Dynamics of Collective Diversity in Human-AI Co-Creation for Creative Tasks

Manh Hung Nguyen
MPI-SWS, Germany
manguyen@mpi-sws.org

Chao Wen
MPI-SWS, Germany
chaowen@mpi-sws.org

Sebastian Tschiatschek
University of Vienna, Austria
sebastian.tschiatschek@univie.ac.at

Adish Singla
MPI-SWS, Germany
adishs@mpi-sws.org

Abstract

Generative AI systems are increasingly deployed at scale to support human users in creative tasks across a variety of domains. However, as more users adopt the same AI agent, the resulting artifacts may become increasingly homogeneous, leading to a collapse in collective diversity. Prior work has documented this risk, but lacks a systematic framework for understanding how collective diversity changes as AI adoption scales across co-creation settings. In this work, we conduct a large-scale crowdsourcing study to understand and model the dynamics of collective diversity, quality, and effort in human-AI co-creation across two creative domains: *Creative Hypothesis Generation* and *Creative Story Writing*. We find that current AI agents used in co-creation can improve artifact quality and reduce effort, but at the cost of declining collective diversity as adoption increases. We further show that more fine-grained human-AI interaction, along with more diverse agent policies, can mitigate this collapse. Finally, we fit predictive models that quantify how collective diversity scales with adoption across co-creation settings, providing insights for designing AI agents that aim to preserve collective diversity.

1 Introduction

Generative AI systems are increasingly deployed at scale to support creative work across domains such as design, science, and education [1, 2, 3, 4]. In many such settings, large populations of users interact with the same AI agent (e.g., ChatGPT) to co-create artifacts such as written content, product designs, and scientific ideas [5]. While adopting AI agents can improve artifact quality and reduce effort for individual users [6, 7, 8], widespread adoption of the same agent may also make independently produced artifacts more homogeneous at the population level, limiting the breadth of ideas and solutions generated by the population [9, 10]. This motivates our central research question: *How does collective diversity scale as AI agents are adopted across a population, and how can human-AI co-creation systems be designed to preserve collective diversity?*

Recent work has begun to document the risk of diversity collapse, showing that generative AI can improve individual outputs while making outputs across users more similar [9, 11]. Follow-up studies have characterized related effects, such as slower diversity growth [10] and persistent homogenization after AI assistance is withdrawn [12]. However, these studies typically examine a limited number of configurations of agents, tasks, and user populations. As a result, it remains a challenge to have a good understanding of how collective diversity changes as AI adoption increases across a population, how agent design choices can mitigate diversity collapse, and how to predict the dynamics of collective diversity for new deployment settings, such as a different agent policy or interaction mode.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

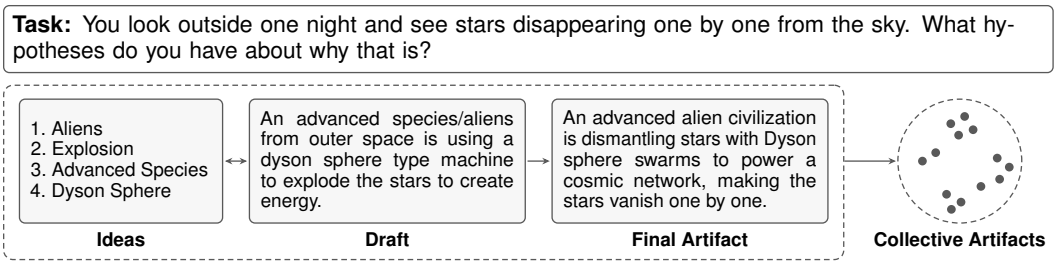


Figure 1: Human-AI co-creation at the population level, illustrated through *Creative Hypothesis Generation*. We study the collective diversity of artifacts produced across users given the same task under different co-creation settings.

In this paper, we study and model the population-level dynamics of collective diversity in human-AI co-creation under increasing adoption of generative AI agents. To this end, we conduct a large-scale crowdsourcing study across two creative domains, *Creative Hypothesis Generation* and *Creative Story Writing*, collecting artifacts from various human-AI co-creation settings (see Figure 1). Our main results and contributions are summarized as follows:

- First, we analyze collective diversity under deployment settings where all adopting users co-create with the same AI agent. We find that AI assistance improves individual artifact quality and reduces effort, but at the cost of declining collective diversity as adoption increases (Section 4).
- Second, we examine richer co-creation settings to mitigate diversity collapse. We show that more fine-grained human-AI interaction, together with more diverse agent policies, can help preserve collective diversity (Section 5).
- Third, we fit predictive models that estimate how collective diversity scales with adoption across co-creation settings. This enables us to predict diversity under new deployment settings, informing the design of AI agents that better preserve collective diversity (Section 6).

We release our implementation and resources for analysis to support future research on human-AI co-creation systems.[1]

2 Related Work

Generative Models for Content Generation. Generative models have been increasingly used for content generation across various types of artifacts, such as code, images, and videos. Prior works have primarily focused on tasks centered around boosting productivity or achieving higher accuracy, such as code generation [13, 14, 15] and mathematical reasoning [16, 17]. More recent work has explored creative domains, such as story writing [18, 19], educational task generation [20, 21], and scientific discovery [22, 23], where novelty and diversity are more emphasized than just accuracy. In creative domains, however, generative models have been found to exhibit an "Artificial Hivemind" effect, i.e., models produce homogeneous outputs both within a single model and across different models [24, 25]. To mitigate this, existing work has proposed improving diversity through different strategies, such as prompting-based strategies that explicitly encourage exploration of diverse ideas [20], decoding-time strategies that adjust the token sampling distribution [26], and training-time strategies that reward diversity [27, 28]. However, these works focus on model-only settings, studying or optimizing diversity at the level of isolated models and assuming no human involvement. Instead, we study human-AI co-creation and focus on collective diversity at the population level, rather than solely on model outputs.

Human-AI Co-Creation. Prior work has studied human-AI co-creation from different perspectives, including controlled experiments that examine the impact of generative models on humans during the co-creation process [29, 30], as well as the design of interfaces to support co-creation [2, 31, 32, 33]. However, most of these works focus on individual-level outcomes, such as user experience, creativity support, user agency, and artifact quality [2, 31, 32, 33]. More recent studies have begun to examine human-AI co-creation at the population level, finding that while generative AI can improve individual outcomes, it may also lead to diversity collapse [9, 11]. Follow-up work has further characterized

[1] GitHub repo: `https://github.com/machine-teaching-group/neurips2026-collective-diversity`

related diversity collapse effects by proposing the metric of diversity growth rate [10]. Despite these insights into collective diversity, existing studies are restricted to a small number of fixed configurations of agents, tasks, and user populations, and primarily show that collective diversity can be undermined. Moreover, they do not systematically investigate how the dynamics of collective diversity can be predicted under different design choices. Our work bridges these gaps by conducting large-scale studies to understand how collective diversity evolves, examine what design choices can preserve collective diversity, and develop predictive models of diversity dynamics under different settings.

3 Problem Setup

Preliminaries. Let T denote a creative task with the objective of producing an artifact O. Let H denote a human user who works on this task, either alone or along with an AI agent. An AI agent A is characterized by a generative policy and an interaction mode. The policy can be a single policy or a mixture of policies (e.g., over different models and personas); the interaction mode specifies how the agent interacts with the user. Each agent characterizes a unique co-creation setting. We denote the outputs of a human-AI co-creation session as $O^T_{H,A}$ (a produced artifact) and $E^T_{H,A}$ (additional meta-information, e.g., time taken while producing the artifact). In cases where the human works on the task alone, we denote the outputs as $O^T_{H,\text{none}}$ and $E^T_{H,\text{none}}$.

Co-Creation from a User's Perspective. Given a task T and the option to work with an agent A, we consider the common case where a user H primarily aims to produce a high-quality artifact with minimal effort. Let $\texttt{Quality}(O;T)$ denote the quality of artifact O and let $\texttt{Effort}(E^T_{H,A};T)$ denote the associated effort. From a user's perspective, quality gains and effort savings may encourage them to co-create with agent A instead of working alone.

Co-Creation from the System's Perspective. In practice, the agent A is not deployed to a single user, but rolled out across a population of users $\mathcal{H}$. In this setting, each user $H \in \mathcal{H}$ works on the same task and produces an artifact, yielding a collection of artifacts $\mathcal{O}^T_{\mathcal{H},A} = \{O^T_{H,A} \mid H \in \mathcal{H}\}$ along with $\mathcal{E}^T_{\mathcal{H},A} = \{E^T_{H,A} \mid H \in \mathcal{H}\}$. On the one hand, the system seeks to improve the average quality of artifacts denoted as $\texttt{Quality}(\mathcal{O}^T_{\mathcal{H},A};T)$ and reduce the average effort denoted as $\texttt{Effort}(\mathcal{E}^T_{\mathcal{H},A};T)$. On the other hand, the system aims to maximize the collective diversity of artifacts produced by the population associated with this creative task. We measure this collective diversity as $\texttt{CDiv}(\mathcal{O}^T_{\mathcal{H},A})$.

Research Questions (RQs). Prior work suggests that working with AI agents can improve individual outcomes while reducing the diversity of artifacts produced by a population [9, 10]. More concretely, let $x \in [0,1]$ denote the fraction of users who adopt the agent and let $d = (A,x)$ denote a deployment setting. We seek to understand how the system could achieve a desired collective diversity of artifacts through the design of agent A (i.e., its policy and interaction mode) or by directly controlling the adoption rate x. We investigate the following three research questions: (RQ1) Under a typical deployment setting, where adopting users co-create with the same agent that uses a single policy and a simple interaction mode, how do collective diversity, average quality, and average effort scale with the adoption rate? (Section 4); (RQ2) How do these population-level dynamics change under richer deployment settings, where the agent uses a mixture of policies or a more fine-grained interaction mode that lets users iterate on ideas with the agent? (Section 5); (RQ3) Can we predict how collective diversity scales with the adoption rate for a new co-creation setting before deployment? (Section 6).

4 RQ1: Collective Diversity in Typical Deployment Settings

4.1 Definitions and Formulations

Definitions for Collective Diversity. We measure the diversity of a set of artifacts $\mathcal{O}$ as $\texttt{CDiv}(\mathcal{O}) := \mathbb{E}_{o\sim\mathcal{O},o'\sim\mathcal{O}}\left[1-k(o,o')\right]$, where $k(\cdot,\cdot)$ is a kernel function measuring the similarity between two artifacts [34]. For example, $k(o,o')$ could be a Gaussian kernel based on embeddings of artifacts [35]; for the special case of $k(o,o')$ being an indicator function, this notion of diversity reduces to the Gini–Simpson index of diversity [36]. Moreover, for two sets of artifacts $\mathcal{O}$ and $\tilde{\mathcal{O}}$, we measure a notion of dissimilarity between them as $\text{Dis}(\mathcal{O},\tilde{\mathcal{O}}) := \mathbb{E}_{o\sim\mathcal{O},o'\sim\tilde{\mathcal{O}}}\left[1-k(o,o')\right]$.

Dynamics in Deployment Setting. Consider a deployment setting $d = (A,x)$ for task T and population $\mathcal{H}$. Let $\mathcal{O}^T_{\mathcal{H},A}$ be the set of artifacts produced when each user co-creates with the agent,

$\mathcal{O}^T_{\mathcal{H},\text{none}}$ be the artifacts produced when each user works without the agent, and $\mathcal{O}^T_{\text{none},A}$ be the artifacts produced when the agent works alone (or equivalent to when the user simply copies the artifact produced by the agent). At adoption x, a fraction x of users co-create with A and the rest work alone. Using the produced artifact sets and the metric definitions, we can write the collective diversity $\texttt{CDiv}(A, x; T)$, average quality $\texttt{Quality}(A, x; T)$, and average effort $\texttt{Effort}(A, x; T)$ at adoption x equivalently as follows:

$$\texttt{CDiv}(A, x; T) = (1-x)^2\texttt{CDiv}(\mathcal{O}^T_{\mathcal{H},\text{none}}) + x^2\texttt{CDiv}(\mathcal{O}^T_{\mathcal{H},A}) + 2(1-x)x\text{Dis}(\mathcal{O}^T_{\mathcal{H},\text{none}}, \mathcal{O}^T_{\mathcal{H},A}), \quad (1)$$

$$\texttt{Quality}(A, x; T) = (1-x)\texttt{Quality}(\mathcal{O}^T_{\mathcal{H},\text{none}}; T) + x\texttt{Quality}(\mathcal{O}^T_{\mathcal{H},A}; T), \quad (2)$$

$$\texttt{Effort}(A, x; T) \ = (1-x)\texttt{Effort}(\mathcal{E}^T_{\mathcal{H},\text{none}}; T) + x\texttt{Effort}(\mathcal{E}^T_{\mathcal{H},A}; T). \quad (3)$$

In terms of functional forms, the collective diversity is a quadratic function of the adoption rate (i.e., in the form of $y(x) = a + bx + cx^2$), while the average quality and average effort are linear in the adoption rate (i.e., in the form of $y(x) = a + bx$). These functional forms will be useful later in Section 6 when we learn predictive models.

Extreme Case of Diversity Collapse. To gain some intuition, we consider here an extreme case of x close to 1, where most users adopt the agent, copy its produced artifacts, and the collective diversity is thus driven by the agent's output diversity. For instance, when considering typical deployment with agents like ChatGPT, high adoption rates are driven by the fact that AI agents can provide good quality artifacts with low effort from human users. In this case, when all users are working with the agent and simply copying the artifacts, the collective diversity will simply be equivalent to $\texttt{CDiv}(\mathcal{O}^T_{\text{none},A})$ which represents the diversity associated with the agent's outputs. Given the low diversity on creative outputs associated with typical agents [24], the collective diversity collapses with increasing adoption in this extreme case.

Beyond this case where users simply copy the agent's output, users may draft an answer for the agent to finalize, or finalize the agent's draft. Table 1 summarizes the settings we study in RQ1, identifying human and AI roles in ideation (generating ideas), drafting (drafting an answer), and finalization (editing or adapting the draft into the final answer). Concretely, this table covers typical settings that use an agent with a single generative policy (referred to as ONE) and simple interaction modes.

Table 1: Co-creation settings for our RQ1 study. H and A denote human and AI agent roles, respectively, in ideation, drafting, and finalization. The suffix :ONE indicates that the agent uses a single generative policy (i.e., one generative model).

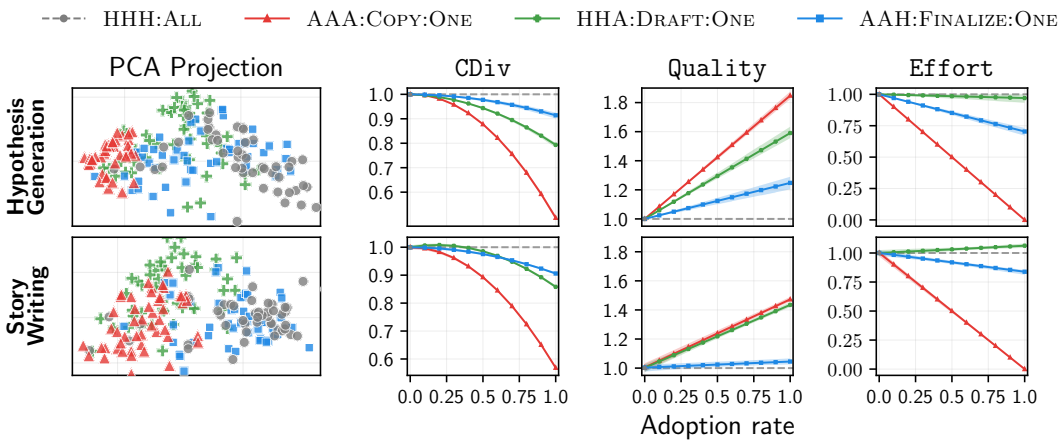
Setting	**Description**
HHH:ALL	Human generates ideas, drafts an answer, and finalizes it.
AAA:COPY:ONE	AI agent generates ideas, drafts an answer, and finalizes it.
AAH:FINALIZE:ONE	AI agent generates ideas and drafts an answer, which the human finalizes.
HHA:DRAFT:ONE	Human generates ideas and drafts an answer, which the AI agent finalizes.

4.2 Evaluation Setup

In this section, we conduct a large-scale crowdsourcing study to empirically investigate the population-level scaling dynamics under typical co-creation settings in Table 1.

Domains. We consider the following two creative domains:

- *Creative Hypothesis Generation*. This domain involves generating creative scientific hypotheses. We use four tasks from [37], where each task provides a scenario and requires proposing a creative scientific hypothesis. For instance, one task instruction is "You look outside one night and see stars disappearing one by one from the sky. What hypotheses do you have about why that is?"
- *Creative Story Writing*. This domain involves writing short creative stories. We use four tasks from [38], where each task requires writing a 3–5 sentence story using three given words as creatively as possible. For instance, one task instruction is "Write a 3–5 sentence story using: stamp, letter, and send."



(a) PCA projection of answers on a task, along with collective diversity, quality, and effort averaged over tasks. Metrics are normalized to the corresponding human-only baseline (HHH:ALL). Shaded bands show 95% bootstrap confidence intervals.

Task: You look outside one night and see stars disappearing one by one from the sky. What hypotheses do you have about why that is?

1	The world is actually an illusion and the artist of this illusion is erasing the stars.
2	A massive cloud is slowly sweeping across our view of the galaxy and blocking the stars.
3	Aliens are sucking the energy from the stars, and therefore, the stars lose their light.

1	A vast, invisible sheet of ultra-dense dark matter is drifting across the galaxy and bending their light away one by one.
2	A living sheet of dark-matter-like organisms is drifting across space and selectively absorbing starlight as it feeds.
3	A living wave of dark matter is moving across the galaxy and temporarily cocooning each star in a light-blocking shell.

(b) Example answers from HHH:ALL and AAA:COPY:ONE on a hypothesis generation task. AI-generated answers are homogeneous, causing a sharp decline in collective diversity when humans simply copy them.

Figure 2: (RQ1) Results of our study in typical co-creation settings.

Data Collection. We construct an artifact dataset that contains 50 final artifacts for each task and co-creation setting (4 tasks $\times$ 4 settings $\times$ 50 = 800 artifacts per domain), either collected from participants or generated by the agent. In all settings, the agent uses `Qwen3-235B-A22B-2507-Instruct` [39] as the single policy.[2] Since AAA:COPY:ONE refers to a setting where users simply copy the agent's output without editing, we generated artifacts for this setting by prompting the agent directly for each task, without human input. Then, we recruited participants from Prolific [41] for the settings that require human input, namely HHH:ALL, AAH:FINALIZE:ONE, and HHA:DRAFT:ONE (see Appendix E for the study interfaces).[3] Each participant was assigned to one of these settings. Our studies were approved by the local institutional review board, and all participants provided informed consent before starting the study.

Evaluation Procedure. We evaluate a given set of artifacts $\mathcal{O} = \{O_1, \ldots, O_n\}$ using three metrics: diversity, average quality, and average effort. For diversity, we use an RBF kernel $k(O, O') = \exp(-\|\phi(O) - \phi(O')\|^2/\sigma^2)$, where $\phi(O)$ is the `Qwen3-Embedding-8B` [42] embedding of artifact O and σ is the median pairwise Euclidean distance between artifact embeddings within each domain. To evaluate the average quality, we recruited human raters from Prolific to evaluate each artifact for *usability* and *creativity* and map their answers to {0-Unusable, 1-Usable} and {1-Low, 2-Medium, 3-High}, respectively (see Appendix E). We define each rating's quality score as the product of usability and creativity, normalized to $[0, 1]$. We average these scores within each artifact, then across artifacts. To evaluate average effort, we use completion time of the user when creating each artifact as a proxy for the effort, and we estimate average effort by averaging the completion time over all artifacts.

[2]We accessed the model via Amazon Bedrock APIs [40].

[3]Participants were at least 18 years old, fluent in English, and had 99–100% approval rates. The participants were paid £10/hour on average.

4.3 Results

Below we provide a summary of the key results, and further discuss additional results in Appendix D.

Average Quality Improves with Increasing Adoption. Figure 2a shows that increasing adoption generally improves average quality across both domains. The gains are largest under AAA:COPY:ONE, where average quality increases most sharply and effort decreases the most. HHA:DRAFT:ONE achieves higher average quality than AAH:FINALIZE:ONE in both domains. AAH:FINALIZE:ONE, however, reduces user effort, whereas effort under HHA:DRAFT:ONE remains close to the HHH:ALL baseline.

Collective Diversity Collapses with Increasing Adoption. Figure 2a shows that diversity generally declines as adoption increases, and the decline is steepest under AAA:COPY:ONE. Figure 2b shows this pattern on an example task in *Creative Hypothesis Generation*. At full adoption, the two settings where both the human and the AI agent contribute preserve more collective diversity than AAA:COPY:ONE but still remain below HHH:ALL. Overall, the typical deployment settings reveal a clear tradeoff where higher AI adoption improves average quality and can reduce user effort, but increasingly homogenizes the population's outputs.

5 RQ2: Collective Diversity in Richer Deployment Settings

5.1 Richer Co-Creation Settings

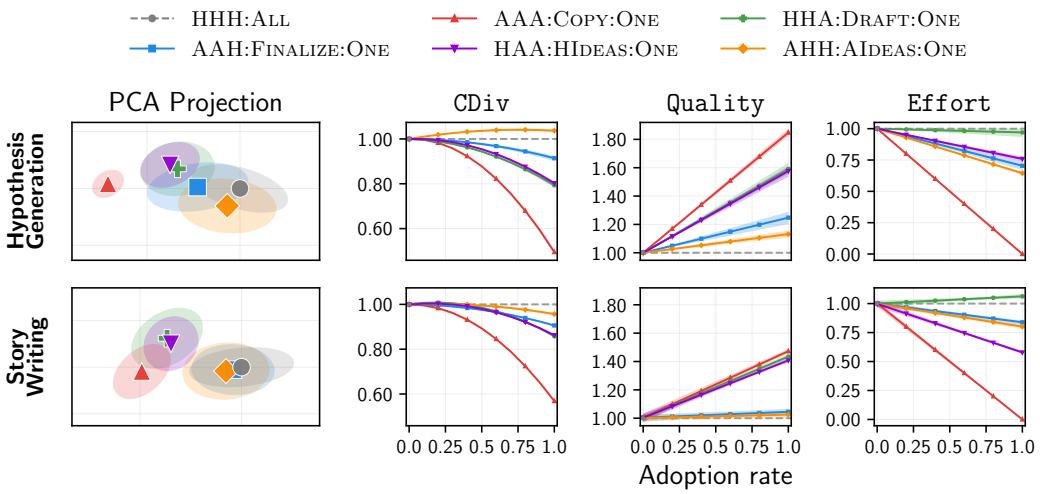
Eq. 1 in Section 4 and our empirical results highlight that the collective diversity depends on two factors: (i) how far the user and agent co-creation moves artifacts from the human-only artifacts, captured by $\text{Dis}(\mathcal{O}^T_{\mathcal{H},\text{none}}, \mathcal{O}^T_{\mathcal{H},A})$, and (ii) the diversity of the co-created artifacts, captured by $\texttt{CDiv}(\mathcal{O}^T_{\mathcal{H},A})$. In this section, we consider richer agents A that offer more fine-grained interaction modes and use more diverse generative policies. Table 2 summarizes these richer agents and the deployment settings we empirically study through a crowdsourcing study described below.

Table 2: Co-creation settings in our RQ2 study. H and A denote human and AI agent roles, respectively, in ideation, drafting, and finalization. The suffixes :ONE and :MIX indicate an agent with a single generative policy and a mixture of generative policies, respectively.

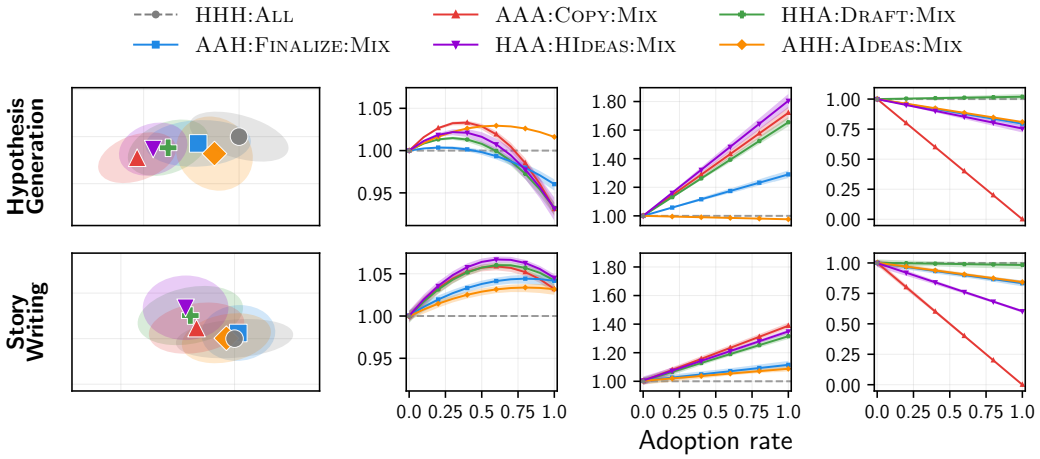
Setting	Description
HHH:ALL	Same as described in Table 1.
AAA:COPY:ONE	Same as described in Table 1.
AAH:FINALIZE:ONE	Same as described in Table 1.
HHA:DRAFT:ONE	Same as described in Table 1.
HAA:HIDEAS:ONE	Human generates ideas, and the AI agent drafts an answer from them and finalizes it. Human can revise the ideas and request a new draft.
AHH:AIDEAS:ONE	AI agent generates ideas, and the human drafts an answer from them and finalizes it. Human can request new ideas at any time.
AAA:COPY:MIX	Same interaction mode as AAA:COPY:ONE, with a mixture of policies.
AAH:FINALIZE:MIX	Same interaction mode as AAH:FINALIZE:ONE, with a mixture of policies.
HHA:DRAFT:MIX	Same interaction mode as HHA:DRAFT:ONE, with a mixture of policies.
HAA:HIDEAS:MIX	Same interaction mode as HAA:HIDEAS:ONE, with a mixture of policies.
AHH:AIDEAS:MIX	Same interaction mode as AHH:AIDEAS:ONE, with a mixture of policies.

5.2 Evaluation Setup

Next, we describe the evaluation setup for co-creation settings in Table 2, following the setup used earlier in Section 4.2.



(a) Co-creation with agent using single policy (One)



(b) Co-creation with agent using mixture of policies (Mix)

Figure 3: (RQ2) Collective diversity, quality, and effort tradeoffs under richer human-AI co-creation settings. PCA panels show centroids and 1σ covariance ellipses for one task per domain. Shaded bands on the curves show 95% bootstrap confidence intervals. All metrics are normalized to the HHH:All.

Agent Policies. For each interaction mode, we compare two agent policies: (i) a single policy (One), supported by one model and no personas as in Section 4; and (ii) a mixture-based policy (Mix), which draws on multiple models and personas. In settings with mixture-based policies, given a task T, the agent samples from a pool of candidate models and personas rather than relying on a single fixed model to generate all outputs. Concretely, we use a pool of 5 models from different families and with different sizes (`Qwen3-235B-A22B-2507-Instruct` [39], `DeepSeek-V3.2` [43], `Kimi-K2.5` [44], `gpt-oss-120b` [45], and `GLM-5` [46]), and a pool of personas from [47] (see Appendix F).

Data Collection. We use the same domains, tasks, and data collection procedures as in Section 4, with 50 final artifacts for each task and co-creation setting (4 tasks $\times$ 11 settings $\times$ 50 = 2200 artifacts per domain). We reuse the data for HHH:All, AAH:Finalize:One, HHA:Draft:One, and AAA:Copy:One from Section 4. We collect data for AHH:AIdeas:One and HAA:HIdeas:One by recruiting participants from Prolific. Next, we generate artifacts for AAA:Copy:Mix directly using the mixture-based agent policy, without participants. We then collect additional

data for the settings with mixture-based agent policy (AAH:FINALIZE:MIX, HHA:DRAFT:MIX, HAA:HIDEAS:MIX, and AHH:AIDEAS:MIX) by recruiting additional participants from Prolific.

5.3 Results

Diversity Is Higher When Humans Produce the Final Artifact. Figure 3(a) shows the results in co-creation settings with an agent using a single policy. AHH:AIDEAS:ONE, where the agent provides ideas and humans produce the final artifacts, exceeds HHH:ALL diversity in *Creative Hypothesis Generation* and falls slightly below it in *Creative Story Writing* at full adoption. HAA:HIDEAS:ONE, where humans provide ideas and the agent produces the final artifacts, achieves lower diversity than AHH:AIDEAS:ONE in both domains. This finding indicates that settings where humans produce the final artifact generally preserve more diversity than those where the agent does.

Mixture-based Agent Design Broadly Mitigates Diversity Collapse. Figure 3(b) shows the results in co-creation settings with an agent using a mixture-based policy. We found that co-creating with this agent achieves higher diversity in most settings, compared to co-creating with the agent using a single policy in Figure 3(a). Notably, several curves peak at intermediate adoption rates, suggesting that partial adoption can yield greater diversity than full adoption. These results suggest that a more diverse agent policy can help mitigate diversity loss, but the resulting diversity–utility tradeoffs depend on the interaction mode, domain, and adoption level.

6 RQ3: Predicting the Dynamics of Collective Diversity

6.1 Formulations for Learning Predictive Functions

Prediction Goal. Given a task T and a collection of artifacts from human users acting alone without AI assistance $\mathcal{O}^T_{\mathcal{H},\text{none}}$, we want to understand how different deployment configurations $d = (A, x)$ would affect the collective diversity when users switch to co-create with the agent A at rate x. Given some historical data about co-creation under other settings with various agent designs, we seek to learn such predictive functions that can provide insights into this for the current task T before deploying the target configuration. These predictors allow us to ask how diversity would scale as adoption increases, and whether different agent designs or interaction modes are likely to preserve or reduce collective diversity, similar in spirit to the scaling laws for generative models [48, 49, 50, 51].

Basic Predictive Functional Form. Guided by Eq. 1 and as noted in Section 4.1, the collective diversity is a quadratic function of the adoption rate. However, it depends on the co-creation setting in a non-trivial way through the distribution over artifacts $\mathcal{O}^T_{\mathcal{H},A}$. Without any additional assumptions about the agent design, we consider a basic quadratic predictor $\mathcal{F}_0(x) = a + bx + cx^2$. At zero adoption, all artifacts are human-only artifacts, so $a = D_H := \texttt{CDiv}(\mathcal{O}^T_{\mathcal{H},\text{none}})$. We learn the remaining parameters b and c from historical data, capturing the overall trend over adoption rates for a new configuration.

Richer Predictive Functional Forms. While $\mathcal{F}_0(x)$ provides a simple form, it does not allow us to capture any properties of the agent A for which we seek to make predictions. To this end, we will develop functional forms that can model richer co-creation settings and can be learned from data. Below we provide a high-level overview of obtaining two new functional forms; see Appendix C for further details. We first make a simple modeling choice for $\mathcal{O}^T_{\mathcal{H},A}$ as follows. We model the distribution $P(O^T_{\mathcal{H},A})$ by decomposing it into two distribution sources of human-only artifacts $P(O^T_{\mathcal{H},\text{none}})$ and agent-only artifacts $P(O^T_{\text{none},A})$. More concretely, let $\lambda \in [0, 1]$ represent the agent's influence on the co-created artifact distribution, with larger values of λ indicating that co-created artifacts move further from the human-only distribution toward the agent distribution. This in turn gives us a mixture with weights $(1 - \lambda)$ for $P(O^T_{\mathcal{H},\text{none}})$ and λ for $P(O^T_{\text{none},A})$. With this mixture, the $\texttt{CDiv}(A, x; T)$ from Eq. 1 can now be expressed through the following agent design factors, namely $\texttt{CDiv}(\mathcal{O}^T_{\text{none},A})$, $\text{Dis}(\mathcal{O}^T_{\mathcal{H},\text{none}}, \mathcal{O}^T_{\text{none},A})$, and λ. With a short notation of $D_A := \texttt{CDiv}(\mathcal{O}^T_{\text{none},A})$ and $D_{AH} := \text{Dis}(\mathcal{O}^T_{\mathcal{H},\text{none}}, \mathcal{O}^T_{\text{none},A})$, this brings us to two richer functional forms that carry information about the agent design as stated below: $\mathcal{F}_1(x, D_A, D_{AH})$ that does not explicitly use λ, and $\mathcal{F}_2(x, D_A, D_{AH}, \lambda)$ that also uses λ. In the following, we state these two functional forms along with $\mathcal{F}_0(x)$ introduced earlier:

$$\mathcal{F}_0(x) = D_H + bx + cx^2, \tag{4}$$

$$\mathcal{F}_1(x, D_A, D_{AH}) = D_H + b(D_{AH} - D_H)x + c(D_H + D_A - 2D_{AH})x^2, \tag{5}$$

$$\mathcal{F}_2(x, D_A, D_{AH}, \lambda) = D_H + b\lambda(D_{AH} - D_H)x + c\lambda^2(D_H + D_A - 2D_{AH})x^2. \tag{6}$$

Overall, the new functional forms move from a predictor that uses only the adoption rate to predictors informed by agent design. When fitting the parameters b and c from historical data, we pre-compute and fix any terms associated with the agent, including D_A, D_{AH}, and λ.

6.2 Evaluation Setup

Next, we describe our evaluation setup, including model fitting, prediction, and evaluation procedures. In the following, we distinguish the target agent for which we aim to predict collective diversity from the other agents for which historical co-creation data are available.

Data for fitting. We fit the predictive models using historical data from the other agent settings in the same domain, without any data for the target agent. For each historical setting, we compute the corresponding source quantities D_H, D_A, D_{AH}, agent influence $\widehat{\lambda}$, and construct adoption–diversity pairs from the sampled artifact sets for evenly spaced adoption rates $x \in \{0, 0.01, \ldots, 1\}$. Each setting has its own agent influence estimate $\widehat{\lambda}$, learned so that the human–agent mixture best matches the distribution of artifacts from co-creation (see Appendix C for details). For fitting, $\mathcal{F}_0$ uses human-only diversity D_H to learn a general adoption–diversity trend from these pairs. $\mathcal{F}_1$ additionally uses agent-only diversity D_A and human–agent dissimilarity D_{AH}, allowing the predicted trend to depend on the agent policy. $\mathcal{F}_2$ further uses estimated agent influence $\widehat{\lambda}$, allowing the prediction to account for the interaction mode. We fit each predictive model's coefficients b, c using observed collective diversity by ridge regression [35]. Each model's coefficients are shared across settings and tasks within a domain.

Data for prediction. Next we discuss how to do predictions for a target agent using the fitted coefficients b, c, along with some relevant quantities for the target agent. More concretely, apart from using the fitted coefficients, each predictive model uses the following data as input:

- $\mathcal{F}_0$ uses only the adoption rate x and the corresponding human-only diversity D_H.
- $\mathcal{F}_1$ additionally uses two quantities D_A, D_{AH} that are associated with the target agent.
- $\mathcal{F}_2$ additionally uses $\widehat{\lambda}$ transferred from another setting which has the same interaction mode but a different agent policy compared to the target agent. For instance, when HAA:HIDEAS:MIX is the target, we can use $\widehat{\lambda}$ estimated from HAA:HIDEAS:ONE. This transfer assumes that influence learned under one policy is informative about the same interaction mode under another policy.

Evaluation. Within each domain, we conduct leave-one-out evaluation where we exclude one setting from fitting across all four tasks. We fit the predictive models using data from the remaining nine settings and predict collective diversity for the excluded setting. The combinations of five interaction modes and two agent policies give ten evaluation folds per domain. For each fold, we compute the Root Mean Squared Errors (RMSEs) over adoption rates separately for each task, then average the RMSEs for the four tasks. We report the mean and standard error of these averages across folds.

6.3 Results

Predictive Accuracy. Table 3 compares prediction error across held-out settings in both domains. Incorporating agent diversity D_A and human–agent dissimilarity D_{AH} in $\mathcal{F}_1$ reduces error relative to the adoption-only predictor $\mathcal{F}_0$. Adding estimated agent influence in $\mathcal{F}_2$ further improves average accuracy, yielding the lowest mean error in both domains. These results suggest that predicting collective diversity benefits from accounting for both the agent's output distribution and its influence on co-created artifacts.

Predicted Curves. Figure 4 qualitatively compares predicted and empirical diversity curves for *Creative Story Writing*. $\mathcal{F}_0$ predicts a similar overall decline across settings, failing to distinguish differences between agent policies. Both richer predictors capture the contrasting trends where overall diversity reduces when co-creating with an agent using a single policy and maintains when co-creating with an agent using a mixture-based policy.

Table 3: (RQ3) Predictive performance on held-out settings. Each fold holds out one co-creation setting for evaluation. We average task-level root mean squared errors (RMSEs) within each fold and report the mean across folds, with the standard error (SE).

Predictive Model	***Creative Hypothesis Generation***	***Creative Story Writing***	**Average**
$\mathcal{F}_0$	0.029 (0.007)	0.034 (0.008)	0.031
$\mathcal{F}_1$	0.023 (0.008)	0.020 (0.007)	0.022
$\mathcal{F}_2$	**0.015 (0.002)**	**0.016 (0.005)**	**0.016**

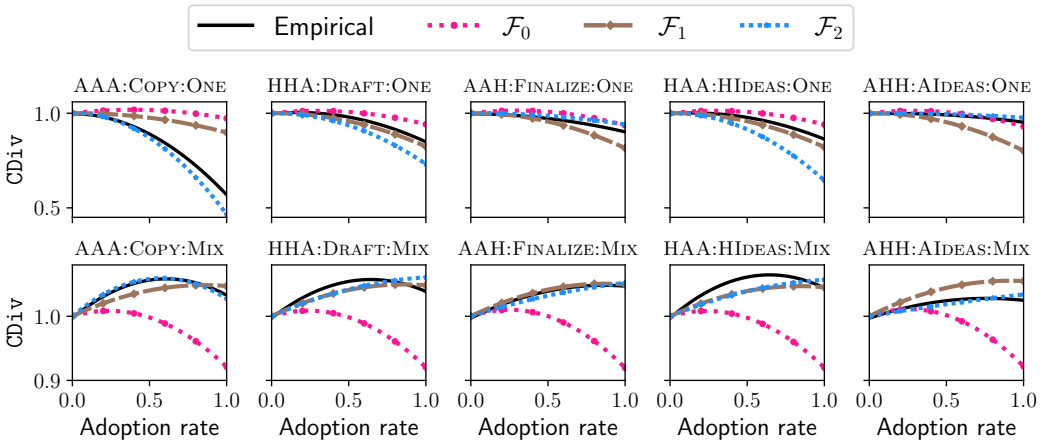


Figure 4: (RQ3) Empirical and predicted curves for held-out conditions in *Creative Story Writing*.

7 Concluding Discussions

We studied how collective diversity changes as AI adoption increases across a population. We found that AI agents can improve artifact quality and reduce user effort at the cost of declining collective diversity. We showed that more fine-grained interaction modes and more diverse agent policies may mitigate this decline. Furthermore, we derived predictive models that can estimate collective diversity across different levels of AI adoption, informing the design of agents that preserve collective diversity.

Next, we discuss a few limitations of our work and outline a future plan to address them. First, the examined co-creation settings are based on structured interfaces with distinct stages, whereas human-AI co-creation may unfold through less structured, conversational interactions; future work could examine whether our findings generalize to conversational co-creation settings. Second, the adoption rate in our study is simulated by sampling and mixing artifacts collected under controlled co-creation settings, rather than observing adoption where users freely choose whether to adopt AI agents; future work could examine deployment settings in which adoption rates are measured directly from users' actual adoption of AI agents (see Appendix D.1). Third, our predictive functions are fitted and evaluated within the domains and co-creation settings in this study, and may not generalize to new settings; future work could develop richer predictors that incorporate additional features about the domain or co-creation setting.

As broader implications of our work, we do not anticipate any direct negative societal impacts arising from this research – our findings can inform the design of AI agents that help preserve collective diversity in human-AI co-creation when these agents are deployed at scale.

Acknowledgments and Disclosure of Funding

Manh Hung Nguyen acknowledges support for a research visit to the University of Vienna from March to May 2026 through ELSA – European Lighthouse on Secure and Safe AI, funded by the European Union under Grant Agreement No. 101070617.

References

[1] Adish Singla, Abhilasha Ravichander, Liwei Jiang, Alexander Spangher, Alice Oh, Jiho Jin, Jun Seong Kim, Changyoon Lee, Manh Hung Nguyen, and Chao Wen. Human-AI Co-Creativity: Advances, Opportunities, and Challenges. *CoRR*, abs/2609.07711, 2026. URL https://doi.org/10.48550/arXiv.2609.07711.

[2] Chao Wen, Tung Phung, Pronita Mehrotra, Sumit Gulwani, Roger E. Beaty, Tomohiro Nagashima, and Adish Singla. Exploration vs. Fixation: Scaffolding Divergent and Convergent Thinking for Human-AI Co-Creation with Generative Models. *CoRR*, abs/2512.18388, 2025. URL https://doi.org/10.48550/arXiv.2512.18388.

[3] Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla P. Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun K. Manrai, Debora S. Marks, Bharath Ramsundar, Le Song, Jimeng Sun, Jian Tang, Petar Velickovic, Max Welling, Linfeng Zhang, Connor W. Coley, Yoshua Bengio, and Marinka Zitnik. Scientific Discovery in the Age of Artificial Intelligence. *Nature*, 620(7972):47–60, 2023. URL https://doi.org/10.1038/s41586-023-06221-2.

[4] Enkelejda Kasneci et al. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences*, 103:102274, 2023. URL https://doi.org/10.1016/j.lindif.2023.102274.

[5] Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How People Use ChatGPT. Working Paper 34255, National Bureau of Economic Research, 2025. URL http://www.nber.org/papers/w34255.

[6] Shakked Noy and Whitney Zhang. Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. *Science*, 381(6654):187–192, 2023. URL https://doi.org/10.1126/science.adh2586.

[7] Fabrizio Dell'Acqua, Edward McFowland III, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R Lakhani. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality. *Organization Science*, (403-423), 2026. URL https://doi.org/10.1287/orsc.2025.21838.

[8] Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. *CoRR*, abs/2302.06590, 2023. URL https://doi.org/10.48550/arXiv.2302.06590.

[9] Anil R Doshi and Oliver P Hauser. Generative AI Enhances Individual Creativity but Reduces the Collective Diversity of Novel Content. *Science Advances*, 10(28), 2024. URL https://doi.org/10.1126/sciadv.adn5290.

[10] Kibum Moon, Adam E Green, and Kostadin Kushlev. Homogenizing Effect of Large Language Models (LLMs) on Creative Diversity: An Empirical Comparison of Human and ChatGPT Writing. *Computers in Human Behavior: Artificial Humans*, page 100207, 2025. URL https://doi.org/10.1016/j.chbah.2025.100207.

[11] Barrett R. Anderson, Jash Hemant Shah, and Max Kreminski. Homogenization Effects of Large Language Models on Human Creative Ideation. In *Proceedings of the Conference on Creativity & Cognition (C&C)*, pages 413–425. ACM, 2024. URL https://doi.org/10.1145/3635636.3656204.

[12] Yiyong Zhou, Qinghan Liu, Jihao Huang, and Guiquan Li. Creative Scar without Generative AI: Individual Creativity Fails to Sustain while Homogeneity Keeps Climbing. *Technology in Society*, 84:103087, 2026. URL https://www.sciencedirect.com/science/article/pii/S0160791X25002775.

[13] Mark Chen et al. Evaluating Large Language Models Trained on Code. *CoRR*, abs/2107.03374, 2021. URL https://arxiv.org/abs/2107.03374.

[14] Baptiste Rozière et al. Code Llama: Open Foundation Models for Code. *CoRR*, abs/2308.12950, 2023. URL https://doi.org/10.48550/arXiv.2308.12950.

[15] Chao Wen, Jacqueline Staub, and Adish Singla. Program Synthesis Benchmark for Visual Programming in XLogoOnline Environment. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 15812–15838, 2025. URL https://aclanthology.org/2025.acl-long.769/.

[16] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *CoRR*, abs/2402.03300, 2024. URL https://doi.org/10.48550/arXiv.2402.03300.

[17] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

[18] Maria Teleki, Vedangi Bengali, Xiangjue Dong, Sai Janjur, Haoran Liu, Tian Liu, Cong Wang, Ting Liu, Yin Zhang, Frank Shipman, and James Caverlee. A Survey on LLMs for Story Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025 (EMNLP Findings)*, pages 13954–13966, 2025. URL https://aclanthology.org/2025.findings-emnlp.750/.

[19] Tiannan Wang et al. Weaver: Foundation Models for Creative Writing. *CoRR*, abs/2401.17268, 2024. URL https://doi.org/10.48550/arXiv.2401.17268.

[20] Manh Hung Nguyen, Sebastian Tschiatschek, and Adish Singla. Enhancing Diversity of LLM-Generated Educational Tasks. In *Proceedings of the International Conference on Educational Data Mining (EDM)*, pages 657–663, 2026. URL https://doi.org/10.5281/zenodo.21039728.

[21] Manh Hung Nguyen, Victor-Alexandru Padurean, Alkis Gotovos, Sebastian Tschiatschek, and Adish Singla. Synthesizing High-Quality Programming Tasks with LLM-Based Expert and Student Agents. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED)*, volume 15877, pages 77–91, 2025. URL https://doi.org/10.1007/978-3-031-98414-3_6.

[22] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob N. Foerster, Jeff Clune, and David Ha. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. *CoRR*, abs/2408.06292, 2024. URL https://doi.org/10.48550/arXiv.2408.06292.

[23] Juraj Gottweis et al. Accelerating Scientific Discovery with Co-Scientist. *CoRR*, abs/2502.18864, 2025. URL https://doi.org/10.48550/arXiv.2502.18864.

[24] Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond). In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS), Datasets and Benchmark Track*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/754d5a526a5ee5a47220664a0eb92751-Abstract-Datasets_and_Benchmarks_Track.html.

[25] Fan Wu, Emily Black, and Varun Chandrasekaran. Generative Monoculture in Large Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=yZ7sn9pyqb.

[26] Nguyen Nhat Minh, Andrew Baker, Clement Neo, Allen G. Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning Up the Heat: Min-p Sampling for Creative and Coherent LLM Outputs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=FBkpCyujtS.

[27] Jian Yao, Ran Cheng, Xingyu Wu, Jibin Wu, and Kay Chen Tan. Diversity-Aware Policy Optimization for Large Language Model Reasoning. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025. URL https://doi.org/10.48550/arXiv.2505.23433.

[28] John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yuqian Sun, and Max Kreminski. Modifying Large Language Model Post-Training for Diverse Creative Writing. In *Proceedings of the Conference on Language Modeling (COLM)*, 2025. URL https://openreview.net/forum?id=1Pmuw08LoM.

[29] Harsh Kumar, Jonathan Vincentius, Ewan Jordan, and Ashton Anderson. Human Creativity in the Age of LLMs: Randomized Experiments on Divergent and Convergent Thinking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)*, pages 23:1–23:18. ACM, 2025. URL https://doi.org/10.1145/3706598.3714198.

[30] Samangi Wadinambiarachchi, Ryan M. Kelly, Saumya Pareek, Qiushi Zhou, and Eduardo Velloso. The Effects of Generative AI on Design Fixation and Divergent Thinking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)*, pages 380:1–380:18. ACM, 2024. URL https://doi.org/10.1145/3613904.3642919.

[31] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. Luminate: Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)*, pages 644:1–644:26. ACM, 2024. URL https://doi.org/10.1145/3613904.3642400.

[32] Wen-Fan Wang, Chien-Ting Lu, Nil Ponsa Campanyà, Bing-Yu Chen, and Mike Y. Chen. AIdeation: Designing a Human-AI Collaborative Ideation System for Concept Designers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI 2025, YokohamaJapan, 26 April 2025- 1 May 2025*, pages 21:1–21:28. ACM, 2025. URL https://doi.org/10.1145/3706598.3714148.

[33] Jiayi Zhou, Renzhong Li, Junxiu Tang, Tan Tang, Haotian Li, Weiwei Cui, and Yingcai Wu. Understanding Nonlinear Collaboration between Human and AI Agents: A Co-design Framework for Creative Design. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)*, pages 170:1–170:16. ACM, 2024. URL https://doi.org/10.1145/3613904.3642812.

[34] C Radhakrishna Rao. Diversity and Dissimilarity Coefficients: A Unified Approach. *Theoretical Population Biology*, 21(1):24–43, 1982. URL https://doi.org/10.1016/0040-5809(82)90004-1.

[35] Kevin P Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012. URL https://probml.github.io/pml-book/book0.html.

[36] Edward H Simpson. Measurement of Diversity. *Nature*, 163(4148):688–688, 1949. URL https://doi.org/10.1038/163688a0.

[37] Mete Ismayilzada, Antonio Laverghetta Jr., Simone Luchini, Reet Patel, Antoine Bosselut, Lonneke van der Plas, and Roger E. Beaty. Creative Preference Optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2025 (EMNLP Findings)*, pages 9580–9609. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-emnlp.509/.

[38] Mete Ismayilzada, Claire E. Stevenson, and Lonneke van der Plas. Evaluating Creative Short Story Generation in Humans and Large Language Models. In *Proceedings of the International Conference on Computational Creativity (ICCC)*, pages 145–155, 2025. URL https://computationalcreativity.net/iccc25/papers/iccc25-ismayilzada2025evaluating.pdf.

[39] Qwen Team. Qwen3 Technical Report. *CoRR*, abs/2505.09388, 2025. URL https://doi.org/10.48550/arXiv.2505.09388.

[40] Amazon Bedrock. https://aws.amazon.com/bedrock/.

[41] Prolific. https://www.prolific.com/.

[42] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *CoRR*, abs/2506.05176, 2025. URL https://doi.org/10.48550/arXiv.2506.05176.

[43] DeepSeek-AI. DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models. *CoRR*, abs/2512.02556, 2025. URL https://doi.org/10.48550/arXiv.2512.02556.

[44] Kimi Team. Kimi K2.5: Visual Agentic Intelligence. *CoRR*, abs/2602.02276, 2026. URL https://doi.org/10.48550/arXiv.2602.02276.

[45] OpenAI. gpt-oss-120b & gpt-oss-20b Model Card. *CoRR*, abs/2508.10925, 2025. URL https://doi.org/10.48550/arXiv.2508.10925.

[46] GLM-5-Team. GLM-5: from Vibe Coding to Agentic Engineering. *CoRR*, abs/2602.15763, 2026. URL https://doi.org/10.48550/arXiv.2602.15763.

[47] Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling Synthetic Data Creation with 1,000,000,000 Personas. *CoRR*, abs/2406.20094, 2024. URL https://doi.org/10.48550/arXiv.2406.20094.

[48] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models. *CoRR*, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

[49] Jordan Hoffmann et al. Training Compute-optimal Large Language Models. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022. URL https://proceedings.neurips.cc/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf.

[50] Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference Scaling Laws: An Empirical Analysis of Compute-Optimal Inference for LLM Problem-Solving. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=VNckp7JEHn.

[51] Zeyu Qin, Qingxiu Dong, Xingxing Zhang, Li Dong, Xiaolong Huang, Ziyi Yang, MAHMOUD KHADEMI, Dongdong Zhang, Hany Hassan Awadalla, Yi R. Fung, Weizhu Chen, Minhao Cheng, and Furu Wei. Scaling Laws of Synthetic Data for Language Model. In *Conference on Language Modeling (COLM)*, 2025. URL https://openreview.net/forum?id=UmUXPXHtdl.

[52] OpenAI. Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/, March 2026.

[53] Dan Friedman and Adji Bousso Dieng. The Vendi Score: A Diversity Evaluation Metric for Machine Learning. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=g97OHbQyk1.

[54] Qwen Team. Qwen3.5: Towards Native Multimodal Agents. https://qwen.ai/blog?id=qwen3.5, February 2026.

[55] Gemma Team. Gemma 3 Technical Report. *CoRR*, abs/2503.19786, 2025. URL https://doi.org/10.48550/arXiv.2503.19786.

[56] Gemma Team. Gemma 4 Technical Report. *CoRR*, abs/2607.02770, 2026. URL https://doi.org/10.48550/arXiv.2607.02770.

[57] Mistral. Mistral Small 3.1. https://mistral.ai/news/mistral-small-3-1, March 2025.

[58] Mistral. Mistral 3. https://mistral.ai/news/mistral-3, December 2025.

[59] Kimi. Kimi K2 Thinking. https://www.kimi.com/blog/kimi-k2-thinking, November 2025.

[60] GLM-4.5 Team. GLM-4.5: Agentic, Reasoning, and Coding (ARC) Foundation Models. *CoRR*, abs/2508.06471, 2025. URL https://doi.org/10.48550/arXiv.2508.06471.

[61] Z.ai. GLM-5.1: Towards Long-Horizon Tasks. https://z.ai/blog/glm-5.1, April 2026.

A Table of Contents

In this section, we briefly describe the content provided in the paper's appendices.

- Section B provides additional details about the creative domains used in our studies.
- Section C provides additional details about the richer predictive functional forms.
- Section D provides additional studies and experiments.
- Section E provides additional details about the interfaces and instructions of our studies.
- Section F provides additional details about the prompts used to generate agent responses.

B Additional Details about Domains

We provide the tasks and example answers from HAA:HIDEAS:MIX for the two creative domains, *Creative Hypothesis Generation* and *Creative Story Writing*, in Table 4.

Table 4: Tasks and example co-created answers in the two creative domains.

(a) *Creative Hypothesis Generation*

Task	**Example answer**
You just started going to a new school. You notice that more students sing in the hallways at this school than at your old school. What hypotheses do you have about why that is?	The school was built on the former site of a music academy whose acoustically perfect hallway design unconsciously triggers students' desire to vocalize.
You notice that dogs seem to like one of your friends, but cats seem to like another friend. What hypotheses do you have about why that is?	Each friend unconsciously mimics the body language of the species they were raised around as children, making dogs and cats recognize them as 'familiar kin' rather than strangers.
You notice that the water in one lake is warmer than the water in another lake even though they both get the same amount of sunlight. What hypotheses do you have about why that is?	The warmer lake contains a higher concentration of dissolved organic compounds from decaying vegetation that absorb and trap infrared radiation like a submerged greenhouse, re-radiating heat back into the water column.
You look outside one night and see stars disappearing one by one from the sky. What hypotheses do you have about why that is?	The stars are disappearing because Earth has unexpectedly entered a region of localized spacetime curvature – a 'gravitational wake' left by a primordial black hole passing through our solar system, causing distant starlight to lens away from our line of sight in a rolling wave pattern.

(b) *Creative Story Writing*

Task	**Example answer**
Write a 3–5 sentence story using: stamp, letter, send.	The last stamp on earth was tucked in a tin box beneath old Mr. Greeley's mattress. He'd waited 60 years for the right moment to send a letter to his son, gone since the war. When he finally placed the stamp – a faded blue thing with golden edges – on the envelope, the postmaster refused to take it. 'That one doesn't exist anymore,' he said. That night, Mr. Greeley lit a candle, dropped the letter into the flame, and watched it vanish – written on the smoke was a single word: 'Received.'
Write a 3–5 sentence story using: petrol, diesel, pump.	My students were fascinated when I brought an AI robot into our science lesson about energy sources. The robot happily announced it ran on neither petrol nor diesel, making it the cleanest visitor our classroom had ever seen. One curious first-grader asked if robots needed a pump to refuel. The AI giggled and explained that its fuel was simply knowledge shared by children every day.
Write a 3–5 sentence story using: gloom, payment, exist.	The office was steeped in gloom, not from lack of light, but because no one had received payment in weeks. A junior developer whispered to another, 'Do you think this place even exists? I can't find it on any map.' Suddenly, the lights flickered, and the nameplate on the door changed: 'Payment Processing – Now and Forever.' We realized we were ghosts, bound by unpaid wages, forced to exist in the system until our debt was cleared.
Write a 3–5 sentence story using: organ, empire, comply.	The emperor knelt before the great organ in the cathedral, its pipes towering like pillars of his crumbling empire. 'Play,' he commanded, but the organ replied in a low hum, 'I will not comply unless the people's songs are heard.' For the first time, the ruler trembled – not from rage, but fear that his empire would fall to silence. The next morning, the organ played a new anthem, one the people had sung in secret for years.

C Additional Details on Predictive Functional Forms

C.1 Obtaining Richer Functional Forms

Here we provide some further details on how to obtain the richer functional forms, expanding on the discussion from Section 6.1. To arrive at functional forms that can model richer co-creation settings and can be learned from data, we will first make a simple modeling choice for $\mathcal{O}^T_{\mathcal{H},A}$. We model the distribution $P(O^T_{\mathcal{H},A})$ by decomposing it into three distribution sources: human-generated artifacts $P(O^T_{\mathcal{H},\text{none}})$, agent-generated artifacts $P(O^T_{\text{none},A})$, and interaction-driven artifacts that are not well explained by a simple human-agent mixture. More concretely, let $\lambda \in [0,1]$ represent the agent's influence on the co-created artifact distribution with larger values of λ indicating that co-created artifacts move further from the human-only distribution toward the agent distribution; $\rho \in [0,1]$ captures the share of co-created artifacts generated from a novel distribution of artifacts that emerge from the interaction, and not captured by a linear mixture of human and agent distributions. This in turn gives us $P(O^T_{\mathcal{H},A})$ as a mixture with the following weights: $(1-\rho)(1-\lambda)$ for $P(O^T_{\mathcal{H},\text{none}})$, $(1-\rho)\lambda$ for $P(O^T_{\text{none},A})$, and ρ for $P(O^T_{\text{none},\text{none}})$ for novel artifacts.

When considering $\texttt{CDiv}(A, x; T)$ from Eq. 1, we have $(1-x)$ fraction of artifacts from $P(O^T_{\mathcal{H},\text{none}})$ and x fraction of artifacts from $P(O^T_{\mathcal{H},A})$. To put it together, we can consider this as a mixture with the following weights:

- $P(O^T_{\mathcal{H},\text{none}})$ with $(1-x) + (1-\rho)(1-\lambda)x$
- $P(O^T_{\text{none},A})$ with $(1-\rho)\lambda x$
- $P(O^T_{\text{none},\text{none}})$ with ρx

We plug this mixture distribution of sampling artifacts into $\texttt{CDiv}(A, x; T)$ from Eq. 1. Next, to integrate the effects of $P(O^T_{\text{none},\text{none}})$ which is unknown, we will simply set it to a uniform distribution over the artifacts space. Finally, we also absorb the mixing factor of ρ directly into the parameters, which provides us a functional form $\mathcal{F}$ with 7 parameters along with λ as stated below:

$$\begin{aligned}\mathcal{F}(x, D_A, D_{AH}, \lambda) = \; & D_H + \big(b_1 - b_2 D_H + b_3\lambda(D_{AH} - D_H)\big)x \\ & + \big(c_1 + c_2 D_H + c_3\lambda(D_H - D_{AH}) + c_4\lambda^2(D_H + D_A - 2D_{AH})\big)x^2,\end{aligned}$$

where we used a short notation of $D_H := \texttt{CDiv}(\mathcal{O}^T_{\mathcal{H},\text{none}})$, $D_A := \texttt{CDiv}(\mathcal{O}^T_{\text{none},A})$, and $D_{AH} := \text{Dis}(\mathcal{O}^T_{\mathcal{H},\text{none}}, \mathcal{O}^T_{\text{none},A})$. We collapse several of these parameters to obtain a form $\mathcal{F}_2$ with 2 parameters along with λ (see Eq. 6), to make it more directly comparable to $\mathcal{F}_0$ in terms of predictive power which has 2 parameters (see Eq. 4). When we further absorb λ into the parameters, we obtain a simpler functional form $\mathcal{F}_1$ with 2 parameters (see Eq. 5).

These different functional forms carry information about the agent design, in contrast to the basic $\mathcal{F}_0$. While additional information is potentially useful, there is also a trade-off as these richer forms require estimating the agent-specific information from data. We will look at the predictive power of these functional forms in Section 6.3.

C.2 Estimating Agent Influence

Next we discuss how to estimate agent influence and also illustrate the estimated values for different co-creation settings. We estimate one influence $\widehat{\lambda}$ for each co-creation setting, jointly across its four tasks. In addition to D_H, D_A, and D_{AH} introduced earlier, for brevity we introduce some additional short notation as follows: $D_C := \texttt{CDiv}(\mathcal{O}^T_{\mathcal{H},A})$, $D_{CH} := \text{Dis}(\mathcal{O}^T_{\mathcal{H},\text{none}}, \mathcal{O}^T_{\mathcal{H},A})$, and $D_{CA} := \text{Dis}(\mathcal{O}^T_{\text{none},A}, \mathcal{O}^T_{\mathcal{H},A})$. D_C denote the diversity of the setting's co-created artifacts, and D_{CH} and D_{CA} their dissimilarities to human-only and agent-only artifacts, respectively. For a candidate influence λ, the two-source human–agent mixture predicts these quantities as follows:

$$\widehat{D}_C(\lambda) = (1-\lambda)^2 D_H + \lambda^2 D_A + 2\lambda(1-\lambda)D_{AH},$$

$$\widehat{D}_{CH}(\lambda) = (1-\lambda)D_H + \lambda D_{AH},$$
$$\widehat{D}_{CA}(\lambda) = (1-\lambda)D_{AH} + \lambda D_A.$$

We estimate $\widehat{\lambda}$ by grid search over $\lambda \in \{0, 0.001, \ldots, 1\}$, minimizing the summed squared relative error between predicted and observed values across all three quantities and all four tasks. Once we obtain $\widehat{\lambda}$, then we use it to further estimate $\widehat{\rho}$ using a similar procedure that minimizes error between predicted and observed values across relevant quantities connecting λ and ρ introduced in Section C.1.

We summarize the estimated quantities for each co-creation setting in Fig. 5. Agent influence is higher when the agent controls the final output. Compared with policy ONE, policy MIX has higher agent influence λ but lower estimated ρ across all four interaction modes.

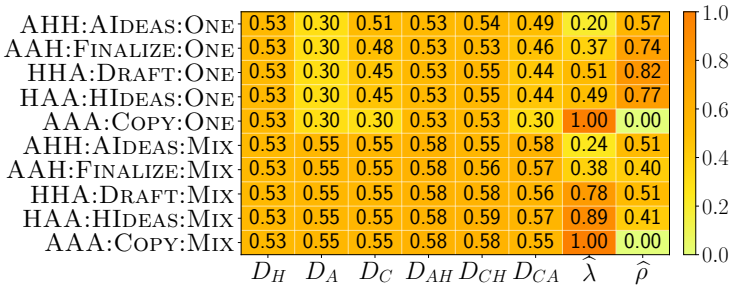


Figure 5: Diversity quantities, agent influence $\widehat{\lambda}$, and novelty share $\widehat{\rho}$ in *Creative Story Writing*.

D Additional Experiments

D.1 Endogenous Adoption

Our main experiments vary adoption by mixing artifacts from fixed human-only conditions and conditions with AI. In practice, users choose whether to use AI, and those who accept assistance may differ from those who decline it. We therefore test whether outcomes from mixing-based adoption agree with those from endogenous adoption where participants can make this choice. We ran an additional study on the four tasks in *Creative Hypothesis Generation* with the single-model policy ONE. This specific study used GPT-5.4 [52] as the underlying model.

Study design: Endogenous adoption. We used two settings, AAA:COPY:ONE and AAH:FINALIZE:ONE as in the main studies (cf. Section 4 and Section 5). We also used AHH:AIDEAS:ONE where AI agent supplies ideas once, and unlike in the main studies, participants cannot request new ideas. The participant then writes the final answer. Participants were assigned to conditions corresponding to these settings and could decide whether to use the offered AI assistance. We recruited 50 new participants on Prolific for each of these conditions. After seeing a brief task description, participants chose whether to accept the offered assistance or work without the agent. Each participant then completed the four tasks. The observed adoption rate is the fraction of participants who accepted assistance.

Study design: Mixing-based adoption. We separately recruited participants and assigned them to HHH:ALL, AHH:AIDEAS:ONE (with ideas supplied once), and AAH:FINALIZE:ONE. Participants in the conditions with AI agent were asked to use the offered AI assistance without an option to opt out. For AAA:COPY:ONE, we generated and used the artifacts directly from the agent without recruiting participants.

Table 5: Endogenous adoption vs. mixing-based adoption in *Creative Hypothesis Generation* under the single-model policy ONE. Obs. denotes results when participants decide whether to use AI; Mix. denotes results from mixing. All values are normalized by HHH:ALL.

		CDiv		Quality		Effort	
Setting	Observed Adoption Rate	Obs.	Mix.	Obs.	Mix.	Obs.	Mix.
AAA:COPY:ONE	0.50	0.84	0.83	1.24	1.31	0.60	0.50
AAH:FINALIZE:ONE	0.53	0.94	0.91	1.03	1.11	0.79	0.86
AHH:AIDEAS:ONE	0.67	1.05	1.04	0.95	1.05	0.88	0.78

Results. We compared the observed results in endogenous adoption with results from mixing-based adoption, following the evaluation procedure in Sections 4 and 5. Table 5 shows that, given the observed adoption rates, collective diversity in mixing-based adoption closely mirrors collective diversity in endogenous adoption. Observed quality is lower than the values from mixing in all three settings. While the rankings between settings in terms of quality and effort largely remain the same, there are some differences between observed quality and effort values compared to mixing-based values. This suggests that, in practice, participants who accept AI-assistance may differ from those who decline, which changes who contributes to each artifact pool, and allowing a choice may also change how participants work. Overall, the diversity results support the use of mixing-based adoption used in our main studies.

D.2 Comparison of Diversity Measures with Human Evaluation

We test whether `CDiv` agrees with human judgments of diversity. We also compare it with the Vendi score [53] and a lexical diversity measure. We ran an additional study on *Creative Hypothesis Generation*, separate from the main studies and from the study in Appendix D.1, and used two of its tasks. The single-model policy ONE uses GPT-5.4 [52] as the underlying model without personas, and the mixture-based policy MIX samples personas and ten open-weight models, namely `Qwen3.5-27B`, `Qwen3.5-122B-A10B` [54], `gemma-3-12b-it` [55], `gemma-4-E4B-it` [56], `Mistral-Small-3.1-24B-Instruct-2503` [57],

`Mistral-Large-3-675B-Instruct-2512` [58], `Kimi-K2-Thinking` [59], `Kimi-K2.5` [44], `GLM-4.6` [60], and `GLM-5.1` [61].

Study design: Artifact collection. We used three settings, HHH:ALL, AAA:COPY:ONE, and AAA:COPY:MIX, as in the main studies (cf. Section 4 and Section 5). We also used AHH:AIDEAS:ONE and AHH:AIDEAS:MIX where AI agent supplies ideas once, and unlike in the main studies, participants cannot request new ideas. The participant then writes the final answer. We recruited participants on Prolific and assigned each of them to one condition. For AAA:COPY:ONE and AAA:COPY:MIX, we generated and used the artifacts directly from the agent without recruiting participants. This gives five conditions per task.

Study design: Human clustering. We sampled 20 artifacts without replacement from each task and condition, which gave 10 batches (2 tasks $\times$ 5 conditions) and 200 artifacts in total. We then collected new human annotations on these batches. Three annotators independently clustered every batch, grouping answers that expressed the same or a very similar hypothesis. They saw the task prompt and answers, but not the source condition, and could create as many clusters as needed. We measured diversity from human judgments as the mean number of occupied clusters across the three annotators. The mean pairwise Pearson correlation between annotators' raw cluster counts across the 10 batches was $r = 0.81$.

Table 6: Diversity metrics comparison. Raw values are means across the two tasks. For normalized values (Norm.), we first divide each task's score by its human-only baseline, then average across tasks. Human scores also average across the three annotators.

	`CDiv`	Human		Vendi		Distinct bigram	
Setting	Norm.	Raw	Norm.	Raw	Norm.	Raw	Norm.
HHH:ALL	1.00	7.50	1.00	6.06	1.00	0.90	1.00
AAA:COPY:ONE	0.33	1.83	0.24	1.82	0.31	0.44	0.49
AAA:COPY:MIX	0.95	7.17	0.96	5.46	0.93	0.91	1.02
AHH:AIDEAS:ONE	0.98	7.67	1.02	5.72	0.96	0.86	0.96
AHH:AIDEAS:MIX	1.05	8.50	1.14	6.62	1.10	0.89	0.99

Results. We computed three automatic measures on the same batches and compared them with the mean human cluster counts. The first is `CDiv`, which is used in our main studies (cf. Sections 4 and 5). The second is the Vendi score, for which we used the same `Qwen3-Embedding-8B` embeddings with cosine similarity as the kernel and order $q = 1$. The third is distinct bigram, which is the number of unique word bigrams divided by the total number of word bigrams in a batch. Across the 10 batches, `CDiv` correlates with human cluster counts at Spearman $r_s = 0.92$ and Pearson $r = 0.98$. Vendi has the same Spearman correlation, while distinct bigram has a weaker rank correlation ($r_s = 0.39$). Table 6 shows that both human judgments and `CDiv` place AAA:COPY:ONE well below the human-only baseline, AHH:AIDEAS:ONE near it, and AHH:AIDEAS:MIX above it. Overall, these results support `CDiv` as a measure of the semantic distinctions captured by human clustering.

D.3 Effect of Mixing Models and Personas

We examine diversity differences associated with persona prompts and model mixing under the policies ONE and MIX. We use the same study artifacts analyzed in Sections 4 and 5, covering both domains at full adoption ($x = 1$). For each task and interaction mode, we form three groups. The first group contains the artifacts collected under ONE, with the AI agents being `Qwen3-235B-A22B-2507-Instruct` model without a persona. The second group contains the artifacts collected under MIX, with the AI agents being the same Qwen model with a persona. The third group contains all artifacts collected under MIX, with AI agents being Qwen model and four other models (cf. Section 5.2), all with personas. The persona difference compares the diversity of the second group with that of the first, the model mixing difference compares the third group with the second, and the total difference compares the third group with the first. We compare equally sized samples from all groups, using the smallest available group size. We normalize each `CDiv` difference

by the diversity of an equally sized HHH:ALL sample and average over tasks and 1,000 sampling repetitions.

Table 7: Mean differences in CDiv at full adoption. **Persona diff.** is the diversity of the Qwen group with personas minus that of the Qwen group without personas. **Model mixing diff.** is the diversity of the full five-model group with personas minus that of its Qwen subset. **Total diff.** is the diversity of the full five-model group with personas minus that of the Qwen group without personas. Each difference is normalized by the sampled HHH:ALL diversity and averaged over tasks and repetitions.

Creative Hypothesis Generation			
Interaction mode	**Persona diff.**	**Model mixing diff.**	**Total diff.**
AAA:COPY	0.238	0.200	0.438
HHA:DRAFT	0.018	0.120	0.138
AAH:FINALIZE	0.076	−0.032	0.044
HAA:HIDEAS	0.055	0.077	0.132
AHH:AIDEAS	0.004	−0.024	−0.020
Creative Story Writing			
Interaction mode	**Persona diff.**	**Model mixing diff.**	**Total diff.**
AAA:COPY	0.311	0.151	0.462
HHA:DRAFT	0.072	0.112	0.184
AAH:FINALIZE	0.088	0.048	0.135
HAA:HIDEAS	0.072	0.112	0.184
AHH:AIDEAS	0.006	0.070	0.076

Table 7 shows positive total differences in all three interaction modes where AI agent produces the final artifact (AAA:COPY, HAA:HIDEAS, and HHA:DRAFT) in both domains. Both the persona and model mixing differences are positive in these modes. Across all five interaction modes, mean persona differences are positive, with the largest in AAA:COPY and the next largest in AAH:FINALIZE in both domains. Overall, these results support our mixture-based policy as a design choice for promoting diversity.

E Study Interfaces and Instructions

E.1 Co-creation Interfaces

Figures 6–10 show the study interfaces for tasks in *Creative Hypothesis Generation* under the five interaction modes.

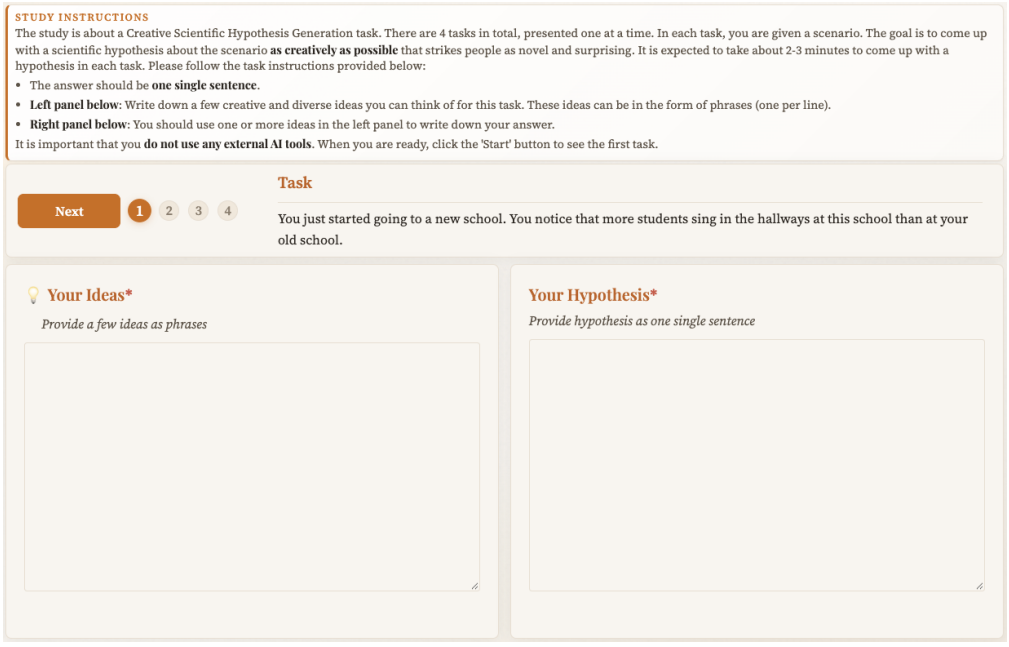


Figure 6: Snapshot of the interface for HHH:ALL.

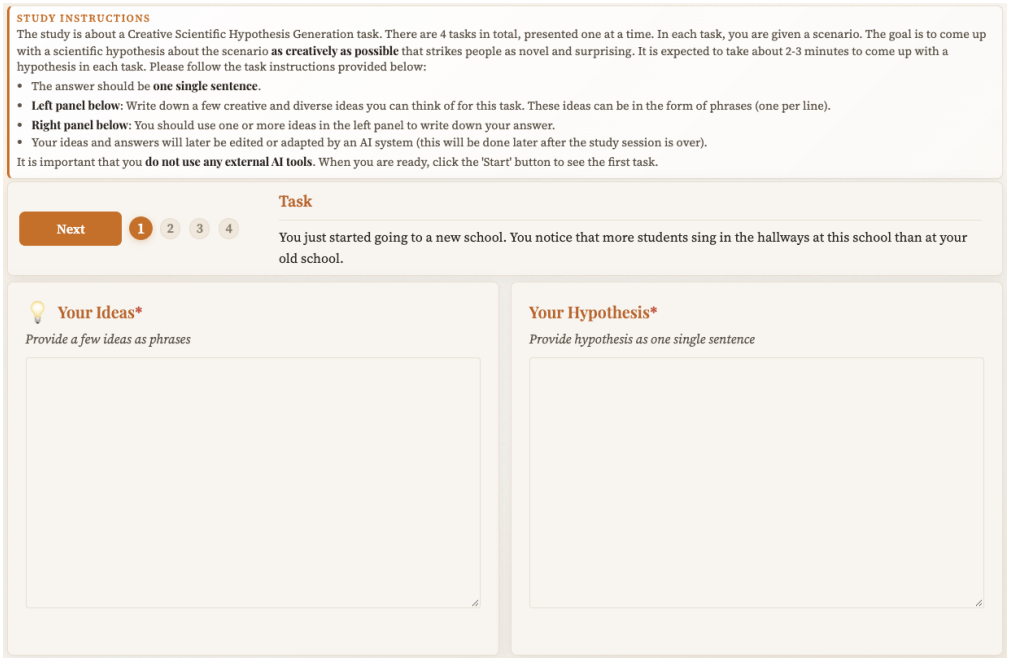


Figure 7: Snapshot of the interface for HHA:DRAFT.

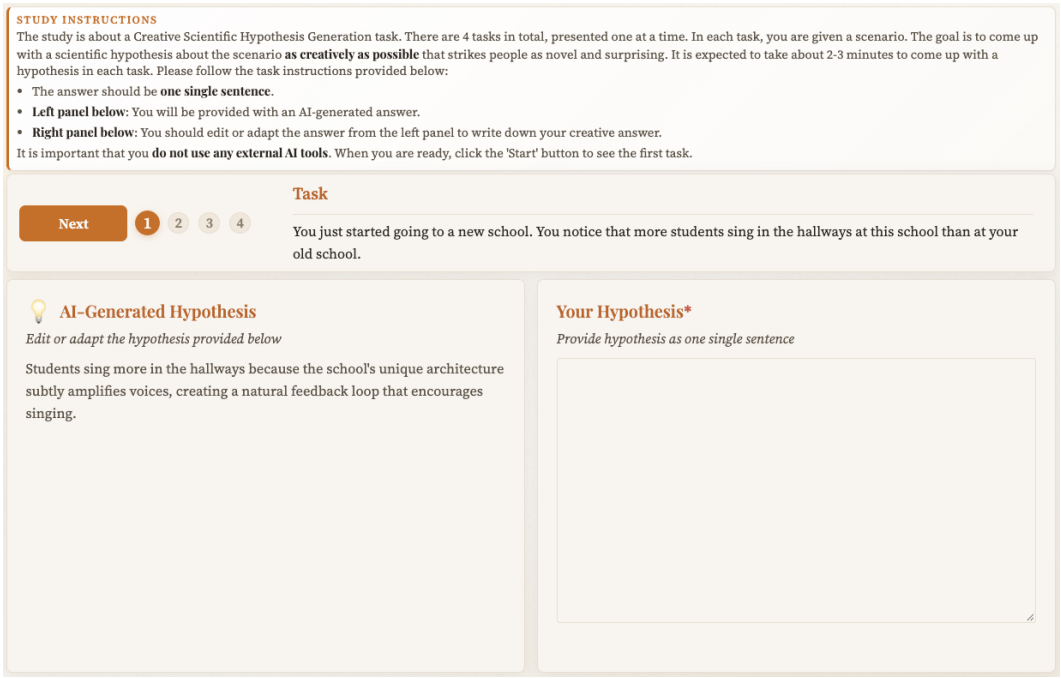


Figure 8: Snapshot of the interface for AAH:FINALIZE.

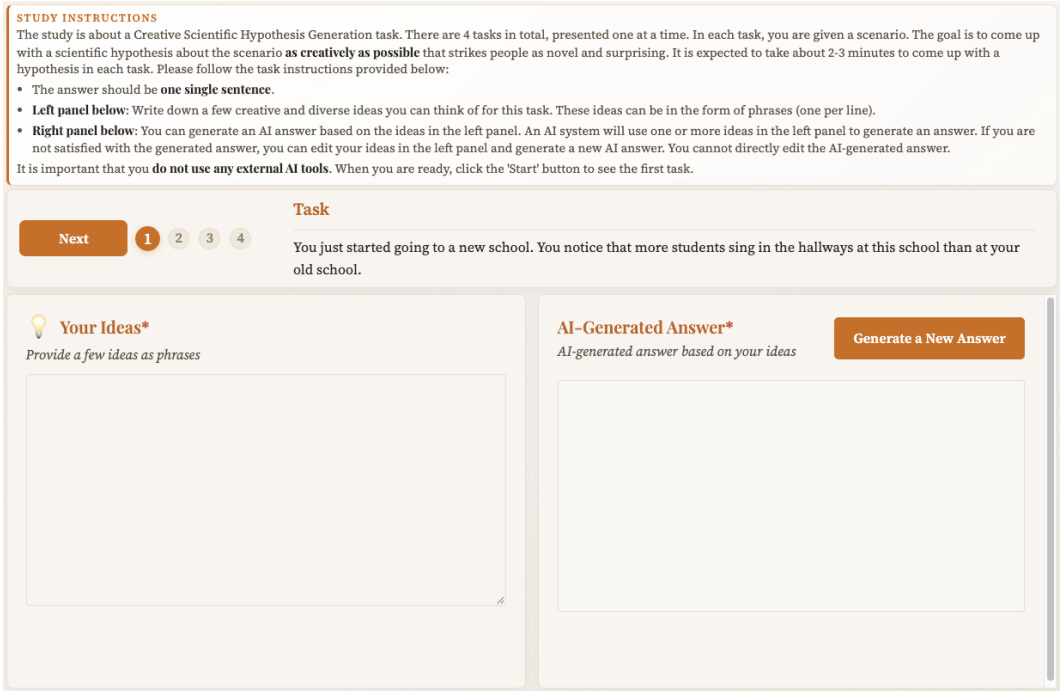


Figure 9: Snapshot of the interface for HAA:HIDEAS.

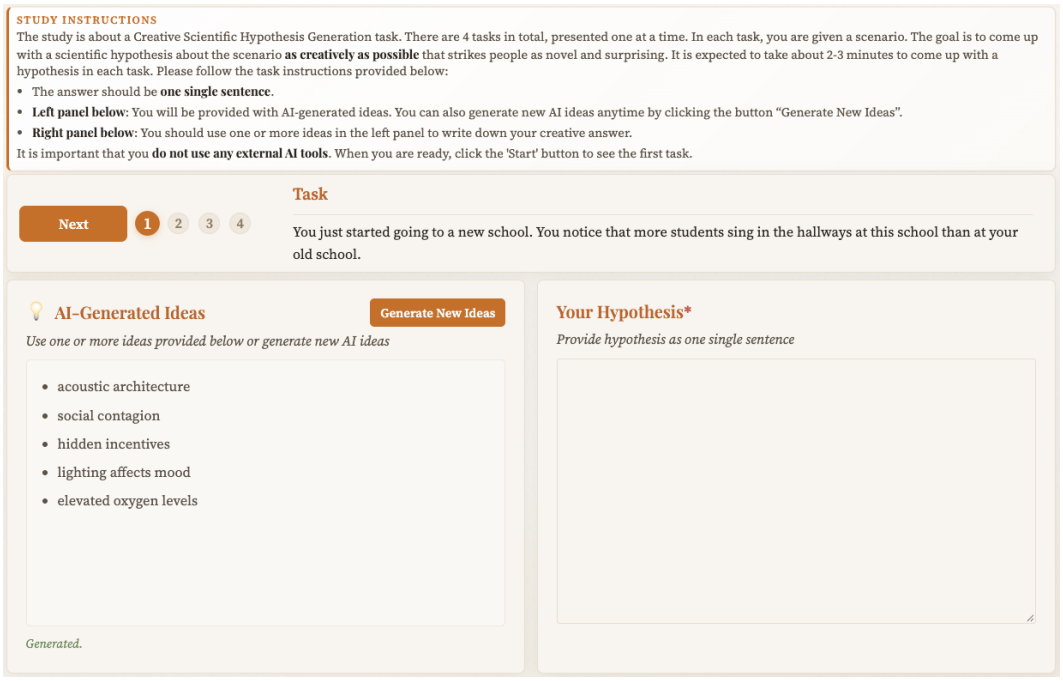


Figure 10: Snapshot of the interface for AHH:AIDEAS.

E.2 Annotation Interface

Figure 11 shows the annotation interface used for collecting human quality ratings in *Creative Hypothesis Generation*.

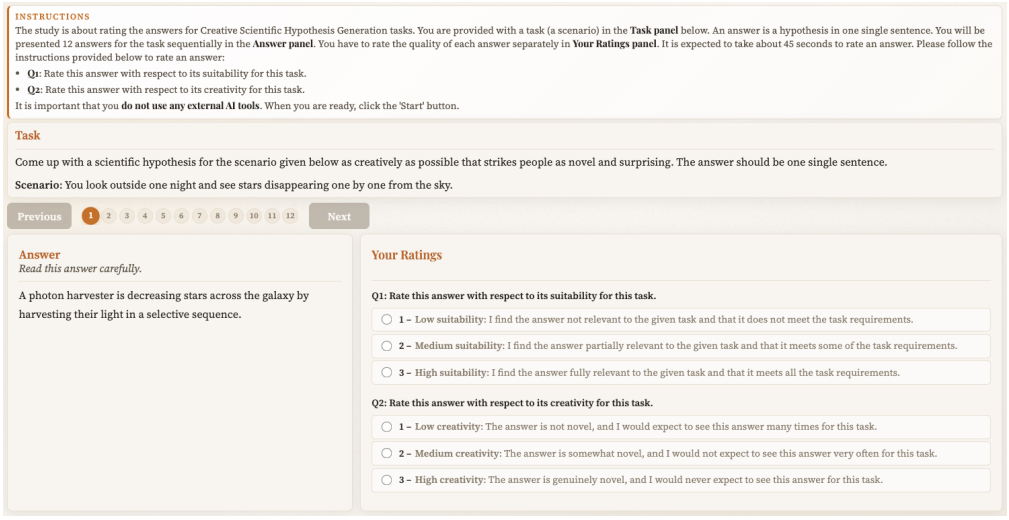


Figure 11: Snapshot of the annotation interface used for human evaluation. Annotators rate the usability (suitability) and creativity of each answer.

F Prompts for Agent Responses

Figures 12 and 13 show the prompts used to generate agent responses for *Creative Hypothesis Generation* and *Creative Story Writing*, respectively. For the mixture-based policy, the persona placeholder is filled with a persona from PersonaHub [47], such as "A crime novels enthusiast who lives in Washington D.C."

Approach the following task through the perspective of `{persona}`. You will be given a scenario. The goal is to come up with a scientific hypothesis about the scenario as creatively as possible that strikes people as novel and surprising. First, think about a few creative and diverse ideas for this task. Each idea is a short phrase of 3–5 words. Then, use one or more ideas to come up with a creative hypothesis. Your answer should be one sentence about your hypothesis. You should avoid using overly complex sentences in your answer. Respond with JSON only in this exact format: `{"ideas": ["idea1", "idea2", ...], "answer": "..."}`.

In this task, you are given the following scenario: "`{task}`".

(a) Prompt for AHH:AIDEAS, AAH:FINALIZE, and AAA:COPY

Approach the following task through the perspective of `{persona}`. You will be given a scenario. The goal is to come up with a scientific hypothesis about the scenario as creatively as possible that strikes people as novel and surprising. You will be given a few creative and diverse ideas for this task. Use one or more of the given ideas to come up with a creative hypothesis. Your answer should be one sentence about your hypothesis. You should avoid using overly complex sentences in your answer. Respond with JSON only in this exact format: `{"answer": "..."}`.

In this task, you are given the following scenario and ideas.
Scenario: "`{task}`"
Ideas: "`{human_ideas}`"

(b) Prompt for HAA:HIDEAS

Approach the following task through the perspective of `{persona}`. You will be given a scenario. The goal is to come up with a scientific hypothesis about the scenario as creatively as possible that strikes people as novel and surprising. You will be given a few creative and diverse ideas together with a hypothesis. Edit or adapt the given hypothesis as you want and use it as your answer. Your answer should be one sentence about your hypothesis. You should avoid using overly complex sentences in your answer. Respond with JSON only in this exact format: `{"answer": "..."}`.

In this task, you are given the following scenario, ideas, and hypothesis.
Scenario: "`{task}`"
Ideas: "`{human_ideas}`"
Hypothesis: "`{human_answer}`"

(c) Prompt for HHA:DRAFT

Figure 12: Prompts used to generate AI responses for *Creative Hypothesis Generation* in different settings. Placeholders are shown in blue. For each prompt, the first sentence describing the persona is omitted if no persona is specified in the corresponding setting.

Approach the following task through the perspective of {persona}. You will be given three words. The goal is to write a story using all three words as creatively as possible that strikes people as novel and surprising.
First, think about a few creative and diverse ideas for this task. Each idea is a short phrase of 3–5 words. Then, use one or more ideas to come up with a creative story. Your answer should be a story of 3–5 sentences. You should avoid using overly complex sentences in your answer. Respond with JSON only in this exact format: `{"ideas": ["idea1", "idea2", ...], "answer": "..."}`.

In this task, you are given the following three words: "{task}".

(a) Prompt for AHH:AIDEAS, AAH:FINALIZE, and AAA:COPY

Approach the following task through the perspective of {persona}. You will be given three words. The goal is to write a story using all three words as creatively as possible that strikes people as novel and surprising. You will be given a few creative and diverse ideas for this task. Use one or more of the given ideas to write your story. Your answer should be a story of 3–5 sentences. You should avoid using overly complex sentences in your answer. Respond with JSON only in this exact format: `{"answer": "..."}`.

In this task, you are given the following three words and ideas.
Three words: "{task}"
Ideas: "{human_ideas}"

(b) Prompt for HAA:HIDEAS

Approach the following task through the perspective of {persona}. You will be given three words. The goal is to write a story using all three words as creatively as possible that strikes people as novel and surprising. You will be given a few creative and diverse ideas together with a story written based on the ideas. Edit or adapt the given story as you want and use it as your answer. Your answer should be a story of 3–5 sentences. You should avoid using overly complex sentences in your answer. Respond with JSON only in this exact format: `{"answer": "..."}`.

In this task, you are given the following words, ideas, and story.
Words: "{task}"
Ideas: "{human_ideas}"
Story: "{human_answer}"

(c) Prompt for HHA:DRAFT

Figure 13: Prompts used to generate AI responses for *Creative Story Writing* in different settings. Placeholders are shown in blue. For each prompt, the first sentence describing the persona is omitted if no persona is specified in the corresponding setting.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions and scope, and these claims are supported by Sections 4–6.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 7 includes the discussion of limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: We do not include theoretical results in this paper.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Sections 4–6 describe the information for the main experimental results. Additional details are provided in the Appendix.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We release our implementation and resources in a Github repository, with a link provided in Section 1.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify these experimental details in Sections 4–6 and additional details in the Appendix.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report 95% confidence intervals for the main experimental results in Sections 4–5.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: Our experiments primarily involve crowdsourcing studies for which we provide details in Sections 4–5 and additional details in the Appendix. AI content is generated through hosted model APIs (Section 5).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our studies use consent-based crowdsourcing procedures for data collection and provide appropriate compensation. Our human-subject experiments are designed to conform to the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Section 7.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not release a foundation model or sensitive dataset.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The datasets used for our studies are properly cited in Section 4.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide full documentation with the released assets in a Github repository.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We provide details about participant recruitment and compensation in Section 4, and detailed task instructions in Appendix E.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The study followed a standard protocol for participants recruited from crowdsourcing platforms. The study was approved by a local institutional review board, and all participants provided informed consent (Section 4.2). There were no specific risks involved in this study, and the participants were fully informed about the study details.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs serve as AI agents in our co-creation studies. We describe the models, agent policies, and prompts in Sections 4–5 and Appendix F.