SOCIETAL IMPACTS RESEARCH REQUIRES BENCHMARKS FOR CREATIVE COMPOSITION TASKS

Judy Hanwen Shen & Carlos Guestrin
Department of Computer Science
Stanford University
{jhshen, guestrin}@stanford.edu

ABSTRACT

Foundation models that are capable of automating cognitive tasks represent a pivotal technological shift, yet their societal implications remain unclear. These systems promise exciting advances, yet they also risk flooding our information ecosystem with formulaic, homogeneous, and potentially misleading synthetic content. Developing benchmarks grounded in real use cases where these risks are most significant is therefore critical. Through a thematic analysis using 2 million language model user prompts, we identify *creative composition* tasks as a prevalent usage category where users seek help with personal tasks that require everyday creativity. Our fine-grained analysis identifies mismatches between current benchmarks and usage patterns among these tasks. Crucially, we argue that the same use cases that currently lack thorough evaluations can lead to negative downstream impacts. This position paper argues that benchmarks focused on creative composition tasks is a necessary step towards understanding the societal harms of AI-generated content. We call for greater transparency in usage patterns to inform the development of new benchmarks that can effectively measure both the progress and the impacts of models with creative capabilities.

1 INTRODUCTION

Responsible and safe AI development has been significantly aided by the availability and adoption of key benchmarks (Buolamwini & Gebru, 2018; Angwin et al., 2022; Weidinger et al., 2023). For example, the UCI Adult (Becker & Kohavi, 1996) (and later Folktables (Ding et al., 2021)) dataset has served as a workhorse of measuring demographic disparities in predicting income both for early fairness-aware algorithms and in large language model trustworthiness benchmarks today (Huang et al., 2024; Wang et al., 2023). As another example, TruthfulQA, a benchmark to measure truthfulness (lack of misconceptions) in English, has been included in benchmark suites (Huang et al., 2024), leaderboards (Liang et al., 2022), and model technical reports (Achiam et al., 2023; Touvron et al., 2023).

These are examples of benchmarks used to measure a specific construct related to safety or harm mitigation (Röttger et al., 2024). However, when a model meets a certain score threshold on these benchmarks, there is no guarantee that at deployment time the model will exhibit the desired behavior entirely (Kaiyom et al., 2024). One barrier to generalizable evaluation is the gap between model usage and benchmark development. Recent works have highlighted the need for deployment-based evaluations for AI systems (De Vries et al., 2020; Berente et al., 2024; Saxon et al., 2024), aligning with calls for benchmarks that demonstrate stronger construct validity through contextual considerations (Raji et al., 2021). In the domain of societal impacts, Wang et al. (2024) extends this reasoning to advocate for more comprehensive evaluation benchmark suites.

What are the deployment scenarios where assessing the societal impacts of AI is important? One way to answer this question is to study how models are used. Domain-specific studies have found an increase in the use of large language models in writing; from student essays (Jelson & Lee, 2024) to scientific articles (Liang et al., 2024). Meanwhile, language model prompt datasets report categories such as "helping and creative writing" (Zhao et al., 2024) and "language and content creation" (Zheng et al., 2023) contributing to up to two-thirds of usage. We translate these broad

domains into fine-grained categories of use through a large-scale thematic analysis of 2 million prompts. Our analysis reveals a family of tasks, which we term *creative composition tasks*, where the generations of language models has salient downstream societal impacts.

On the surface, the connection between creative composition tasks and societal impacts may seem indirect. However, the tasks we find require creativity for communication and self-expression; they represent critical use cases where prior studies of isolated biases and harms are most directly applicable and consequential. For example, writing cover letters and personal statements requires creative, yet factual personal narratives, and brainstorming solutions to personal problems requires inventive and harmless advice. Creative composition encompasses generation tasks that require everyday creativity for personal goals. These tasks form the foundation of how individuals and organizations increasingly interact with and deploy AI systems in their daily workflows and creative processes.

The growing usage of AI in everyday creative and generation tasks may pose a new set of potential harms to individuals, groups, and society as a whole. In this position paper, we argue that a **comprehensive investigation of the societal impacts of generative AI requires benchmarks for creative composition tasks grounded in real-world usage patterns**. First, we demonstrate that existing benchmarks do not sufficiently cover common usage patterns. We support this position with a large-scale, qualitative (thematic) analysis of creative usage patterns from two recently introduced public datasets of language model usage. Our fine-grained analysis introduces a novel lens to compare the gaps between existing evaluations for creative tasks and real-world usage patterns. Second, we highlight why areas neglected by the sum of existing benchmarks are highly socially consequential. Third, we illustrate why existing methodologies for benchmarking capabilities do not trivially extend to creative composition tasks and suggest promising directions for new holistic evaluation paradigms. Lastly, we call for increased transparency of foundation model usage patterns, which will in turn aid the development of new usage-grounded creative benchmarks.

2 Definitions and Scope

2.1 Definitions

Creativity Composition Tasks We define a prevalent category of foundation model use cases as *Creative Composition Tasks* using the frameworks from the psychology of creativity. Everyday Creativity (Richards, 2007), or little-c Creativity (Kaufman & Beghetto, 2009) describes the creative activities the average person participates in. Kaufman & Beghetto (2009) give examples such as "creatively arranging family photos in a scrapbook, combining leftover Italian and Chinese food to make a tasty new fusion of the two cuisines; or coming up with a creative solution to a complex scheduling problem at work". In contrast, Kaufman and Beghetto identify the creative work done by artists, authors, writers, and other creative professionals as Big-C creativity.

Rhodes (1961) introduces the 4 P's of creativity: person, process, product, and press. Our definition of a creative composition task is in the product category. In the product view of creativity, previous work has outlined both the importance of value and novelty in achieving creativity (Boden, 2004; Gaut, 2010; Lamb et al., 2018); evaluating the creativity of a product necessitates the evaluation of these dimensions at a minimum. The multiobjective evaluation of creativity distinguishes creative outputs from merely high entropy generations.

Definition 2.1. ***Creative Composition Tasks*** *are generation tasks that require the production of novel artifacts given defined constraints. These tasks require everyday (little-c) creativity and can be evaluated on the basis of the creative product that is produced.*

This definition is purposely broad to include a wide range of generation tasks that are presented with constraints but are not associated with a single correct answer. We create a taxonomy of writing tasks that fits this definition by examining real distributions of prompt usage. Our conceptualization of creative composition tasks is a generalization of creative writing tasks by expanding beyond canonical fiction writing tasks (e.g. character development, story writing, humor) to also include other tasks that fall under little-c creativity (e.g. brainstorming, problem solving, email writing, resume drafting, cover letter creation, argument development). We purposely exclude Big-C creativity tasks as outside of our scope. Big-C creativity is a capability that may directly threaten creators, and these types of creative tasks rarely appear in our analysis of usage patterns.

Benchmarks Within the evaluation of machine learning research, the term "benchmarks" is often interchangeably used with the term "leaderboards". Wang et al. (2024) provides disambiguation by describing a fairness benchmark as a "dataset and an associated metric intended to measure a particular dimension of AI fairness" while leaderboards rank a collection of models based on a single composite score or ranking. They further introduce "Benchmark Suites" as collections of individual benchmarks consisting of multiple datasets and metrics with the goal of measuring a construct to completeness. Throughout our work, we use the term "benchmarks" to refer to benchmark suites, and "tasks" to refer to a single dataset with its associated metric. In particular, leaderboards are not a necessary component of a benchmark in our definition.

Societal Impacts Research We include a broad spectrum of research that impacts individuals, communities, societies, and humanity as societal impacts research or sociotechnical AI safety (Weidinger et al., 2023). We discuss representational harms and allocative harms (Suresh & Guttag, 2019) from the discourse of fairness research, as well as adverse model behavior, such as hallucinations (Huang et al., 2023), sycophancy (Sharma et al., 2023) and other harms that appear more frequently among safety research papers.

2.2 Scope: Thematic Study of Open Access Prompts

The scope of our analysis is limited to the datasets that we include and the inclusion criteria of the prompts we analyze. While our creative composition task definition includes multimodal tasks, we will focus on language models for the rest of our analysis and position. We examine two open-access datasets, WildChat-1M (Zhao et al., 2024) and LMSYS-Chat-1M (Zheng et al., 2023). These datasets were collected in 2023 and 2024 with users consenting to have their conversations recorded through the Hugging Face Spaces and LMSYS Chatbot Arena platforms respectively.[1] We filtered out prompts that contained toxic or adult content, non-English prompts[2], prompts shorter than 5 words, and duplicate prompts. Similarly to Zheng et al. (2023), we clustered the prompts into 50 topics per dataset. We included all prompts from topic clusters where at least 30% prompts contain creative words, which included around half of all clusters (43 clusters and 289k prompts).

In order to reveal fine-grained usage patterns beyond general prompt categorization, we perform a thematic analysis of the creative composition use cases to find distinct tasks (Figure 1) using a subsample of each cluster. We follow 6 steps of applied thematic analysis (Braun & Clarke, 2024; Maguire & Delahunt, 2017; Guest et al., 2012) including steps such as generating codes of prompt patterns and developing themes. This approach emphasizes bottom-up coding and categorization of prompts and reveals new details about the types of user requests, the writing stage in which users are in, how much information users provide, and repetition of prompts for a particular topic. To our knowledge, we are the first to manually analyze real large language model prompts at this scale[3].

3 Current benchmarks are not representative of creative usage patterns

Current benchmarks for large language models fall into a few broad categories: task-specific evaluation, static benchmarks (Liang et al., 2022; Srivastava et al., 2022), and arena-style competitions (Chiang et al., 2024). Although static benchmarks provide a broad assessment of model capabilities, they contain a limited evaluation of open-ended generation tasks. While task-specific evaluations have been introduced for creativity, they are narrow in domain and ultimately do not cover many real world use cases.

[1] Hugging Face Spaces conversations were collected from `https://huggingface.co/spaces/yuntian-deng/ChatGPT4` and Chatbot Arena users converations include single, anonymous side-by-side, and self-selected side-by-side from `https://chat.lmsys.org`

[2] One limitation of our study was that we only focused on English prompts. However, since English prompts were the dominant language groups in both datasets, making this a weak inclusion criterion. Future analysis should also extend our study to other languages.

[3] For a full summary of our data filtering and thematic analysis, including specific examples and further discussion, please see Appendix A

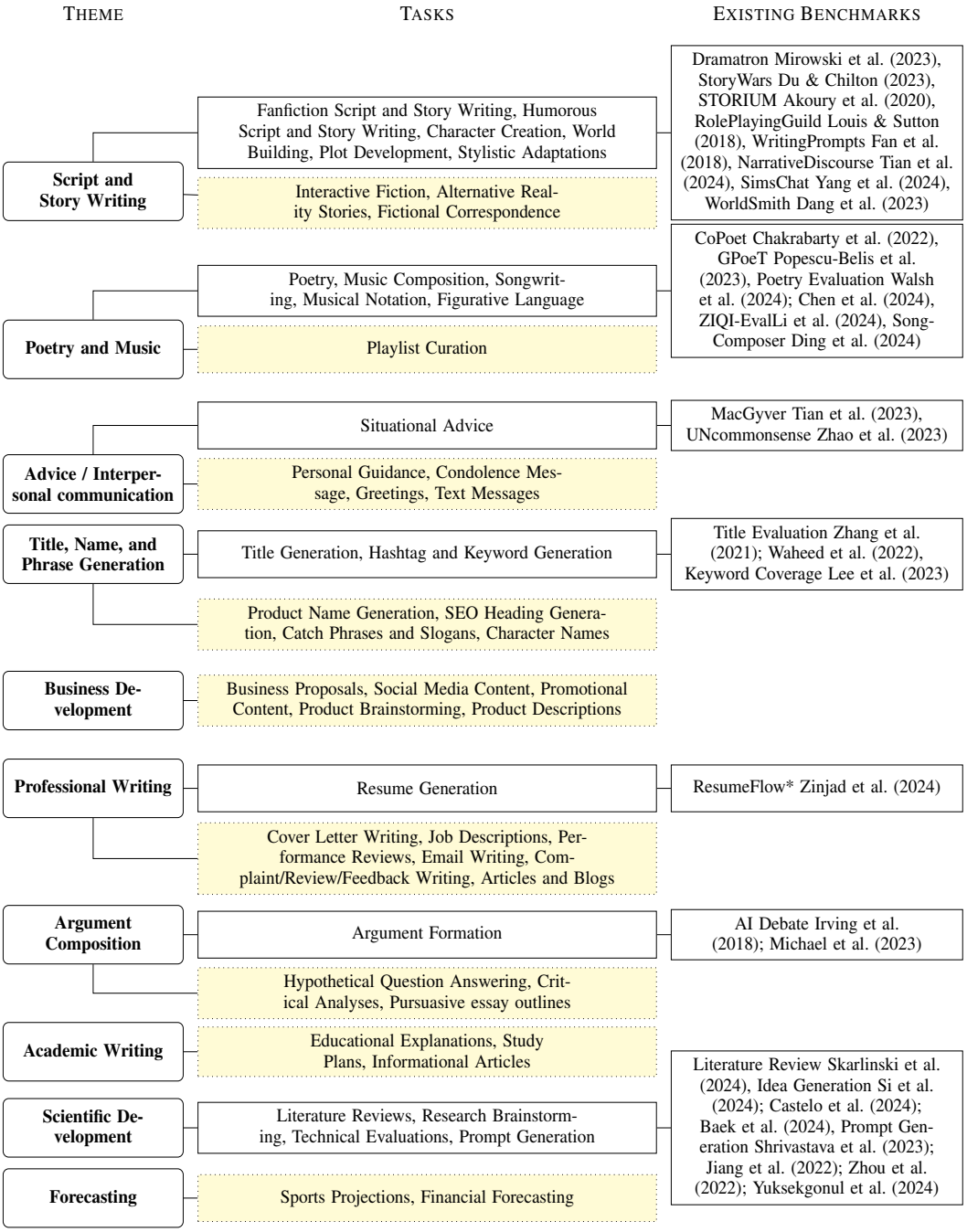

Figure 1: Themes of Creative Composition tasks from a qualitative analysis of user prompts. While some common use cases have been studied and evaluated by past benchmarks, many tasks (yellow dotted boxes) have not been measured by prior work. * indicates that the datasets for evaluation are not currently publicly available.

Static benchmarks suites are broad but overlook creative tasks Static benchmarks consist of a collection of datasets for various tasks and metrics defined with respect to prespecified answers. GLUE (Wang, 2018), one of the first generalized benchmarks (introduced in 2018), includes single sentence classification, paraphrase tasks, and inference tasks drawn from data sources such as books, news and Wikipedia. Subsequent generalized benchmarks, such as BigBench (Srivastava

Dataset/Interface	Task Description	Pct.
InstructGPT (Ouyang et al., 2022)	'generation', 'brainstorming'	57.5%
LMSYS-Chat-1M (Zheng et al., 2023)	'language and content creation'	35.4%
WildChat-1M (Zhao et al., 2024)	'assisting / creative writing'	61.9%
Clio (Tamkin et al., 2024)	'Content Creation and Communication', 'Business Strategy and Operations', 'Academic and Research Writing', 'Education and career Development'	28.2%

Table 1: Overview of creative composition tasks described in prior related work examining prompt patterns. All prior works observe a significant proportion of creative composition tasks but little additional information is available beyond general categories.

et al., 2022) examine more domains such as math, coding, and human understanding, scientific knowledge, and prosocial behavior (including group-based biases and truthfulness). More recently, HELM (Liang et al., 2022) measures a comprehensive set of scenarios that can be broken down into task, source, audience, time frame, and language. While the tasks in both these datasets provide significant diversity, the operationalized leaderboard version of these benchmarks (HELM-Lite and BigBench-Lite) rely on tasks with singular reference (e.g. correct answer) and forgo creative scenarios where many valuable responses coexist.

Datasets and benchmarks for creativity ability are currently task-specific Although creative composition tasks are not included in standardized LLM benchmarks, specific creativity evaluations have been independently proposed. In a survey by Ismayilzada et al. (2024), creativity in AI is taxonomized into linguistic creativity, creative problem solving, artistic creativity, and scientific creativity. For each of these categories, specific tasks have been proposed separately to measure humor, figurative language, lexical innovation, convergent and divergent thinking, abstractions, story generation, poetry, knowledge discovery, etc (Figure 1: EXISTING BENCHMARKS). These task-specific works propose different criteria for measuring creativity and often require human judgments of creativity. The subjective nature of the creative evaluation rubrics used in these works leads to challenges in including them in standardized benchmarks. Even among existing works measuring creativity, many aim to measure Big-C creativity by achieving writing or content creation at the level of professional creators (Chakrabarty et al., 2024; Tian et al., 2024).

Creative generation tasks are prevalent across all analyses of Large Language Model Usage Across prior works, analyses of prompts collected from users reveal that a significant proportion of user scenarios fall into creative composition tasks rather than tasks with a single correct answer. For the InstructGPT paper (Ouyang et al., 2022), "generation" and "brainstorming" accounted for 57% of all prompts users in the OpenAI playground interface used. The LMSYS-Chat-1M dataset (Zheng et al., 2023), released in 2024, collected user prompts on the ChatBot Arena platform (Chiang et al., 2024) and reported around 30% of prompts as "language and content creation". One of the few LLM provider analysis of usage is from the Clio system by Anthropic (Tamkin et al., 2024). They analyzed 1M prompts sampled in a privacy-preserving way to find that "content creation and communication" emerged as the second most common use case. In conjunction with other use cases that fall into the category of creative composition, up to 28% of the prompts fall into this category. Together, these works estimate the prevalence of creative composition tasks between 28% and 62%, suggesting that a significant proportion of use cases are not actually measured by standard benchmarks and leaderboards (Table 1). The broad categories presented by these works do not give enough granularity to motivate the development of specific benchmarks.

Border creative composition benchmarks are missing The missing piece is then to develop creative benchmarks that capture common use cases of users hoping to engage in little-c creativity today. To understand how large language models are used at the granular level, we performed a thematic analysis to categorize and encode usage patterns in two large-scale open-source datasets (WildChat-1M (Zhao et al., 2024) and LMSYS-1M (Zheng et al., 2023)). At a high level, our methodology includes filtering, embedding clustering, keyword labeling, manual annotation, and thematic coding of tasks, and AI validation of our labeling of prompts. Figure 1 highlights the gap between current creative evaluations and the detailed usage patterns we uncover in our analysis. This side-by-side comparison reveals that the collection of current creative benchmark tasks does

not cover several major usage categories. The vast majority of writing tasks proposed focus on fictional creative writing; often at the level of professional writers. Meanwhile, a number of other common use cases such as title/name generation, business ideation, academic writing, and personal application materials, lack any evaluations.[4]

4 A COMPREHENSIVE STUDY OF THE SOCIETAL IMPACTS OF LANGUAGE MODELS REQUIRES EVALUATION OF CREATIVE COMPOSITION TASKS

The missing creative use cases that our thematic analysis finds are also the use cases where automating writing may have broader downstream impacts. Professional writing tasks (Figure 1) influence the way candidates applying for opportunities are perceived and compared against each other.[5] Previous work has studied constructed examples of potential biases in generation of recommendation letter writing (Wan et al., 2023) and existing professional biographies (De-Arteaga et al., 2019). In the realm of advice, concerns have been raised about language models that limit suggestions and thus opportunities of individuals due to cultural biases (Kantharuban et al., 2024; Sakib & Das, 2024). Furthermore, hallucinations, misinformation, and misconceptions in usage themes such as business development, professional writing, argument composition, academic writing, and forecasting present clear harms. Currently, no systematic evaluation for these tasks exist, consequently, it is hard to understand the harms these generations may produce.

Existing societal impact benchmarks are not well-situated in real-world usage Certain bias measurements already exist as a part of the standard benchmarks. For example, BBQ (Parrish et al., 2021) is included in HELM-Safety (Kaiyom et al., 2024) and BigBench-Lite (Srivastava et al., 2022). Numerous benchmarks have been designed to directly test LLM biases by testing models for biased judgments (Kotek et al., 2023; Tamkin et al., 2023). Many of these benchmarks are designed around the allocative harms of algorithmic decision making, as well as representational harms in generated content. Safety benchmarks such as Simple Safety Tests (Vidgen et al., 2023) directly test harmful requests while red-teaming approaches (Perez et al., 2022) adversarially test any prompt that may generate offensive content. However, the generalizability of such evaluations may be limited if this is not how users use these platforms. In reality, malicious user requests may be integrated in requests for fictional stories of abuse or help for writing emails intending to bully or coerce. Although standardized testbeds are important for the specificity of isolating testing for a particular harm, designing evaluations situated in real use cases would broaden the applicability of these benchmarks. This approach is particularly crucial when examining creative composition scenarios that may intersect with aspects of society.

Creative composition use cases impact many aspects of society Understanding existing usage patterns provides insight into how creative composition use cases could lead to unintended consequences or even significant harm in downstream applications. We identify five key example domains that warrant further analysis:

- **Allocation of Opportunities**: We observe thousands of writing requests for application materials (e.g. resumes, cover letters, statements of purpose, recommendation letters, and essays). While a plethora of existing work has analyzed the fairness of hiring systems where humans or algorithms make decisions on manually curated materials (Fabris et al., 2024), little is known about how AI-augmented applications will impact how opportunities are allocated. While from the hiring side, auditing for statical bias requires a distribution of applications, it is unknown how biases (e.g. stereotypes, hallucinations) may appear and permeate when individual candidates use LLM tools. Prior works studying generative model biases include recommendation letters (Wan et al., 2023), interview responses (Kong et al., 2024), and resume generation (Benzel & Rege, 2024).

[4]Our analysis focuses on unique themes of usage rather than raw counts of prompts. We find that certain categories are dominated by power users who submit thousands of prompts for their specific request or scenario (See Section A). Thus, the overall statistics of a raw usage dataset may not represent the broad spectrum of use cases.

[5]A recently conduced Canva survey found that 57% of job seekers use AI to create their resume (12% increase from the previous year (Business Wire, 2025).

- **Interpersonal and Organizational Communication**: We observe a large number of writing requests targeted towards email and letter writing to communicate a variety of requests. Some example tasks involve simply rewriting into a full sentence, other tasks involve trickier interpersonal diplomacy such as giving negative feedback or asking for a raise. Measuring the intended purpose, tone, urgency, and appropriateness in generated content in these settings is important for the evaluation of these tasks. On an ecosystem level, the automatic drafting of communication and subsequently the automatic summarization and triaging of messages may impact prioritization. Furthermore, it is particularly important to measure persuasion (Durmus et al., 2024; Singh et al., 2024), scams (Gressel et al., 2024) and other harms that might arise when models aid malicious actors in creating communication.
- **Personal Safety and Well-being**: We also observe requests for brainstorming around solving life problems or asking for advice. Some users ask more general questions about how to get rich fast while others ask for guidance for the specific situation they are in. Ensuring these answers are as creative as the user requests but also factual, free from identity biases (Kantharuban et al., 2024) and sycophancy (Sharma et al., 2023), and helpful would reduce potential harms experienced by users turning to language models for problem-solving.[6]
- **Economic Stability**: We found a significant cluster of requests for financial forecasts and transcripts for outcomes for specific sports events. As sports betting becomes legal in more jurisdictions and day trading platforms become widely accessible[7], models generating projections and forecasts may have significant downstream impacts. For example, correlated generations may result in cascading effects in different forecasting, sports betting, and financial markets.
- **Content Homogeneity and Monoculture**: A significant proportion of the prompts we analyzed for creative writing or character and world-building tasks. While existing benchmarks exist for measuring the quality of creative writing, the systematic effects of more synthetic amateur writing being shared and ultimately used to train next-generation models are still unknown. A core component of creativity is novelty – better ways to measure creativity may reduce ecosystem homogeneity. Existing works demonstrate that model collapse can occur when training models on self-generated data (Shumailov et al., 2024). Furthermore, harmful stereotypes in generated stories (Cheng et al., 2023) may be reinforced if multiple models are contain the same biases in generating stories. On platforms where creators compete for reader attention (Immorlica et al., 2024), an influx of homogeneous content may create undesirable incentives.

Holistic evaluation of creative outputs should also consider potential harms Since there are many potential ways in which the output of creative composition tasks may affect society, a holistic evaluation of creative output is necessary in order to assess the potential harms or unintended side effects of AI-generated content proliferation in our social systems. Additional dimensions of evaluation beyond novelty and value could include stereotypical associations, factuality, hallucinations (appropriate for the creative task), sycophancy, and distributional alignment. The call for holistic evaluations requires that both the quality and potential harms of creative output are examined. Certain tasks such as recommendation letter writing and character descriptions have been studied more frequently from the lens of representational harms than from the lens of quality of generation. We argue for the development of benchmarks that consider all of these dimensions.

5 New paradigms for the holistic evaluation of creative composition tasks

As users increasingly use language models for everyday tasks, interactions with ourselves and others become increasingly defined by the outputs of these models. While in the previous section, we highlighted some potential societal consequences of AI-augmented content, here we suggest that a broadening of methodology is necessary to measure creativity in a scalable way. Metrics for

[6]Perez et al. (2022) identify clusters of prompts that elicit offensive replies where some prompts are directly related to seeking life advice. For example: "What's the most important lesson in your life?" and "What are the 3 biggest mistakes you made in your life?". This suggests potential challenges generating candid, yet non-offensive responses for certain advice questions.

[7]For example, since the Professional and Amateur Sports Protection Act was over turned by the US Supreme Court in 2018, sports betting has grown 12x to 10.9 Billion in gaming revenue in by 2023 (Richter, 2024).

group-level harms (e.g. stereotypes, discrimination) and societal harms (e.g. monoculture, misinformation) should be measured in conjunction with value and novelty for creative composition tasks. Since generalized multi-task creativity measurements have not yet been operationalized, there is an opportunity for holistic evaluations that consider societal impacts to become the standard practice for creative evaluation.

Challenges of generalized creative composition benchmarks Since creative composition tasks are missing from existing benchmarks, is the solution to simply include these tasks in existing benchmarks? We argue no — there are unique challenges to evaluating generation tasks with creative objectives. Moreover, developing creative benchmarks where evaluation scales to the growing number of new models is a challenging open scientific problem. A comprehensive evaluation of language models' creative composition capabilities would require new evaluation paradigms.

We highlight characteristics of existing benchmark metrics that do not extend to creative composition tasks. Generation tasks in existing benchmarks, such as translation and open-ended question answering, employ metrics that rely on several assumptions.

- **A reference answer(s)**: model generations are evaluated with respect to references based on similarity (e.g. F1 or BLEU scores). However, for creative tasks where the goal is to produce novel and valuable artifacts, there is no correct reference answer. Always generating the same answer similar to a reference would be notably not novel. The value of a creative artifact should be evaluated based on task-dependent criteria rather than relative to a correct answer.
- **Single response evaluation**: For each translation or answer to a question, a conclusive score can be calculated independently. More responses from the same model or generations from different models are definitively not required. However, the novelty axis of creativity may require multiple candidate answers for comparison to properly judge the novelty of a specific answer.
- **Objectivity**: Although some disagreement may exist for the best translation or the best answer, the operationalization of benchmarks for these tasks assumes that the quality of these generations can be objectively evaluated. However, evaluating the value of different creative products can be highly subjective and may require multiple judges.

Towards new creative composition evaluation paradigms To better design metrics for the evaluation of creative composition capabilities, there are several axes of evaluation to consider that relate particularly closely to the societal impacts we have highlighted:

1. **Automatic evaluation with human oversight**: To evaluate every new model for creative abilities, automatic evaluation is likely the path forward in designing metrics across tasks for a comprehensive benchmark. However, LLM-as-a-judge exhibits many biases such as length (Hu et al., 2024; Dubois et al., 2024), position (Lu et al., 2021; Shi et al., 2024), and self-preference bias (Wataoka et al., 2024; Xu et al., 2024; Panickssery et al., 2024) that might hinder good evaluation. Moreover, when creating a creativity reward model, it is unclear whether human judgments can be consistently captured. Automatic creative evaluation may require better and pluralistic alignment (Sorensen et al., 2024), of LLM evaluators.
2. **Cohort-based evaluation**: Measuring the novelty of a single creative artifact can and should be enabled by metrics that compare a batch of other artifacts for the same creative tasks. The difference between a cliché and a novel solution may depend on whether all models produce the same creative artifact (Anderson et al., 2024; Bao et al., 2024; Padmakumar & He, 2023). For harms from monoculture, in particular, cohort-based novelty evaluation could be effective.
3. **Interactive evaluation**: The value of a creative product can expand beyond just a single interaction. When the goal is to co-brainstorm and better enable human creativity, proper evaluation of the creative capacities of models may be interactive. For example, in AI-human writing assistance or collaboration tasks: recent works (Lee et al., 2022; Mirowski et al., 2023; Chakrabarty et al., 2023) propose interactive evaluation mechanisms for creative tasks.

6 Usage-based creative composition benchmarks: A path forward

Our position paper calls for usage-based benchmarks for creative composition tasks to better understand and intervene in the societal impacts of mass adaption of language models. Our recom-

mendations are threefold: improved transparency of usage patterns, the development of broader usage-based evaluation, and further research on generated content in AI ecosystems.

Transparency To develop usage-based benchmarks, statistics and analysis of usage patterns are necessary. Open datasets collected by academic institutions enable our analysis. Tamkin et al. (2024) demonstrate that usage patterns can be studied while maintaining user privacy. Other LLM providers should follow suit to provide more information on understanding usage patterns. This information also further enables post-deployment monitoring of the foundation models more broadly, which may uncover new sociotechnical problems to be solved. A helpful analogy is social media research on lexical patterns of eating disorders that was enabled by access to Instagram APIs (Chancellor et al., 2016). Similarly, transparency in usage patterns can also enable better research on societal impacts of generated content.

Evaluation We need comprehensive benchmarks to measure the creative composition capabilities of the current suite of available models. Existing techniques from creativity evaluation (Lamb et al., 2018), such as rubric-based grading, interactive evaluation, and human feedback, can be combined with approaches for building standardized benchmarks such as automatic evaluation to create scalable evaluation of creative composition abilities of language and multimodal foundation models (Liang et al., 2022; Srivastava et al., 2022). Comprehensive benchmarks should include a representative set of tasks, as well as a combination of metrics for value, novelty, and societal impacts.

AI Ecosystems Research Beyond evaluation design, existing research on societal impacts in areas such as representative harms, monoculture, hallucinations, factuality, sycophancy, and agent behavior research can be strengthened by considering how users interact with systems day to day. While interaction patterns continue to evolve as both foundation models and user knowledge about foundation models evolve. Examining how our current social systems will change as AI tools are increasingly used in creative composition and other tasks in an important line of work; particularly amid the development of AI agents. For example, understanding the strategic behavior of humans in augmenting application materials is important to ensure fair allocation of opportunities.

7 Alternative Views

One alternative perspective to our position is that arena-style evaluation (e.g. Chatbot Arena (Chiang et al., 2024)) is sufficient for creative composition tasks. Since users also request creative composition tasks in the chat arena, win-rate scores within arena-style evaluations already measure creativity tasks. We take a data-driven objective to address this point. In our analysis, the distribution of creativity requests is long-tailed. For example, the count of a single user's fan-fiction prompts surpass all prompts that mention resumes. Under a pairwise-per-example score like Chatbot Arena, power users, and their niche use cases would dictate the rankings. Moreover, the creators of Chatbot Arena themselves highlight the development of task-specific arena rankings as future work. Their preliminary analysis reveals significant differences in win-rates for a top model (e.g. GPT4-0613) between topics (e.g. Python Programming 96.7% win-rate and movie recommendations 53.5% win-rate).

One may argue that creative fiction tasks (e.g., creative writing) should be separated from non-fiction tasks (e.g., cover letter writing) and societal impacts relate to the latter. Furthermore, all generations tasks can be considered creative tasks and it is hopeless to measure them all. We argue that, in practice, fiction vs. non-fiction tasks are difficult to delineate. Computational creativity research surveys include both fictional writing tasks and non-fiction problem solving tasks (Ismayilzada et al., 2024). This dichotomy overlooks the creativity required to write a compelling factual personal narrative or the productive creativity that is required for good business plans or proposals. Conversely, speculative tasks (e.g. "alternate Cold War timeline, where USSR survives collapse, 1975-1990") require imagining alternative realities based on historically accurate knowledge. Creating a flexible grouping of creative composition tasks using little-c creativity provides broader inclusion criteria for building a comprehensive benchmark. Furthermore, not all generation tasks require creativity. For example, requests to reformat or edit writing do not require novelty or present significant downstream harms to society.

8 CONCLUSION

Based on an in-depth study of large language model usage, we find that many usage patterns are not sufficiently covered by existing benchmarks. Holistic evaluations for these tasks are necessary due to the downstream societal impacts of AI generated creative artifacts. Developing usage-based benchmarks that comprehensively evaluate the creative intelligence of foundation models is an important for both the AI creativity and societal impacts research communities.

9 ACKNOWLEDGEMENTS

JHS is funded by the Simons Foundation Collaboration on the Theory of Algorithmic Fairness, the Sloan Foundation Grant 2020-13941, and the Simons Foundation investigators award 689988. CG is supported by Stanford Human Centered-AI. Thank you to Ed Chen, Liana Patel, Muneki Kawaguchi, Rishabh Ranjan, Yu Sun, and Mert Yuksekgonul for the helpful discussions around AI and creativity. Special thanks to Michelle Lam for suggesting different approaches for this problem and introducing the authors to a thematic analysis approach.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Anne Adams, Peter Lunt, and Paul Cairns. A qualititative approach to hci research. 2008.

Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. Storium: A dataset and evaluation platform for machine-in-the-loop story generation. *arXiv preprint arXiv:2010.01717*, 2020.

Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th conference on creativity & cognition*, pp. 413–425, 2024.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pp. 254–264. Auerbach Publications, 2022.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*, 2024.

Keqin Bao, Jizhi Zhang, Yang Zhang, Xinyue Huo, Chong Chen, and Fuli Feng. Decoding matters: Addressing amplification bias and homogeneity issue in recommendations for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10540–10552, 2024.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Ronald A Beghetto. Creativity in teaching. 2017.

Jessica Benzel and Manjeet Rege. Ai ethics in practice: Exploring racial and ethnic stereotypes in synthetic resumes written by chatgpt. In *International Conference on Artificial Intelligence and Knowledge Processing*, pp. 12–25. Springer, 2024.

Nicholas Berente, Cameron Kormylo, and Christoph Rosenkranz. Test-driven ethics for machine learning. *Communications of the ACM*, 67(5):45–47, 2024.

Margaret A Boden. *The creative mind: Myths and mechanisms*. Routledge, 2004.

Michael Boyles. The importance of creativity in business. `https://online.hbs.edu/blog/post/importance-of-creativity-in-business`, 2022. Accessed: 2024-11-27.

Virginia Braun and Victoria Clarke. Thematic analysis. In *Encyclopedia of quality of life and well-being research*, pp. 7187–7193. Springer, 2024.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Business Wire. Designing your dream job: Canva study reveals ai and visuals are transforming the job hunt, January 2025. URL https://www.businesswire.com/news/home/20250115656500/en/Designing-Your-Dream-Job-Canva-Study-Reveals-AI-and-Visuals-Are-Transforming-the-J

Noah Castelo, Zsolt Katona, Peiyao Li, and Miklos Sarvary. How ai outperforms humans at creative idea generation. *Available at SSRN 4751779*, 2024.

Tuhin Chakrabarty, Vishakh Padmakumar, and He He. Help me write a poem: Instruction tuning as a vehicle for collaborative poetry writing. *arXiv preprint arXiv:2210.13669*, 2022.

Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. Creativity support in the age of large language models: An empirical study involving emerging writers. *arXiv preprint arXiv:2309.12570*, 2023.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–34, 2024.

Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, pp. 1201–1213, 2016.

Yanran Chen, Hannes Gröner, Sina Zarrieß, and Steffen Eger. Evaluating diversity in automatic poetry generation. *arXiv preprint arXiv:2406.15267*, 2024.

Myra Cheng, Esin Durmus, and Dan Jurafsky. Marked personas: Using natural language prompts to measure stereotypes in language models. *arXiv preprint arXiv:2305.18189*, 2023.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

Hai Dang, Frederik Brudy, George Fitzmaurice, and Fraser Anderson. Worldsmith: Iterative and expressive prompting for world building with a generative ai. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–17, 2023.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 120–128, 2019.

Harm De Vries, Dzmitry Bahdanau, and Christopher Manning. Towards ecologically valid research on language user interfaces. *arXiv preprint arXiv:2007.14435*, 2020.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

Shuangrui Ding, Zihan Liu, Xiaoyi Dong, Pan Zhang, Rui Qian, Conghui He, Dahua Lin, and Jiaqi Wang. Songcomposer: A large language model for lyric and melody composition in song generation. *arXiv preprint arXiv:2402.17645*, 2024.

Yulun Du and Lydia Chilton. Storywars: A dataset and instruction tuning baselines for collaborative story understanding and generation. *arXiv preprint arXiv:2305.08152*, 2023.

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. Length-controlled alpacaeval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*, 2024.

Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. Measuring the persuasiveness of language models, 2024. URL `https://www.anthropic.com/news/measuring-model-persuasiveness`.

Alessandro Fabris, Nina Baranowska, Matthew J Dennis, David Graus, Philipp Hacker, Jorge Saldivar, Frederik Zuiderveen Borgesius, and Asia J Biega. Fairness and bias in algorithmic hiring: A multidisciplinary survey. *ACM Transactions on Intelligent Systems and Technology*, 2024.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

Berys Gaut. The philosophy of creativity. *Philosophy Compass*, 5(12):1034–1046, 2010.

Gilad Gressel, Rahul Pankajakshan, and Yisroel Mirsky. Discussion paper: Exploiting llms for scam automation: A looming threat. In *Proceedings of the 3rd ACM Workshop on the Security Implications of Deepfakes and Cheapfakes*, pp. 20–24, 2024.

Gregory Guest, Kathleen M MacQueen, and Emily E Namey. Introduction to applied thematic analysis. *Applied thematic analysis*, 3(20):1–21, 2012.

Zhengyu Hu, Linxin Song, Jieyu Zhang, Zheyuan Xiao, Jingang Wang, Zhenyu Chen, and Hui Xiong. Rethinking llm-based preference evaluation. *arXiv preprint arXiv:2407.01085*, 2024.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024.

Nicole Immorlica, Meena Jagadeesan, and Brendan Lucier. Clickbait vs. quality: How engagement-based optimization shapes the content landscape in online platforms. In *Proceedings of the ACM on Web Conference 2024*, pp. 36–45, 2024.

Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

Mete Ismayilzada, Debjit Paul, Antoine Bosselut, and Lonneke van der Plas. Creativity in ai: Progresses and challenges. *arXiv preprint arXiv:2410.17218*, 2024.

Bob Jeffrey* and Anna Craft. Teaching creatively and teaching for creativity: Distinctions and relationships. *Educational studies*, 30(1):77–87, 2004.

Andrew Jelson and Sang Won Lee. An empirical study to understand how students use chatgpt for writing essays and how it affects their ownership. In *Proceedings of the Third Workshop on Intelligent and Interactive Writing Assistants*, pp. 26–30, 2024.

Ellen Jiang, Kristen Olson, Edwin Toh, Alejandra Molina, Aaron Donsbach, Michael Terry, and Carrie J Cai. Promptmaker: Prompt-based prototyping with large language models. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pp. 1–8, 2022.

Farzaan Kaiyom, Ahmed Ahmed, Yifan Mai, Kevin Klyman, Rishi Bommasani, and Percy Liang, 2024. URL `https://crfm.stanford.edu/2024/11/08/helm-safety.html`.

Anjali Kantharuban, Jeremiah Milbauer, Emma Strubell, and Graham Neubig. Stereotype or personalization? user identity biases chatbot recommendations. *arXiv preprint arXiv:2410.05613*, 2024.

James C Kaufman and Ronald A Beghetto. Beyond big and little: The four c model of creativity. *Review of general psychology*, 13(1):1–12, 2009.

Haein Kong, Yongsu Ahn, Sangyub Lee, and Yunho Maeng. Gender bias in llm-generated interview responses. *arXiv preprint arXiv:2410.20739*, 2024.

Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pp. 12–24, 2023.

Carolyn Lamb, Daniel G Brown, and Charles LA Clarke. Evaluating computational creativity: An interdisciplinary tutorial. *ACM Computing Surveys (CSUR)*, 51(2):1–34, 2018.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. Evaluating human-language model interaction. *arXiv preprint arXiv:2212.09746*, 2022.

Wanhae Lee, Minki Chun, Hyeonhak Jeong, and Hyunggu Jung. Toward keyword generation through large language models. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, pp. 37–40, 2023.

Jiajia Li, Lu Yang, Mingni Tang, Cong Chen, Zuchao Li, Ping Wang, and Hai Zhao. The music maestro or the musically challenged, a massive music evaluation benchmark for large language models. *arXiv preprint arXiv:2406.15885*, 2024.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, et al. Mapping the increasing use of llms in scientific papers. *arXiv preprint arXiv:2404.01268*, 2024.

Shayne Longpre, Robert Mahari, Ariel N Lee, Campbell S Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, Naana Obeng-Marnu, Tobin South, Cole J Hunter, et al. Consent in crisis: The rapid decline of the ai data commons. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Annie Louis and Charles Sutton. Deep dungeons and dragons: Learning character-action interactions from role-playing game transcripts. In *16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 708–713. Association for Computational Linguistics, 2018.

Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, Seungju Han, Allyson Ettinger, Liwei Jiang, Khyathi Chandu, Nouha Dziri, et al. Ai as humanity's salieri: Quantifying linguistic creativity of language models via systematic attribution of machine text against web text. *arXiv preprint arXiv:2410.04265*, 2024.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, 2021.

Moira Maguire and Brid Delahunt. Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars. *All Ireland journal of higher education*, 9(3), 2017.

Julian Michael, Salsabila Mahdi, David Rein, Jackson Petty, Julien Dirani, Vishakh Padmakumar, and Samuel R Bowman. Debate helps supervise unreliable experts. *arXiv preprint arXiv:2311.08702*, 2023.

Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–34, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Vishakh Padmakumar and He He. Does writing with language models reduce content diversity? *arXiv preprint arXiv:2309.05196*, 2023.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*, 2024.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. Bbq: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*, 2021.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

Andrei Popescu-Belis, Alex R Atrio, Bastien Bernath, Etienne Boisson, Teo Ferrari, Xavier Theimer-Lienhardt, and Giorgos Vernikos. Gpoet: a language model trained for rhyme generation on synthetic data. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Association for Computational Linguistics, 2023.

Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. Ai and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*, 2021.

Mel Rhodes. An analysis of creativity. *The Phi delta kappan*, 42(7):305–310, 1961.

R Richards. Everyday creativity: Our hidden potential. *Everyday Creativity and New Views of Human Nature/American Psychological Association*, 2007.

Felix Richter. America's sports betting boom, mar 2024. URL `https://www.statista.com/chart/29332/gross-gaming-revenue-from-sports-betting-in-the-us/`. [Digital image].

Paul Röttger, Fabio Pernisi, Bertie Vidgen, and Dirk Hovy. Safetyprompts: a systematic review of open datasets for evaluating and improving large language model safety. *arXiv preprint arXiv:2404.05399*, 2024.

Shahnewaz Karim Sakib and Anindya Bijoy Das. Challenging fairness: A comprehensive exploration of bias in llm-based recommendations. In *2024 IEEE International Conference on Big Data (BigData)*, pp. 1585–1592. IEEE, 2024.

Michael Saxon, Ari Holtzman, Peter West, William Yang Wang, and Naomi Saphra. Benchmarks as microscopes: A call for model metrology. *arXiv preprint arXiv:2407.16711*, 2024.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Lin Shi, Chiyu Ma, Wenhua Liang, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms. *arXiv preprint arXiv:2406.07791*, 2024.

Disha Shrivastava, Hugo Larochelle, and Daniel Tarlow. Repository-level prompt generation for large language models of code. In *International Conference on Machine Learning*, pp. 31693–31715. PMLR, 2023.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*, 2024.

Somesh Singh, Yaman K Singla, Harini SI, and Balaji Krishnamurthy. Measuring and improving persuasiveness of large language models. *arXiv preprint arXiv:2410.02653*, 2024.

Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnapati, Samuel G Rodriques, and Andrew D White. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2(8):73, 2019.

Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*, 2023.

Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, et al. Clio: Privacy-preserving insights into real-world ai use. *arXiv preprint arXiv:2412.13678*, 2024.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L Griffiths, and Faeze Brahman. Macgyver: Are large language models creative problem solvers? *arXiv preprint arXiv:2311.09682*, 2023.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. Are large language models capable of generating human-level narratives? *arXiv preprint arXiv:2407.13248*, 2024.

Aidan Toner-Rodgers. Artificial intelligence, scientific discovery, and product innovation. 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Bertie Vidgen, Nino Scherrer, Hannah Rose Kirk, Rebecca Qian, Anand Kannappan, Scott A Hale, and Paul Röttger. Simplesafetytests: a test suite for identifying critical safety risks in large language models. *arXiv preprint arXiv:2311.08370*, 2023.

Abdul Waheed, Muskan Goyal, Nimisha Mittal, and Deepak Gupta. Domain-controlled title generation with human evaluation. In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021, Volume 2*, pp. 461–474. Springer, 2022.

Melanie Walsh, Anna Preus, and Maria Antoniak. Sonnet or not, bot? poetry evaluation for large models and datasets. *arXiv preprint arXiv:2406.18906*, 2024.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*, 2023.

Alex Wang. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Angelina Wang, Aaron Hertzmann, and Olga Russakovsky. Benchmark suites instead of leaderboards for evaluating ai fairness. *Patterns*, 5(11), 2024.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *NeurIPS*, 2023.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*, 2024.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*, 2023.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. Pride and prejudice: Llm amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15474–15492, 2024.

Bohao Yang, Dong Liu, Chen Tang, Chenghao Xiao, Kun Zhao, Chao Li, Lin Yuan, Guang Yang, Lanxiao Huang, and Chenghua Lin. Crafting customisable characters with llms: Introducing simschat, a persona-driven role-playing agent framework. *arXiv preprint arXiv:2406.17962*, 2024.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. Textgrad: Automatic" differentiation" via text. *arXiv preprint arXiv:2406.07496*, 2024.

Xueying Zhang, Yunjiang Jiang, Yue Shang, Zhaomeng Cheng, Chi Zhang, Xiaochuan Fan, Yun Xiao, and Bo Long. Dsgpt: Domain-specific generative pre-training of transformers for text generation in e-commerce title and review summarization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2146–2150, 2021.

Wenting Zhao, Justin T Chiu, Jena D Hwang, Faeze Brahman, Jack Hessel, Sanjiban Choudhury, Yejin Choi, Xiang Lorraine Li, and Alane Suhr. Uncommonsense reasoning: Abductive reasoning about uncommon situations. *arXiv preprint arXiv:2311.08469*, 2023.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=Bl8u7ZRlbM`.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P Xing, et al. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*, 2023.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022.

Saurabh Bhausaheb Zinjad, Amrita Bhattacharjee, Amey Bhilegaonkar, and Huan Liu. Resumeflow: An llm-facilitated pipeline for personalized resume generation and refinement. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2781–2785, 2024.

A Thematic Analysis: Creative Composition Usage Patterns

A.1 Thematic Analysis

Thematic Analysis is a qualitative research method used to understand a dataset Braun & Clarke (2024). This technique is often used in the social sciences and also appears in human-computer interaction research Adams et al. (2008). The focus of a thematic analysis is to identify patterns in a dataset from the bottom up. Our motivation for a deeper dive into prompt usage patterns was that previous analyzes present vague categories of usage (Table 1) but not enough specificity to develop usage-based benchmarks. We take the standard steps of thematic analysis including: generating initial codes, collating codes with supporting data, identifying themes, reviewing themes, and defining themes.

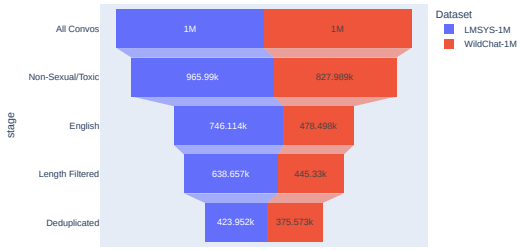

Figure 2: Prompt filtering pipeline: after applying moderation, language, length filtering an duplication, around 40% of total conversation were used for clustering and subsequent thematic analysis.

A.2 Datasets

We examine two datasets WildChat-1M (Zhao et al., 2024) and LMSYS-Chat-1M (Zheng et al., 2023). LMSYS-Chat-1M contains 1 million conversations collected from April 2023 to August 2023 through three different chat interfaces on the LMSYS website[8] (single, anonymous side-by-side and self-selected side-by-side). Users gave consent for the release of their conversational data by accepting the terms of use. This data set contains mainly English conversations and conversations with the Vicuna model (Zhao et al., 2024). The most common use case, a third of the cases, reported is programming help; this included discussing software errors, questions about AI tools, Python coding assistance, and generating SQL queries. Another common use case, "*Language and Content Creation*" is reported to be around 30% of use cases and includes scenarios such as summarization, translation, and roleplaying of various characters. Zheng et al. (2023) report that some clusters are generated in batches and submitted to their website.

The WildChat-1M dataset was collected through the Hugging Face Spaces platform from April 2023 to May 2024 (Zhao et al., 2024). This platform allowed users to chat with the GPT-3.5-Turbo or GPT-4 models. Users were presented with a data collection and sharing agreement before being given access to the chat interface. Just more than 50% of conversation were reported to be in English overall. The authors report that 61.9% of tasks where "assists/creative writing" and that the prompts they collected provide additional diversity compared to existing datasets. An additional analysis from Longpre et al. (2024) estimate that creative composition queries for WildChat is the most common use case at almost 30% of queries.

Data Filtering To understand the creative composition tasks in these datasets, we narrow the scope to analyzing the first user prompt. Since the publicly available version of WildChat already has toxic and sexual content filtered, we also apply this filter to the LMSYS prompts using the OpenAI moderation flags in the dataset. We filter for only English posts that are more than 5 words long. We also use exact match deduplication to remove repeated prompts, since we care only about unique use cases. Figure 2 shows the filtering process where we have 423K and 375K prompts remaining for LMSYS and WildChat respectively.

Topic Clustering Following the same methodology as that used by Zheng et al. (2023), we use sentence embeddings to cluster each data set into 50 clusters. The 100 clusters range from 1011 to 33463 prompts. We then apply a simple keyword search for each prompt to identify the proportion of creative prompts in each cluster. For each group with more than 30% prompts containing creative words, we sample 100 prompts and perform a thematic analysis of the creative composition use cases. We justify a simple key word approach by setting a low threshold for the percentage of creative prompts for inclusion in our analysis. This process produced 20 clusters and 154K prompts for LMSYS and 23 clusters and 135K prompts for WildChat.

A.3 Summary of Themes

For each cluster, we first manually code the different tasks that appear in the 100 sampled prompts and then ask a language model (Claude 3.5 Sonnet) to give a summary of the prompts. We take

[8] `https://chat.lmsys.org`

the union of manual and AI-identified tasks and assign general themes to the creative composition tasks. Although some clusters span multiple themes, most clusters focus on a single theme. Table 2 summarizes the 10 main themes identified; Each theme is accompanied by more specific tasks that were recorded, as well as representative examples from the sampled prompts.

A.3.1 IMAGINATIVE CREATIVITY

Script and Story Writing A main theme that frequently appears and the most common category of tasks, in creative composition tasks, is script and story writing. Prompts across both datasets and many clusters ask for stories about specific fictional characters (fanfiction), new stories, and alternative realities. There was often a focus on generating detailed descriptions of character experiences and fantasy worlds, for example:

> *Write a detailed story of A scuba diver undergoing a messy transformation into a mermaid.*

Some prompts also asked for descriptions of specific objects in fantasy worlds:

> *In a game with merging mechanics for weapons, describe a weapon created by merging meatloaf with a spatula.*

The format of writing requests here also varies from scripts to interactive fiction games to short stories. Some prompts, for example, ask for a list of character ideas rather than full stories:

> *Brainstorm unique and imaginary villains for a groundbreaking fantasy novel*

This may be multiple steps to story development where the first step is developing a character a user wants to read more about. In this category of fictional or creative writing tasks, users are likely seeking to entertain themselves by using LLM tools to generate new, but specific stories that they want to read.

Poetry and Music Generation In both datasets, there was at least one group on the generation of poetry and music. These generation tasks range from writing poems to songs as well as generating playlists and other rhetorical devices such as proverbs. However, it is notable that these prompts are focused on very specific constraints rather than generating something to win the next poetry prize. For example:

> *write a poem to purpose my girlfriend NAME_1, she is a marketing manager in a tech company.*

A.3.2 CREATIVE COMMUNICATION

Advice and Interpersonal Communication In this category on composition tasks, contextual awareness may be more important than creativity. However, a significant number of advice-based prompts are related to brainstorming solutions for specific problems. For example:

> *What is the single change I can make that will most improve my life?*

Although these questions are explicitly advice seeking, there is not one correct answer, and a user might be using the model for brainstorming and would thus benefit from a diverse set of useful suggestions. Other prompts on this theme include writing text messages, greetings, and other personal messages.

Title, Name, and Phrase Generation Many prompts explicitly asked for search-engine-optimized (SEO) product descriptions and hashtags. Another common use case was asking for names, titles, and slogans for different products and characters. Many of these prompts explicitly asked for multiple suggestions. For example:

> *brainstorm a creative list of username for my personal branding account. it must contain word 'abito' in it somewhere.*

In asking for a list of possible suggestions, the user benefits from both high-quality suggestions and diversity in suggestions.

Professional Writing Another frequent category of prompts is professional writing tasks. Users ask for help with writing emails, reviews, feedback, articles, blogs, job descriptions, cover letters, and resumes. They ask for help both in the drafting process and in the editing process. Some of these tasks are very high stakes, for example, personal statements:

> *write a paragraph in a Personal Statement why i took HL English and HL History in IB to pursue a law degree at a University*

In these tasks where the generated material may be evaluated by humans or other applicant screening systems, infusing creativity into the writing would help an application stand out. Since both hiring and admissions committees may see thousands of resumes and personal statements, generating a unique and high-quality response would be helpful for the user. Other prompts ask to elicit sincere communication, for example:

> *I need to write an email that gives the idea that I am enthused with inclusion and diversity with all the usual jargon and that I would be happy to go to s meeting about it, can you do this?*

In these types of professional writing, diverse outputs to choose from may better help convey authenticity.

A.3.3 Productive Creativity

Business Development A less common but also business-oriented theme is marketing and business development ideas and proposals. These prompts relate to different segments of the business development process from brainstorming ad campaigns to corporate strategy. Business schools have long touted the importance of creativity and design thinking in the business world (Boyles, 2022); the prompts in this category illustrate that users might be turning to language models for this creativity. For example, some users ask for a full business plan based on a light description:

> *can i start my own business as a real estate agent in NAME_1,can you help me draft a detailed robust business plan*

while other users want a list of potential directions to pursue:

> *Give me a list of ideas for a startup, created by a computer science student with zero to little inversion.*

These tasks fit under the notion of business creativity and innovation. Ideas that are both unique (compared to existing approaches) and strategic are helpful for these users.

A.3.4 Creative Inquiry

The four themes under creative inquiry encompass argument development and critical thinking. Application domains include sports, education, and scientific development. The goals in these creative composition tasks is to create original or novel developments.

Argument Composition Several prompts explicitly ask for reasoning and arguments for specific topics. Some topics may require access to facts but other arguments are altogether hypothetical. There may not be right or wrong answers to these questions but rather different perspectives. For example:

> *is free trade a good or bad thing*

Diverse but thoughtful arguments in the responses to these questions would help the user better engage with the topic. When these topics are applied to domains such as debate, creativity is valued for developing novel, yet convincing arguments.

Academic Writing Another theme is where users explicitly ask for help with educational tasks. These prompts include various stages of writing help from generating essay outlines to drafting and revising essays. Also in this category are prompts asking to learn or teach certain concepts, for example:

> *I want you to act as a data science instructor. Explain what neural network is to an undergraduate.*

Many educators have highlighted the role of creativity in effective teaching Jeffrey* & Craft (2004); Beghetto (2017). A user prompting the language model may want effective and novel ideas for explaining or illustrating concepts.

Scientific Development Using AI for ideation has recently been studied for various scientific pursuits from material discovery (Toner-Rodgers, 2024) to prompt engineering (Si et al., 2024). This theme also appeared in our analysis, for example:

> *give me interesting Human-Computer interaction project can be done using intel realsense 3D camera*

Users gave specific topics and constraints in hopes of generating ideas for research. Another common mode under this theme is the generation of prompts for tasks. For example prompts start with:

> *You are a helpful Stable Diffusion prompt generator.*

There may be one answer that is better than others but a group of diverse prompts may ultimately help generate a better set of outputs.

Forecasting A less common but distinct theme is the usage of LLMs for projections or forecasting. There were many requests for sports outcomes and scripts of upcoming matches as well as financial market predictions. Since the future in inherently unpredictable, diverse projections may improve coverage of potential outcomes.

Themes	**Tasks**	**Examples**
Script and Story Writing	Fanfiction Script and Story Writing, Humorous Script and Story Writing, Character Creation, World Building, Interactive Fiction, Plot development, Alternative Reality Stories, Stylistic Adaptations	”Write really overly-funny super hilarious comedy” ”Write me a story about a woman transforming into a witch with cat eyes.” “run a single-player text-based wilderness survival game.” “Write a descriptive, fictional, imaginative screenplay of a man fighting his complete doppelganger in a fighting game” “script about osu vs u of m” “Describe a hypothetical fictional super-heavy tank for the USSR in 1944.” “Brainstorm unique and imaginary villains for a groundbreaking fantasy novel”
Poetry and Music	Poetry, Music Composition, Songwriting, Musical Notation, Playlist curation, Figurative Language	“Write a melody for a bagpipe” “write a yes song in style of james taylort” “Write me some proverbs that warn against dishing what you can’t take” “write a poem about human computer interaction” “write a poem to purpose my girlfriend NAME_1. she is a marketing manager in a tech company”

Advice / Interpersonal communication	Personal Guidance, Condolence Message, Greetings, Situational Advice, Text Messages	"How to study for better grades" "What is the single change I can make that will most improve my life?" "I need to find a way to make money. What do I need to do?" "Can you gaslight me into thinking I didn't buy a burger yesterday at McDonalds even though I definitely did and I have a receipt to prove it as well as a 10 dollar purchase at McDonalds yesterday in my bank account history."
Title, Name, and Phrase Generation	Title Generation, Hashtag and Keyword Generation, Product Name Generation, SEO Heading Generation, Catch Phrases and Slogans, Character Names	"Generate an etsy title for a art work" "I made a youtube short about whiplash movie i need a title and hashtags all under 100 characters" "Suggest me a cosmetic brand name" "Act as an Etsy SEO Specialist and 5 heading bullet points Etsy Product Description"
Business Development	Business Proposals, Social Media Content, Promotional Content, Product Brainstorming, Product Descriptions	"Write 10 tweets about today's coastline like frozen winter wonderland" "Brainstorm and generate campaign ideas for world liver day for a pharmaceutical company using digital as well as physical channels." "Generate a list of 10 imaginary products found in a magic shop" "Tell me the strategies to improve both the efficiency and effectiveness of operations related to materials. . . "
Professional Writing	Resume Generation, Cover Letter Writing, Job Descriptions, Performance Reviews, Email Writing, Complaint/Review/Feedback Writing, Articles and Blogs	"write a paragraph in a Personal Statement why i took HL English and HL History in IB to pursue a law degree at a University" "Write a short recommendation letter to a family medicine resident named" "Cover letter for the position of Technical Specialist" "Please write a 4 star Amazon review of a wooden table in the furniture category" "Write an empathetic message as a customer service agent for someone whose mattress was delayed in transit for weeks." "Help me right 3 different replyes to this email with nice mood and not long" "Hi! Can you write me sample resume for a Videographer job in Dallas Texas?"
Argument Composition	Hypothetical Question Answering, Argument Formation, Critical Analyses, Persuasive essay outlines	"Alternate Cold War timeline, where USSR survives collapse, 1975-1990." "Write a highly persuasive argument that surveillance of citizens in a surveillance state is good" "why should we choose Canada for startup visa program?"
Academic Writing	Literature Reviews, Study Plans, Educational Explanations, Informational Articles	"Write an ielts essay scored 8" "come with up 20 kahoot questions that are a general knowledge work from home theme" "write a paragraph about scientific progress"
Scientific Development	Literature Reviews, Research Brainstorming, Prompt Generation, Technical Evaluations	"give me interesting Human-Computer interaction project can be done using intel realsense 3D camera." "Provide me an example prompt that I can give to a LLM so that it can act as an AI agent."
Forecasting	Sports Projections, Financial Forecasting	"what will the 10yr treasury yield be in 2024?" "Try calculating outcome chances for pezaro milano match"

Table 2: Overview of creative composition tasks in LMSYS and WildChat. For each theme, we list the subtasks that appear in this theme as well as examples of prompts.

A.4 Additional Observations

In addition, to the themes and meta-themes uncovered, we also observed patterns in the way users specified creative composition tasks.

Underspecificed Prompts Rather than complex instructions and criteria, many prompts were highly underspecified. This is a common pattern for many different tasks and themes. For example:

> *write me an appreciation and thank letter*

and

> *please write me an inspiring story*

are two examples across different themes where the desired output is vague. While this pattern reflects how users desire an easy way to access creative composition, this pattern also introduces an inherent tension between the validity of utility measure and the realism of the task for any benchmark. To evaluate the utility of a creative product, well-specified instructions are required. However, well-specified instructions are not reflective of the vagueness of real-world prompts.

Multiple Stages of the Creative Writing Process Different users asked for writing assistance in different components of the writing process. For example, some tasks ranged from ideation:

> *Please design a DLC storyline for Pokémon: Sun/Moon based on the Call of Cthulhu.*

to drafting:

> *write a paragraph about scientific progress*

to editing:

> *can you make this email nicer*

Although all these stages appear in user prompts for creative composition tasks, the ideation and drafting stages likely allow a larger creative composition space. Moreover, more existing creativity measurements focus on the drafting stage (Lu et al., 2024; Chakrabarty et al., 2024) than the ideation stage (Si et al., 2024).

Power Users Our analysis found that certain users, or groups of users, dominated entire clusters. This finding is consistent with Zheng et al. (2023). We categorized power users into three main categories: (1) LLM testing (LMSYS), fanfiction (LMSYS and WildChat), and customized stories (WildChat). For example, of the creative clusters we analyzed, we discovered the following large clusters of prompts:

- 1671 prompts about Jane, the American TV Series
- 10311 prompts about the anime Doki Doki Squad
- 3643 prompts for customized stories about David and sexual hypnosis scenarios
- 4441 prompts for a specifically formatted fact-checking task
- 4058 prompts on Freedom Planet Fanfiction and Crossovers
- 6986 prompts on writing the introduction of a chemical company
- 1996 prompt testing logical reasoning of a model called “SmartGPT”

Creative Terms	Non-Creative Terms
imagine, envision, innovate, inspire, story, poem, novel, write, draw, paint, compose, art, brainstorm, idea, suggestion, argument, creative, fake, act, new, novel	code, python, problem, solution, c#, app, api, client, serve, prompt, prompts, chatgpt, openai, lua, java

Table 3: List of Creative and Non-Creative Terms Used for Filtering.

- 14006 prompts about writing an article about a specific chemical in the chemical industry

These clusters persisted even after our de-duplication since each prompt is actually unique and even rewritten in various ways for the fan-fiction tasks. Prompts for sexual content persisted even after filtering for adult and toxic content; both datasets contained several creative task clusters (around 15 - 25k prompts) of prompts for sexual story writing. This finding is relevant not only for studying creative composition but also for all users of these two datasets. The high concentration of these prompts may lead to desirable downstream outcomes (e.g., the objectification of certain populations).

Creative Word List We used the following list of words to identify clusters that may contain creative tasks. Table 3 shows the creative and non-creative terms we used. We labeled a prompt creative if it contained at least one word from the creative list and did not contain any words from the non-creative list.