

Schema-Guided Culture-Aware Complex Event Simulation with Multi-Agent Role-Play

Sha Li¹, Revanth Gangi Reddy¹, Khanh Duy Nguyen¹, Qingyun Wang¹, May Fung¹,
Chi Han¹, Jiawei Han¹, Kartik Natarajan², Clare R. Voss³, Heng Ji¹

¹University of Illinois Urbana-Champaign

²The Private Sector Humanitarian Alliance ³DEVCOM Army Research Laboratory
{shal2, jih}@illinois.edu

Abstract

Complex news events, such as natural disasters and socio-political conflicts, require swift responses from the government and society. Relying on historical events to project the future is insufficient as such events are sparse and do not cover all possible conditions and nuanced situations. Simulation of these complex events can help better prepare and reduce the negative impact. We develop a controllable complex news event simulator¹ guided by both the event schema representing domain knowledge about the scenario and user-provided assumptions representing case-specific conditions. As event dynamics depend on the fine-grained social and cultural context, we further introduce a geo-diverse commonsense and cultural norm-aware knowledge enhancement component. To enhance the coherence of the simulation, apart from the global timeline of events, we take an agent-based approach to simulate the individual character states, plans, and actions. By incorporating the schema and cultural norms, our generated simulations achieve much higher coherence and appropriateness and are received favorably by participants from a humanitarian assistance organization.

1 Introduction

History repeats itself, sometimes in a bad way, underscoring the importance of recognizing patterns and taking proactive measures to mitigate or ideally eliminate potential natural or man-made disasters. The necessity of this approach is evident in the context of emerging crises such as the COVID-19 pandemic and the Ukraine Crisis. Addressing these situations effectively demands a comprehensive, time-sensitive understanding to inform appropriate decision-making and prompt responses (Reddy et al., 2024). These pressing situations highlight the need for advanced tools capable of scenario simulation to provide predictive insights and facilitate

preemptive planning, thereby enhancing preparedness and response strategies.

In developing such a simulator, we define several desiderata: (1) the simulator should be **controllable**, allowing the user to manage and set the conditions under which the simulation will occur; (2) it must be **knowledgeable**, meaning it should adhere to and incorporate domain-specific knowledge relevant to the scenario being simulated; (3) the simulator should be **realistic**, ensuring that each event within the simulation is believable and aligns with commonsense principles; (4) the generated events must be **coherent**, avoiding any internal conflicts or contradictions; (5) the simulator should exhibit **sociocultural awareness**, being sensitive to and accurately reflecting diverse geographical contexts and societal norms.

In this context, we introduce MIRIAM, a novel news event simulator designed to function as an intelligent prophetess. By leveraging “What-if” conditions and assumptions provided by domain experts regarding disaster scenarios, MIRIAM generates a complex event simulation that describes future events with character-centric narratives, while catering to the geo-cultural diversity inherent in the scenario assumptions. Effectively, our event simulator system that has the following characteristics:

- User-defined assumptions that can steer the direction of the simulation.
- Event schemas as input that can be used to constrain the global structure and inject domain knowledge.
- Entity-level agent-based simulation which promotes coherence over long simulations.
- Norm-aware knowledge enhancement for more culturally appropriate simulations.

Figure 1 shows an overview of MIRIAM, our proposed system for complex event simulation. By effectively simulating disaster scenarios in both event graph and natural language formats, MIRIAM aims

¹Demo: <https://duynguyen2001.github.io/newssimulator/>

to assist humanitarian workers and policymakers in conducting reality checks, ultimately aiding in the prevention and management of future disasters.

2 Related Work

Language Model Agents: Language models are adept at “roleplaying”: given the description of a character, the language model can produce responses in character. Notably, this ability can be used to enable multi-agent collaboration on tasks such as solving logical puzzles (Wang et al., 2024c), writing complex software (Hong et al., 2024; Wang et al., 2024b), reviewing papers (Zeng et al., 2024), proposing hypothesis (Qi et al., 2023; Wang et al., 2024a), machine translation (Bi et al., 2019), question answering (Puerto et al., 2023), causality explanation generation (He et al., 2023), and radiology report summarization (Karn et al., 2022). Another line of work is using LMs to create social simulations (Suo et al., 2021; Park et al., 2023; Sun et al., 2023), either to improve LM alignment (Liu et al., 2024) or to create synthetic user data for user studies (Aher et al., 2023). However, previous papers concentrate on the feasibility of LM-based social simulation and their alignment with social behaviors. In comparison, we explore using LM agents to assist scenario simulation and story generation. Moreover, unlike existing approaches (Qiu et al., 2022; Miceli Barone et al., 2023) relying on dialogue to simulate social interactions, our framework generates a comprehensive scenario story that encompasses interactions among various agents, the environment, and the scenario itself. Yang et al. (2023) conducts a multi-agent simulation to explore residents’ consumption behavior under various government regulations. Our work is also the first to leverage scenario-specific event schemas induced from historical events and culture-specific norms.

Neural Story Generation: Due to the complex nature of story generation, controllable story generation has been proposed to address the causality of story events. Existing story generation mainly focuses on two aspects (Goldfarb-Tarrant et al., 2020): story planning and character modeling. Previous improvements for story planning can be divided into several categories: keywords planning (Xu et al., 2020; Kong et al., 2021), coarse-to-fine planning (Fan et al., 2019; Yao et al., 2019), commonsense reasoner (Wang et al., 2022; Peng et al., 2022a,b), event graphs (Zhai et al., 2020; Chen et al., 2021; Lu et al., 2023), and interper-

sonal relationships (Vijjini et al., 2022). In contrast, we generate stories in a two-level way, conditioned on event schemas, user-provided assumptions, and commonsense norms. Our work also relates to character modeling in story generation (Liu et al., 2020; Zhang et al., 2022). However, instead of generating character descriptions based on existing stories (Brahman et al., 2021), we generate character profiles dynamically based on existing events and event schemas. Furthermore, we assign each character as a language agent to simulate his/her interactions with the scenario.

3 MIRIAM: Complex Event Simulator

3.1 Overview

Our event simulator takes as input a set of **assumptions** and an **event schema**. Assumptions, provided as free text, can be scenario-specific, such as the infection rate for disease outbreaks, or scenario-agnostic, such as the (source) location of the event. An event schema is a graph representation of the typical events that occur in a scenario. The nodes are atomic events and edges may include temporal edges, hierarchical edges, and logical gates (AND, OR, XOR). The event schema typically encodes prior knowledge about the event scenario (restricting the simulation to parts relevant to the use case). Figure 1 shows an example of the scenario assumptions and corresponding event schema provided as input to MIRIAM.

For the output, the system provides the generated simulation in the form of an **event log** and an **overview document**. The event log is a list of event records and profiles of the characters involved in the events. When the event can be grounded to the schema, it has an event type and arguments according to the event ontology. The overview document is derived from the event log and is a more concise free-text version of the simulation.

3.2 System Design: Bi-level Simulation

Our simulator contains two levels: the global level and the character level. We will first introduce the two different types of controllers before providing more details (in §3.3 and §3.4) for the lifecycle of how an event is generated. The global level is defined by the `Global Controller` object, which takes the event schema and user assumptions as input. We leverage the open-domain schema library induced from our state-of-the-art event schema induction techniques (Li et al., 2023)

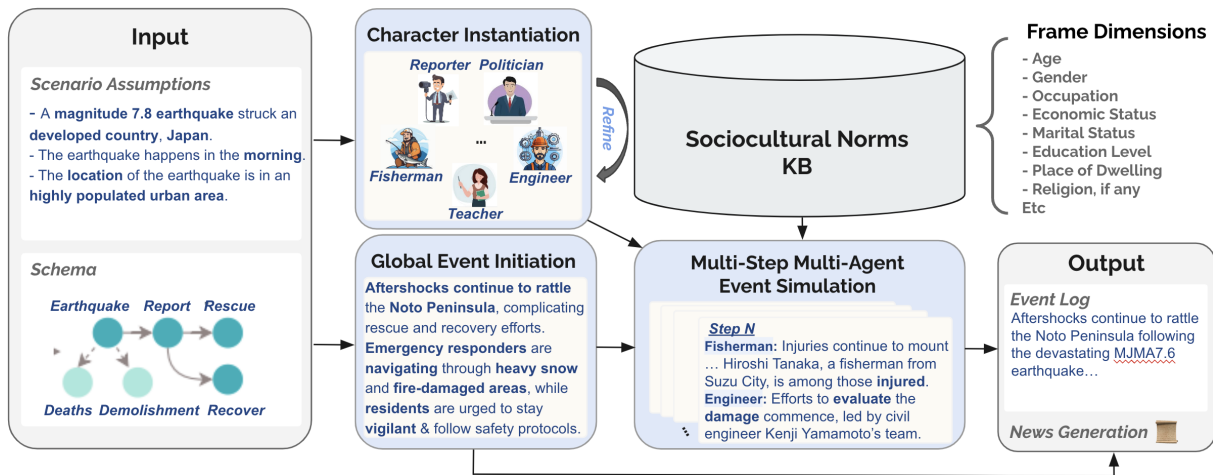


Figure 1: A Simplified Overview of our proposed MIRIAM System for Complex Event Simulation.

which covers 41 newsworthy scenarios. The global controller maintains pointers to the active characters (their `Character Controller` objects), entities that have appeared in the simulation, the event history, an event queue, and a message queue. Figure 2 shows the bi-level simulation framework with global and character-level controllers.

The event queue is filled by events from the schema and character controllers. For each time step, once the event queue is filled, the global controller will start to execute the events in temporal order and add the simulated result to the event history. Message passing in our simulation is implemented by the message queue with the global controller routing the message to the recipient. The character controllers are more simple in design as they only take care of a single agent. Each controller has its profile and history and is prompted to make plans based on the limited information it acquires. Initially, there are no character controllers and the characters are generated on-the-fly during the simulation.

3.3 Simulating Events

Events go through the cycle of (1) (optionally) event assignment, (2) event planning, (3) event execution, and (4) event reaction. There are two ways of initiating events, either proposed by the schema or by characters. Events proposed by the schema might undergo the optional event assignment stage, where the `Schema Event` is assigned to an existing character or creates a new character. This decision is presented to the language model as a multiple-choice question, given the context of the previous simulated events. Note that some events do not involve any character (such as the mutation

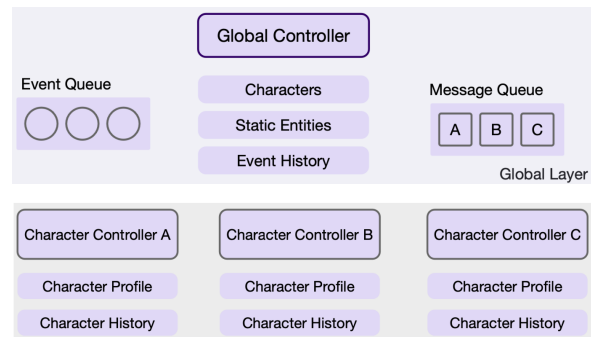


Figure 2: Figure depicting the global and character-level controllers during our simulation generation.

of a virus strain) and are directly handled by the global controller. In the event planning stage, given the candidate events from the schema, the character controller (or global controller) generates a list of planned events. Each planned event is accompanied by a timestamp that falls between the current time and the next time step of the simulation. This timestamp determines the initial execution order of events but may be affected by event reactions. In this step, character controllers also have the liberty of including events not present in the schema. These events will be represented by short text descriptions instead of event types. Finally, after the planning is complete, each event will be represented as a triple (timestamp, event description, controller name). Figure 3 illustrates the generated planned events from three different controllers.

In the event execution stage, the planned events will be sorted by their timestamp and executed in order. Executing an event involves filling in the arguments (including person, location, instrument etc.), and generating a detailed description of the event. Once executed, the event will be added to

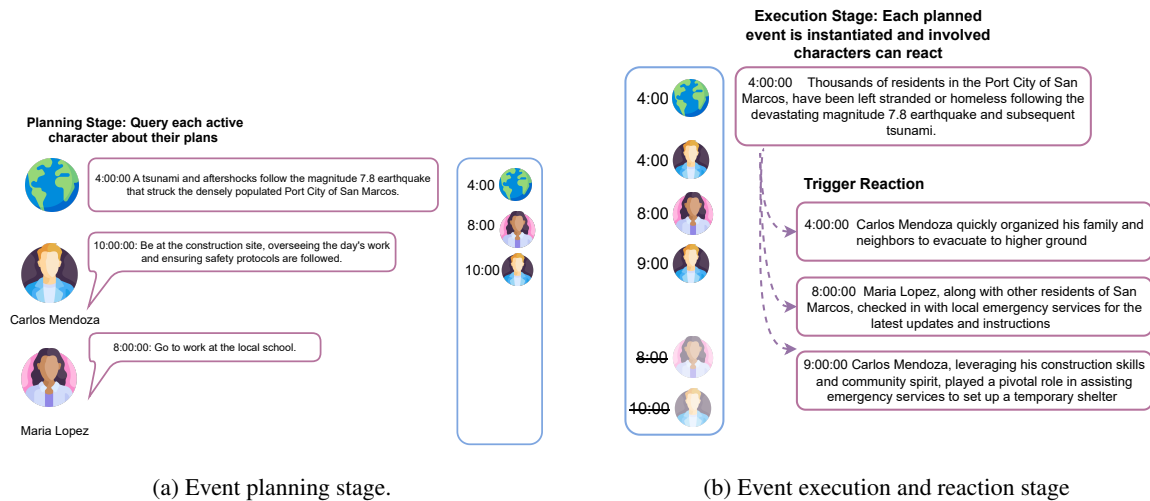


Figure 3: In the event lifecycle, the global controller and each one of the characters plans its own events for the next time step. Then all of the plans are centralized and executed in temporal order. If the executed event involves other characters, the other character will be informed and replan its actions.

the character history and the global event history.

Events that are executed earlier might affect later events. This is handled through reactions (of characters to events). Concretely, an event proposed by character A but also involves character B will trigger a reaction from character B. Character B can then make alternative plans and change the event queue. For example, A could be a doctor who performs a medical test on a patient B. If the test result is positive, patient B might cancel the remaining plans for the day and become hospitalized. In Figure 3 we see that the two characters Carlos and Maria originally made work plans for their day, but after the execution of the earthquake event, Carlos and Maria react by evacuating and assisting emergency services.

Cultural Enhancement Additionally, event simulation should be dependent on the geodiverse sociocultural situation in order to cater to globally interconnected audience. For example, a simulation of an earthquake scenario in Western communities valuing individualism may showcase parents prioritizing their children’s safety over their daily professional activities. In contrast, a simulation of an earthquake scenario in China, reflecting communities that generally value collectivism more greatly, may showcase parents first committing to societal rescue efforts before checking on the safety of their own children. To address this, Miriam integrates sociocultural knowledge across the event simulation pipeline to enrich the realisticness and insightfulness of the event story generation, as well

as to ensure that the simulated responses are culturally appropriate.

- **Character Profile Initialization:** When simulating event scenarios, the fine-grained background information (e.g., age, gender, occupation, marriage/family status, economic status, education level, ethnicity, religious beliefs, etc.) of each simulated individual really matters, but an LLM may often miss important social profile dimensions while generating the initial character descriptions. We leverage the social theory grounded formulation in (Ziems et al., 2023) and ask LLM to enhance the initial character profile descriptions for any important missing social profile dimensions.
- **Per-Character Event Description:** To better tailor event descriptions towards the cultural norms of a particular society being simulated, we leverage the concept of norm discovery on-the-fly (Fung et al., 2023). Specifically, we discover relevant social norms through LLM self-retrieval augmented generation grounded on the concept of internal knowledge elicitation, and further supplement the norms with the set of pre-existing norms from (Fung et al., 2024), which covers massively multi-cultural norm for 1000+ sub-country regions and 2000+ ethnolinguistic groups (discovered through web documents via ShareGPT), to dynamically construct and enrich the NormKB relevant for the scenario context. Then, we rank the social norms by relevance and insightful to the situation context, and condition on these so-

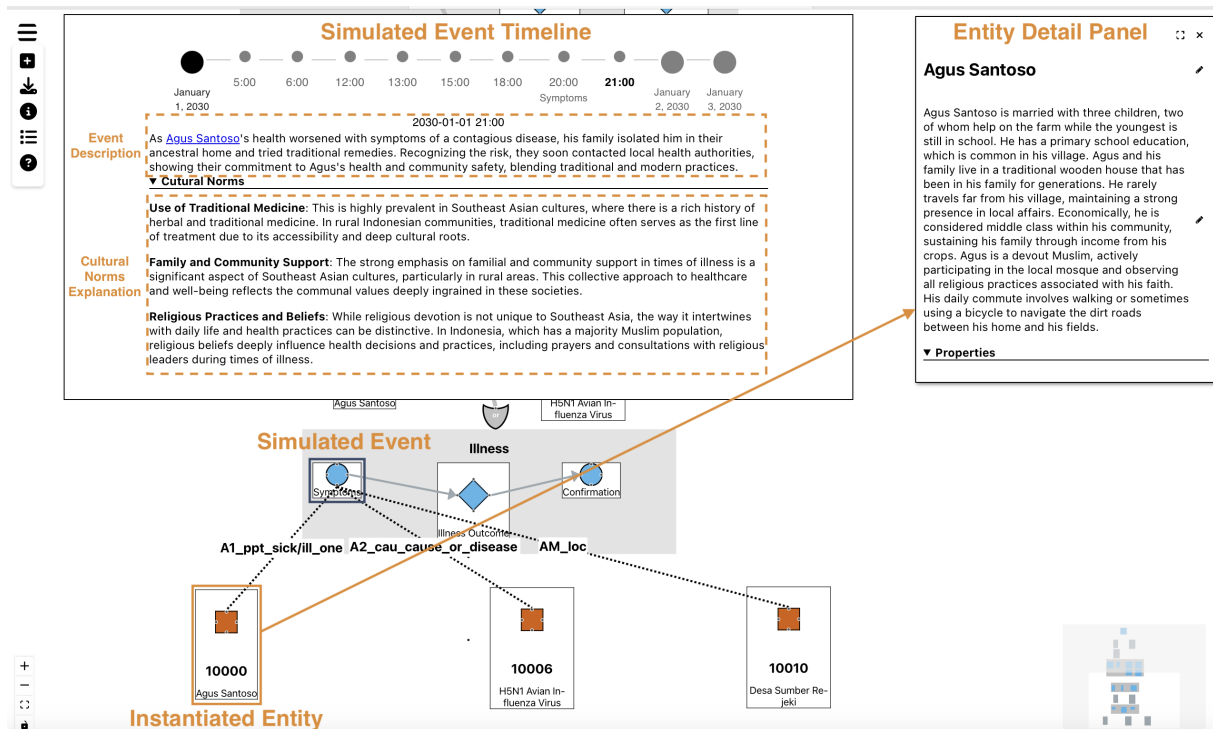


Figure 4: Screenshot of the MIRIAM interface showing an example simulation for a disease outbreak in Indonesia. The simulation is visualized in the form of an event timeline, with each event provided with a detailed description including related socio-cultural norms, along with background details of the characters involved in the event.

cial norms as additional context for refining news simulation with greater cultural detail. Specifically, we refine the event descriptive by a LLM prompting mechanism that takes as input the original event description, as well as the relevant sociocultural norms for auxiliary context, followed by the task instruction of *"Revise the event description to be more tailored to the unique cultural norms, while keeping the overall event description a similar length"*, to derive the norm-enhanced event description.

We refer the reader to Table 2 in the appendix for an example comparing event simulation with and without cultural norm enhancement.

3.4 Simulating Agent Behavior

We can also inspect our simulation on a character level. Each character in the system is defined by a name, age, profession, backstory, and plotline. Different from prior work, the characters in our system are created dynamically by the global controller. The attributes of the character are generated upon creation based on the global assumptions and the event that the character participates in.

At the beginning of each time step of the simulation, every active character(agent) will be polled for their upcoming planned events. Each character

will keep track of the events that he/she has been involved in. These memories will be part of the input when the agent makes up the plan.

In particular, we introduce a **self-critique loop** to the planning stage. The theory-of-mind inspires this self-critique loop: the model is required to infer what the agent will do so that the plot is fulfilled (while the agent does not know about the plot). To model this second-order relationship, we first ask the model to role-play as the character and generate a draft plan based on the character profile and character history. Then the model is instructed to behave as a critic and check if the actions agree with the plot. The critic will give detailed feedback on which actions should be kept, removed, or revised, along with the reasoning for adjustments ("you are feeling unwell today so you should not go out"). In our system, this self-critique will stop when the critic does not have any suggestions or when we reach a maximum of 3 rounds.

Since the efficiency of the simulation is heavily influenced by the number of active agents, we set a threshold for the maximum number of characters active at each time step. If the current number of characters exceeds that threshold, we retire the least recently used character.

4 Experiments

Our experiments aim to investigate the impact of various components integrated into our system, alongside assessing the overall utility of the tool. First, §4.1 outlines the automatic evaluations to determine the benefit of leveraging the event schemas and cultural norms in simulation generation. Then, §4.2 studies the perceived utility of our tool, based on feedback from participants affiliated with a humanitarian assistance organization. The GPT-4o MINI model serves as the underlying LLM in the simulation generation process.

4.1 Automatic Evaluation

To demonstrate the benefit of incorporating the event schemas and cultural norms into our system, Table 1 presents a comparative analysis of simulations generated by different variants of our system. Our approach, designated as *Schema + Norms*, is evaluated against (a) *Schema Only*, which does not utilize cultural norms, and (b) *W/O Schema*, which employs the LLM directly to generate simulations without schema guidance. The evaluation criteria include a range of metrics: (i) *coherence*, assessing the overall flow of the simulation, (ii) *entailment*, determining whether the simulation aligns with given assumptions, (iii) *realism*, evaluating the plausibility of the simulation in the given scenario, and (iv) *cultural appropriateness*. We employed GPT-4o for the automatic evaluation of simulation quality, with detailed prompts provided in Table 3 in the Appendix. The evaluation covered 47 simulations generated for scenarios including 'Earthquake,' 'Disease Outbreak' and 'Chemical Spill,' across five distinct regions (cultures): the United States, France, China, Peru, and Indonesia. The results demonstrate that incorporating both cultural norms and event schemas significantly enhances the quality of the generated simulations across all metrics, with notable improvements in cultural appropriateness and entailment with assumptions.

4.2 Human Utility Evaluation

We conducted a human evaluation to assess the perceived utility of the tool. The study involved five participants from a humanitarian assistance organization who navigated the generated simulations using the interface depicted in Figure 4. Participants provided qualitative feedback during the study and completed a post-study questionnaire for quantitative evaluation. The results, presented in Figure 5,

Metric	Schema + Norms	Schema Only	W/O Schema
Coherent	7.49	6.94	6.57
Entailment	8.36	8.11	7.23
Realistic	7.61	7.09	6.79
Appropriate	8.57	7.02	6.60

Table 1: Automatic evaluation (rated by GPT-4o) of simulations generated by different variants of our system.

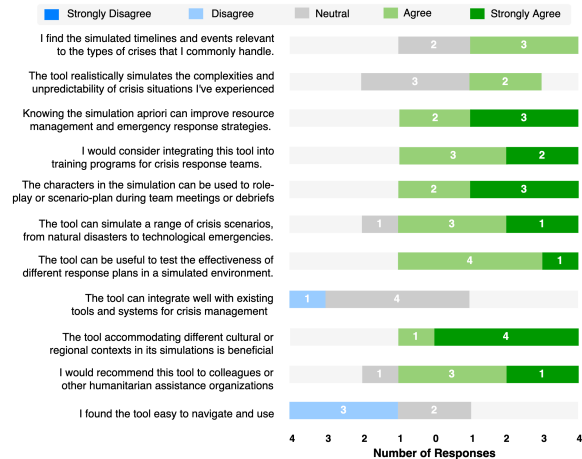


Figure 5: Results from utility evaluation by participants from a humanitarian assistance organization.

indicate that participants found the system promising and useful for training crisis response teams. However, feedback highlighted significant areas for improvement in the interface, suggesting that the current version may be limiting when integrating the system into existing workflows. We plan to address these issues in future iterations based on the qualitative feedback received.

5 Conclusion

We introduce MIRIAM, a controllable complex news event simulator designed to improve preparation and response to events like natural disasters and socio-political conflicts. Using event schemas for domain knowledge and incorporating user assumptions, Miriam offers global control over event dynamics. It enhances realism by integrating geodiverse commonsense and cultural norm awareness. The system generates a coherent global timeline and employs a large language model to simulate the states, plans, and actions of individual agents, enabling detailed and realistic character-based stories. This agent-based approach outperforms traditional schema-only methods, providing a valuable tool for training, preparedness, and societal resilience.

Acknowledgement

This research is based upon work supported by DARPA KAIROS Program No. 18 FA8750-19-2-1004, DARPA SemaFor Program No. HR001120C0123, DARPA CCU Program No. HR001122C0034, DARPA ITM Program No. FA8650-23-C-7316 and DARPA INCAS Program No. HR001121C0165. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

- Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. [Using large language models to simulate multiple humans and replicate human subject studies](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR.
- Tianchi Bi, Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. [Multi-agent learning for neural machine translation](#). 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 856–865, Hong Kong, China. Association for Computational Linguistics.
- Faeze Brahman, Meng Huang, Oyvind Tafjord, Chao Zhao, Mrinmaya Sachan, and Snigdha Chaturvedi. 2021. [“let your characters tell their story”: A dataset for character-centric narrative understanding](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1734–1752, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Hong Chen, Raphael Shu, Hiroya Takamura, and Hideki Nakayama. 2021. [GraphPlan: Story generation by planning with event graph](#). In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 377–386, Aberdeen, Scotland, UK. Association for Computational Linguistics.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2019. [Strategies for structuring story generation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2660, Florence, Italy. Association for Computational Linguistics.
- Yi Fung, Tuhin Chakrabarty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. 2023. [NORM-SAGE: Multi-lingual multi-cultural norm discovery from conversations on-the-fly](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15217–15230, Singapore. Association for Computational Linguistics.
- Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. [Massively multi-cultural knowledge acquisition & lm benchmarking](#). *arXiv preprint arXiv:2402.09369*.
- Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. [Content planning for neural story generation with aristotelian rescoring](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338, Online. Association for Computational Linguistics.
- Zhitao He, Pengfei Cao, Yubo Chen, Kang Liu, Ruopeng Li, Mengshu Sun, and Jun Zhao. 2023. [LEGO: A multi-agent collaborative framework with role-playing and iterative feedback for causality explanation generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9142–9163, Singapore. Association for Computational Linguistics.
- Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiwu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. [MetaGPT: Meta programming for a multi-agent collaborative framework](#).
- Sanjeev Kumar Karn, Ning Liu, Hinrich Schuetze, and Oladimeji Farri. 2022. [Differentiable multi-agent actor-critic for multi-step radiology report summarization](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1542–1553, Dublin, Ireland. Association for Computational Linguistics.
- Xiangzhe Kong, Jialiang Huang, Ziquan Tung, Jian Guan, and Minlie Huang. 2021. [Stylized story generation with style-guided planning](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2430–2436, Online. Association for Computational Linguistics.
- Sha Li, Ruining Zhao, Manling Li, Heng Ji, Chris Callison-Burch, and Jiawei Han. 2023. [Open-domain hierarchical event schema induction by incremental prompting and verification](#). In *Proceedings of The 61st Annual Meeting of the Association for Computational Linguistics (ACL2023)*.
- Danyang Liu, Juntao Li, Meng-Hsuan Yu, Ziming Huang, Gongshen Liu, Dongyan Zhao, and Rui Yan. 2020. [A character-centric neural model for automated story generation](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1725–1732.
- Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Diyi Yang, and Soroush Vosoughi. 2024. [Training socially](#)

- aligned language models on simulated social interactions. In *Proceedings of the Twelfth International Conference on Learning Representations*.
- Zhicong Lu, Li Jin, Guangluan Xu, Linmei Hu, Nayu Liu, Xiaoyu Li, Xian Sun, Zequn Zhang, and Kaiwen Wei. 2023. Narrative order aware story generation via bidirectional pretraining model with optimal transport reward. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6274–6287, Singapore. Association for Computational Linguistics.
- Antonio Valerio Miceli Barone, Craig Innes, and Alex Lascarides. 2023. Dialogue-based generation of self-driving simulation scenarios using large language models. In *Proceedings of the 3rd Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics (SpLU-RoboNLP 2023)*, pages 1–12, Singapore. Association for Computational Linguistics.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST ’23*, New York, NY, USA. Association for Computing Machinery.
- Xiangyu Peng, Siyan Li, Sarah Wiegrefe, and Mark Riedl. 2022a. Inferring the reader: Guiding automated story generation with commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7008–7029, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Xiangyu Peng, Kaige Xie, Amal Alabdulkarim, Harshith Kayam, Samihan Dani, and Mark Riedl. 2022b. Guiding neural story generation with reader models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7087–7111, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Haritz Puerto, Gözde Şahin, and Iryna Gurevych. 2023. MetaQA: Combining expert agents for multi-skill question answering. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3566–3580, Dubrovnik, Croatia. Association for Computational Linguistics.
- Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. 2023. Large language models are zero shot hypothesis proposers.
- Liang Qiu, Yizhou Zhao, Yuan Liang, Pan Lu, Weiyan Shi, Zhou Yu, and Song-Chun Zhu. 2022. Towards socially intelligent agents with mental state transition and human value. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 146–158, Edinburgh, UK. Association for Computational Linguistics.
- Revanth Gangi Reddy, Daniel Lee, Yi R. Fung, Khanh Duy Nguyen, Qi Zeng, Manling Li, Ziqi Wang, Clare Voss, and Heng Ji. 2024. Smartbook: Ai-assisted situation report generation for intelligence analysts. *Computation and Language Repository*, arXiv:2303.14337.
- Chenkai Sun, Jinning Li, Yi Fung, Hou Chan, Tarek Abdelzaher, ChengXiang Zhai, and Heng Ji. 2023. Decoding the silent majority: Inducing belief augmented social graph with large language model for response forecasting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 43–57, Singapore. Association for Computational Linguistics.
- Simon Suo, Sebastian Regalado, Sergio Casas, and Raquel Urtasun. 2021. Trafficsim: Learning to simulate realistic multi-agent behaviors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10400–10409.
- Anvesh Rao Vijjini, Faeze Brahman, and Snigdha Chaturvedi. 2022. Towards inter-character relationship-driven story generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8970–8987, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- PeiFeng Wang, Jonathan Zamora, Junfeng Liu, Filip Ilievski, Muhao Chen, and Xiang Ren. 2022. Contextualized scene imagination for generative commonsense reasoning. In *International Conference on Learning Representations*.
- Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024a. Scimon: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL2024)*.
- Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. 2024b. Executable code actions elicit better llm agents. In *Proceedings of the Forty-first International Conference on Machine Learning (ICML2024)*.
- Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024c. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. 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 257–279, Mexico City, Mexico. Association for Computational Linguistics.
- Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models. In *Proceedings of the*

2020 *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2831–2845, Online. Association for Computational Linguistics.

Menghua Yang, Hong Chen, Ruyin Long, and Jiahui Yang. 2023. [How does government regulation shape residents' green consumption behavior? a multi-agent simulation considering environmental values and social interaction.](#) *Journal of Environmental Management*, 331:117231.

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 Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'19/IAAI'19/EAAI'19*. AAAI Press.

Qi Zeng, Mankeerat Sidhu, Hou Pong Chan, Lu Wang, and Heng Ji. 2024. [Scientific opinion summarization: Paper meta-review generation dataset, methods, and evaluation.](#) In *1st AI4Research Workshop*.

Fangzhou Zhai, Vera Demberg, and Alexander Koller. 2020. [Story generation with rich details.](#) In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2346–2351, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhexin Zhang, Jiaxin Wen, Jian Guan, and Minlie Huang. 2022. [Persona-guided planning for controlling the protagonist's persona in story generation.](#) In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3346–3361, Seattle, United States. Association for Computational Linguistics.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. [NormBank: A knowledge bank of situational social norms.](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7756–7776, Toronto, Canada. Association for Computational Linguistics.

	w/o Norms	w/ Norms
character profile	Li Wei is a 32-year-old marketing executive living in Shanghai. He is known for his outgoing personality and enjoys attending social events and networking gatherings. Li Wei is health-conscious but often finds himself in crowded places due to his job.	Li Wei is a 32-year-old marketing executive living in Shanghai. He is known for his outgoing personality and enjoys attending social events and networking gatherings. Li Wei is health-conscious but often finds himself in crowded places due to his job. <i>He is single and lives in a modern high-rise apartment in the bustling Jing'an District of Shanghai. He holds a MBA from Fudan University. Originally from a smaller city in Jiangsu province, he moved to Shanghai ten years ago to pursue his career. Economically, Li Wei is well-off, earning a comfortable salary that allows him to indulge in his interests and maintain a cosmopolitan lifestyle. He usually commutes to work primarily by metro, which is efficient and fits his environmentally conscious values. Although culturally rooted in Confucian values, Li Wei is not particularly religious, focusing more on personal and professional growth.</i>
per-character event plan	Hispanic Single-Mother Amidst Covid Outbreak in the US: wake up early to get prepared for morning shift as part-time shopping mall cashier. Check news and see gov't announces pandemic lockdown, which causes her shift to be cancelled. Prepares healthy breakfast for daughter and helps her prepare for remote class.	Hispanic Single-Mother Amidst Covid Outbreak in the US: wake up early to get prepared for morning shift as part-time shopping mall cashier. Check news and see gov't announces pandemic lockdown, doesn't have a job now and searches for gov't subsidy options. Prepares healthy omelette breakfast for daughter and helps her prepare for remote class over zoom .
per-character event description	Unwind at Home: Despite the ongoing outbreak in Jakarta, Andi Pratama decided to go for a morning jog in the park, taking extra precautions to avoid crowded areas and maintain personal hygiene.	Unwind at Home: During a disease outbreak in Jakarta, Andi Pratama, a devout Muslim, performed the Tahajjud prayer at night in his apartment. As the new year began, he prayed earnestly for his community's well-being. In the morning, after performing Fajr prayer at home, Andi Pratama jogged in a nearby park, embracing the " gotong royong " spirit by carefully avoiding crowded areas and keeping distance from others.

Table 2: Comparison of event simulations with and w/o knowledge enhancement from culture-specific social norms.

<u>Evaluation Prompt</u>
<pre> You are an automatic quality evaluator. You will be provided with some simulations and you will need to evaluate them based on the criteria that is mentioned. -- You are provided with some simulations corresponding to the scenario: {scenario_name} -- The simulations were generated based on the following assumptions in no specific order: -- Assumptions: {list_of_assumptions} -- The simulations are below. Each simulation is from the future in the form of a listwise log of events. Each log item has the time and a description of the event. -- Simulation 1: {list_of_events} -- Simulation 2: {list_of_events} -- Simulation 3: {list_of_events} -- Metric: For each of the simulations, you need to evaluate how coherent the simulation is and provide a single score in the range of 1-10, where a higher score indicates better coherence. -- DO NOT bias your judgment based on the length of the simulation. You should only respond in the JSON format as described below. You SHOULD ensure that the provided output can be directly parsed into json using python json.loads -- Response Format: {{ "thoughts": "Your step-by-step reasoning for the evaluation scores you will provide", "simulation_1": "Score for simulation 1. Just provide a number here in the range of 1 to 10", "simulation_2": "Score for simulation 2. Just provide a number here in the range of 1 to 10", "simulation_3": "Score for simulation 3. Just provide a number here in the range of 1 to 10", }} </pre>

Table 3: Prompts for automatic evaluation of the simulations.