompting (Grattafiori et al., 2024; Wang et al., 2022), execute in isolation and often converge on familiar patterns, struggling to explore vast open-ended spaces. Unlike monolithic systems, a MAS comprises multiple autonomous entities: software agents, robots, or human-AI hybrids. This structure enables emergent collaboration and richer exploration of open-ended creative spaces (Park et al., 2023).

Here, *computational creativity* denotes the production of artifacts—ideas, behaviors, or solutions—that are both novel and valuable, showing meaningful utility or appeal rather than randomness (Wiggins, 2006; Veale and Cardoso, 2019). In MAS, creativity emerges through various dynamics—critique loops, competitive incentives, or coalition-forming. Together, these processes can yield outcomes designers never anticipated. For example, conversational agents can automate screenwriting: one agent as Writer drafting character profiles and outlines, another as Editor offering revision suggestions, and multiple Actors engaging in role-playing to improvise dialogues (Chen et al., 2024).

Although recent surveys examine LLM-based MAS architectures (Li et al., 2024; Han et al., 2024b), collaboration mechanisms (Tran et al., 2025; Zhang et al., 2024c; Mu et al., 2024), autonomy and alignment (Händler, 2023), communication protocols (Yan et al., 2025), and environment/simulation platforms (Guo et al., 2024a; Gao et al., 2024), they concentrate on infrastructure. However, they overlook evaluating creative outputs, the impact of agent personas and workflow integration on creativity, and the specific techniques that drive ideation. We present the first survey on creativity in LLM-based MAS to bridge this gap. Our paper systematically maps techniques, datasets, evaluation metrics, and remaining challenges, offering researchers a unified framework to assess and amplify creativity across multi-agent pipelines.

This survey focuses on systems whose inputs and outputs span text and images, and whose participants range from LLM-based chatbots to human agents, as in Fig. 1. We aim to map the current landscape of techniques, datasets, evaluations, and challenges to foster and measure creativity in such multimodal and heterogeneous systems. By analyzing how different agents interact, we reveal how collaborative structures can unlock creative potentials that exceed what isolated LLMs or individuals can achieve.

*These authors contributed equally.

[1] https://github.com/MiuLab/MultiAgent-Survey

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 27584–27607
November 4-9, 2025 ©2025 Association for Computational Linguistics

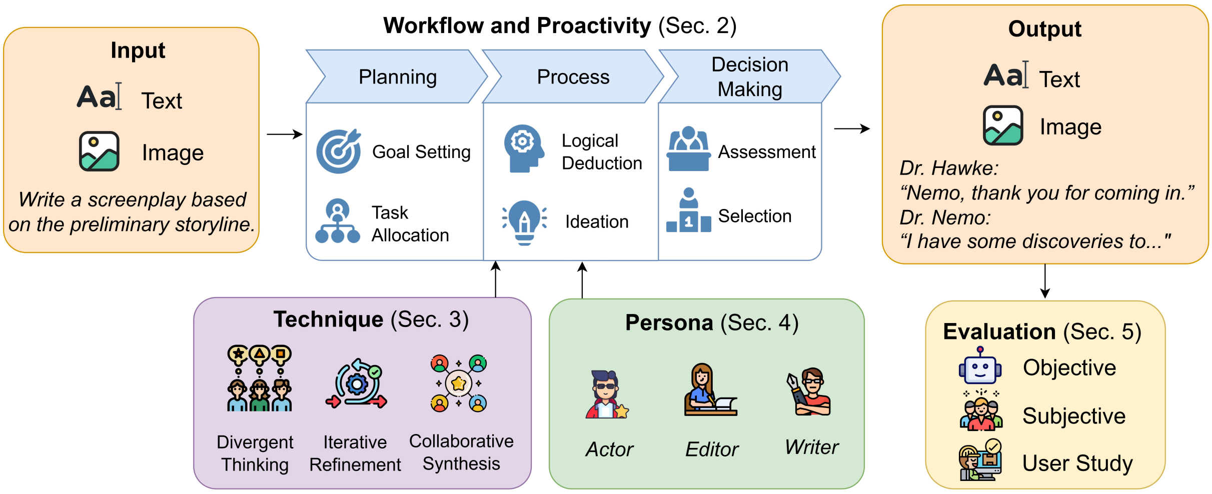


Figure 1: Overview of multi-agent creativity systems. Given user inputs in text or image form, agents engage in a three-stage process: Planning, Process, and Decision Making, using a variety of techniques (Sec. 3) and persona (Sec. 4), with outputs evaluated both subjectively by humans and objectively by automated metrics (Sec. 5).

2 MAS Workflow and Proactivity

2.1 MAS Workflow

Recent work shows that LLMs can generate novel content, yet a clear creativity gap exists between human designers and agents based on LLM (He et al., 2024). Therefore, most existing creativity support systems keep humans "in the loop", asking users to critique or complement machine-generated ideas (Shaer et al., 2024; Radensky, 2024; Lin et al., 2022; Lataifeh et al., 2024; Zhang et al., 2022). This also reflects on a focus that utilizes agents to imitate human behavior and replace their role in MAS (Xu et al., 2024; Sun et al., 2024). As human-agent collaboration becomes more sophisticated, it becomes increasingly important to consider *when* and *how* agents should be involved within the system's workflow. To reason about this question, we decompose the creative workflow of MAS into three key phases: *Planning*, *Process*, and *Decision Making* (Xie and Zou, 2024; Mukobi et al., 2023).

- *Planning*: where Agents formulate objectives and structure task execution.
- *Process*: where Agents implement tasks and coordinate through interaction.
- *Decision Making*: where Agents evaluate options and determine outcome

Real-world LLM-based MAS often interleave these steps. For instance, **StoryVerse** combines author-defined outlines with emergent character simulations through iterative narrative planning loops (Wang et al., 2024b), while **Generative Agents** integrate observation, planning, and reflection in overlapping processes (Park et al., 2023). In contrast, we keep these steps distinct to ensure our framework remains clear and easy to follow.

2.2 Spectrum of Agent Proactivity

We define an LLM agent's *proactivity* as the degree to which it initiates, guides, and owns creative actions within a MAS. Proactivity combines two facets—*initiative* (who starts or extends an action) and *control* (who judges whether the action is satisfactory)—and lies on a continuum from *reactive* agents, which wait for explicit prompts and follow specified instructions, to *proactive* agents, which formulate sub-goals, dispatch subtasks, and self-evaluate without human cues.

Planning In the *Planning* phase, the system defines *what* needs to be done before any content is generated. This typically involves (1) setting high-level objectives, (2) decomposing the overall goal into subtasks, and (3) configuring the downstream generation pipeline. To ensure predictability, most MAS frameworks delegate these responsibilities to humans because of their natural reliability. (Fan et al., 2024; Zhang et al., 2022; Ge et al., 2025; Zhang and Arawjo, 2025).

However, a few studies have set about addressing *Planning* subtasks through agents to alleviate the burden on human users (Venkadesh et al., 2024; Zhai et al., 2025; Venkatesh et al., 2025). For example, **VirSci** (Su et al., 2025) uses an autonomous "team leader" agent to select collaborators, define research topics, and orchestrate task distribution based on a researcher database. These agent-driven planning frameworks lean toward the proactive end of our spectrum, empowering agents to au-

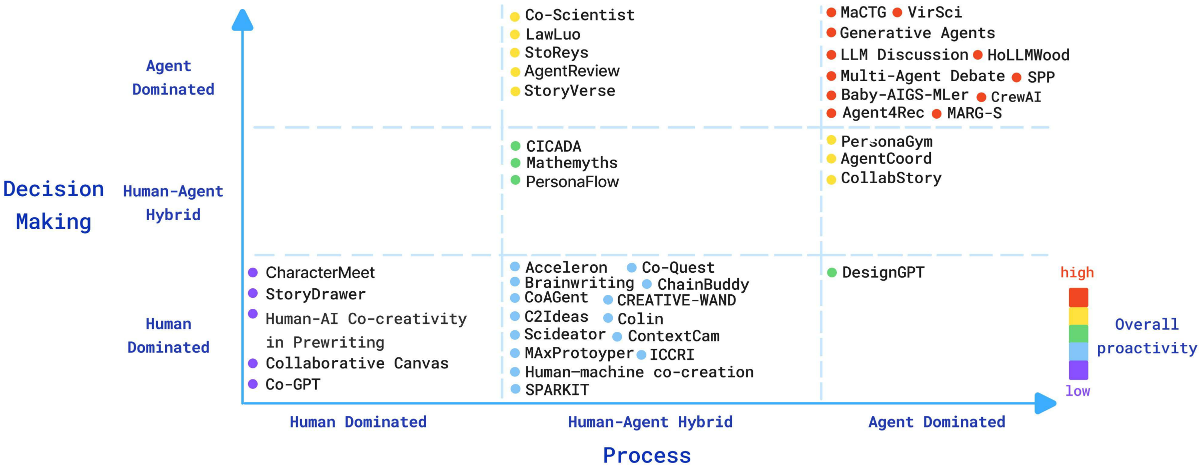


Figure 2: A discrete categorization distinguishes MAS frameworks by agent proactivity across the *Process* and *Decision-Making* phases, where the *Planning* phase is omitted due to consistently low proactivity in existing studies. Otherwise, the color gradient represents overall proactivity, with criteria detailed in Appendix A.

tonomously formulate and allocate tasks while humans retain only the overarching goal-setting.

Process In the *Process* phase, agents execute the generation pipeline by creating intermediate artifacts, utilizing methods such as peer sharing or refining them in response to feedback. Highly proactive systems instantiate multiple agents that drive every step without human steering: they launch subtasks, critique each other's outputs, and merge the results into a cohesive artifact. For example, **LLM Discussion** (Lu et al., 2024) assigns distinct personas to agents that autonomously activate the commands of others, debate ideas, and converge to final proposals.

Conversely, low-proactivity systems require humans to inject prompts or corrective instructions at each stage, with agents simply executing the specified commands (Hou et al., 2024). Fig. 2 visualizes this continuum: from fully autonomous, agent-only pipelines to human-in-the-loop workflows where agents act in a strictly supportive role. The grading change is conveyed through color labels, unfolding in a rainbow gradient. Each color label clusters frameworks according to their level of proactivity, revealing the trade-off between controllability and contextual load.

Decision Making The *Decision Making* phase evaluates and selects among the artifacts produced in the *Process* phase, thus revealing who ultimately controls the creative outcome. At the low-proactivity end, humans retain full evaluative authority. For example, **Scideator** (Radensky, 2024) presents users with candidate hypotheses and allows them to iteratively review, modify, and validate each idea against the literature. Moving toward higher proactivity, some systems embed a dedicated evaluator agent: Liang et al. (2024) introduces a "judge" agent that scores outputs on creativity and quality, only forwarding those that exceed a predefined threshold. Finally, purely loss-driven selection such as **CICADA** (Ibarrola et al., 2024), a co-creative agent proposed in **Drawing with Reframer** (Lawton et al., 2023), automates decision-making via implicit LLM optimization. Although loss-based metrics help ease the burden on humans, we still classify such methods as *low–mid proactivity* because they lack explicit, actor-driven assessment by an independent agent.

2.3 Creativity Analysis on Proactivity

Empirical studies reveal a trade-off between agent proactivity, creative diversity, and user trust. **Collaborative Canvas** (He et al., 2024) shows that excessive AI-initiated suggestions can collapse the idea space, producing homogeneous outputs. The **Co-Quest** interface (Liu et al., 2024b) demonstrates that boosting agent initiative increases idea volume but erodes user satisfaction and trust, highlighting the need for transparent, interpretable agents. Furthermore, precision-critical tasks (e.g. automated theorem proving) demand low proactivity to ensure correctness (Song et al., 2025), with humans retaining evaluative authority to guarantee reliability and accountability. Overall, agent proactivity accelerates ideation without undermining user agency, whereas sustained high proactivity risks over-reliance, reduced creative independence, and trust deficits (Chakrabarty et al., 2024). Future MAS should therefore adaptively calibrate proactivity to task demands and user preferences.

3 MAS Techniques for Creativity

MAS enhances creativity by dividing the cognitive workload, such as idea generation, evaluation, and coordination, across specialized agents. For example, some agents focus on quickly generating a wide range of ideas, others evaluate the feasibility and coherence of those ideas, and another set of agents helps guide the overall workflow through multiple iterations. Unlike single-LLM models like GPT-3 (Brown et al., 2020), which typically generate outputs in a single step, MAS frameworks achieve greater novelty and higher-quality solutions by enabling structured and collaborative processing. For example, **CoQuest** (Liu et al., 2024b) integrates multiple agents into an interactive workflow that combines wide idea exploration, focused deepening of promising directions, and organized feedback. This coordinated setup significantly enhances user creativity and their sense of control.

Below, we outline three core MAS techniques—*Divergent Exploration*, *Iterative Refinement*, and *Collaborative Synthesis* by explaining the cognitive rationale and algorithmic structure behind them, with brief references to detailed case studies provided in Appendix B.

3.1 Divergent Exploration

Divergent exploration emphasizes generating various ideas without applying early filters or judgment (Guilford, 1950; Wallach and Kogan, 1965). MAS supports this process by giving each agent a distinct perspective, prompt style, or domain of knowledge, allowing them to explore different creative directions independently. This helps avoid early narrowing and encourages novel outcomes.

One example is the **Group-AI Brainwriting** (Shaer et al., 2024). First, people come up with their own ideas. Then, GPT-3 adds new versions and expands on those ideas. After that, the team brings together both the human and AI ideas, and works on improving them. Finally, GPT-4 gives feedback by judging how original and insightful the ideas are. In this setup, GPT-3 helps with idea generation, and GPT-4 helps with checking quality. This shows how different agents can do different jobs, and how working together with AI can improve creative results.

Other systems take similar approaches. **Co-GPT Ideation** (Lim and Perrault, 2024) broadens idea diversity in fast-paced brainstorming. **ICCRI** (Ali et al., 2025) supports co-creation between children and robots over multiple sessions. Meanwhile, Kumar et al. (2025) raised concerns that repeated LLM use may reduce long-term originality if users become too reliant on automated suggestions.

3.2 Iterative Refinement

Iterative refinement involves progressively enhancing ideas through repeated feedback and revision cycles. In MAS, this process is facilitated by assigning distinct roles to agents, such as proposer, reviewer, and implementer, who work together in cycles to improve initial drafts into polished results.

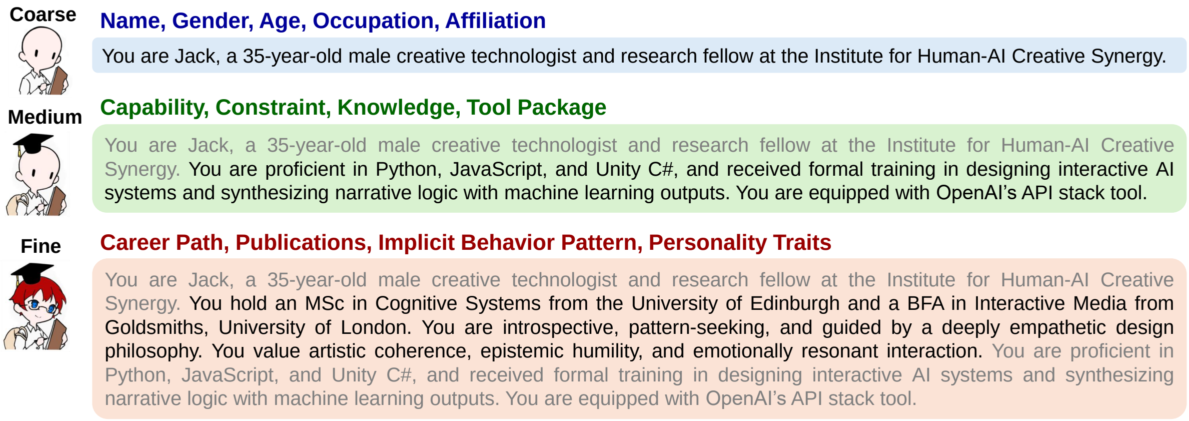
An example is **HoLLMwood** (Chen et al., 2024), a system for collaborative screenwriting. It defines three agent roles: a *Writer* generates the script, an *Editor* offers suggestions, and an *Actor* simulates character behavior to check tone and consistency. The process continues iteratively until agents either converge on a shared solution or satisfy a predefined stopping condition, such as a fixed number of iterations or convergence in output. This collaborative loop results in richer character development and a more coherent story structure compared to outputs from a single LLM.

Refinement strategies differ across systems. Some impose a fixed number of cycles, while others adapt dynamically based on user input or inter-agent consensus. These strategies have shown effectiveness beyond creative writing. For instance, **DesignGPT** (Ding et al., 2023) applies iterative refinement to product design, **Baby-AIGS-MLer** (Zijun Liu, 2024) uses it to construct machine learning pipelines, and the **Multi-agent Debate Framework** (Liang et al., 2024) leverages it for logical reasoning tasks.

3.3 Collaborative Synthesis

Collaborative synthesis focuses on integrating diverse agent perspectives into coherent, high-level outputs. Agents are often given roles like planner, critic, or synthesizer, and they work together in structured conversations or workflows. This approach is beneficial for tasks requiring both creative exploration and logical organization.

A prime example is **MaCTG** (Zhao et al., 2025), built for software engineering at scale. The system organizes agents into two levels: some are assigned to functional modules (horizontal), while others handle planning and integration across the system (vertical). Responsibilities are divided:


Coarse
Name, Gender, Age, Occupation, Affiliation

You are Jack, a 35-year-old male creative technologist and research fellow at the Institute for Human-AI Creative Synergy.

Medium
Capability, Constraint, Knowledge, Tool Package

You are Jack, a 35-year-old male creative technologist and research fellow at the Institute for Human-AI Creative Synergy. You are proficient in Python, JavaScript, and Unity C#, and received formal training in designing interactive AI systems and synthesizing narrative logic with machine learning outputs. You are equipped with OpenAI's API stack tool.

Fine
Career Path, Publications, Implicit Behavior Pattern, Personality Traits

You are Jack, a 35-year-old male creative technologist and research fellow at the Institute for Human-AI Creative Synergy. You hold an MSc in Cognitive Systems from the University of Edinburgh and a BFA in Interactive Media from Goldsmiths, University of London. You are introspective, pattern-seeking, and guided by a deeply empathetic design philosophy. You value artistic coherence, epistemic humility, and emotionally resonant interaction. You are proficient in Python, JavaScript, and Unity C#, and received formal training in designing interactive AI systems and synthesizing narrative logic with machine learning outputs. You are equipped with OpenAI's API stack tool.

Figure 3: **Categories of Persona Granularity:** A conceptual framework illustrated with selected attributes, accompanied by a concise example representing each defined persona.

some agents write or test code, others resolve logic gaps or coordinate dependencies. A planning module manages task flow, and outputs are validated at different levels, from individual functions to the full project. It assigns high-level reasoning to **DeepSeek-V3** (DeepSeek-AI, 2024), while code generation is handled by the more lightweight **Qwen2.5-Coder-7B** (Hui et al., 2024). This setup ensures results are not only creative but also feasible and efficient.

Other frameworks reflect similar goals. **CollabStory** (Venkatraman et al., 2024) supports multi-agent storytelling. **CoQuest** (Liu et al., 2024b) helps researchers collaboratively explore academic questions. **Human-AI Co-creativity** (Wan et al., 2024) assists with early stage writing. These examples show how agent collaboration can result in well-rounded and structurally sound creative work.

3.4 Limitations in Creativity Methods

The scope of *Divergent Exploration*, *Iterative Refinement*, and *Collaborative Synthesis* is inherently constrained by computing resources and task complexity. For example, Kumar et al. (2025) showed that LLM support can broaden idea space in divergent exploration, but may reduce independent performance in subsequent tasks. In iterative refinement, limited resources restrict the number of revisions, and excessive iterations may lead to unproductive fixation or overly narrow search paths. Collaborative synthesis, while useful for aligning ideas, can dilute originality when too many contributors are involved, making it difficult to ensure both creativity and coherence in execution.

4 Persona and Agent Profile

Drawing from the idea of chain-of-thought (Wei et al., 2022), persona-based approaches seek to harness the logical coherence of intelligent agents to diversify idea generation, while maintaining consistency within specific configurations of designed persona and avoiding whimsical deviations. Conceptually, this extends from chain of thought reasoning in event logic to behavioral reasoning within character modeling. The appropriate design of agent profiles efficiently improves complex problem-solving (Gabriel, 2020; Hu and Collier, 2024) and innovation by supporting collaborative synthesis (Samuel et al., 2024). Within the established interaction framework, multiple agents can participate in collaborative processes such as social simulation and hierarchical collaboration (Park et al., 2023). These approaches also integrate knowledge in diverse domains and strategic interaction, thereby fostering creative and adaptive problem-solving. The overview of references with respect to the persona can be found in Table 3.

4.1 Granularity of Persona

We characterize persona design along a *granularity spectrum* that indicates how much detail is embedded in the agent's profile (Fig. 3). Granularity governs both controllability and diversity: coarse profiles favor breadth and spontaneous idea generation, whereas fine-grained profiles offer precise, predictable behavior at the expense of flexibility.

Coarse-Grained Persona Agents carry only high-level identity or expertise labels (e.g. "marketing strategist," "data analyst"). This minimal specification tolerates ambiguity, fostering diverse idea generation across fewer constraints. For example, **Solo Performance Prompting** (Wang et al., 2024c) assigns expert-role tags and later merges independent outputs into a unified solution, capitalizing on varied perspectives without prescribing detailed behavior. However, coarse profiles can produce shallow or inconsistent contributions. Under-specified roles may generate irrelevant or incoherent ideas when finer guidance is needed (Cemri et al., 2025).

Medium-Coarse Persona Medium-coarse profiles enhance basic role labels with concise, domain-relevant knowledge or tools, giving agents enough context to break down tasks strategically without requiring deep psychological detail. In **HoLLMwood**, each agent knows specific screenwriting functions (e.g., plot structuring, dialogue crafting), allowing them to focus on tailored narrative subtasks (Chen et al., 2024). Similarly, **TRIZ Agents** (Szczepanik and Chudziak, 2025) assign each agent a single TRIZ innovation principle (e.g., “contradiction resolution”), guiding systematic idea generation in engineering contexts. This intermediate granularity improves task focus and collaboration but still requires coordination to integrate specialized outputs into a coherent whole.

Fine-Grained Persona Agents receive detailed psychometric or demographic profiles, such as academic backgrounds or the *Big Five personality traits* (Digman, 1990), yielding stable, human-like decision patterns. For example, **PersonaFlow** mines scholarly CVs to form interdisciplinary research teams that adaptively ideate and evaluate concepts (Liu et al., 2024c). Similarly, Big-Five-driven agents demonstrate how nuanced traits (e.g., openness, conscientiousness) enhance idea synthesis (Serapio-García et al., 2025; Jiang et al., 2024; Duan et al., 2025). Yet, the high specificity increases design complexity, reduces adaptability to new domains, and risks reinforcing bias or overfitting to narrow behavioral patterns.

4.2 Agent Profiling Methods

Agent profiling methods vary according to the level of persona granularity they support. We group these methods into three paradigms: *Human-Defined, Model-Generated,* and *Data-Derived* approaches (Guo et al., 2024b; Wang et al., 2024a).

The *Human-Defined* approach relies on explicit, manually crafted descriptions to specify each agent’s role and behavior. This method is straightforward but demands extensive domain knowledge to maintain coherent coordination in MAS. In particular, **PersonaGym** (Samuel et al., 2024) provides concise role definitions and directs agents to emulate the prescribed persona’s skills and knowledge.

The *Model-Generated* approach proposes an automated pipeline that rapidly produces large sets of fictional agent profiles. These profiles are produced or refined using state-of-the-art large language models (LLMs), presenting a notable trade-off between profile quality and the scalability of large-scale generation. Additionally, model-generated methods are frequently integrated into dynamic prompt adjustment mechanisms, enabling iterative refinement of agent profiles in real time. **LLM Discussion** (Lu et al., 2024) exemplifies this: it begins with structured role descriptions and then leverages LLMs to produce a wide array of detailed, varied profiles.

Finally, *Data-Derived* methods construct personas grounded in real-world behavior patterns. Agent profiles reconstructed from demographic data emphasize stability compared to purely fictional descriptions, helping to avoid contradictions in implicit traits. In contrast, data-driven methods tend to capture latent behavioral patterns that are difficult to articulate through text. These methods also enable efficient tracing of an individual’s experience and expertise, supporting more accurate reconstruction and alignment within expert models. **VirSci** (Su et al., 2025) illustrates this paradigm by mining scientific publication data to build “digital twins” of researchers. Each agent thus operates with a persona rooted in authentic scientific expertise, enabling more realistic and diverse collaborative interactions in MAS.

5 Evaluation

Evaluating creativity in MAS, including human–agent collaborations, presents unique challenges. Unlike tasks with clear correctness criteria, creativity are inherently subjective and multifaceted, lacking a universally accepted assessment framework. To address this, researchers typically employ two complementary evaluation approaches:

- **Artifact Evaluation**: This approach focuses on assessing the creative content generated by MAS, either through the system’s processes or its final outputs. It encompasses:
  - **Objective, Metric-Based Measures** use

formulas such as cosine similarity and statistics methods to evaluate creativity.
  - **Subjective, Natural Language Instructed Assessments** ask experts, crowds, or LLMs to rate creativity.
- **Interaction Evaluation**: Beyond evaluating generated content, this method assesses the interaction processes between users and MAS. **User studies** are primarily employed here, focusing on criteria such as satisfaction.

The subsequent sections will first review the evaluation methods for text and image artifacts from both objective and subjective perspectives, discuss their practical applications, and then concentrate on assessing creative interactions between users and systems, emphasizing the role of user studies.

5.1 Objective Measurements

For text generation tasks, several metrics evaluate lexical richness and diversity. *Distinct-n* (Li et al., 2016) computes the proportion of unique n-grams, while *Entropy-n* (Shannon, 1948) measures the Shannon entropy over n-gram distributions, both serving as proxies for creative variety. In the screenwriting application (Chen et al., 2024), researchers routinely report 4-gram repetition rates alongside *Distinct-3* and *Entropy-3* to detect redundancy in long-form outputs. At the sentence level, **Self-BLEU** score (Zhu et al., 2018) treats each generated sentence as a hypothesis and the remainder as references to quantify internal diversity. Beyond surface counts, vector-based metrics capture deeper semantic variation. **Sentence-BERT (SBERT)** (Reimers and Gurevych, 2019) embeddings enable pairwise cosine similarity or Euclidean distance comparisons, where lower similarity or greater distance indicates broader exploration. Building on this, **Semantic Entropy** (Kuhn et al., 2023) clusters embeddings and computes the entropy over the categories, revealing a level of semantic diversity that goes beyond surface lexical patterns.

For image generation, **Fréchet Inception Distance (FID)** (Heusel et al., 2018) compares feature-space statistics between generated and real images and a lower score implies closer alignment in quality and diversity, while **Truncated Inception Entropy (TIE)** (Ibarrola et al., 2024) calculates the Shannon entropy of image features in the Inception latent space, with higher values reflecting richer variation. These metrics are particularly valuable for tasks such as silhouette generation (Lataifeh et al., 2024), offering standardized evaluation.

5.2 Subjective Assessments

Torrance Tests of Creative Thinking (TTCT) (Torrance, 1966) is a common standard for subjectively assessing creativity. Agents' artifacts are scored along four primary dimensions:

- *Fluency*: Total count of meaningful, relevant responses.
- *Flexibility*: Number of distinct categories or conceptual shifts among responses.
- *Originality*: Statistical rarity of each response versus a normative sample.
- *Elaboration*: Degree of detail or development added to each idea, measured by descriptive richness beyond the base concept.

Beyond the traditional **TTCT**, there are still other general criterion schemas such as **Boden's Criteria** (Boden, 2004) and **Creative Product Semantic Scale (CPSS)** (Besemer and Treffinger, 1981) used to evaluate different aspects of creative artifacts. Nowadays, researchers often invoke additional subjective criteria tailored to specific text generation tasks. **Insightfulness** (Shaer et al., 2024) is used to quantify how deeply ideas engage with underlying problem structures rather than merely diverging from norms. **Interestingness** (Chen et al., 2024) captures the entertainment value of narrative artifacts such as emotional resonance, and is commonly assessed through viewer ratings in screenwriting and storytelling studies.

For tasks in the image domain, researchers augment those general-purpose criteria with specific dimensions such as **Inspiring** (Hou et al., 2024). Beyond mere variety, this criterion assesses whether the generated images spark new ideas for designers or artists. For example, a system that produces a variety of color schemes, layouts, or conceptual motifs is diverse and inspiring, guiding users toward unexpected creative directions.

Building on the aforementioned subjective criteria, researchers often conduct user studies, arrange expert panels, or employ LLMs to evaluate artifact creativity. Subjective dimensions are typically rated on Likert scales, yielding interval-level scores suitable for statistical analysis. Expert panels may engage in structured discussions to reach consensus on feasibility and coherence. More recently, **LLM-as-a-judge** approaches have gained popularity, leveraging LLMs to assign scores on predefined scales (Zheng et al., 2023a).

5.3 Interaction Evaluation with User Study

In addition to evaluating creative artifacts, assessing the interaction between users and MAS through user studies is an equally essential dimension of the system evaluation. A general-purpose tool in this context is **Creative Support Index (CSI)** (Cherry and Latulipe, 2014), which captures user experience across dimensions such as *Collaboration* (ease of working with others), *Engagement* (enjoyment and willingness to repeat the activity), and *Expressiveness* (freedom to be creative). Also, researchers often develop specific evaluations for targeted tasks. For instance, **Colin** (Ye et al., 2024) evaluates children's narrative skills before and after using a storytelling system, focusing on engagement, understanding of cause-and-effect relationships, and the quality of their new story creations. More details can be found in Appendix G.

5.4 Discussion on Evaluation Methods

Evaluating creativity in MAS presents unique challenges. For artifact evaluation, objective metrics are scalable and reproducible but often capture only limited aspects of creativity, overlooking qualities such as emotional resonance and surprise. Subjective assessments, by contrast, can capture these nuances but suffer from bias, variability, and higher costs in time and effort, particularly at scale. As a result, researchers often combine both approaches to achieve a more balanced understanding of generated artifacts. Beyond artifacts, interaction studies reveal how users experience, steer, and trust system outputs. Taken together, artifact and interaction evaluations form two complementary pillars, offering a holistic view of MAS performance throughout its lifecycle. A categorization of existing methods is provided in the Appendix F and Table 4.

6 Challenges and Future Work

Balancing Agent Proactivity and Human Trust While high agent proactivity can spark more ideas, it can also overwhelm users, flatten idea diversity, erode perceived agency, and undermine trust (Lee and See, 2004). A major challenge is designing systems that intelligently adapt to the specific task and the individual user. Simple "proactivity thresholds" fail to account for context changes: What feels helpful in a brainstorming session can become intrusive during refinement (Houde et al., 2025). Furthermore, users differ significantly in their comfort with AI taking initiative; domain experts might embrace bold suggestions, whereas newcomers might feel distrustful (Naiseh et al., 2020). Future work can focus on *mixed-initiative* systems that continuously monitor both the task state and explicit or implicit user feedback (e.g., acceptance rates, signs of hesitation, or direct ratings) to calibrate the agent's level of initiative in real time, ensuring a more intuitive and supportive interaction.

Fairness and Profile Bias Agent personas can carry hidden stereotypes and preferences into the creative process when drawn from narrow or unbalanced data. This bias acts like a filter on the idea stream (Wan and Kalman, 2025). Agents with skewed profiles will repeatedly surface familiar, mainstream perspectives, crowding out novel angles from less-represented backgrounds (Liu et al., 2024a; Huot et al., 2025; Gupta et al., 2024). In recent works, **MALIBU Benchmark** (Vasista et al., 2025) quantifies how persona-based interactions risk amplifying biases and reinforcing stereotypes in creativity, while **Towards Implicit Bias Detection and Mitigation** (Borah and Mihalcea, 2024) investigates how implicit bias escalates during MAS interactions. **Argumentative Experience** (Shi et al., 2024) examines using diverse personas to reduce user confirmation bias in debates. Despite their contributions, these works share limitations. They often focus narrowly on specific types of bias (e.g., gender) or simplified tasks, and precisely measuring subtle persona bias remains challenging. Moreover, many studies examine only the final output rather than the interaction dynamics, and their experimental designs tend to oversimplify the complex processes involved in genuine multi-agent collaboration. As a result, these narrow scopes and simplified setups leave us with an incomplete understanding of how bias truly affects creative and equitable multi-agent systems.

Managing and Leveraging Creative Conflicts Conflicts between agents in MAS are typically seen as detrimental to efficiency and are often resolved through negotiation or central control (Kai et al., 2010; Yan et al., 2025). However, for creative MAS, controlled conflict or clashing perspectives can drive novelty and innovation, similar to human brainstorming or debate. Recent research explores multi-agent debate to leverage such "creative conflicts." **Multi-agent Debate** (Lin et al., 2024b) propose using multi-agent debate to interpret and mitigate hallucinations in multi-modal LLMs while promoting divergent thinking. **MAD** framework (Liang et al., 2024) demonstrates how agents debate under a judge can improve performance on counter-intuitive tasks and potentially aid creative ideation. Despite these advances, existing debate-based methods have some key limitations: they work with small groups of agents and offer no protocols for scaling to large populations or managing emerging coalitions; they lack mechanisms for continual learning that would allow agents to adapt their conflict strategies based on past outcomes; and they provide no mixed-initiative controls that let human users tune conflict intensity, or timing to keep interactions productive rather than chaotic.

Unified, Scalable Evaluation Frameworks Most LLM-based creative generation methods today focus on specific tasks: story writing, poem completion, ad copy, or code snippets, each with its own data and custom evaluations. That patchwork approach makes it impossible to tell which method drives progress. **MultiAgentBench** (Zhu et al., 2025) represents a first step toward a common suite of benchmarks and shared LLM-based evaluators, but significant challenges remain: devising a unified scoring rubric that balances novelty, coherence and utility across diverse domains; extending evaluation to real-time, interactive scenarios; and ensuring reproducible human judgments with standardized instructions.

7 Conclusion

This survey examines the rise of LLM-based multi-agent systems for creative tasks. We propose a unified framework for collaborative workflows and analyze how agent proactivity influences idea generation. We identify three key techniques that reliably enhance creative performance and review current evaluation methods. We then discuss overarching challenges, such as adaptive initiative control, bias mitigation, scalable interaction protocols, and the lack of standardized benchmarks, and outline promising directions for future research. Our goal is to clarify this rapidly evolving field and support the development of transparent, effective systems that augment human creativity.

Limitations

While this survey aims to provide a comprehensive overview of LLM-based creative multi-agent systems, several limitations remain that offer opportunities for future refinement.

First, our focus on text and image modalities was intended to ensure depth of analysis, but it necessarily excludes other important interaction channels, such as audio (Wu et al., 2024; Kuan et al., 2024), video (Huang et al., 2024), and embodied robotics (Duan and Zou, 2025), which may bring distinct challenges and opportunities for creative MAS.

Second, while we briefly discuss persona-related biases, we do not delve into broader ethical considerations. These include issues such as data licensing and provenance (e.g., the use of proprietary or copyrighted corpora), user privacy when agents log interactions or generate persistent memory traces, informed consent in human-agent data collection, and the environmental costs associated with large-scale multi-agent deployments.

Third, the majority of systems reviewed in this survey are developed and evaluated in English and rely heavily on Western-centric datasets. We do not cover how cultural norms, multilingual settings (Lin et al., 2024a), or low-resource languages affect agent design, creative expression, or evaluation standards. Addressing these dimensions is critical to building more inclusive, globally relevant systems that reflect diverse forms of creativity and collaboration.

Acknowledgements

We thank the reviewers for their insightful comments. This work was financially supported by the National Science and Technology Council (NSTC) in Taiwan, under Grants 111-2222-E-002-013-MY3 and 112-2223-E002-012-MY5, and the Center of Data Intelligence: Technologies, Applications, and Systems, National Taiwan University under the Grant 114L900901, from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) of Taiwan.

References

Safinah Ali, Ayat Abodayeh, Zahra Dhuliawala, Cynthia Breazeal, and Hae Won Park. 2025. Towards Inclusive Co-creative Child-robot Interaction: Can Social Robots Support Neurodivergent Children's Creativity? In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '25.

Alfonso Amayuelas, Jingbo Yang, Saaket Agashe, Ashwin Nagarajan, Antonis Antoniades, Xin Eric

Wang, and William Wang. 2025. Self-resource allocation in multi-agent LLM systems. *Preprint*, arXiv:2504.02051.

Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W. Ayers, Dragomir Radev, and Jeremy Avigad. 2023. ProofNet: Autoformalizing and Formally Proving Undergraduate-Level Mathematics. *Preprint*, arXiv:2302.12433.

Stephanie P. Besemer and Donald J. Treffinger. 1981. Analysis of creative products: Review and synthesis. *Journal of Creative Behavior*, 15(3):158–178.

Margaret A Boden. 2004. *The creative mind: Myths and mechanisms*. Psychology Press.

Angana Borah and Rada Mihalcea. 2024. Towards implicit bias detection and mitigation in multi-agent LLM interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9306–9326. Association for Computational Linguistics.

Tom B. Brown et al. 2020. Language Models are Few-Shot Learners. *Preprint*, arXiv:2005.14165.

Erika K Carlson. 2020. Artificial intelligence can invent but not patent–for now. *Engineering*, 6(11):1212–1213.

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025. Why Do Multi-Agent LLM Systems Fail? *Preprint*, arXiv:2503.13657.

Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2024. Creativity Support in the Age of Large Language Models: An Empirical Study Involving Emerging Writers. *Preprint*, arXiv:2309.12570.

Jing Chen, Xinyu Zhu, Cheng Yang, Chufan Shi, Yadong Xi, Yuxiang Zhang, Junjie Wang, Jiashu Pu, Rongsheng Zhang, Yujiu Yang, and Tian Feng. 2024. HoLLMwood: Unleashing the Creativity of Large Language Models in Screenwriting via Role Playing. *CoRR*.

Erin Cherry and Celine Latulipe. 2014. Quantifying the Creativity Support of Digital Tools through the Creativity Support Index. *ACM Trans. Comput.-Hum. Interact.*, 21(4).

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. *Preprint*, arXiv:2403.04132.

Niall Creech, Natalia Criado Pacheco, and Simon Miles. 2021. Resource allocation in dynamic multiagent systems. *Preprint*, arXiv:2102.08317.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. Marg: Multi-agent review generation for scientific papers. *Preprint*, arXiv:2401.04259.

Allegra De Filippo, Michela Milano, et al. 2024. Large language models for human-AI co-creation of robotic dance performances. In *IJCAI*, pages 7627–7635.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

John M Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual review of psychology*.

Shiying Ding, Xinyi Chen, Yan Fang, Wenrui Liu, Yiwu Qiu, and Chunlei Chai. 2023. DesignGPT: Multi-Agent Collaboration in Design. In *2023 16th International Symposium on Computational Intelligence and Design (ISCID)*, pages 204–208.

Kangkang Duan and Zhengbo Zou. 2025. Enhancing construction robot collaboration via multiagent reinforcement learning. *Journal of Intelligent Construction*, 3(2):1–16.

Yifan Duan, Yihong Tang, Xuefeng Bai, Kehai Chen, Juntao Li, and Min Zhang. 2025. The Power of Personality: A Human Simulation Perspective to Investigate Large Language Model Agents. *Preprint*, arXiv:2502.20859.

Xianzhe Fan, Zihan Wu, Chun Yu, Fenggui Rao, Weinan Shi, and Teng Tu. 2024. ContextCam: Bridging Context Awareness with Creative Human-AI Image Co-Creation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. Association for Computing Machinery.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines*.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1–24.

Shijun Ge, Yuanbo Sun, Yin Cui, and Dapeng Wei. 2025. An Innovative Solution to Design Problems: Applying the Chain-of-Thought Technique to Integrate LLM-Based Agents With Concept Generation Methods. *IEEE Access*, 13:10499–10512.

Juraj Gottweis et al. 2025. Towards an AI co-scientist. *Preprint*, arXiv:2502.18864.

Aaron Grattafiori et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

J. P. Guilford. 1950. Creativity. *American Psychologist*, 5(9):444–454.

J. P. Guilford. 1967. *The Nature of Human Intelligence*, chapter 8. McGraw-Hill, New York.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024a. Large Language Model Based Multi-agents: A Survey of Progress and Challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8048–8057. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024b. Large language model based multi-agents: A survey of progress and challenges. *Preprint*, arXiv:2402.01680.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. In *The Twelfth International Conference on Learning Representations*.

Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Tak Yeon Lee, So-Yeon Ahn, and Alice Oh. 2024a. RECIPE4U: Student-ChatGPT interaction dataset in EFL writing education. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13666–13676, Torino, Italia. ELRA and ICCL.

Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, Zhaozhuo Xu, and Chaoyang He. 2024b. LLM multi-agent systems: Challenges and open problems. *Preprint*, arXiv:2402.03578.

Jessica He, Stephanie Houde, Gabriel E. Gonzalez, Darío Andrés Silva Moran, Steven I. Ross, Michael Muller, and Justin D. Weisz. 2024. Ai and the future of collaborative work: Group ideation with an llm in a virtual canvas. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work*, CHIWORK '24, New York, NY, USA. Association for Computing Machinery.

Jie He, Tao Wang, Deyi Xiong, and Qun Liu. 2020. The box is in the pen: Evaluating commonsense reasoning in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3662–3672, Online. Association for Computational Linguistics.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2018. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Preprint*, arXiv:1706.08500.

Yihan Hou, Manling Yang, Hao Cui, Lei Wang, Jie Xu, and Wei Zeng. 2024. C2ideas: Supporting creative interior color design ideation with a large language model. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24.

Stephanie Houde, Kristina Brimijoin, Michael Muller, Steven I. Ross, Dario Andres Silva Moran, Gabriel Enrique Gonzalez, Siya Kunde, Morgan A. Foreman, and Justin D. Weisz. 2025. Controlling ai agent participation in group conversations: A human-centered approach. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, page 390–408, New York, NY, USA. Association for Computing Machinery.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in llm simulations. *Preprint*, arXiv:2402.10811.

Kaiyi Huang, Yukun Huang, Xuefei Ning, Zinan Lin, Yu Wang, and Xihui Liu. 2024. Genmac: Compositional text-to-video generation with multi-agent collaboration. *Preprint*, arXiv:2412.04440.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Kai Dang, et al. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Fantine Huot, Reinald Kim Amplayo, Jennimaria Palomaki, Alice Shoshana Jakobovits, Elizabeth Clark, and Mirella Lapata. 2025. Agents' room: Narrative generation through multi-step collaboration. In *The Thirteenth International Conference on Learning Representations*.

Thorsten Händler. 2023. Balancing autonomy and alignment: A multi-dimensional taxonomy for autonomous LLM-powered multi-agent architectures. *Preprint*, arXiv:2310.03659.

Francisco Ibarrola, Tomas Lawton, and Kazjon Grace. 2024. A Collaborative, Interactive and Context-Aware Drawing Agent for Co-Creative Design. *IEEE Transactions on Visualization and Computer Graphics*.

Naomi Imasato, Kazuki Miyazawa, Takayuki Nagai, and Takato Horii. 2024. Creative agents: Simulating the systems model of creativity with generative agents. *Preprint*, arXiv:2411.17065.

Masaki Ishizaka, Akihito Taya, and Yoshito Tobe. 2024. Sparkit: A mind map-based mas for idea generation support. In *Engineering Multi-Agent Systems*, pages 1–22, Cham. Springer Nature Switzerland.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. *Preprint*, arXiv:2305.02547.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. AgentReview: Exploring peer review dynamics with LLM agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1208–1226, Miami, Florida, USA. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611. Association for Computational Linguistics.

Huang Kai, Zhang Zhonghua, Qin Zheng, and Liu Bing. 2010. Conflict resolution in multi-agent systems based on negotiation and arbitrage. In *2010 2nd IEEE International Conference on Information Management and Engineering*, pages 304–307.

Junseok Kim, Nakyeong Yang, and Kyomin Jung. 2024. Persona is a double-edged sword: Mitigating the negative impact of role-playing prompts in zero-shot reasoning tasks. *Preprint*, arXiv:2408.08631.

Chun-Yi Kuan, Chih-Kai Yang, Wei-Ping Huang, Ke-Han Lu, and Hung yi Lee. 2024. Speech-copilot: Leveraging large language models for speech processing via task decomposition, modularization, and program generation. *Preprint*, arXiv:2407.09886.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. *Preprint*, arXiv:2302.09664.

Harsh Kumar, Jonathan Vincentius, Ewan Jordan, and Ashton Anderson. 2025. Human Creativity in the Age of LLMs: Randomized Experiments on Divergent and Convergent Thinking. In *International Conference on Human Factors in Computing Systems*.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. Propublica compas recidivism risk score data and analysis. Data and analysis for "Machine Bias" investigation.

Mohammad Lataifeh, Xavier A Carrasco, Ashraf M Elnagar, Naveed Ahmed, and Imran Junejo. 2024. Human–machine co-creation: a complementary cognitive approach to creative character design process using GANs. *The Journal of Supercomputing*, 80(11):16574–16610.

Tomas Lawton, Francisco J Ibarrola, Dan Ventura, and Kazjon Grace. 2023. Drawing with Reframer: Emergence and Control in Co-Creative AI. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23.

John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1(1):9.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Gionnieve Lim and Simon T. Perrault. 2024. Rapid AIdeation: Generating Ideas With the Self and in Collaboration With Large Language Models. *arXiv preprint arXiv:2403.12928*.

Yi-Cheng Lin, Tzu-Quan Lin, Chih-Kai Yang, Ke-Han Lu, Wei-Chih Chen, Chun-Yi Kuan, and Hung-Yi Lee. 2024a. Listen and speak fairly: a study on semantic gender bias in speech integrated large language models. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 439–446.

Zheng Lin, Zhenxing Niu, Zhibin Wang, and Yinghui Xu. 2024b. Interpreting and mitigating hallucination in mllms through multi-agent debate. *Preprint*, arXiv:2407.20505.

Zhiyu Lin, Rohan Agarwal, and Mark Riedl. 2022. Creative wand: a system to study effects of communications in co-creative settings. In *Proceedings of the Eighteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, AIIDE'22. AAAI Press.

Andy Liu, Mona Diab, and Daniel Fried. 2024a. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9832–9850, Bangkok, Thailand. Association for Computational Linguistics.

Yiren Liu, Si Chen, and Haocong et al. Cheng. 2024b. How AI Processing Delays Foster Creativity: Exploring Research Question Co-Creation with an LLM-based Agent. In *Proc. CHI '24*.

Yiren Liu, Pranav Sharma, Mehul Jitendra Oswal, Haijun Xia, and Yun Huang. 2024c. PersonaFlow: Boosting Research Ideation with LLM-Simulated Expert Personas. *Preprint*, arXiv:2409.12538.

Tao Long, Katy Ilonka Gero, and Lydia B Chilton. 2024. Not just novelty: A longitudinal study on utility and customization of an ai workflow. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, DIS '24, page 782–803, New York, NY, USA. Association for Computing Machinery.

Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung yi Lee, and Shao-Hua Sun. 2024. Llm discussion: Enhancing the creativity of large language models via discussion framework and role-play. *Preprint*, arXiv:2405.06373.

Alaeddine Mellouli et al. 2024. Storeys: A neurosymbolic approach to human-ai co-creation of novel action-oriented narratives in known story worlds. In *Proceedings of the 15th International Conference on Computational Creativity, ICCC 2024, Jönköping, Sweden, June 17-21, 2024*.

Chunjiang Mu, Hao Guo, Yang Chen, Chen Shen, Die Hu, Shuyue Hu, and Zhen Wang. 2024. Multi-agent, human–agent and beyond: A survey on cooperation in social dilemmas. *Neurocomputing*, 610:128514.

Anirban Mukherjee and Hannah Hanwen Chang. 2025. Stochastic, dynamic, fluid autonomy in agentic ai: Implications for authorship, inventorship, and liability. *Preprint*, arXiv:2504.04058.

Gabriel Mukobi, Hannah Erlebach, Niklas Lauffer, Lewis Hammond, Alan Chan, and Jesse Clifton. 2023. Welfare diplomacy: Benchmarking language model cooperation. In *Socially Responsible Language Modelling Research*.

Mohammad Naiseh, Nan Jiang, Jianbing Ma, and Raian Ali. 2020. Personalising explainable recommendations: Literature and conceptualisation. In *Trends and Innovations in Information Systems and Technologies*, pages 518–533, Cham. Springer International Publishing.

Harshit Nigam, Manasi Patwardhan, Lovekesh Vig, and Gautam Shroff. 2024. An interactive co-pilot for accelerated research ideation. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 60–73, Mexico City, Mexico. Association for Computational Linguistics.

U.S. Copyright Office. 2025. *Copyright and Artificial Intelligence, Part 2: Copyrightability Report*. U.S. Copyright Office.

Bo Pan, Jiaying Lu, Ke Wang, Li Zheng, Zhen Wen, Yingchaojie Feng, Minfeng Zhu, and Wei Chen. 2024. Agentcoord: Visually exploring coordination strategy for llm-based multi-agent collaboration. *Preprint*, arXiv:2404.11943.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186, Bangkok, Thailand. Association for Computational Linguistics.

Hua Xuan Qin, Shan Jin, Ze Gao, Mingming Fan, and Pan Hui. 2024. CharacterMeet: Supporting Creative Writers' Entire Story Character Construction Processes Through Conversation with LLM-Powered Chatbot Avatars. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. Association for Computing Machinery.

Yansheng Qiu, Haoquan Zhang, Zhaopan Xu, Ming Li, Diping Song, Zheng Wang, and Kaipeng Zhang. 2025. AI Idea Bench 2025: AI Research Idea Generation Benchmark. *Preprint*, arXiv:2504.14191.

Marissa Radensky. 2024. Mixed-Initiative Methods for Co-Creation in Scientific Research. In *Proceedings of the 16th Conference on Creativity & Cognition*, C&C '24, page 1–7. Association for Computing Machinery.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, and Hao Sun. 2024. Liveideabench: Evaluating llms' scientific creativity and idea generation with minimal context. *arXiv preprint arXiv:2412.17596*.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2024. Personagym: Evaluating persona agents and llms. *Preprint*, arXiv:2407.18416.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. Personality Traits in Large Language Models. *Preprint*, arXiv:2307.00184.

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*, CHI '24, New York, NY, USA. Association for Computing Machinery.

C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Li Shi, Houjiang Liu, Yian Wong, Utkarsh Mujumdar, Dan Zhang, Jacek Gwizdka, and Matthew Lease. 2024. Argumentative Experience: Reducing Confirmation Bias on Controversial Issues through LLM-Generated Multi-Persona Debates. *arXiv preprint arXiv:2412.04629*.

Aakriti Singh, Shipra Saraswat, and Neetu Faujdar. 2017. Analyzing Titanic disaster using machine learning algorithms. In *2017 International Conference on Computing, Communication and Automation (ICCCA)*.

Peiyang Song, Kaiyu Yang, and Anima Anandkumar. 2025. Lean copilot: Large language models as copilots for theorem proving in lean. *Preprint*, arXiv:2404.12534.

Aarohi Srivastava et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, Philip Torr, Bowen Zhou, and Nanqing Dong. 2025. Many Heads Are Better Than One: Improved Scientific Idea Generation by A LLM-Based Multi-Agent System. *Preprint*, arXiv:2410.09403.

Jingyun Sun, Chengxiao Dai, Zhongze Luo, Yangbo Chang, and Yang Li. 2024. LawLuo: A Multi-Agent Collaborative Framework for Multi-Round Chinese Legal Consultation. *Preprint*, arXiv:2407.16252.

Yuqian Sun and Stefano Gualeni. 2025. Between puppet and actor: Reframing authorship in this age of ai agents. *Preprint*, arXiv:2501.15346.

Kamil Szczepanik and Jarosław Chudziak. 2025. TRIZ Agents: A Multi-Agent LLM Approach for TRIZ-Based Innovation. In *17th International Conference on Agents and Artificial Intelligence*. SciTePress.

Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. ArnetMiner: Extraction and Mining of Academic Social Networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 990–998. ACM.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L. Griffiths, and Faeze Brahman. 2024. MacGyver: Are Large Language Models Creative Problem Solvers? In *Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

E. Paul Torrance. 1966. *Torrance Tests of Creative Thinking*. Personnel Press, Princeton, NJ.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D. Nguyen. 2025. Multi-Agent Collaboration Mechanisms: A Survey of LLMs. *Preprint*, arXiv:2501.06322.

Ishwara Vasista, Imran Mirza, Cole Huang, Rohan Rajasekhara Patil, Aslihan Akalin, Kevin Zhu, and Sean O'Brien. 2025. MALIBU Benchmark: Multi-Agent LLM Implicit Bias Uncovered. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Tony Veale and F Amílcar Cardoso. 2019. *Computational Creativity: The Philosophy and Engineering of Autonomously Creative Systems*. Springer.

P Venkadesh, SV Divya, and K Subash Kumar. 2024. Unlocking AI Creativity: A Multi-Agent Approach with CrewAI. *Journal of Trends in Computer Science Smart Technology*.

Kavana Venkatesh, Connor Dunlop, and Pinar Yanardag. 2025. CREA: A Collaborative Multi-Agent Framework for Creative Content Generation with Diffusion Models. *Preprint*, arXiv:2504.05306.

Saranya Venkatraman, Nafis Irtiza Tripto, and Dongwon Lee. 2024. Collabstory: Multi-llm collaborative story generation and authorship analysis. In *Proc. NAACL '25*.

M. A. Wallach and N. Kogan. 1965. *Modes of Thinking in Young Children: A Study of the Creativity-Intelligence Distinction*. Holt, Rinehart & Winston, New York.

Qian Wan, Siying Hu, Yu Zhang, Piaohong Wang, Bo Wen, and Zhicong Lu. 2024. "It Felt Like Having a Second Mind": Investigating Human-AI Co-creativity in Prewriting with Large Language Models. *Proc. ACM Hum.-Comput. Interact.*, 8.

Yun Wan and Yoram M Kalman. 2025. Using Generative AI Personas Increases Collective Diversity in Human Ideation. *Preprint*, arXiv:2504.13868.

Boshi Wang, Xiang Deng, and Huan Sun. 2022. Iteratively prompt pre-trained language models for chain of thought. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2714–2730. Association for Computational Linguistics.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024a. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Yi Wang, Qian Zhou, and David Ledo. 2024b. Storyverse: Towards co-authoring dynamic plot with llm-based character simulation via narrative planning. In *Proceedings of the 19th International Conference on the Foundations of Digital Games*, FDG '24, New York, NY, USA. Association for Computing Machinery.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024c. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *Preprint*, arXiv:2307.05300.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Geraint A Wiggins. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-based systems*, 19(7):449–458.

Haibin Wu, Xuanjun Chen, Yi-Cheng Lin, Kai wei Chang, Ho-Lam Chung, Alexander H. Liu, and Hung yi Lee. 2024. Towards audio language modeling – an overview. *Preprint*, arXiv:2402.13236.

Zhikun Wu, Thomas Weber, and Florian Müller. 2025. One Does Not Simply Meme Alone: Evaluating Co-Creativity Between LLMs and Humans in the Generation of Humor. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, page 1082–1092. ACM.

Chengxing Xie and Difan Zou. 2024. A human-like reasoning framework for multi-phases planning task with large language models. In *ICML 2024 Workshop on LLMs and Cognition*.

Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z. Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Jang, Yiqing Xie, Shuyan Zhou, and Graham Neubig. 2024. TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks. *Preprint*, arXiv:2412.14161.

Bingyu Yan, Xiaoming Zhang, Litian Zhang, Lian Zhang, Ziyi Zhou, Dezhuang Miao, and Chaozhuo Li. 2025. Beyond self-talk: A communication-centric survey of llm-based multi-agent systems. *Preprint*, arXiv:2502.14321.

Kaixun Yang, Yixin Cheng, Linxuan Zhao, Mladen Raković, Zachari Swiecki, Dragan Gašević, and Guanliang Chen. 2024. Ink and algorithm: Exploring temporal dynamics in human-ai collaborative writing. *Preprint*, arXiv:2406.14885.

Lyumanshan Ye, Jiandong Jiang, Yuhan Liu, Yihan Ran, Pengfei Liu, and Danni Chang. 2024. Colin: A Multimodal Human-AI Co-Creation Storytelling System To Support Children's Multi-Level Narrative Skills. *Preprint*, arXiv:2405.06495. Version 4, last revised 17 Mar 2025.

Junwei Yu, Yepeng Ding, and Hiroyuki Sato. 2025. DynTaskMAS: A dynamic task graph-driven framework for asynchronous and parallel llm-based multi-agent systems. *Preprint*, arXiv:2503.07675.

Mingyue Yuan, Jieshan Chen, and Aaron Quigley. 2024. Maxprototyper: A multi-agent generation system for interactive user interface prototyping. *Preprint*, arXiv:2405.07131.

J.D. Zamfirescu-Pereira, Eunice Jun, Michael Terry, Qian Yang, and Bjoern Hartmann. 2025. Beyond code generation: Llm-supported exploration of the program design space. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, page 1–17. ACM.

Lidong Zhai, Zhijie Qiu, Xizhong Guo, and Jiaqi Li. 2025. The Athenian Academy: A Seven-Layer Architecture Model for Multi-Agent Systems. *arXiv preprint arXiv:2504.12735*.

An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. 2024a. On generative agents in recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1807–1817, New York, NY, USA. Association for Computing Machinery.

Chao Zhang, Xuechen Liu, Katherine Ziska, Soobin Jeon, Chi-Lin Yu, and Ying Xu. 2024b. Mathemyths: Leveraging large language models to teach mathematical language through child-ai co-creative storytelling. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Chao Zhang, Cheng Yao, Jiayi Wu, Weijia Lin, Lijuan Liu, Ge Yan, and Fangtian Ying. 2022. StoryDrawer: A Child–AI Collaborative Drawing System to Support Children's Creative Visual Storytelling. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22. Association for Computing Machinery.

Guibin Zhang, Luyang Niu, Junfeng Fang, Kun Wang, Lei Bai, and Xiang Wang. 2025. Multi-agent architecture search via agentic supernet. *Preprint*, arXiv:2502.04180.

Jingyue Zhang and Ian Arawjo. 2025. ChainBuddy: An AI-assisted Agent System for Generating LLM Pipelines. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, page 1–21. ACM.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024c. Exploring collaboration mechanisms for LLM agents: A social psychology view. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14544–14607, Bangkok, Thailand. Association for Computational Linguistics.

Zixiao Zhao, Jing Sun, Zhe Hou, Zhiyuan Wei, Cheng-Hao Cai, Miao Qiao, and Jin Song Dong. 2025. Mactg: Multi-agent collaborative thought graph for automatic programming. *arXiv preprint arXiv:2410.19245*.

Lianmin Zheng et al. 2023a. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*.

Qingxiao Zheng, Zhongwei Xu, Abhinav Choudhry, Yuting Chen, Yongming Li, and Yun Huang. 2023b. Synergizing Human-AI Agency: A Guide of 23 Heuristics for Service Co-Creation with LLM-Based Agents. *Preprint*, arXiv:2310.15065.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You. 2025. Multiagentbench: Evaluating the collaboration and competition of llm agents. *Preprint*, arXiv:2503.01935.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. *Preprint*, arXiv:1802.01886.

Hanxi Zhu Zijun Liu, Cheng Gao. 2024. Systematic Idea Refinement for Machine Learning Research Agents. In *Submitted to Tsinghua University Course: Advanced Machine Learning & Machine Learning*. Under review.

A Proactivity Spectrum Supplementary

This section details how we classify the proactivity levels shown in Fig. 2. We classify agent proactivity along two dimensions—*Process* and *Decision Making*—using a rainbow scale from red (highest) to purple (lowest). We also note that the *Planning* phase remains under-explored, likely due to LLM agents' low confidence in autonomous planning.

The highest level of agent proactivity, marked in red, embodies a fully agent-driven pipeline. At this level, agents autonomously perform all tasks, including discussion, idea sharing, and peer review, without any human guidance or feedback. In **MaCTG** (Zhao et al., 2025), agents are assigned individual tasks and kick off the project on their own—they come up with ideas, write code, assemble components, validate results, and refine the output. From start to finish, the entire creative process runs without any human input.

At the second-tier level of proactivity, represented in yellow and green, human intervention is slightly enhanced and helps shape the agent's output, resulting in more stable and predictable behavior (Zhang et al., 2024b; Mellouli et al., 2024). **Co-Scientist** (Gottweis et al., 2025) enables human users to inject additional ideas into a shared workspace among agents, stimulating agents' divergent thinking during the creativity process. **Collab-Story** (Venkatraman et al., 2024) attempts to build a large-scale story creation database with minimal human effort. It provides LLM agents with brief human instructions and storylines from previous agents. These iterative human inputs have a latent influence on idea generation from agents, despite no direct outlines.

Human-Agent synergy, otherwise, leads to medium proactivity, characterized by peer-level collaboration. Both parties jointly engage in the *Process* phase to enhance the diversity and feasibility of creative outputs. However, to prevent potential ethical hazards and unexpected outcomes, these frameworks tend to entrust the final evaluation to human users, thereby inevitably exhibiting low agent proactivity at *Decision Making* (Hou et al., 2024; Lin et al., 2022).

In line with our definition of proactivity in Section 2, systems that follow data-driven processes or act beyond direct human instructions demonstrate a higher level of proactivity. In our classification, such cases are highlighted in blue, and this increase of proactivity is particularly prominent in the *Process* phase. For example, **ContextCam** (Fan et al., 2024) not only receives iterative user requests during refinement but also incorporates environmental data collected from its sensors, such as weather conditions, camera input, and audio input. **Colin** (Ye et al., 2024) exhibits the agent proactivity through a different way. The system initiates the interaction questions to trace the understanding and idea of users, rather than relying on reactive prompt-based communication like typical Human-Agent synergy frameworks.

The purple-marked work shows relatively low proactivity in both phases. Humans mainly use the LLM agent to generate ideas from alternative perspectives, helping to fill in where human thinking might be limited. The system keeps its creative output strong by leaning on a solid human-driven backbone and manual evaluation. While the results are good, it also imposes an excessive load on designers and creators (Wan et al., 2024; He et al., 2024; Lim and Perrault, 2024).

B Detailed MAS Technique Case Studies

Section 3 talked about three key ways MAS boosts creativity: *Divergent Exploration*, *Iterative Refinement*, and *Collaborative Synthesis*.

The next parts of this appendix give more examples of real systems that use each approach. If you want a quick snapshot of all the systems and the tasks they work on, check Table 1.

B.1 Divergent Exploration Case Studies

Co-GPT Ideation Lim and Perrault (2024) compared individual ideation with co-ideation using GPT-3.5. Participants working with the LLM generated more diverse and detailed ideas, though top-rated ideas still tended to come from humans. The system expanded the idea space without replacing human creativity, supporting LLMs as useful collaborators during early brainstorming.

Group-AI Brainwriting Shaer et al. (2024) proposed a framework that guides students through four steps: (1) independent human ideation, (2) idea expansion using GPT-3, (3) collaborative refinement, and (4) evaluation by GPT-4. LLMs help widen the scope of creative ideas and serve as both contributors and evaluators. Many final proposals were co-developed with GPT-3, showing strong MAS potential for guided creativity.

ICCRI The Inclusive Co-Creative Child-Robot Interaction (ICCRI) system (Ali et al., 2025) was tested in a special education setting. Across five sessions, children worked with a robot agent to co-create stories and drawings. Creativity was significantly enhanced during ICCRI-supported sessions (S1–S3) and remained above baseline even after its removal, suggesting that MAS can leave a lasting creative imprint.

Long-Term Impact Study Kumar et al. (2025) explored how repeated use of LLMs might affect human creativity. While AI assistance improved short-term performance, it reduced originality and diversity in unassisted follow-ups. The results raise concerns about long-term over-reliance on AI, emphasizing the importance of maintaining human autonomy in divergent thinking.

B.2 Iterative Refinement Case Studies

DesignGPT DesignGPT (Ding et al., 2023) simulates a design firm by assigning LLM agents to roles such as product manager and materials expert. These agents iteratively develop product proposals through structured feedback and refinement. Compared to one-shot generation tools, this MAS achieved higher completeness, novelty, and practicality in product outcomes.

Baby-AIGS-MLer This MAS (Zijun Liu, 2024) tackles machine learning research by splitting tasks into ideation, coding, testing, and evaluation. Each role is handled by a specialized agent. Tested on benchmarks like Chatbot Arena (Chiang et al., 2024) and Titanic (Singh et al., 2017), the system showed improved predictive accuracy and generalization, illustrating the benefit of multi-step refinement.

HoLLMwood HoLLMwood (Chen et al., 2024) supports screenwriting using three roles: Writer (script generation), Editor (content review), and Actor (dialogue simulation). Scripts improve through repeated rounds of agent interaction, enhancing plot structure and character depth. It outperforms single-LLM systems in relevance, fluency, and coherence.

Multi-agent Debate Framework Liang et al. (2024) applied debate as a refinement tool for reasoning tasks. Agents take on the roles of proponent, opponent, and judge, and they take turns debating an issue across several rounds. The debate ends automatically when no new ideas or arguments are being introduced. This approach helps the system perform much better on challenging reasoning tasks, such as Commonsense Machine Translation and the CIAR benchmark (He et al., 2020).

B.3 Collaborative Synthesis Case Studies

MaCTG MaCTG (Zhao et al., 2025) organizes agents into horizontal layers (modules) and vertical layers (management). It combines DeepSeek-V3 (DeepSeek-AI, 2024) for planning with Qwen2.5-Coder-7B (Hui et al., 2024) for coding. Agents are assigned roles like planner, tester, or integrator, and outputs are validated across multiple levels. This MAS delivers scalable and cost-efficient software design.

CollabStory Multiple LLMs sequentially write paragraphs of a shared story (Venkatraman et al., 2024). GPT-4o evaluated the coherence of transitions and found over 75% to be consistent. The study shows that decentralized authorship can still maintain narrative coherence and readability, illustrating collaborative synthesis through turn-based coordination.

Human-AI Co-creativity Involving 15 creative students, Wan et al. (2024) tested a three-stage writing pipeline: LLM-led ideation, human-guided elaboration, and final authoring. MAS agents helped inspire new ideas and filled in missing details. Participants described the system as feeling like a "second mind", demonstrating the supportive role of LLMs in collaborative writing.

CoQuest CoQuest (Liu et al., 2024b) assists researchers in formulating meaningful questions. It features tools like a flow editor, a visual citation graph, and an AI agent that suggests direction. By combining broad exploration with deeper follow-up, the system balances creativity and structure for interdisciplinary research planning.

C Supplement of Persona Design

MAS is eligible to portray various roles, participating in the collaborative framework and significantly demonstrating divergent innovation and convergent perspectives. Recent studies have dissected the impact of persona on reasoning and inference, exhibiting the prominent enhancement in the exploration of diverse ideas through a variety of interaction frameworks. However, Persona design serves as a double-edged sword, an inappropriate agent profile accidentally results in erosion of performance (Kim et al., 2024).

As shown in Table 3, we summarized the comprehensive survey on MAS persona design related to the cultivation of creativity, to provide future researchers with immediate access to relevant papers in this field. We also acknowledge that some architectures invoke hybrid methods to delineate the personalities of agents. The profile design incorporates the aforementioned techniques, such as iterative refinement and collaboration synthesis, dynamically adjusting the prompt context to the expected performance. For a clear and intuitive statement, we only depict the initialization strategy in the construction of the agent profile in Table 3. Furthermore, as we stated in Section 4, the prescribed agent profile helps establish the basic description and positioning of the role, thereby excluding the purely task-oriented prompt from Table 3.

D Datasets

The datasets used to evaluate creativity in multi-agent systems are highly diverse and vary based on the specific creativity task. In this section, we categorize them into two groups: (1) psychological test datasets, which incorporate established creativity assessments from the field of psychology, and (2) task-specific datasets, which are either custom-designed or adapted from other domains to align with the requirements of the target creativity task.

D.1 Psychological Test Datasets

Psychological test datasets comprise a collection of established tasks originally designed for humans but adapted for evaluation in multi-agent systems. For example, to assess divergent thinking, some studies use the Wallach Kogan Creativity Tests (Wallach and Kogan, 1965), which involve open-ended tasks measuring originality and flexibility; the Alternative Uses Task (Guilford, 1967), where participants are asked to think of as many uses as possible for a common object; and the Torrance Tests of Creative Thinking (Torrance, 1966), a widely used battery assessing creative potential through both verbal and figural prompts. These adapted psychological tests provide a standardized foundation for evaluating creative capacities in multi-agent systems, enabling consistent comparisons across different studies.

D.2 Task Specific Datasets

In addition to using established psychological tests to assess creativity, different creativity targets also require task specific datasets, which are either self-constructed or adapted from existing works. For example, Chakrabarty et al. (2024) released the first dataset of co-written stories by multi-agent systems and humans. Similarly, AI Idea Bench 2025 (Qiu et al., 2025) was introduced to foster the development of research idea generation. In Table 2, we offer a detailed dataset collection with respect to each task. We hope these collections facilitate standardized evaluation and support future work in creativity-oriented MAS research.

E Challenges and Future Directions

Authorship of Creativity Output Another significant set of challenges revolves around the complex and often ambiguous authorship question. Establishing who or what holds the “author” status for collaboratively generated artifacts presents a fundamental open problem. A primary challenge stems from traditional copyright doctrine, which intrinsically links authorship to human creativity. This is illustrated by the consistent stance of the U.S. Copyright Office, which has repeatedly denied copyright protection to works generated solely by AI (Office, 2025). The office maintains that such works lack the requisite human authorship, placing the onus on applicants to demonstrate, on a case-by-case basis, that significant human intervention was involved in the creation process.

The complexity extends to the creators and proprietors of the AI tools themselves. Legal practitioners caution that neither the developers nor the owners of these sophisticated AI systems typically possess the level of direct creative control over individual outputs necessary to assert authorship (Carlson, 2020). This lack of direct creative input for any specific artifact generated by the system underscores an urgent challenge: establishing clearer guidelines and legal frameworks to govern ownership, attribution, and royalty distribution in the rapidly expanding field of AI-augmented creativity.

Beyond whether AI-generated works qualify for copyright, we also need to decide how to apportion authorship among agents in creative MAS, as this determines their legal and moral credit. In practice, one might imagine a collaborative novel-writing system where Agent A (a planning module) generates the story outline, Agent B (a stylistic refiner) polishes prose, and a human "editor" selects, tweaks, and sequences the final chapters. Which of these agents "holds" authorship? A recent study reframes AI agents as lying between "puppets" and "actors," arguing that an agent's level of autonomy, not just its technical role, should inform its claim to authorship (Sun and Gualeni, 2025). Others point out that, when creative contributions are stochastic, dynamic, and fluidly intertwined, disentangling individual inputs is often infeasible; in such cases, human and machine contributions may need to be treated as functionally equivalent for attribution purposes (Mukherjee and Chang, 2025).

Future research can develop quantitative metrics that capture an agent's decision-making depth and creative originality. Empirical validation across domains—such as text generation, music composition, and visual art—would demonstrate whether these metrics reliably predict when an agent's contribution merits standalone authorial credit. Also, given the stochastic interplay of multi-agent pipelines, there is a need for algorithms that can disentangle and visualize each agent's creative "fingerprint." Described AI techniques could be adapted to highlight which components of an output were most influenced by which agent, thus attributing based on measurable statistical contributions.

Resource-Efficient Orchestration While MAS promise remarkable creative capabilities through parallel specialization, they also introduce substantial computational overhead, making resource-efficient orchestration an urgent challenge (Creech et al., 2021). Naively spawning dozens of agents can lead to prohibitive latency, high cloud costs, and unsustainable energy consumption. **Self-Resource Allocation** (Amayuelas et al., 2025) mechanism lets each agent budget its own compute, achieving near-optimal cost–performance tradeoffs. **DynTaskMAS** (Yu et al., 2025) leverages dynamic task graphs to asynchronously decompose workflows, reducing execution time by up to 33% and improving utilization by 35%. **MaAS**'s agentic supernet adapts architecture to each query, slashing inference costs to 6–45 % of static systems (Zhang et al., 2025).

Future research can explore adaptive agent pruning and distillation techniques that dynamically identify and deactivate or compress agents whose incremental contributions to a creative task fall below a meaningful threshold, yielding a leaner ensemble that retains quality while dramatically lowering computational overhead. Complementing this, meta-learning for orchestration policies could train a higher-order controller via meta-reinforcement learning to rapidly specialize scheduling strategies to new creative domains, such as narrative generation versus musical composition, using only a handful of trial interactions, thereby minimizing costly exploration in production. Finally, integrating human-in-the-loop orchestration channels will allow lightweight, real-time user feedback to signal when an intermediate creative draft meets subjective standards of "good enough," enabling the system to halt or redirect further agent invocations and align resource consumption with human satisfaction rather than arbitrary performance metrics.

Longitudinal User Studies In contrast to the abundance of controlled single-session evaluations, understanding how users engage with multi-agent creative systems over extended periods remains a significant hurdle. Longitudinal investigations have revealed that users undergo an initial novelty phase before stabilizing their expectations and customizing AI workflows (Long et al., 2024). In educational settings, semester-long dialogues with ChatGPT demonstrated evolving revision strategies and satisfaction levels. This underscores that early positive impressions can change as learners develop mental models of AI partners (Han et al., 2024a). Temporal pattern analysis in collaborative writing revealed distinct AI reliance phases, where users gradually transition from exploratory interactions to purpose-driven selective assistance as trust and competence grow (Yang et al., 2024).

Future work can focus on three directions. First, we need longitudinal studies that measure how users' creative abilities develop when collaborating with multi-agent systems. Through automated analysis and expert evaluation, these studies would track improvements in specific skills like narrative coherence or compositional technique. Second, researchers should investigate how users' and LLMs' understanding of different specialized agents (such as "plot architects" or "style editors") evolves. This research would examine how these evolving perceptions affect which helpers humans or AI systems choose to collaborate with during different parts of the creative process. Third, systems that can customize agent teams for individual users should be developed: automatically introducing new agents, removing unhelpful ones, or adjusting existing agents based on the user's preferences and performance. This would create personalized creative partnerships that support each user's ongoing artistic development.

F Additional Artifact Evaluation Criterion

In addition to the assessment methods discussed in Section 5, various other criteria have been employed to evaluate different aspects of generated content, such as:

- **Helpfulness**: Assesses the extent to which the artifact provides useful and informative content that effectively addresses the user's query or task
- **Relevance**: Measures how well the generated content aligns with the input prompt and context
- **Clarity**: Evaluates the ease of understanding the artifacts, focusing on the use of clear, concise, and unambiguous language

To provide a comprehensive overview, Table 4 summarizes and categorizes the evaluation approaches utilized in the cited works. It is important to note that this summary focuses exclusively on studies where the primary objective is to assess the creativity of the generated content, deliberately excluding those centered on accuracy, precision, or similar metrics. Furthermore, systems that do not include an evaluation of the generated content are also omitted from this overview.

Some abbreviations in the table are explained in the follows:

- **AUT (Alternative Uses Test)**: A divergent thinking task where participants list as many alternative uses as possible for a common object
- **RAT (Remote Associates Test)**: A creativity assessment where individuals are presented with three seemingly unrelated words and must identify a fourth word that connects them all, evaluating associative thinking and creative potential
- **MICSI (Mixed-Initiative Creative Support Index)**: A framework assessing systems that facilitate collaborative creativity between humans and computers, emphasizing interactive co-creation processes

G Methods of User Studies

In evaluating interactions between users and MAS in creative tasks, researchers commonly employ two primary methods: **Self-Report Instruments and Interviews**, and **Researcher Observation and Analysis**. The former involves collecting assessments and feedback directly from users, while the latter entails researchers analyzing user interactions to derive insights.

Self-Report Instruments and Interviews

- **Self-Report Instruments** involve participants completing surveys or questionnaires to provide personal assessments of their creative experiences and outputs. These tools often utilize Likert scales or other quantitative measures to gauge aspects such as exploration, expressiveness, perceived creativity, and enjoyment during creative tasks. Standardized questionnaires like the **Creativity Support Index (CSI)** (Cherry and Latulipe, 2014) are commonly used for this purpose.
- **Interviews**, on the other hand, offer a qualitative approach to understanding user experiences. They can be structured, semi-structured, or unstructured, allowing researchers to delve deeper into participants' thoughts, feelings, and behaviors. Interviews can provide rich, narrative feedback that complements quantitative data, offering a more

comprehensive view of user interactions with MAS.

Researcher Observation and Analysis This method involves researchers directly examining how users interact with MAS, and the artifacts they produce, to gain insights into the creative process and system usability. Observations can be conducted in real experiment scenarios (live observation) or through the analysis of video and audio recordings, system interaction logs, or textual transcriptions of the interactions. The analysis may focus on interaction patterns, problem-solving approaches, expressions of creativity, and usability issues.

By combining these methods, often within a mixed-methods framework, researchers can obtain a comprehensive understanding of both the subjective experiences and observable behaviors of users interacting with MAS in creative contexts. The subsequent section will discuss specific studies that utilize these methodologies to evaluate their systems.

ContextCam (Image Generation) Both the methods are used in this work (Fan et al., 2024). *Self-Report Instruments* captured users' subjective feedback, indicating positive engagement and enjoyment with the system's creative inspiration. *Interviews* delved deeper, exploring how users perceived and utilized context-aware features and their influence on the creative process. Findings from interviews highlighted users' insights into contextual data's role in image themes, behaviors, and inspiration. *Researcher Observation and Analysis* involved examining user interactions and analyzing system log data.

Virtual Canvas (Idea Generation) The user study investigated how groups generate ideas with an LLM in a virtual environment (He et al., 2024). *Self-Report Instruments* are not implemented. *Interviews* explored participants' perceptions of the AI's contribution to their ideation process, how it influenced group collaboration, and the challenges or benefits they encountered. *Researcher Observation and Analysis* focus on analyzing the group's interaction patterns within the virtual canvas, observing how the LLM's input was utilized, and identifying novel user needs of the system.

CoQuest (Research Ideation) The user study was conducted to investigate the impact of AI processing delays on the co-creative process (Liu et al., 2024b). *Self-Report Instruments* measured participants' subjective experiences, such as their perception of the degree of control of the system, how much they trust the system, and the inspiration from the help of the system. *Interviews* were used to gain deeper qualitative insights into participants' thought processes during co-creation with the system, exploring the differences between breadth-first search and depth-first search they felt during the experiments. *Researcher Observation and Analysis* focused on analyzing the interaction dynamics between the human and the LLM agent within the co-creation task. This would involve observing how participants reacted to delays, how they utilized the time during delays, their interaction patterns with the virtual environment and the agent.

Human-AI Co-creativity (Creative Writing) The user study explored the dynamics of human-LLM collaboration in prewriting (Wan et al., 2024). *Self-Report Instruments* are not implemented. *Interviews* were a primary method to investigate human-LLM collaboration patterns and dynamics during prewriting. These explored participants' experiences across the identified three-stage co-creativity process (Ideation, Illumination, and Implementation), delving into their thoughts on the LLM's role, initiative, and contributions, as well as uncovering collaboration breakdowns and user perceptions of using LLMs for prewriting. *Researcher Observation and Analysis* involved analyzing the co-creative process through screen recordings or analysis of interaction logs. This observation focused on identifying the iterative nature of the collaboration, how the human and AI took initiative, and how ideas were developed and refined across the different stages of prewriting.

MAS Technique	Task Domain	Framework
Divergent Exploration	AUT and RAT	Long-Term Guidance (2025)
	Character Design	PersonaGym (2024)
	Creative Writing	Creativity Support (LLMs) (2024)
	Humor Co-Creation	Meme Alone (2025)
	Idea Generation	Co-GPT Ideation (2024)
	Idea Generation	Group-AI Brainwriting (2024)
	Idea Generation	Virtual Canvas (2024)
	Image Generation	ContextCam (2024)
	Interior Color Design Ideation	C2Ideas (2024)
	Research Ideation	Ideation Co-Pilot (2024)
	Research Ideation	PersonaFlow (2024c)
	Scientific Research Co-creation	VirSci (2025)
	Sketches Generation	StoryDrawer (2022)
	Story Generation	ICCRI (2025)
	Story Generation	SPARKIT (2024)
Iterative Refinement	Character Design	CharacterMeet (2024)
	Debating (Fairness)	Multi-Agent Debate (2024)
	Hallucination Mitigation	Hallucination Mitigation (2024b)
	Legal Consultation	LawLuo (2024)
	LLM Pipeline Generation	ChainBuddy (2025)
	Product Design	DesignGPT (2023)
	Scientific Research Co-creation	Baby-AIGS-MLer (2024)
	Screenwriting	HoLLMwood (2024)
	Sketches Generation	CICADA (2024)
	Social Simulation	Generative Agents (2023)
Collaborative Synthesis	Agent Benchmarking	TheAgentCompany (2024)
	Agent Collab. Visualisation	AgentCoord (2024)
	Cognitive Synergy	Solo Performance Prompting (2024c)
	Creative Writing	Human-AI Co-creativity (2024)
	Creativity Simulation	Creative Agents (2024)
	Formal Proof	ProofNet (2023)
	Program Design	Beyond Code Generation (2025)
	Recommendation	Agent4Rec (2024a)
	Research Ideation	CoQuest (2024b)
	Research Peer Review	MARG (2024)
	Scientific Peer Review	AgentReview (2024)
	Scientific Research Co-creation	CrewAI (2024)
	Scientific Research Co-creation	Co-Scientist (2025)
	Silhouette Generation	Human-Machine Co-Creation (2024)
	Software Engineering	ChatDev (2024)
	Software Engineering	MaCTG (2025)
	Story Generation	CollabStory (2024)
	Story Generation	Colin (2024)
	Story Generation	Mathemyths (2024b)
	Story Generation	StoReys (2024)
	Story Generation	StoryVerse (2024b)
	UI Prototyping	MAxPrototyper (2024)

Table 1: Overview of representative MAS frameworks categorized by their core creative techniques—Divergent Exploration, Iterative Refinement, and Collaborative Synthesis.

Task	**Task Description**	**Available Datasets**
Problem-Solving in Physically Grounded Scenarios	Test multi-agent's ability to think resourcefully and act creatively in novel physical situations.	MacGyver (Tian et al., 2024)
Creative Writing	Evaluate the writing skills and collaborative abilities of multi-agents.	Human-AI co-writing Stories (Chakrabarty et al., 2024), Collab-Story (Venkatraman et al., 2024)
Music Genre	Evaluate the multi-agents in robot dance creation.	(De Filippo et al., 2024)
Character Design	Evaluating the creativity of multi-agent systems in visualizing and generating new characters.	(Lataifeh et al., 2024)
Question Answering	Open-ended question task.	TriviaQA (Joshi et al., 2017), QA in Game simulation (Park et al., 2023), GPQA Diamond Set (Rein et al., 2024)
Codenames Task	Evaluate the models' ability to identify words associated with a given word.	(Srivastava et al., 2023)
Mathematical Formal Proof Generation and Verification	Test the model's ability of autoformalization and formal proving of undergraduate-level mathematics.	(Azerbayev et al., 2023)
Idea Generation	Quantitatively evaluate and compare the ideas generated by LLMs.	AI Idea Bench 2025 (Qiu et al., 2025), AMiner Computer Science Dataset (Tang et al., 2008), LiveIdeaBench (Ruan et al., 2024)
Fairness-Aware Debating	Evaluate the ethical and practical implications of automated decision-making systems in the justice system.	COMPAS dataset (Larson et al., 2016)

Table 2: Creative tasks along with their associated datasets. Tasks that lack datasets and rely primarily on user studies are not included.

Granularity	**Framework**	**Method**	**Persona Example**
Coarse	Solo Performance Prompting (2024c)	Model-Generated	Self-Defined
	LLM Discussion (2024)	Model-Generated	Self-Defined
	PersonaGym (2024)	Human-Defined	Self-Defined
	Baby-AIGS-MLer (2024)	Human-Defined	Assistant
	SPARKIT (2024)	Human-Defined	Self-Defined
	Multi-Agent Debate (2024)	Human-Defined	Debater
	Acceleron (2024)	Human-Defined	Mentor & Colleague
	ChainBuddy (2025)	Human-Defined	Mentor & Planner
Medium-Coarse	TheAgentCompany (2024)	Model-Generated	Company Employee
	HoLLMwood (2024)	Human-Defined	Artist
	TRIZ Agents (2025)	Human-Defined	Problem Solver
	Co-Scientist (2025)	Human-Defined	Researcher
	MaCTG (2025)	Human-Defined	Programmer
	DesignGPT (2023)	Human-Defined	Self-Defined
	CoQuest (2024b)	Human-Defined	Researcher
	LawLuo (2024)	Human-Defined	Lawyer
	MARG (2024)	Human-Defined	Expert
Fine	PersonaFlow (2024c)	Data-Derived	Researcher
	VirSci (2025)	Data-Derived	Researcher
	Agent4Rec (2024a)	Data-Derived	Media User
	PersonaLLM (2024)	Model-Generated	Self-Defined
	The Power of Personality (2025)	Model-Generated	Self-Defined
	Creative Agents (2024)	Model-Generated	Artist
	CoAGent (2023b)	Model-Generated	Self-Defined
	Generative Agents (2023)	Human-defined	Sandbox Character
	AgentReview (2024)	Human-defined	Reviewer

Table 3: Summary of agent profile granularity and generation methods in MAS, with each paradigm's role definition and paper citation. 'Self-defined' personas grant agents the freedom to adopt diverse characters, promoting flexible collaboration and creative innovation.

Paper	Task	Subjective	Objective
Kumar et al. (2025)	AUT and RAT	TTCT, Boden's Criterion, and others	Semantic similarity
Duan et al. (2025)	AUT and others	TTCT	-
Lu et al. (2024)	AUT and others	TTCT	-
Ge et al. (2025)	Conceptual Design	TTCT	-
Lim and Perrault (2024)	Idea Generation	TTCT	-
Shaer et al. (2024)	Idea Generation	Innovation, Insightfullness, and others	Semantic Similarity
Sun et al. (2024)	Legal Consultation	Personalization and Professionalism	-
Azerbayev et al. (2023)	Mathematical Proving	-	BLEU Score
Ding et al. (2023)	Product Design	Novelty, Completeness, and Feasibility	-
Wan and Kalman (2025)	Plot Generation	-	Semantic Similarity
Chakrabarty et al. (2024)	Poem Writing	Fluency and Creativity	-
D'Arcy et al. (2024)	Paper Review Generation	Specificity and Overall Rating	-
Liu et al. (2024b)	Research Ideation	Boden's criterion	-
Liu et al. (2024c)	Research Ideation	Creativity, Usefulness, and Helpfulness	-
Chen et al. (2024)	Screenwriting	Interestingness, Relevance and others	Entropy-n, Self-BLEU and others
Gottweis et al. (2025)	Scientific Research Co-creation	Novelty and Impact	-
Radensky (2024)	Scientific Research Co-creation	Novelty, Specificity, and others	Semantic Similarity
Su et al. (2025)	Scientific Research Co-creation	Novelty, Clarity, Feasibility	Semantic Euclidean Distance
Zhang et al. (2024b)	Story Generation	Readability, Perceived Creativity, and others	-
Mellouli et al. (2024)	Story Generation	Interactivity, Coherence, and others	Self-BLEU
Lin et al. (2022)	Story Generation	Goal completion and Satisfication	-
Venkatraman et al. (2024)	Story Generation	Creativity	Entropy and others
Ali et al. (2025)	Story Generation	TTCT	-
Hou et al. (2024)	Interior Color Design Ideation	Inspiring, Reasonableness, and others	-
Venkatesh et al. (2025)	Image Editing & Generation	Expressiveness, Aesthetic appeal, and others	CLIP scores and others
Wu et al. (2025)	Meme Generation	Funniness, Creativity, and Shareability	-
Zhang et al. (2022)	Sketches Generation	TTCT	-
Ibarrola et al. (2024)	Sketches Generation	TTCT	FID, TIE ,and Semantic loss
Lawton et al. (2023)	Sketches Generation	Novelty and Surprise within MICSI	-
Lataifeh et al. (2024)	Silhouette Generation	Designer's review	FID
Yuan et al. (2024)	UI Prototype Generation	-	FID and Generation Diversity

Table 4: Output evaluation methods employed across various tasks. The upper section details evaluations for text generation tasks, while the lower section focuses on image generation tasks.