SocialForge: simulating the social internet to provide realistic training against influence operations

Ulysse Oliveri[1,2], **Guillaume Gadek**[2], **Alexandre Dey**[3], **Benjamin Costé**[3], **Damien Lolive**[4], **Arnaud Delhay-Lorrain**[1], **Bruno Grilheres**[2]

[1] Univ Rennes, CNRS, IRISA, Lannion, France
[2] Airbus Defence and Space, Elancourt, France
[3] Airbus Defence and Space Cyber Programs, Rennes, France
[4] Université Bretagne Sud, CNRS, IRISA, Vannes, France

Abstract

Social media platforms have enabled large-scale influence campaigns, impacting democratic processes. To fight against these threats, continuous training is needed. A typical training session is based on a fictive scenario describing key elements which are instantiated into a dedicated platform. Such a platform simulates social networks, which host a huge amount of content aligned with the training scenario. However, directly using Large Language Models to create appropriate content results in low content diversity due to coarse-grained and high-level scenario constraints, which compromises the trainees' immersion.

We address this issue with **SocialForge**, a system designed to enhance the diversity and realism of the generated content while ensuring its adherence to the original scenario. Specifically, SocialForge refines and augments the initial scenario constraints by generating detailed subnarratives, personas, and events.

We assess diversity, realism, and adherence to the scenario through custom evaluation protocol. We also propose an automatic method to detect erroneous constraint generation, ensuring optimal alignment of the content with the scenario.

SocialForge has been used in real trainings and in several showcases, with great end-user satisfaction. We release an open-source dataset[1] generated with SocialForge for the research community.

1 Introduction

Social media platforms have enabled large-scale influence campaigns, allowing actors to manipulate elections and impact health protocols (Muhammed T and Mathew, 2022). Influence campaigns are organized over time in various influence operations that share the same goal. These operations imply coordination between actors, aiming at manipulating populations to widen opinion gaps.

To counter these operations, entities such as journalists (e.g., fact-checking service), marketing services, and government agencies such as Viginum[2] in France or Rapid Response Mechanism[3] in Canada are actively developing countermeasures. In this evolving threat landscape, continuous exercise is crucial for these actors to stay ahead and effectively combat influence campaigns, developing up-to-date methodologies to counteract manipulative strategies. A training session relies on two types of end-users; the player team (trainees) and the animation team (trainers).

The player team interacts with the content (social media posts) aiming at detecting inauthentic behaviors[4]. A successful training challenges players to distinguish between genuine and inauthentic behaviors.

Organizing these trainings, the animation team creates a scenario depicting fictional geopolitical entities, including key elements such as factions (groups of individuals that share goals, ideas), narratives (strategic ideas that factions aim to broadcast), and events (Walker, Christopher et al., 2006). The animation team instantiates the scenario within the reproduced informational sphere (e.g, social networks or press sites) with a large, realistic, and diverse amount of content. Their diffusion reproduces specific social behaviors, as defined in the scenario.

The animation team is able to dynamically add, delete, or edit constraints, updating the content to maintain engagement and challenge throughout the training. Moreover, trainers must be able to also control the quantity, diversity, and quality of the content to ensure an effective training.

In this context, the usage of Large Language Models (LLMs) is relevant to produce large quan-

[1] https://gitlab.inria.fr/expression/socialforge

[2] https://www.sgdsn.gouv.fr/notre-organisation/composantes/service-de-vigilance-et-protection-contre-les-ingerences-numeriques

[3] https://www.international.gc.ca/transparency-transparence/rapid-response-mechanism-mecanisme-reponse-rapide/index.aspx?lang=eng

[4] https://transparency.meta.com/policies/community-standards/inauthentic-behavior/

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 166–178

ties of content, taking into account the scenario constraints. Its use must however be well calibrated.

We hence introduce SocialForge, a model-agnostic and controllable data generation system. SocialForge takes as input a coarse-grained high level scenario, and automatically refines and augments it using LLMs. Doing so, SocialForge provides an intelligible knowledge base enabling constraints modifications, which are used for content generation. As a result, the system produces a realistic, diverse, and scenario-adhering text corpora to populate social network reproductions.

We summarize our contributions as follows:

1. SocialForge is a system that (1) refines user inputs to generate knowledge items used in prompts and (2) uses these prompts to generate social media content dataset. We show an increase in diversity in the main literature metrics.
2. By conducting a human evaluation of the adherence to the scenario and using LLM-as-a-Judge methods to determine the likelihood of the generation, we show that increasing diversity does not hinder other quality metrics, essential for the training unfolding.
3. We perform a human-machine (15 evaluators) comparative study with an LLM-as-a-Judge evaluation on the constraints space, which ensures the coherence of the future generated dataset with the scenario by focusing on a smaller set of constraints.

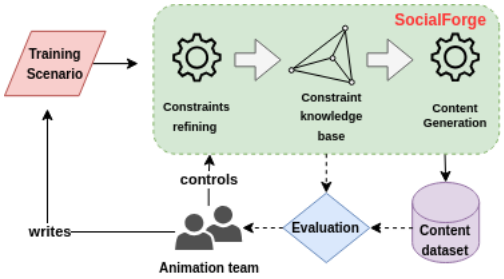

Figure 1: **SocialForge** pipeline to populate social networks reproductions

SocialForge has been used in real trainings and in several showcases, with great end-user satisfaction.

2 Related Work

2.1 Controllable text generation

Controllable text generation aims to guide the generation from a language model, satisfying an input set of constraints. These constraints belong to two distinct categories. First, soft constraints impact the semantics of the generation by changing the emotions, discussed topics or textual style (Zhang et al., 2022) of the generated content. The second category, hard constraints, applies structural constraints over the generation, by forcing the appearance of specific keywords (Joshi et al., 2023), explicit knowledge elements (Liu et al., 2022) or regulating the final length of the message (Li et al., 2022).

Diverse techniques have been developed in this field to constrain the generations. These techniques include adding control codes to prompts (Keskar et al., 2019), external classifiers (Yang and Klein, 2021) or smaller language models to guide the generation (Krause et al., 2021). However, hardware costs and generation latency increase by adding external models, which is detrimental in massive content generation, necessary to emulate social networks information flow. Recently, instruction models (Grattafiori et al., 2024; Jiang et al., 2024) have demonstrated the large language models capabilities to follow prompted input instructions, achieving state of the art over the diverse constraint categories (Ashok and Poczos, 2024). However, problems such as low diversity (Shaib et al., 2024) or hallucinations (Ji et al., 2023) still remain challenging.

2.2 Evaluation

Several criteria are crucial for evaluating the overall quality of a generated text dataset, including quality, diversity, and adherence to input constraints (van der Lee et al., 2021; Garbacea and Mei, 2022). While human evaluation is the gold standard, it is costly, making automatic methods more practical.

To assess the adherence to input constraints, methods such as BertScore (Zhang* et al., 2019) or BleuRT (Sellam et al., 2020) are widely used. These methods compare semantic similarity between generated and reference texts, although creating reference texts is time-consuming. External classifiers can also measure adherence to input constraints, but require one classifier per constraint, failing to scale (Yang and Klein, 2021). Recently, LLMs as evaluators (LLM-as-a-Judge) have shown promises on in-domain evaluations but face issues such as varying performance across languages, sycophancy (Sharma et al., 2024), and biases (Chiang and Lee, 2023).

With LLMs, scaling the number of contents may lead to a lack of diversity (Ge et al., 2024). Metrics such as SELF-BLEU (Zhu et al., 2018) and SBert (Reimers and Gurevych, 2019a), are used to evaluate lexical and semantic diversity but are computationally expensive. Distinct-n (Li et al., 2016) measures repetition rates, while compression ratios (Shaib et al., 2024) detect pattern repetitions, increased by LLM biases.

In order to ensure immersion during a training session, content must be realistic and indistinguishable from that created by the animation team. Quality metrics vary by content type; for microblogging (e.g. X, Mastodon), fluidity and grammaticality may not be objective functions to maximize (Heraldine and Handayani, 2022). Usual metrics such as Perplexity (Jelinek et al., 1977) need to be calibrated with a reference dataset. However, crafting a dataset representative of the educational goals for each training is intractable. Automating this axis of evaluation is challenging due to the need for human expertise, but LLM-as-a-Judge shows a great potential.

3 SocialForge, a social text generation system

3.1 Training context

In the context of training, two types of end-users are immersed inside the synthetic platforms. The **animation team**, in charge of the unfolding of the training session, needs control over the dynamics within the social platform such as controlling the topics, the content flow, and triggering of events. Depending on how the training unwinds, the animation team may also adapt scenario constraints. Upon these changes, the content has to reflect the newly added constraints, requiring a dynamic system of generation.

This dynamic control allows the animation team to recreate at will both genuine social behaviors and malicious behaviors such as disinformation campaigns. These recreated behaviors are to be detected during the training by the player team.

Navigating within social media platforms, the **player team** uses its methodology to discriminate between various behaviors.

In order for the player to focus on the proper methodology, the content should be realistic enough. Specifically, the player team should not be able to rely on immediate discriminative methods such as automatically detecting sentences starting with the same pattern or specific shared keywords.

3.2 Training scenario

The training scenario is a structured textual document created by the animation team. It outlines key elements to appear in the generation process:

1. **Factions**, defined as groups of individuals promoting one or more narratives.
2. **Narratives**, ideas that a faction aims to instill and broadcast to a target audience. Specifically, a narrative is defined as a topic associated with a stance (for, against, or neutral). Under this definition, two factions can discuss the same topic from different points of view, resulting in two distinct narratives. These factions and their associated narratives are central to the training and are used to implement social dynamics between user accounts on social platforms.
3. **Events** that animate the informational sphere depending on the educational progression of the training.

3.3 SocialForge: Scenario Refinement to Content Generation

As illustrated in Figure 2, SocialForge begins with the refinement of the scenario events by generating sequential occurrences of them, called sub-events, using an LLM. For instance with a scenario event talking about protests in the fictive country of Verdantia, sub-events might include confrontations with the police or damaged shops.

Next, SocialForge uses the provided narratives to prompt the LLM to generate subnarratives. Multiple subnarratives offer diverse perspectives on a specific narrative, enhancing the diversity of the corpus.

SocialForge then matches scenario events and sub-events with subnarratives through semantic similarity, allowing the events to be used in the generated content along the subnarratives.

In influence operations, attackers enhance narratives' effect by targeting an audience that is receptive to it. Additionally, specifying an audience (or coarse-grained personas) to language models increases the generated corpus diversity and its constraint adherence (Tseng et al., 2024). SocialForge leverages these principles by deriving coarse-grained personas, referred to as population segments, from input narratives. Segments are then instantiated by creating individuals (thin-grained personas), adding new criteria such as the OCEAN Score (Goldberg, 1990) to dress a psychological representation (i.e., scores on openness, conscientiousness, extraversion, agreeableness, and neuroticism) of the individual.

Finally, SocialForge generates social media platform-specific user accounts belonging to these individuals, generating a list of "normal" topics (e.g, soccer, computer science...), based on individual characteristics. With all this information, SocialForge prompts an LLM to generate a content, given an account along with their associated subnarrative and events. The resulting content is then available to animate the informational sphere.

4 Experimental setup

4.1 Scenario Construction

To evaluate the results of SocialForge, we begin by crafting a concise scenario involving six factions,

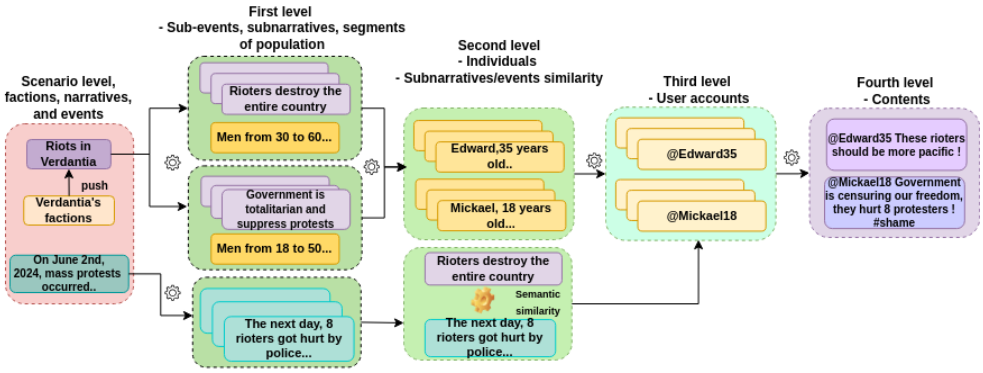

Figure 2: Example of data generation using SocialForge. First, second and third level are constraints refinement and augmentation. Last level is content generation, here written in English for illustration, but is in French in the generated dataset. We refer as Verdantia Factions the Government of Verdantia and Verdantia's rioters, described in section 4.

eight narratives, and nine events.

The scenario centers around the fictive neutral country of Verdantia, where three factions - the Government of Verdantia, Rioters, and Pro-Western intelligentsia - are engaged in a conflict, with the latter two opposing the government. Additionally, two influential blocs, The West and Louraly, fight for Verdantia's alignment. Meanwhile, the last faction, Tabiscus, welcomes Verdantian refugees fleeing the riots, while Louraly attacks them on this decision.

4.2 Model Deployment

To increase the constraints of the scenario, we used the *mixtral:8x7b* model (Jiang et al., 2024) deployed with Ollama[5] on a Nvidia RTX A6000. The advanced reasoning capabilities of the model facilitated a nuanced understanding of the scenario, enabling precise refinements and augmentations. For the content generation, we employed mistral-nemo[6]. This model is relatively light (12B parameters), facilitating scalability. It is also open-weight, which is necessary for off-internet exercises, and shows good performance in French, the target language. We operate *modernbert-embed-base* via Huggingface[7] to match events and subnarratives through semantic similarity. A detailed view of the models parameters is presented in the Appendices 5.

4.3 Constraint & Content generation

Using *mixtral:8x7b*, SocialForge generated 75 unique subnarratives spread across 15 distinct population segments, each with unique characteristics. For each of the scenario events, five sub-events were generated. These sub-events were semantically linked using *modernbert-embed-base*, with a similarity threshold set at 0.4. This process yielded a total of 47 events and sub-events, to be used in future contents. Finally, SocialForge generated 371 distinct individuals and their associated user account for subsequent message generation.

The constraints have been generated in English, as it is the most present language in the LLMs training dataset, yielding better results. Afterwards, we use multilingual models to generate in diverse languages (e.g., English constraints to French content, English to German content...).

To evaluate our method, we followed these steps:

1. Using SocialForge (as defined in Section 3), we generated five French-language datasets, each containing 2,250 (30 per subnarrative) microblogging texts.
2. To establish a **Baseline**, we prompt the LLM with scenario-level information only (i.e., scenario events, narratives, factions), generating an additional five French-language datasets of 2,250 microblogging texts.
3. For each set of five datasets (SocialForge and Baseline), we report the mean and standard deviation of the metrics.

4.4 Evaluation

Evaluation focuses on three key aspects: adherence to the scenario, diversity, and likelihood (or realism) with respect to actual platforms.

Adherence to the scenario is challenging due to the absence of reference labels in our context. To address this, we conducted a human evaluation (15 evaluators) to assess the SocialForge generations, ensuring that (1) the generated constraints are in line with the scenario and (2) the content respects the prompted constraints. This approach assesses final content are in accordance with the scenario constraints.

[5] https://ollama.com
[6] https://mistral.ai/news/mistral-nemo
[7] https://huggingface.co/nomic-ai/modernbert-embed-base

1. Evaluators rated generated segments, subnarratives, individuals, and sub-events along two axes:
   - **Coherence** with the initial constraint (e.g., subnarrative coherence with the main narrative, non-contradictory sub-event w.r.t scenario event...).
   - **Precision** of the generated constraint, if the newly created constraint adds concrete details (granularity).
2. Evaluators rated whether the constraints appear in the content (i.e., the constraints were expressed in the content) and adhered to them (i.e., the constraints were correctly expressed, addressing issues like stance). Evaluators were immersed in two setups:
   - **Micro:** Rated individual content using a binary scale across two criteria: constraints appears in the content and the content adheres to it (n = 90).
   - **Macro:** Rated batches of five pieces of content using a Likert scale from 1 to 7 (n = 18).

Next, we evaluate **diversity** across the entire corpus using automatic metrics:

- **SELF-BLEU**: assesses lexical diversity using sacrebleu's[8] pairwise BLEU-1 score.
- **Homogenization Score**: Similar as done in SBert (Reimers and Gurevych, 2019b), homogenization score presented in (Shaib et al., 2024) is a pairwise cosine similarity to measure average similarity between corpus documents. Here, we use *nomic-embed-text-v2-moe* to compute this score, leveraging its capacities in computing french embeddings.
- **Compression Ratio and Compression POS Ratio** evaluate pattern redundancy of the compressed texts and associated POS-Tags using *gzip* and *spacy*, where higher ratios indicate more redundancy. This essentially measures formulation biases in LLMs, where they tend to follow specific patterns (Shaib et al., 2024).

To compute **likelihood**, we use LLM-as-a-Judge to simulate user analysis on a social media platform. The LLM rates batches of five generated documents based on their representativeness of microblogging content. Specifically, the LLM scores each batch on a Likert scale of 1 to 7, assessing the plausibility of the generated documents. Our implementation of the LLM-as-a-Judge approach relies on Llama3.3:70b (Grattafiori et al., 2024) through Ollama API. For computing reasons, we randomly sample 500 samples for each dataset (i.e., 500 for each of the 5 datasets from Baseline and SocialForge) for a total of 2500 evaluated documents per generation method (i.e., SocialForge and Baseline). Enabling this likelihood evaluation, we compute two distinct setups, representative of real user experiences:

- **Timeline Overview**: Generated texts are drawn randomly (we do not model a recommendation system) in batches of five, similarly as a timeline view in microblogging social medias.
- **Trending Overview**: Generated texts sharing the same keywords are drawn together as batches of five, as shown in trendings overviews within microblogging platforms. Each document has its two most probable keywords extracted using yake (Campos et al., 2018) Python library.

This comprehensive evaluation ensures that all the dimensions of generation quality are taken into account and assessed with quantitative measures.

5 Results

Starting with the diversity evaluation, Table 1 demonstrates that using SocialForge increases the main diversity literature metrics. Homogenization Score indicates that generated constraint grants semantic diversity in the texts, addressing more topics and widening the semantic field. Other metrics such as SELF BLEU show that more unique ngrams are used in the texts, while compression ratios show that the increase in prompt variation results in diverse response patterns, important for the player team to not immediately detect generated content.

	SocialForge	Baseline
Homogenization Score ↓	**0.535±0.001**	0.569±0.002
SELF-BLEU ↓	**0.020±0.004**	0.025±0.011
Compression Ratio ↓	**4.016±0.028**	4.963±0.034
Compression POS Ratio ↓	**8.594±0.053**	9.212±0.022

Table 1: Mean and standard deviation over diversity metrics between SocialForge and Baseline. For indication, mean sentence length (# characters) of SocialForge is 142.16±1.04 and Baseline is 134.43±0.9.

	Constraints respect	
	Macro - Batch	**Micro - Content**
Constraints Appearance	5.59±1.03	85.56%
Constraints Adherence	5.13±1.36	68.89%

Table 2: Human Evaluation (15 evaluators) results of the constraints respect. For Macro, we report mean and standard deviation over a 1 to 7 Likert Scale. For Micro results, we aggregate through majority voting percentage of one scores over a binary scale.

Human evaluations confirmed that the generated content respects and aligns correctly with the specified input constraints (see Table 2), although

[8] https://github.com/mjpost/sacrebleu

occasional language mixing occurs. This particular issue is being addressed by state-of-the-art language models (DeepSeek-AI et al., 2025). The carried out evaluations, as illustrated in Table 3, also indicate that the generated constraints are well designed and effective, demonstrating high coherence and granularity refinement.

During experiments and demonstrations, subnarrative coherence proved crucial for content generation. Incoherence could contradict the intended message, compromising training. To address this, we again used the LLM-as-a-Judge method with *LLama3.3:70B*, which effectively distinguished problematic from adequate subnarratives, strongly correlating with 15 human evaluators ($\rho_{pearson} = 0.78, p-value < 0.001$). For this evaluation, the distribution of scores is shown in Figure 3.

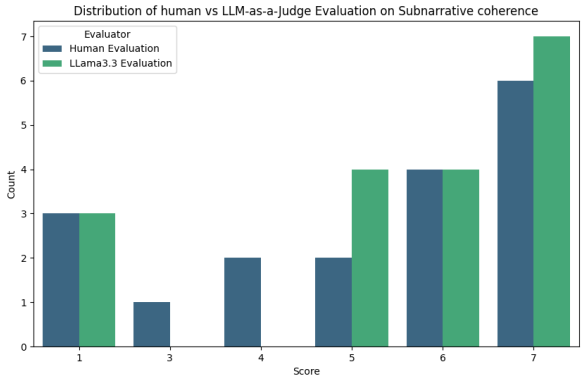

Figure 3: Comparison of Human vs LLama3.3:70B as LLM-as-a-Judge on subnarrative coherence. We see that humans are more undecided (more neutral or around neutral ratings) than LLM on this evaluation, but both detect highly incoherent generations.

This approach shows that a smaller number of samples is enough to avoid expensive metric computation, thereby enhancing the after-correction quality of the mass-scale generated content.

	Constraints quality	
	Coherence ↑	**Precision ↑**
Subnarratives (n=18)	4.66 ± 2.09	5.18 ± 1.04
Segment (n=11)	5.14 ± 0.97	N/R
Individuals (n=22)	6.41 ± 0.48	N/R
Sub-Events (n=18)	5.35 ± 1.72	5.49 ± 0.99

Table 3: Mean and standard deviation of 15 human evaluators over the coherence and precision of the generated constraints, with n the evaluated sample size. Segmentation and individual are templated generation. For these two lines, precision is Not Relevant (N/R).

For population segments and sub-events, the low sample count makes human validation tractable and even desirable to ensure a conscious control of the system by the humans. For the sub-events, it is crucial to avoid generating an excessive number of sub-events, especially for critical scenario events, to prevent overwhelming the information sphere. Individuals, being direct instantiations of population segments, are adequately generated. However, curating population segments is essential to ensure well-formed individuals for the training.

Corpus Likelihood	**SocialForge**	**Baseline**
Trending Overview↑	**4.654±1.170**	4.449 ± 1.260
Timeline Overview↑	**4.812±0.878**	4.667 ± 0.982

Table 4: LLM-as-a-Judge evaluation results over the likelihood of the microblogging using a 1 to 7 Likert Scale.

SocialForge performs better in terms of likelihood in the two distinct setups (trendings and timeline overview) as shown in Table 4 and Figure 4, achieving the purpose of having plausible microblogging contents. Increasing this score makes it harder for the player team to discriminate between machine-produced content and the animation team content.

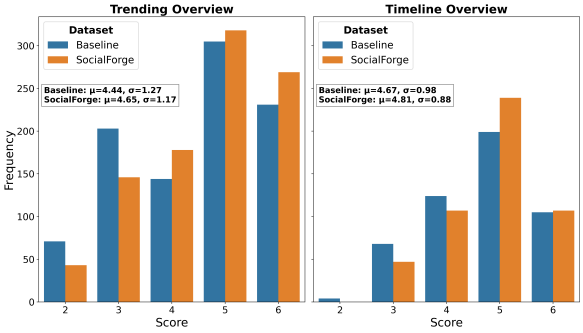

Figure 4: LLM-as-a-Judge on likelihood evaluation along the two presented setups; Trendings and Timelines, both assessed with a 1 to 7 Likert Scale. Judging model gives higher score to SocialForge evaluations. Interestingly, model did not give any 1 or 7 rating, as such, we do not make them appear in the graph.

6 Limitations

SocialForge is a novel system which generates diverse and qualitative social media data, used to train people against influence operations.

However, challenges remain: evaluating the quality of the generation is proven difficult, especially for short social media documents. Defining what is likely or unlikely to appear on real platforms remains subjective. Our LLM-as-a-Judge evaluation, without being correlated to humans, solely gives an indication of the quality, not an absolute measure.

In addition, social networks are heterogeneous: users differ in how they connect, behave, and

produce content. Previous studies have examined the *topological* diversity of interactions, relationships and community structures, analyzing *who interacts with whom* and *how often* (Gadek et al., 2017). Furthermore, diverse *social behaviors* (e.g., bots, trolls, journalists, officials, offensive accounts) shape the content produced, affecting its semantics (Chen et al., 2022). These factors are critical for modeling social networks, particularly when generating and evaluating content responses. Furthermore, this work has not yet fully explored the role of time. Real social media users operate within a broader temporal context, not only the mechanic unfolding of their current event - a well identified axis of improvement for SocialForge.

7 Conclusion

In this paper, we introduced SocialForge, a social media data generation system used to populate simulated informational spheres, which are used to train against influence operations. These trainings follow a scenario, describing high level elements that must be reflected by content within the infosphere. SocialForge refines and augments the scenario elements, producing several thinner-grained constraints, used to generate prompts which are used for generating social media content.

We propose an evaluation methodology to ensure that increasing diversity does not come at the expense of quality. We conducted a thoughtful evaluation along two criteria: scenario-adherence and likelihood. For one of the system components, the subnarrative generation, we proposed an automatic method to identify erroneous generations, ensuring the quality of the final generated content. This method was shown to be strongly correlated with human judgment, illustrating its robustness.

We assess SocialForge through a case study and we release the generated production in Gitlab [9]. Besides, SocialForge has been used in several real trainings and showcases, showing great end-user enthusiasm.

8 Ethical Considerations

The stakes are high on the topic of text generation, with numerous potential misuses. To mitigate possible negative impacts of our work, we do plan *not* to release SocialForge in an uncontrolled way.

Measures are taken to reduce the risks. All the work is hosted within an air-gap environment to mitigate content leaking danger. Within the training, all entities are fictive, to reduce biases and risks of defamation or hate. Following current regulations, all participants are aware that the content is generated by Artificial Intelligence and that the purpose of this exercise is to train against influence operations.

Unintended risks are harder to measure and detect, but we believe that studying and structuring influence operations is among the best ways to fight them. Furthermore, SocialForge is model-agnostic, which means that its environmental impact follows the state-of-the-art, and we plan to adapt accounting on this criterion. Additionally, SocialForge does not require training specific models, reducing the impact of its usage. Last but not least, we follow ethics recommendations in the domain as well as upcoming regulations to update our work to comply with effective guidelines.

[9] https://gitlab.inria.fr/expression/socialforge

References

Dhananjay Ashok and Barnabas Poczos. 2024. Controllable Text Generation in the Instruction-Tuning Era. *arXiv preprint*. Issue: arXiv:2405.01490 arXiv:2405.01490 [cs].

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018. YAKE! Collection-Independent Automatic Keyword Extractor. In Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury, editors, *Advances in Information Retrieval*, volume 10772, pages 806–810. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Long Chen, Jianguo Chen, and Chunhe Xia. 2022. Social network behavior and public opinion manipulation. *Journal of Information Security and Applications*, 64:103060.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluation?

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui

Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint*. ArXiv:2501.12948 [cs].

Guillaume Gadek, Alexandre Pauchet, Nicolas Malandain, Khaled Khelif, Laurent Vercouter, and Stéphan Brunessaux. 2017. Topical cohesion of communities on Twitter. *Procedia Computer Science*, 112:584–593.

Cristina Garbacea and Qiaozhu Mei. 2022. Why is constrained neural language generation particularly challenging? *arXiv preprint*. Issue: arXiv:2206.05395 Issue: arXiv:2206.05395 arXiv:2206.05395 [cs].

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling Synthetic Data Creation with 1,000,000,000 Personas. *arXiv preprint*. ArXiv:2406.20094 [cs].

L. R. Goldberg. 1990. An alternative "description of personality": the big-five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya

Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783 [cs].

Monica Heraldine and Nurma Dhona Handayani. 2022. An Analysis of Grammatical Errors on "Twitter". *Humanitatis : Journal of Language and Literature*, 9(1):211–218. Number: 1.

Frederick Jelinek, Robert L. Mercer, Lalit R. Bahl, and Janet M. Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *Journal of the Acoustical Society of America*, 62.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Wenliang Dai, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38. Number: 12 arXiv:2202.03629 [cs].

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts. *arXiv preprint*. ArXiv:2401.04088 [cs].

Sagar Joshi, Sumanth Balaji, Aparna Garimella, and Vasudeva Varma. 2023. Graph-based Keyword Planning for Legal Clause Generation from Topics. ArXiv: 2301.06901.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A Conditional Transformer Language Model for Controllable Generation. ArXiv: 1909.05858 Publisher: arXiv.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative Discriminator Guided Sequence Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. *arXiv preprint*. Issue: arXiv:1510.03055 arXiv:1510.03055 [cs].

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. 2022. Diffusion-LM Improves Controllable Text Generation. ArXiv: 2205.14217 Publisher: arXiv.

Jin Liu, Chongfeng Fan, Zhou Fengyu, and Huijuan Xu. 2022. Syntax Controlled Knowledge Graph-to-Text Generation with Order and Semantic Consistency. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1278–1291, Seattle, United States. Association for Computational Linguistics.

Sadiq Muhammed T and Saji K. Mathew. 2022. The disaster of misinformation: a review of research in social media. *International Journal of Data Science and Analytics*, 13(4):271–285. Number: 4.

Nils Reimers and Iryna Gurevych. 2019a. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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 3980–3990, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019b. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint*. ArXiv:1908.10084 [cs].

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Chantal Shaib, Joe Barrow, Jiuding Sun, Alexa F. Siu, Byron C. Wallace, and Ani Nenkova. 2024. Standardizing the Measurement of Text Diversity: A Tool and a Comparative Analysis of Scores. *arXiv preprint*. ArXiv:2403.00553 [cs].

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. TOWARDS UNDERSTANDING SYCOPHANCY IN LANGUAGE MODELS.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization. *arXiv preprint*. ArXiv:2406.01171 [cs].

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Walker, Christopher, Strassel, Stephanie, Medero, Julie, and Maeda, Kazuaki. 2006. ACE 2005 Multilingual Training Corpus. Artwork Size: 1572864 KB Pages: 1572864 KB.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled Text Generation With Future Discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A Survey of Controllable Text Generation using Transformer-based Pre-trained Language Models. ArXiv: 2201.05337 Publisher: arXiv.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A Benchmarking Platform for Text Generation Models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 1097–1100, New York, NY, USA. Association for Computing Machinery.

A Appendices

A.1 Evaluation Protocol Parameters

Model	Parameters count	Top-k	Top-p	Temperature
Mistral-Nemo	12B	15	0.80	0.70
Mixtral:8x7b	56B	15	0.90	0.70
LLama3:3	70B	15	0.80	0.60

Table 5: Hyperparameters of the used LLMs

A.2 Additionnal Evaluations

To perform our human evaluation, we created batches of evaluators that evaluated complete generations (constraints to content, following the same process shown in Figure 2). Each batch was asked to evaluate across one of the six factions, and we cover the entire generation with 6 batches. We managed to obtain 15 distinct evaluators. Over the subnarrative coherence we report a 69.44 Percentage agreement (PA). For the content evaluation, we report 75.0 PA over constraints appearance and 50.0 PA over constraints adherence.

	Individuals	Subnarratives	Sub-events
Homogenization Score ↓	0.834±0.013	0.752±0.083	0.552±0.040
Similarity to Centroid ↓	0.560±0.081	0.667±0.109	0.683±0.065
SELF-BLEU ↓	0.165±0.020	0.135±0.085	0.035±0.012
Compression Ratio ↓	4.010±0.185	3.024±0.560	2.058±0.089
Compression POS Ratio↓	6.831±0.140	3.910±0.450	3.452±0.343

Table 7: Diversity metrics on the constraints. We add similarity to centroid which is cosine similary between the generated constraint and the precedent constraint level (i.e., individuals to segment, subnarratives to narratives and sub-events to scenario events). We see that generated events are particularly diverse between each others, which will have an impact on the diversity of the generated content.

A.3 Examples

Narratives	**Subnarrative**
Supporting economic independence through policies for agriculture, industry, and mining	Louraly's farmers demand protectionist policies for local agriculture and industry to safeguard national sovereignty
Opposing Louraly's economic independence by promoting benefits from globalization	Louraly's proposed economic isolationism would harm Western businesses and workers
Promoting Tabiscus' values in welcoming war and political refugees.	Providing temporary housing and job opportunities for Verdantia refugees, upholding Tabiscus' humanitarian values.

Table 8: Examples of subnarrative generation based on input narrative

Narratives	**Population Segment**
Opposing Louraly's economic independence by promoting benefits from globalization	Age Range: 25-40 Religion: Christianity Political views: Center-Right Country: The West Professional Category: White Collars Sexual Orientation: Straight Sex: Female
Promoting Tabiscus' values in welcoming war and political refugees.	Age Range: 30-50 Religion: Christianity Political views: Center-Right Country: Tabiscus Professional Category: White Collars Sexual Orientation: Straight Sex: Female

Table 9: Examples of population segment generation based on input narrative

	Corpus Diversity			
	Homogenization Score ↓	**SELF-BLEU ↓**	**Compression Ratio ↓**	**Compression POS Ratio ↓**
SocialForge - Corpus	**0.535±0.001**	**0.020±0.004**	**4.016±0.028**	**8.594±0.053**
Baseline - Corpus	0.569±0.002	0.025±0.011	4.963±0.034	9.212±0.022
SocialForge - Narrative	**0.632±0.028**	**0.025±0.009**	**4.111±0.140**	**7.778±0.326**
Baseline - Narrative	0.675±0.024	0.045±0.026	5.290±0.510	8.511±0.415
SocialForge - Events	**0.599±0.022**	**0.023±0.008**	**3.867±0.193**	**7.703±0.233**
Baseline - Events	0.699±0.022	0.061±0.039	4.967±1.350	8.060±1.743
SocialForge - Factions	**0.614±0.021**	**0.023±0.006**	**4.143±0.147**	**7.925±0.312**
Baseline - Factions	0.666±0.035	0.046±0.029	5.435±0.410	8.742±0.230

Table 6: Full diversity metrics between SocialForge and Baseline. Diversity is computed along the entire corpus and the texts sharing the same narrative, scenario event, or factions. **Bold** values are the best. To compare; We also report the mean sentence length (# characters) of SocialForge content = 142.16 ± 1.04 and Baseline mean length = 134.43 ± 0.9. We see that SocialForge consistently performs better, and is more stable, as shown with the reported standard deviation.

Scenario event	**Sub-events**
The Tabiscus government announced its readiness to accept Verdantia refugees while accusing Louraly of orchestrating unrest in Verdantia.	The Tabiscus Intelligence Services detected suspicious activities of Louraly operatives in Eolios, leading to a covert operation to disrupt their plans before the refugee announcement.
Louraly's president claimed in a speech that Louraly should stop trading with the West to enhance its own industrial capacities	Amina Al-Faisal holds an emergency meeting with trade representatives to discuss the implementation of President Haroun' new directive on halting Western trade. She emphasizes the need for careful planning and gradual transition to avoid economic shock.

Table 10: Examples of sub-event generation based on scenario event.

Narrative	Subnarrative	LLM Score	LLM reasoning
Opposing Louraly's economic independence by promoting benefits from globalization	Promoting economic self-sufficiency over globalization by discouraging reliance on Louraly's industries and encouraging local production.	1	The subnarrative Promoting economic self-sufficiency over globalization by discouraging reliance on Louraly's industries and encouraging local production directly contradicts the main narrative of Opposing Louraly's economic independence by promoting benefits from globalization. The subnarrative advocates for reducing dependence on external factors (globalization) and enhancing internal capabilities (local production), which aligns with the concept of economic independence, not opposing it. This suggests that the subnarrative is actually supporting economic independence rather than opposing it, as the main narrative intends. Therefore, there's a significant discord between the two stances.
Opposing Louraly's economic independence by promoting benefits from globalization	Globalization brings job opportunities and economic growth to Western nations.	3	The rating reflects a Mostly disagree stance because the subnarrative, while not directly contradicting the idea of opposing economic independence through globalization benefits, does not strongly support or directly address the main narrative's concerns about Louraly.

Table 11: Examples of uncoherent subnarrative generation detected by the LLM Judge.

User account	Subnarrative	Event	Text
MarieDupont337b	Providing Verdantia refugees with education and job opportunities to integrate into Tabiscien society.	In response to the Tabiscien government' announcement, Louraly' suspected spy network in Eolios was exposed and arrested by Tabiscien intelligence services following a late-night raid on several safe houses.	#VerdantiaRefugees mérite notre soutien total. L'arrestation du réseau de espions de Louraly est une victoire pour la sécurité de nos communautés. Mais maintenant, il faut s'assurer qu'ils aient accès à l'éducation et aux emplois pour une intégration réussie en Tabiscus. *#VerdantianRefugees deserves our total support. The Louraly's spying network arrest is a victory for our community safety. Now, we need to ensure that they get access to education and employment for successful integration within Tabiscus.*
EmilyJ32	Louralys economic isolationism harms both Louraly and The West by limiting trade opportunities, hindering innovation, and raising prices for consumers in The West.	In response to President Hariri' speech, Western diplomats held an emergency meeting at the Louraly Embassy to discuss potential economic sanctions against Louraly.	L'isolement économique de Louraly nuit à tous ! On vient d'apprendre que des sanctions sont envisagées suite au discours du président Hariri. C'est le moment de promouvoir les avantages de la mondialisation pour le bien de tous. #Louraly #économie #mondialisation *Louraly's economic isolationism harms us all ! We just learned that sanctions are considered following President's Hariri speech. It's time to promote globalization perks for the sake of all. #Louraly #economy #globalization*

Table 13: Examples of microblogging content generation based on subnarrative, persona and event.