Creative Planning with Language Models: Practice, Evaluation and Applications

Alexander Spangher[1] **Tenghao Huang**[1] **Philippe Laban**[2] **Nanyun Peng**[3]
[1]University of Southern California
[2]Microsoft Research
[3] University of California, Los Angeles
spangher@usc.edu, tenghaoh@usc.edu, philippe.laban@microsoft.com, vnpeng@ucla.edu

Abstract

The use of large language models (LLMs) in human-centered creative domains — such as journalism, scientific writing, and storytelling — has showcased their potential for content generation but highlighted a critical gap: planning. Planning, a fundamental process in many creative domains, refers to higher level decisions writers (or agents) make that influence textual output they produce. Planning is especially hard to perform in creative domains, where human rewards are often unclear or sparsely observed. This tutorial explores how planning has been learned and deployed in creative workflows. We will cover three aspects of creativity: **Problem-Finding** (how to define rewards and goals for creative tasks), **Path-Finding** (how to generate novel creative outputs that meet goals) and **Evaluation** (how to judge). We will also consider three learning settings: *Full Data Regimens* (when observational data for decisions and resulting text exist), *Partial* (when text exists but decisions can be inferred) and *Low* (when neither exist). The tutorial will end with practical demonstrations in computational journalism, web agents, and other creative domains. By bridging theoretical concepts and practical demonstrations, this tutorial aims to inspire new research directions in leveraging LLMs for creative planning tasks.

1 Introduction

LLMs have demonstrated impressive generative capacities across a range of tasks. However, many human creative tasks (e.g. in journalism, scientific writing, video script writing and creative story generation) involve extensive planning. For example, a human journalist typically follows a multi-step process before they are even *ready* to write a news article (e.g. "*find story idea*" → "*develop angle*" → "*find informational sources*" → "*get quotes*" → "*confirm facts*") (Cohen et al., 2011). An emerging body of work has pointed to key short-comings of LLMs and opportunities for progress in domains where planning is required, actions need to be taken and objectives are poorly defined.

Many emerging tasks in NLP can be framed as "planning" tasks: either those that are explicitly using LLMs as planning-agents (e.g. (Zhou et al., 2023)) or those that attempt to infer or learn from the plans guiding human text generation (Spangher et al., 2024a). In this tutorial, we aim to bring tasks in this umbrella into dialogue. Can the ability to plan make LLMs become more useful, more human-like and more attuned to the needs of diverse creative professionals? We aim to consolidate an emerging direction of work that lies in the intersection of: creative generation, agentic planning, and human-centered NLP.

2 Three Aspects of Creativity

The main structure of our tutorial breaks down creative planning into three main stages: Problem-Finding, Path-Finding and Evaluation. Each section is grounded in a large history cognitive science literature. We cover each stage in turn.

2.1 Problem-Finding

> "The formulation of a problem is often more essential than its solution, which may be merely a matter of mathematical or experimental skill."
>
> *Einstein and Infeld (1938), The Evolution of Physics*

In their seminal work, Getzels and Csikszentmihalyi (1976) studied art students, and observed that those who focused most on *defining the problem* produced more creative work. The *problem-finding* domain of creativity research has since expanded to include various ways that creative actors define tasks, goal-states and rewards.

We map *problem-finding* broadly, in NLP, to Learning *complex rewards*. A key question is: how can we build systems that define their own reward

functions, understand fuzzily observed rewards or mix multiple rewards for one cohesive output? We will frame several advances in language modeling in this lens. We will look at approaches that mix multiple rewards (Shi et al., 2024), framing of language modeling as inverse-reinforcement learning (IRL) (Wulfmeier et al., 2024), and explicit *emulation learning* settings.

What is emulation learning? In cognitive science, key work has been done to observe how humans (and chimpanzees) learn rewards (Hopper, 2010). In such *emulation learning* settings, humans attempt to observe and understand the *motivations and rewards* of other humans relying not just on observing the actions of others, but also upon observing the *end-state outputs* of human processes. For example, when we, as scientists, read research papers, we are often able to "read through the lines" to guess actions that were taken, even if they are not explicitly mentioned – e.g. implementation decisions, negative results, or hyperparameter sweeps (without this ability, reproducibility in our field would be nearly impossible). Another domain is shown in Figure 1, where the decision-making other humans employ prior to writing a news article can often be inferred through discourse markers (Spangher et al., a). Computer science work focused on *emulation learning* typically seeks to explicitly uncover human actions from observed text (e.g. in news articles (Spangher et al., 2024b) and in scientific writing (Starace et al., 2025)). The key in these approaches is, after uncovering these actions, is to then use them to learn rewards from human behavior at scale, utilizing frameworks like IRL (Abbeel and Ng, 2004). In this talk, we will explore how these approaches can uncover human values, motivations and rewards.

2.2 Path-Finding

> "Creativity involves breaking out of established patterns to look at things in a different way."
>
> *de Bono (1992), Serious Creativity*

Defining creativity as *how humans develop alternative methods for solving problems* has been another dominant thread in creativity research (Runco, 2001), dating to the 1950s, when J.P. Guilford addressed the American Psychological Association (Guilford, 1950). Guilford and others developed theories of creativity centered on *path-finding*, where humans engage in alternative uses (Guilford et al., 1978), exploration (Finke, 1996) and

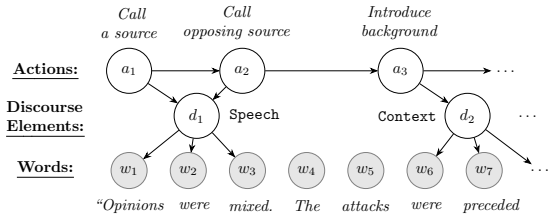

Figure 1: An example of a creative-planning task, in computational journalism: specifically, planning which sources and other forms of background information to use. Actions taken by humans while they write (i.e. a_1, a_2..) are implied through discourse acts (d_1, d_2...), which are inferred from the written text (w_1, w_2,...). This insight, and it's application across domains, allows us to infer higher-level human rewards and train agents to understand human creative processes.

metaphor (Gentner and Wolff, 2014) to come up with more inventive solutions to problems. We will explore ways computer scientists have extended such directions.

Forwards approaches Forward approaches to planning assume that we can directly train or prompt a model to generate sequences of actions. Researchers typically take an approach that involves prompt-engineering and in-context learning (Tian et al., 2024b). We will discuss some of the drawbacks of these approaches, including biases that might be introduced and reasoning failures in modeling. On the other hand, researchers with access to more data usually include enough training data to explicitly train planning agents. This can include directly planning a chain-of-thought reasoner (Chen et al., 2024b) or a environment with clearly defined reward (e.g. a tool-usage platform) (Côté et al., 2018; Shridhar et al., 2020, 2021; Huang et al., 2023; Tian et al., 2024b; Song et al., 2024). These approaches typically fall into an area of reinforcement learning referred to as imitation learning: human actions are observed, and the goal is to infer the motivations behind them in order to predict them in the future.

Backwards approaches Here, state information is available (even if just the end state), and we usually seek to infer the sequence of actions that lead to this state. Theoretically, these approaches call back to earlier domains of modeling: means-ends analysis (Newell and Simon, 1961), backtracking (Golomb and Baumert, 1965) and regression planning (McDermott, 1991; Xu et al., 2019). These methods all assume access to the final state, and

use this information to arrive here. We will discuss recent approaches incorporating these ideas into NLP (Gandhi et al., 2024; Chen et al., 2024a). We will also discuss backwards reasoning in terms of latent variable modeling, for example: discovering in-context learning examples (Min et al., 2022); infer underlying topics by generating and clustering language-modeling responses (Pham et al., 2024); learning form and structure via the Bayesian Wake-Sleep algorithm; and infer chain-of-thought reasoning steps through bootstrapping (Zelikman et al., 2022). We will highlight the overlapping symmetry between variational inference formulas and classical RL formulations. By illustrating how latent variable modeling and imitation learning can be integrated to infer and utilize latent plans, we discuss the benefits of combining these approaches for modeling creative tasks.

2.3 Evaluation Methods for Creative Plans

> "The unexamined life is not worth living."
>
> *Plato, Apology of Socrates*

For the majority of tasks in creative domains, there is no objective metric for when a plan is successful: creative tasks can be ill-defined, with multiple alternative plans being equally preferable. Thus, in this section of the tutorial, we will focus on evaluation methods based around human preference. There are two modes of evaluation:

Offline Evaluation In this evaluation setting, we assume that we cannot conduct human experiments on enough subjects to make meaningful conclusions, either because they are unavailable or too expensive to obtain data from. The goal of evaluations in this setting is to compare *our* plans to what human plans *would have been*. Novel metrics that have emerged in this space and have been used to evaluate planning include: *latent criticism* (Shi et al., 2023) and *conditional perplexity* (Chen et al., 2019). Latent criticism involves modeling and evaluating the underlying reasoning processes in language models, while conditional perplexity assesses the alignment between generated text and the intended plan. These evaluation metrics moves beyond surface-level metrics, e.g. BLEU or ROUGE scores, whose limitations we will discuss, towards structural comparisons of the output. They are appealing because they allow us to validate in a largely offline manner, without recruiting subject participants.

Online Evaluation Evaluation methods in this setting fall more into a Human-Computer Interaction (HCI) framework of evaluation. In this setting, subject participants are recruited and either asked to conduct trials or are allowed to use tools and then observed. HCI approaches to studying human preferences for plans can involve studying human preferences for recommendations (Spangher, 2015; Zhao et al., 2023), suggestions (Clark and Smith, 2021), edits (Laban et al., 2024) and other aides that a model can provide short of generating an entire text. We will not focus too deeply on this area, though, at the risk of being duplicative with other tutorials.

3 Data Regimes: Full, Partial and Low

In order to conceptualize methods required to study creative planning, we divide creative tasks into three categories based on the availability of data: **Full Visibility**, **Partial Visibility** and **Low Visibility**. To frame these categories, we use vocabulary from the field of reinforcement learning: *actions* refers to planning steps or inferences the model can take. *State-space* refers broadly to textual states (e.g. utterances, documents or retrievals) that are caused or influenced by actions.

Low Data Regimens: settings in which little-to-no data is available about the planning process, including either the end-states or any of the actions or states in between. Examples of tasks in this domain, including: OSWorld (Xie et al., 2024b), WebArena (Zhou et al., 2023) and other web-agent tasks (Branavan et al., 2009; Shi et al., 2017; Liu et al., 2018; Deng et al., 2023; Kim et al., 2024; Gur et al.), where the language model is tasked with navigating webpages without any examples of the output.

Partial Data Regiments: settings where end-state information, but no actions, are available to the planning process. Tasks in this planning domain encompass fields like: computational journalism (Spangher et al., 2024a), computational law (Ravichander et al., 2019), scientific writing (Si et al., 2024) and creative fictional writing (Huang et al., 2023; Tian et al., 2024a). In these tasks, it is typically cheap to collect voluminous datasets of finished news articles, for instance, but it is typically too expensive to observe actions leading up to the finished articles.

Partial-to-Full Data Regiments are characterized by situations in which pre-final text *and/or* action sequences are available for the models to train on. We briefly introduce various tasks and domains where datasets have emerged to support these plans plans, such as tool learning (Schick et al., 2023; Patil et al., 2023; Qin et al., 2023; Li et al., 2023), edit prediction (Spangher et al., 2022b; Lee et al., 2024), math problem-solving (Cobbe et al., 2021; Hendrycks et al., 2021) and instruction-learning (Wu et al., 2023, 2022). In these settings, more of a supervised approach can be taken to learn plans.

4 Application Domains of Creative Planning: Demonstrations

Having established a better definition for "plans" and methods for inferring plans from observed text, we close by discussing applications in various domains. We will give live demonstration of creative tools and compare tools that do not formally plan (e.g. those that engineer sequences of prompts) with tools that do.

Computational Journalism (CJ) This field aims to build decision-support tools for journalists to help find stories and sources; verifying facts; and write articles (Cohen et al., 2011). CJ gives us a good example of a domain of tasks where (1) abundant medium-visibility data exists (2) professional standards across organizations dictate regular and formalized planning and (3) outcomes are socially beneficial. Recent tasks in CJ include: "help a journalist find informational sources to support the story" (Huang et al., 2024a; Spangher et al., a,b; Lu et al.), "find newsworthy stories to cover" (Spangher et al., 2024b; Welsh et al.; Diakopoulos et al., 2010), "plan longer-term article structures" (Spangher et al., 2022a, 2021; Choubey et al., 2020). We will showcase tools without formalized planning, such as *AngleKindling*, a tool for angle selection in journalistic writing (Petridis et al., 2023). We then demonstrate tools that learn and utilize latent plans to enhance output quality, such as *NewsSources* (Huang et al., 2024a) and SPINACH (Liu et al., 2024).

Proactive Task-oriented Agents This field aims to build agents to proactively identify and clarify missing or ambiguous information essential to methodical, domain-specific tasks (Lu et al.; Wu et al., 2025; Liu et al., 2025). By systematically examining the effects of partially available information, these methods train models to optimally balance the cost of queries against the improved accuracy and completeness of their outputs. This proactive reasoning capability significantly enhances the practical utility, creativity and reliability of task-oriented agents, particularly in high-stakes, information-sensitive environments, and represents a promising direction for human–AI collaborative systems.

Creative Writing and Editing Planning plays a crucial role in creative language generation, especially in long-form text generation. Content planning, such as sketching out plot points (Yao et al., 2019; Ammanabrolu et al., 2020; Clark and Smith, 2021), has been shown to improve the quality of generated stories and for generating creative outputs like poetry, where form constraints must be adhered to (Tian and Peng, 2022), or metaphor or figurative language (Chakrabarty et al., 2021) must be used. Incorporating knowledge into the planning process can significantly enhance the ability of LLMs to produce more nuanced, creative outputs (Bosselut et al., 2019; Chakrabarty et al., 2024).

5 Broader Relevance: Connection to Existing Fields

We situate creative planning in a broader field of artificial intelligence and natural language processing, with explicit intersections in:

- **Creative Generation**: Although recent tutorials (Chakrabarty et al., 2023) have covered creative generation, prior work has focused more on the "final product" of generation (e.g. longer-form structural output, cohesiveness and evaluation), not the planning steps. However, awareness of creative processes in different fields and the ability of LLMs to understand and use plans have progressed rapidly, necessitating a novel iteration to explicitly focus on planning in creative tasks.

- **Agentic Planning**: Task-oriented planning (Yu et al., 2023; Huang et al., 2024b; Deng et al., 2024; Zhang et al., 2024; Kohli and Sun, 2024; Xie et al., 2024a), agentic workflows (Wang et al., 2023, 2024; Sodhi et al., 2024; Huang et al., 2024c, 2025) likewise is an area that has received tremendous interest. However, we find the focus of planning in *creative* tasks to be notably lacking. As we will show, creative tasks are tantalizing tasks for planners and agents because trajectories must be

developed on the fly in these domains (Côté et al., 2018; Shridhar et al., 2020, 2021; Tian et al., 2024b).

- **Human-Centered NLP**: A large emphasis in prior Human-Centered NLP tutorials (Yang et al., 2024) has been in Human-Computer Interaction (HCI)-focused methodologies. While this is an important component, we will explicitly focus on emerging experimental methodologies that seek to *infer* human preferences in approaches that can often be more generalizable and robust than direct observational studies.

We consider the following skills useful for researchers considering making advances in creative planning:

- **Latent Variable Modeling**: an understanding of classical Bayesian graphical modeling and hierarchical reasoning. Understand how reinforcement learning, specifically imitation learning, forms the basis for human preference learning.
- **Evaluation Methods for Latent Plans**: Evaluation metrics, like latent evaluation techniques like latent criticism and conditional perplexity, that go beyond surface-level assessments.
- **Creative Agentic Workflows**: Explore how inferred plans are applied in creative tasks. Analyze the differences in model performance when optimizing for concrete rewards versus abstract, creative goals (i.e. imitating human preference). Demonstrate of creative tools and compare those that use engineered prompt sequences with those that utilize latent plans.

6 Suggested Reading List Summary

This tutorial will include our own work, notably in the fields of computational journalism, creativity, latent variable modeling and agent modeling (Huang et al., 2024a; Spangher et al., 2024a, b; Welsh et al.; Spangher et al., 2021; Lu et al.; Tian et al., 2024b) and work by other researchers in NLP and machine learning communities, including but not limited to: (Petridis et al., 2023; Shi et al., 2023; Deng et al., 2023; Schick et al., 2023; Shridhar et al., 2020; Chakrabarty et al., 2023; Zelikman et al., 2022).

7 Tutorial Instructors

Our instructors consist of experts who have conducted research in different aspects related to this tutorial topic.

Alexander Spangher Alexander Spangher is a final-year Ph.D. Candidate in the Department of Computer Science at University of Southern California. He is the recipient of a Bloomberg PhD fellowship and an Outstanding Paper awards at EMNLP 2024 and NAACL 2022. His research focuses on planning, with specific applications in Computational Journalism, law and music. Prior to this, he was a data journalist at *The New York Times*.

Tenghao Huang Tenghao Huang is a Ph.D. Candidate in the Department of Computer Science at University of Southern California. Tenghao is a receipt of ISI distinguished graduate researcher fellowship. His research interests lie in agents and information retrieval. His recent work focuses on bridging the gaps between agents and creative tasks through planning and grounding. Prior to this, Tenghao received his bachelor degree from the University of North Carolina at Chapel Hill.

Philippe Laban Philippe Laban is a Research Scientist at Microsoft Research. His research is at the intersection of NLP and HCI, focusing on several tasks within text generation, including text simplification and summarization. He received his Ph.D. in Computer Science from UC Berkeley in 2021. His recent work has focused on expanding the scope of text simplification to the paragraph and document-level and evaluating textediting interfaces.

Nanyun (Violet) Peng Nanyun (Violet) Peng is an Assistant Professor in the Department of Computer Science at the University of California Los Angeles. She received her Ph.D. in Computer Science from Johns Hopkins University. Her research focuses on the generalizability of NLP technologies, with applications to creative language generation, low-resource information extraction, and zero-shot cross-lingual transfer. Her works have won the Outstanding Paper Award at NAACL 2022, the Best Paper Award at AAAI 2022 Deep Learning on Graphs workshop, and have been featured an IJCAI 2022 early career spotlight.

References

Pieter Abbeel and Andrew Y Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1.

Prithviraj Ammanabrolu, Wesley Cheung, Anju Dhamala Dang, William Broniec, Matthew Hausknecht, and Mark O. Riedl. 2020. Story realization: Expanding plot events into sentences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9055–9069.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779.

S.R.K. Branavan, Harr Chen, Luke Zettlemoyer, and Regina Barzilay. 2009. Reinforcement learning for mapping instructions to actions. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 82–90, Suntec, Singapore. Association for Computational Linguistics.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2021. Mermaid: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4250–4261.

Tuhin Chakrabarty, Philippe Laban, and Chien-Sheng Wu. 2024. Can ai writing be salvaged? mitigating idiosyncrasies and improving human-ai alignment in the writing process through edits. *arXiv preprint arXiv:2409.14509*.

Tuhin Chakrabarty, Vishakh Padmakumar, He He, and Nanyun Peng. 2023. Creative natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 34–40, Singapore. Association for Computational Linguistics.

Justin Chih-Yao Chen, Zifeng Wang, Hamid Palangi, Rujun Han, Sayna Ebrahimi, Long Le, Vincent Perot, Swaroop Mishra, Mohit Bansal, Chen-Yu Lee, et al. 2024a. Reverse thinking makes llms stronger reasoners. *arXiv preprint arXiv:2411.19865*.

Xinyun Chen, Changliu Wang, Fisher Yu, and Dawn Song. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1909.12840*.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024b. Self-play fine-tuning converts weak language models to strong language models. *Preprint*, arXiv:2401.01335.

Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Elizabeth Clark and Noah A Smith. 2021. Choose your own adventure: Paired suggestions in collaborative writing for evaluating story generation models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3566–3575.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Sarah Cohen, James T Hamilton, and Fred Turner. 2011. Computational journalism. *Communications of the ACM*, 54(10):66–71.

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, et al. 2018. Textworld: A learning environment for text-based games. In *Workshop on Computer Games*, pages 41–75. Springer.

Edward de Bono. 1992. *Serious Creativity: Using the Power of Lateral Thinking to Create New Ideas*. HarperBusiness.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2024. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. *Preprint*, arXiv:2306.06070.

Nicholas Diakopoulos, Mor Naaman, and Funda Kivran-Swaine. 2010. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *2010 IEEE Symposium on Visual Analytics Science and Technology*, pages 115–122. IEEE.

Albert Einstein and Leopold Infeld. 1938. *The Evolution of Physics*. Simon and Schuster, New York.

Ronald A Finke. 1996. Imagery, creativity, and emergent structure. *Consciousness and cognition*, 5(3):381–393.

Kanishk Gandhi, Denise Lee, Gabriel Grand, Muxin Liu, Winson Cheng, Archit Sharma, and Noah D Goodman. 2024. Stream of search (sos): Learning to search in language. *arXiv preprint arXiv:2404.03683*.

Dedre Gentner and Phillip Wolff. 2014. Metaphor and knowledge change. In *Cognitive dynamics*, pages 295–342. Psychology Press.

Jacob W Getzels and Mihaly Csikszentmihalyi. 1976. The creative vision: A longitudinal study of problem finding in art.

Solomon W Golomb and Leonard D Baumert. 1965. Backtrack programming. *Journal of the ACM (JACM)*, 12(4):516–524.

Joy Paul Guilford. 1950. Creativity. *American psychologist*, 5(9):444–454.

Joy Paul Guilford, Paul R Christensen, Philip R Merrifield, and Robert C Wilson. 1978. Alternate uses.

Izzeddin Gur, Hiroki Furuta, Austin V Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and Aleksandra Faust. A real-world webagent with planning, long context understanding, and program synthesis. In *The Twelfth International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Lydia M Hopper. 2010. 'ghost'experiments and the dissection of social learning in humans and animals. *Biological Reviews*, 85(4):685–701.

Tenghao Huang, Kinjal Basu, Ibrahim Abdelaziz, Pavan Kapanipathi, Jonathan May, and Muhao Chen. 2025. R2d2: Remembering, reflecting and dynamic decision making for web agents. *Preprint*, arXiv:2501.12485.

Tenghao Huang, Yiqin Huang, Lucas Spangher, Sewon Min, Mark Dredze, and Alexander Spangher. 2024a. A novel multi-document retrieval benchmark grounded on journalist source-selection in newswriting.

Tenghao Huang, Dongwon Jung, Vaibhav Kumar, Mohammad Kachuee, Xiang Li, Puyang Xu, and Muhao Chen. 2024b. Planning and editing what you retrieve for enhanced tool learning. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 975–988, Mexico City, Mexico. Association for Computational Linguistics.

Tenghao Huang, Donghee Lee, John Sweeney, Jiatong Shi, Emily Steliotes, Matthew Lange, Jonathan May, and Muhao Chen. 2024c. Foodpuzzle: Developing large language model agents as flavor scientists. *Preprint*, arXiv:2409.12832.

Tenghao Huang, Ehsan Qasemi, Bangzheng Li, He Wang, Faeze Brahman, Muhao Chen, and Snigdha Chaturvedi. 2023. Affective and dynamic beam search for story generation. *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2024. Language models can solve computer tasks. *Advances in Neural Information Processing Systems*, 36.

Harsh Kohli and Huan Sun. 2024. Cleared for takeoff? compositional conditional reasoning may be the achilles heel to (flight-booking) language agents. *Preprint*, arXiv:2404.04237.

Philippe Laban, Jesse Vig, Marti Hearst, Caiming Xiong, and Chien-Sheng Wu. 2024. Beyond the chat: Executable and verifiable text-editing with llms. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–23.

Ryan Lee, Alexander Spangher, and Xuezhe Ma. 2024. Patentedits: Framing patent novelty as textual entailment.

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. 2023. Api-bank: A comprehensive benchmark for tool-augmented llms. *Preprint*, arXiv:2304.08244.

Evan Zheran Liu, Kelvin Guu, Panupong Pasupat, and Percy Liang. 2018. Reinforcement learning on web interfaces using workflow-guided exploration. In *International Conference on Learning Representations*.

Shicheng Liu, Sina J Semnani, Harold Triedman, Jialiang Xu, Isaac Dan Zhao, and Monica S Lam. 2024. Spinach: Sparql-based information navigation for challenging real-world questions. *arXiv preprint arXiv:2407.11417*.

Xingyu Bruce Liu, Shitao Fang, Weiyan Shi, Chien-Sheng Wu, Takeo Igarashi, and Xiang Anthony Chen. 2025. Proactive conversational agents with inner thoughts. *Preprint*, arXiv:2501.00383.

Michael Lu, Hyundong Cho, Weiyan Shi, Jonathan May, and Alexander Spangher. Newsinterview: a dataset and a playground to evaluate llms' ground gap via informational interviews.

Drew McDermott. 1991. Regression planning. *International Journal of Intelligent Systems*, 6(4):357–416.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Allen Newell and Herbert Alexander Simon. 1961. Gps, a program that simulates human thought.

Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. 2023. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334*.

Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023.

Anglekindling: Supporting journalistic angle ideation with large language models. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–16.

Chau Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. Topicgpt: A prompt-based topic modeling framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2956–2984.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4947–4958, Hong Kong, China. Association for Computational Linguistics.

Mark A Runco. 2001. Introduction to the special issue: Commemorating guilford's 1950 presidential address. *Creativity Research Journal*, 13(3-4):245–245.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

Ruizhe Shi, Yifang Chen, Yushi Hu, Alisa Liu, Hanna Hajishirzi, Noah A Smith, and Simon S Du. 2024. Decoding-time language model alignment with multiple objectives. *Advances in Neural Information Processing Systems*, 37:48875–48920.

Tianlin Shi, Andrej Karpathy, Linxi Fan, Jonathan Hernandez, and Percy Liang. 2017. World of bits: An open-domain platform for web-based agents. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3135–3144. PMLR.

Zewei Shi, Wanyu Du, and James Zou. 2023. Large language models as optimizers. *arXiv preprint arXiv:2301.07085*.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2021. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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

Paloma Sodhi, S. R. K. Branavan, Yoav Artzi, and Ryan McDonald. 2024. Step: Stacked llm policies for web actions. *Preprint*, arXiv:2310.03720.

Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, and Bill Yuchen Lin. 2024. Trial and error: Exploration-based trajectory optimization for llm agents. *Preprint*, arXiv:2403.02502.

Alexander Spangher. 2015. Building the next new york times recommendation engine. *The New York Times*, pages 08–26.

Alexander Spangher, Jonathan May, Sz-Rung Shiang, and Lingjia Deng. 2021. Multitask semi-supervised learning for class-imbalanced discourse classification. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 498–517.

Alexander Spangher, Yao Ming, Xinyu Hua, and Nanyun Peng. 2022a. Sequentially controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6848–6866.

Alexander Spangher, Nanyun Peng, Emilio Ferrara, and Jonathan May. a. Identifying informational sources in news articles. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Alexander Spangher, Nanyun Peng, Sebastian Gehrmann, and Mark Dredze. 2024a. Do llms plan like human writers? comparing journalist coverage of press releases with llms. In *Conference on Empirical Methods in Natural Language Processing*.

Alexander Spangher, Xiang Ren, Jonathan May, and Nanyun Peng. 2022b. Newsedits: A news article revision dataset and a novel document-level reasoning challenge. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 127–157.

Alexander Spangher, Serdar Tumgoren, Ben Welsh, Nanyun Peng, Emilio Ferrara, and Jonathan May. 2024b. Tracking the newsworthiness of public documents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14150–14168, Bangkok, Thailand. Association for Computational Linguistics.

Alexander Spangher, James Youn, Matt DeButts, Nanyun Peng, Emilio Ferrara, and Jonathan May. b. Explaining mixtures of sources in news articles. In *The 2024 Conference on Empirical Methods in Natural Language Processing*.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, et al. 2025. Paperbench: Evaluating ai's ability to replicate ai research. *arXiv preprint arXiv:2504.01848*.

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

Yufei Tian and Nanyun Peng. 2022. Zero-shot sonnet generation with discourse-level planning and aesthetics features. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3587–3597.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L. Griffiths, and Faeze Brahman. 2024b. Macgyver: Are large language models creative problem solvers? In *Proceedings of NAACL*.

Zihao Wang, Shaofei Cai, Anji Liu, Xiaojian Ma, and Yitao Liang. 2023. Jarvis-1: Open-world multi-task agents with memory-augmented multimodal language models. In *Second Agent Learning in Open-Endedness Workshop*.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. 2024. Agent workflow memory. *arXiv preprint arXiv:2409.07429*.

Benjamin Welsh, Arda Kaz, Michael Vu, Naitian Zhou, and Alexander Spangher. Newshomepages: Homepage layouts capture information prioritization decisions.

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. 2025. Collabllm: From passive responders to active collaborators. *Preprint*, arXiv:2502.00640.

Te-Lin Wu, Alex Spangher, Pegah Alipoormolabashi, Marjorie Freedman, Ralph Weischedel, and Nanyun Peng. 2022. Understanding multimodal procedural knowledge by sequencing multimodal instructional manuals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4525–4542.

Te-Lin Wu, Caiqi Zhang, Qingyuan Hu, Alexander Spangher, and Nanyun Peng. 2023. Learning action conditions from instructional manuals for instruction understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3023–3043.

Markus Wulfmeier, Michael Bloesch, Nino Vieillard, Arun Ahuja, Jorg Bornschein, Sandy Huang, Artem Sokolov, Matt Barnes, Guillaume Desjardins, Alex Bewley, et al. 2024. Imitating language via scalable inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 37:90714–90735.

Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. 2024a. Travelplanner: A benchmark for real-world planning with language agents. In *Forty-first International Conference on Machine Learning*.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. 2024b. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Preprint*, arXiv:2404.07972.

Danfei Xu, Roberto Martín-Martín, De-An Huang, Yuke Zhu, Silvio Savarese, and Li F Fei-Fei. 2019. Regression planning networks. *Advances in neural information processing systems*, 32.

Diyi Yang, Sherry Tongshuang Wu, and Marti A. Hearst. 2024. Human-AI interaction in the age of LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 34–38, Mexico City, Mexico. Association for Computational Linguistics.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.

Xiao Yu, Maximillian Chen, and Zhou Yu. 2023. Prompt-based Monte-Carlo tree search for goal-oriented dialogue policy planning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7101–7125, Singapore. Association for Computational Linguistics.

Eric Zelikman, Yuhuai Wu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning in language models. *arXiv preprint arXiv:2203.14465*.

Kexun Zhang, Weiran Yao, Zuxin Liu, Yihao Feng, Zhiwei Liu, Rithesh Murthy, Tian Lan, Lei Li, Renze Lou, Jiacheng Xu, et al. 2024. Diversity empowers intelligence: Integrating expertise of software engineering agents. *arXiv preprint arXiv:2408.07060*.

Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, et al. 2023. Recommender systems in the era of large language models (llms). *arXiv preprint arXiv:2307.02046*.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, et al. 2023. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.