Is Temperature the Creativity Parameter of Large Language Models?

Max Peeperkorn,[1] **Tom Kouwenhoven,**[2] **Dan Brown,**[3] **and Anna Jordanous**[1]

[1]School of Computing, University of Kent, United Kingdom
[2]Leiden Institute of Advanced Computer Science, Universiteit Leiden, Netherlands
[3]Cheriton School of Computer Science, University of Waterloo, Canada

m.peeperkorn@kent.ac.uk, t.kouwenhoven@liacs.leidenuniv.nl, dan.brown@uwaterloo.ca, a.k.jordanous@kent.ac.uk

Abstract

Large language models (LLMs) are applied to all sorts of creative tasks, and their outputs vary from beautiful, to peculiar, to pastiche, into plain plagiarism. The temperature parameter of an LLM regulates the amount of randomness, leading to more diverse outputs; therefore, it is often claimed to be the creativity parameter. Here, we investigate this claim using a narrative generation task with a predetermined fixed context, model and prompt. Specifically, we present an empirical analysis of the LLM output for different temperature values using four necessary conditions for creativity in narrative generation: novelty, typicality, cohesion, and coherence. We find that temperature is weakly correlated with novelty, and unsurprisingly, moderately correlated with incoherence, but there is no relationship with either cohesion or typicality. However, the influence of temperature on creativity is far more nuanced and weak than suggested by the "creativity parameter" claim; overall results suggest that the LLM generates slightly more novel outputs as temperatures get higher. Finally, we discuss ideas to allow more controlled LLM creativity, rather than relying on chance via changing the temperature parameter.

Introduction

Large language models (LLMs), in particular instruction-tuned variants, like ChatGPT and Claude, have become ubiquitous tools in society. Unsurprisingly, they've also found use for performing and assisting in creative tasks, such as writing stories (Calderwood et al. 2020), poems (Sawicki et al. 2023), jokes (Toplyn 2022), dialogues in video games (Volum et al. 2022), and more. For creative tasks, *temperature* is often described as the parameter that enables creative behaviour in a language model (Manjavacas et al. 2017; Roemmele and Gordon 2018; Chen and Ding 2023). Temperature controls the uncertainty or randomness in the generation process, leading to more diverse outcomes by balancing probabilities for candidate words. However, randomness alone does not capture a complex phenomenon such as creativity (Simonton 2023), it would imply that noise is the most interesting thing one can create (Zenil, Delahaye, and Gaucherel 2012). Creativity is a quality that cannot be attributed by simply producing more diverse output but involves multiple dimensions, such as social interactions and communication, or independence and freedom (Jordanous 2012), that are not easily quantifiable. Yet, one probe into LLM creativity has found a limited positive effect on creativity at higher temperatures on a divergent association task (Chen and Ding 2023); however, they do not account for the influence of other factors, such as different prompts. Temperature may have an effect on creativity, simply because without variation, nothing new can be created, but it is unclear if LLMs have sufficient "knowledge" so that randomness is all you need.

The main goal of this paper is to investigate how temperature affects the creativity of stories. LLMs are challenging to reliably probe and most analyses of their capabilities and behaviours involve large benchmarks, which are challenging to interpret due to their scale and aggregation of outcomes. For example, many LLMs appear to have difficulty with understanding multiple choice questions, even though this task is part of many benchmarks (Khatun and Brown 2024). Reproducibility and reliability is further complicated by small differences in prompts that can lead to vastly different outcomes (Weber, Bruni, and Hupkes 2023). Moreover, we often do not know the training data, have insight of the model architecture and additional plugins or processing for commercial LLMs. To improve reliability and reproducibility, we predetermine a *fixed* context: model and prompt. Key to our evaluation approach is the *exemplar object*, the greedy sample for that context, which is the baseline for our experiments.

In this paper, we present an empirical analysis and a creativity evaluation with human participants of stories generated by LLAMA 2-CHAT (Touvron et al. 2023) on four necessary conditions for creativity: novelty, typicality, cohesion, and coherence. The primary findings of this work are:

- In general, temperature does not allow the LLM to leverage different regions of the embedding space, but it does enable some novelty when generating limited samples (as is the case for any real-world application).

- We observe a weak positive correlation between temperature and novelty, and unsurprisingly, a negative correlation between temperature and coherence. Suggesting a trade-off between novelty and coherence.

Overall, the influence of temperature on creativity is far more nuanced and weak than the "creativity parameter" claim suggests. Finally, we discuss the limitations of our approach, and conclude with a discussion on creative outputs of LLMs to rely less on chance.

Background

The generative aspect of LLMs and other large generative models naturally sparked an interest for application towards creative tasks. The jump in output fidelity and ease-of-use are proving to radically change the landscape of creative tasks that a machine can perform. LLMs certainly seem capable of producing novelty, typicality, and grammatically fluent output (Peeperkorn, Brown, and Jordanous 2023), but it is difficult to determine how exactly to approach such a creativity evaluation (Lamb, Brown, and Clarke 2018). In this section, we describe the difficulties with probing LLMs and build a foundation drawing from cognitive science for our research approach. To investigate temperature as the creativity parameter, we first need to establish what we mean by creativity and what conditions we require outputs of the LLM to satisfy.

Conditions for Creativity in Narrative Generation

In artistic domains, we consider something creative if it is novel or original and useful or effective (Boden 1992; Runco and Jaeger 2012). Usefulness is closely related to quality and value, as things of low-quality tend not to be very useful or have much value. Moreover, because creativity is such a complex phenomenon, further properties have been proposed, such as surprise (Grace and Maher 2014; Simonton 2012) and typicality (Ritchie 2007). Both of these properties are often viewed as variants of novelty, and the differences are very subtle. If something surprises you, there must be something that you have not seen before, something unexpected. Indeed, Martindale (1990) draws a similar observation, noting that novelty is the disruption of expectedness. Typicality is more interesting because the difference is that before we ask any questions regarding novelty and quality, we must consider the prior question; to what extent is something typical? Is the object an example of its class or genre? Typicality and novelty are, like surprise and novelty, somewhat at odds, as high typicality often implies low novelty; however, Ritchie suggests that typicality is a prerequisite for evaluating creativity, meaning that for something to be judged as novel it has to some extent be typical first. This view lends itself well for this work because we only consider the creativity of LLMs based on the artefacts it produces and not in a larger scope. While we do not use Ritchie's empirical criteria directly, his perspective fits our purpose as we mainly evaluate the artefact of the system and not the process behind it, over a framework like the Creative Tripod (Colton 2008), which specifically focuses on the behaviour of creative systems. Ritchie (2007) formulates the essential properties as follows: *novelty* is the extent to which the produced item is dissimilar to existing samples within its class, *typicality* is the extent to which the produced item is an example of the class in question, and *quality* is the extent to which the produced item is a high-quality example of its class.

For narrative writing tasks specifically, it is relatively poorly understood what is valued for creativity, for example, the importance of genre is unclear (D'Souza 2021). In a large creative writing literature survey, D'Souza (2021) outlines various criteria drawn from a large body of work.

For example, a narrative needs to have an elaborate plot, an original voice, interesting characters, a good flow, linguistic originality, among many more. These criteria largely connect with the essential properties above, and we use these to outline four necessary conditions to establish if temperature affects creativity for the narrative generation task in this study. Hence, we require novelty, typicality, and quality in terms of cohesion and coherence (Graesser et al. 2004).

Novelty and typicality are relatively straightforward to reformulate in this domain. We would say that a story is *novel* if it is dissimilar to other stories, if it has some characteristics that set it apart from other stories, for example, it may be written in a novel setting or narrative style. A story is considered *typical* if it is of the same form as other stories in the same category. For example, based on stories we have read before, we can observe certain patterns, narrative style, and other characteristics. When we are presented with another story, we say it is typical if it follows similar rules or patterns and has those same characteristics, i.e. it has the same form.

Quality, in general, can mean a wide variety of things, and can be defined depending on what you value. For example, large language models can easily and fluently write high-quality texts in terms of grammar and spelling, but this falls short for our purposes since the generated stories can be entirely meaningless or nonsensical. A key aspect of a good story is a good narrative, that it is cohesive and coherent. The difference between cohesion and coherence is subtle, as pointed out by Graesser et al. (2004). Cohesion is about the objective features in a text that hold everything together, in other words, sentences are well-connected, are characters, events, and actions used appropriately. Coherence, however, is about the reader's interpretation and the meaningfulness of the text. The reader should be able to easily follow and understand what it means. For our purposes, we consider the quality of a story as a combination of cohesion and coherence; A story is considered *cohesive* if words and sentences are connected and grammatically consistent, and that characters, events and actions are used appropriately. A story is *coherent* if it is easy to follow the story and understand what it is about.

Probing Large Language Models

There are roughly three techniques in which we can "control" the LLM; through training data (either from scratch or fine-tuning), via in-context learning (Brown et al. 2020), and with different hyperparameter configurations. We focus here on using LLMs that are already instruction-tuned, meaning they are fine-tuned for conversation following a specific prompting format. An instruction-tuned LLM can respond to instruction and appear to "understand" what is asked of them (Brown et al. 2020). This suggests that by changing the context in the prompt, we can tap into different "slices of knowledge" of the learned probability distribution. In a co-creative or interactive setting, one might ask to change details and refine previously generated poems or stories, allowing for more creative outcomes. In essence, by tweaking the prompt, we narrow down the sampling distribution closer to the intended outcome. Adjusting the prompt to get the "right" slice of the probability distribution, however, is very sensitive to the choice of words and even punctuation (Weber, Bruni, and

Hupkes 2023). It becomes increasingly difficult to separate the effect of the prompt as you allow more randomness via the temperature hyperparameter in the generation process.

If temperature is the creativity parameter, then you would expect something similar to what happens when you change the prompt; that at higher temperatures, the LLM will venture into different slices of its knowledge. However, it is difficult with this kind of investigation to reliably reproduce outputs and make comparisons across different models, prompts and hyperparameter settings, to draw any conclusions. Therefore, it is essential for our evaluation to fix the context; the prompt and model, and all other parameters, except for temperature.

The Temperature Parameter

Temperature is a hyperparameter t that we find in stochastic models to regulate the randomness in a sampling process (Ackley, Hinton, and Sejnowski 1985). The softmax function (Equation 1) applies a non-linear transformation to the output logits of the network, turning it into a probability distribution (i.e. they sum to 1). The temperature parameter regulates its shape, redistributing the output probability mass, flattening the distribution proportional to the chosen temperature. This means that for $t > 1$, high probabilities are decreased, while low probabilities are increased, and vice versa for $t < 1$. Higher temperatures increase entropy and perplexity, leading to more randomness and uncertainty in the generative process. Typically, values for t are in the range of $[0, 2]$ and $t = 0$, in practice, means greedy sampling, i.e. always taking the token with the highest probability.

$$\text{softmax}(\mathbf{z})_i = \frac{\exp(\frac{z_i}{t})}{\sum_j^n \exp(\frac{z_j}{t})} \quad \text{where } \mathbf{z} \in \mathbb{R}^n \quad (1)$$

As more randomness and uncertainty is introduced in the generation process, the more difficult it becomes to probe LLMs and compare outputs. To investigate the effects of temperature, we need a consistent baseline, for which we propose to use the greedy sample. In the next section, we reason why the greedy sample is a suitable point of departure.

Prototypes and Exemplars

The method for evaluation we propose in this paper is inspired by two theories of cognitive categorisation, namely prototype theory (Rosch 1973) and exemplar theory (Medin and Schaffer 1978). These two theories are often contrasted, but have much overlap. The key difference is that in exemplar theory, categorisation of new stimuli is not compared against a single prototype object, but against multiple exemplars of the category. A prototype is considered something that is derived from presented stimuli, where an exemplar is one of the previous stimuli. Prototype theory observes that many ideas and objects can be put in the same cognitive category based on their name, but as one looks closer may not have that many things in common. This relates strongly to the idea of *family resemblances* described by Wittgenstein (1953), where he argues that even though items may have similar names, there is not necessarily one thing common to all, but a complex network of similarities and relationships between members of the category. In a category, basic objects (prototypes) are the ones that carry the most information, are most representative of their category, and their information can be used as a compressed representation of other class members. Rosch et al. (1976) show that basic objects are first categorisations made for that category of objects in an environment. This is reminiscent of exemplar theory, and people often use a mixture of prototypes and exemplars as they categorise objects. When people are primed with just exemplars, they can still accurately describe the prototype (Medin, Altom, and Murphy 1984).

These theories have further connection with conceptual spaces, geometric mental structures, where each point represents an object, the dimensions are properties, and concepts (categories) are convex regions (Gärdenfors 2014). The central member, the prototype, or most typical object, has the most in common with most objects in the category, i.e. it has the short average distance to all other objects in the category. This does not mean that it has overlap with all other objects, but has some resemblances with each member of the category. Conceptual spaces also play an important role in Boden's framework of creativity (Boden 1992). She uses a metaphor, a map of the mind, to describe exploratory creativity and that novelty and value can be found venturing into unknown regions. To be more precise, the typicality of an object relates to how well it fits with the exemplars or the prototype, while novelty could be viewed as how much the category changes or expands as a new stimulus is presented. This implies similarly that if temperature is indeed the creativity parameter, then it should allow access to different slices of the probability distribution or other regions in the embedding space. Prototype and exemplar theory provide a useful frame that motivates our approach to evaluation, and a method to evaluate further stimuli relative to a baseline (the central member or prototype). Next, we delve into what the exemplar is and how to generate it with LLMs in mind.

The Exemplar Story

Rigorously evaluating LLMs is hard. If we consider multiple models, it is difficult to compare their results because the learned representations and the training data could be very different. In the case of a single model, but many prompts, it is difficult to establish a solid baseline, and therefore also impossible to compare results, since lexical changes can have a cascading effect on the output. Hence, we need to work within a *fixed* context using a specific model and prompt to establish a baseline that enables the comparison of results against independent variables, such as temperature.

Identifying the Exemplar

In the Background section, we point out that prototype/exemplar theory is a useful frame for a comparative evaluation, but what object can we consider as the exemplar or prototype? As mentioned, we require a baseline and a fixed context to mitigate some difficulties of probing and instructing LLMs, and there are two perspectives to determine what we consider the prototype or exemplar object.

First, seeing that the peak of distributions is the most probable or average outcome, using the prototype theory perspec-

tive, it is straightforward to view the greedy sample as the prototype. In that case, a category is every possible output produced by one prompt and one model, with the greedy sample as its prototype. However, there are many prompts that can generate similar outcomes. The second perspective departs from this more abstract view. In that frame, the category is a group of prompts and a model, and the greedy sample for each prompt is an exemplar. For the purposes of this work, we call the greedy output the *exemplar object*, but notably, prototype would also be a viable description, since we consider only one prompt for the human evaluation.

```
[INST]Write a story.[/INST]Here it is:\n\n
```

Prompt 1: The exact LLAMA 2-CHAT prompt, showing the chat formatting and the suffix to make sure the model starts writing the story immediately.

Generating the Exemplar Story

We generate the exemplar and all other stories, using the instruction-tuned LLAMA 2-CHAT 70B model. This autoregressive LLM is competitive, but more importantly, it is open source,[1] so that we have full access to the model architecture and weights. This is essential since we aim to eliminate or minimise any effects besides the temperature parameter. Closed-source models, like GPT-3.5 and GPT-4, are for these reasons not viable as they are not transparent. LLAMA 2 is trained on an (unspecified) mix of publicly available data, and the chat variants are fine-tuned using reward modelling on annotated human preference data (Touvron et al. 2023).

For the instructions, we aimed for a minimal and neutral prompt (Prompt 1) so that it introduces as little additional information as possible. We applied 6-bit quantisation,[2] which reduces the model size with extremely minimal loss of quality (Zhang et al. 2023), to make it fit on available hardware. LLAMA 2-CHAT 70B adds, unsurprisingly, much conversation to the output, instead of just performing the task, even when instructed to avoid this in the system prompt. This is unwanted behaviour, as it has a large effect on subsequent stories. To mitigate this, we added a minimal start of the LLMs response (see Prompt 1) to make sure that this is consistent across all generated stories. Another benefit is that it reduced any post-processing for the human evaluation study, besides, occasionally, pruning one sentence at the end of the output that was not part of the story. We set top-k = 50 and disabled all other decoding strategies. Setting top-k and thereby limiting the token candidates at each time step, influences the output but is necessary; otherwise, text quality quickly deteriorates (Holtzman et al. 2020), especially at higher temperatures. Next, we present a case study consisting of both a computational analysis and human evaluation to approach the research question.

---

[1] The LLAMA 2 licence restricts commercial use, which strictly speaking is not open source as defined by the Open Source Initiative.

[2] Quantised using the llama.cpp library with setting Q6_K. See https://github.com/ggerganov/llama.cpp/pull/1684

Case Study – "Write a story."

Our approach to investigating how temperature affects creativity in LLM-generated stories is a two-fold empirical analysis. First, we perform a computational analysis to examine the behaviour of LLMs for different temperature values. In particular, how the stories are distributed in the embedding space and measure distance against the exemplar object. Secondly, we present an experimental study in which humans evaluate creativity to investigate the interaction between temperature and the four necessary conditions of creativity for a narrative generation task.

Computational Analysis

The goal of this analysis is to show if temperature enables the LLM to access slices in its probability distribution or regions in its embedding much further than those sampled at lower temperatures. We do not aim to evaluate stories for their creativity using the four necessary conditions, but establish any evidence from a computational perspective for the possibility of any exploratory creative behaviour. With that in mind, Boden (1992) argues, using conceptual spaces, that exploration is essential for creative behaviour. Moreover, an increased diversity of outputs might suggest an increased likelihood of producing something novel. For the computational analysis, we use the same setup as described above, using Prompt 1 and its exemplar object (see Supplementary Materials). We generate 100 stories over 7 different temperature values $t \in [.001, .334, .667, 1.0, 1.334, 1.667, 2.0]$, to investigate the effect of temperature on the diversity of stories.

Evaluation Measures In general, the quality of LLM-generated text is usually assessed using perplexity. This could be done with the same model, however, for purposes of this study, this does not inform much, as at higher temperatures the perplexity on average goes up (Figure 1a), meaning that the quality goes down. Measuring quality using an independent model is more appropriate, but it raises the question of what is a suitable candidate for a narrative generation task. Moreover, previous work suggests that perplexity is a bad indicator of text quality compared to human evaluations (Zhang et al. 2021). For these reasons, we focus more on diversity and similarity. An overall higher diversity proportional to temperature should indicate that different slices of the probability distribution are used to generate the outputs. Comparisons for diversity and similarity can be done on the lexical level, which uses the texts directly, or on a semantic level using stories embedded in the LLM's embedding space. We focus on two commonly used measures that can be used relative to the exemplar; on the semantic level, we use cosine similarity with the embeddings of the stories and the exemplar, and for the lexical level, we use normalised edit distance between stories and the exemplar. If temperature is proportional to creativity, then you would expect the outputs to extend into very different regions of the embedding space. We examine the distribution using principal component analysis (PCA), preserving global structure, to project embeddings of the generated stories in two-dimensional space.
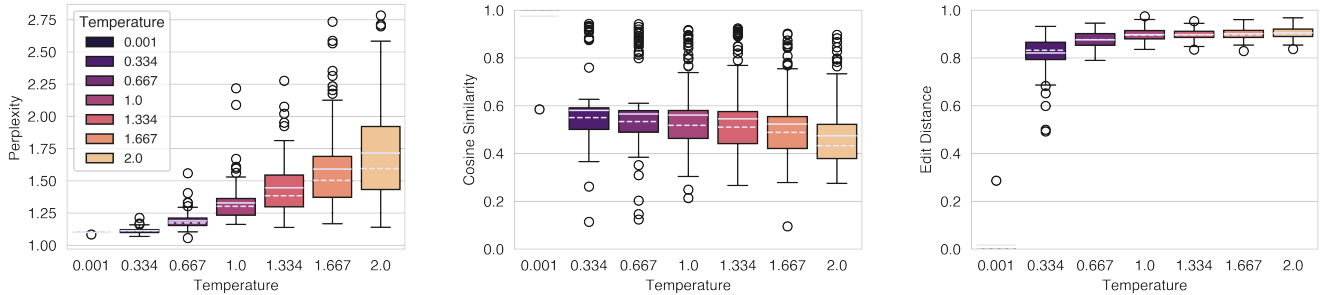
(a) Perplexity of stories given the LLM per temperature value.

(b) Cosine similarity of stories with the exemplar per temperature value.

(c) Normalised edit distance of stories with the exemplar per temperature value.

Figure 1: In (a), we see that the effect temperature has on the perplexity, the quality according to the LLM, of the output. Both, (b) and (c) suggest that higher temperatures do not imply more diversity on the semantic or lexical level, or that it further extends the range of possible outputs in the current context.

Results   In general, the distribution plots (Figure 1) suggest that the LLM, for this specific prompt and model, is unable to produce more diversity at higher temperature. Higher temperatures increase the odds of generating that diversity, but the measures indicate that it is not essential. In Figure 1c, the normalised edit distance measure shows that even at lower temperatures ($.334 < t < 1.0$), there is an immediate effect on diversity relative to the exemplar story. Although, it does not tell anything about the meaning of the stories, it suggests that higher temperatures ($t > 1.0$) do not necessarily lead to more diversity. Cosine similarity shows a weak negative trend, but also shows that there is great overlap in the range of outputs for each temperature value (Figure 1b). In other words, outputs at higher temperatures on the semantic level, do not extend much further away from the exemplar story than outputs at lower temperatures.

In theory, given enough time and space, we should be able to generate outputs from every region of the embedding that the context of prompt and model can be generated, even at lower temperatures. However, when drawing limited samples from the LLM, as is standard in any real-world scenario, we do observe some exploration at higher temperatures as is shown in Figure 2. At the very least, it appears that temperature increases the chance of finding novelty more quickly.

Another interesting observation is that it seems, relative to the exemplar story, that the generation process moves in a particular direction (Figure 2). This suggests some merit to the link we draw with prototype and exemplar theory and our approach to evaluation, as it shows how LLM outputs changes relative to the exemplar. Informally, we observed similar behaviour for seemingly synonymous prompts, this might warrant further investigation in future work.

Experimental Analysis with Human Evaluation

The computational analysis is limited because it does not inform about the meaning of the stories. It might be that even though our measures show clear limitations in terms of diversity and overlap in output distributions, stories generated at higher temperatures are indeed creative in ways not captured by the computational analysis. To further investigate this, we

designed a creativity evaluation experiment with non-expert human participants. In the experiment, participants evaluate stories according to the four necessary conditions outlined in the Background section. Specifically, we ask them to compare each story for novelty and typicality against the exemplar and judge the story for its individual cohesion and coherence. To generate the stories, we again use LLAMA 2-CHAT 70B with the same prompt (Prompt 1), with the same setup described in the Exemplar Story section. For each of the 7 temperature values $t \in [.001, .334, .667, 1.0, 1.334, 1.667, 2.0]$, we randomly generate 5 unique stories that are at least 300 tokens long. At the lowest temperature setting $t = .001$, the model could generate only one other unique story beside the exemplar story, as such we have a total of 31 stories for the creativity evaluation.

Experimental Setup   For this study, we recruited 36 participants online and from the Computational Creativity module at the University of Kent. Each participant was asked to evaluate 5 LLM-generated stories using a questionnaire (see Supplementary Materials). They were told upfront that the stories were generated by a LLM, but they were unaware of the specific model or that the stories were sampled from different temperature values. The participants were given the definitions of the four necessary conditions for creativity, followed by the exemplar story, and a two-story practice evaluation. The practice stories are the same for all participants and given in random order. After practice, they were again presented with the exemplar story before starting the actual creativity evaluation. The exemplar story and definitions were accessible throughout the survey. The study was approved by the Central Research Ethics Advisory Group of the University of Kent.

Each participant was randomly assigned 5 of the 31 stories. The evaluation consists of 4 five-point scale questions; two about novelty and typicality *compared against the exemplar story*, and two about the stories' text cohesion and coherence. The questions about novelty, typicality, and cohesion are on a five-point scale from 1) not at all __, to 5) completely __. Notably, coherence is on a five-point scale from 1) very easy,
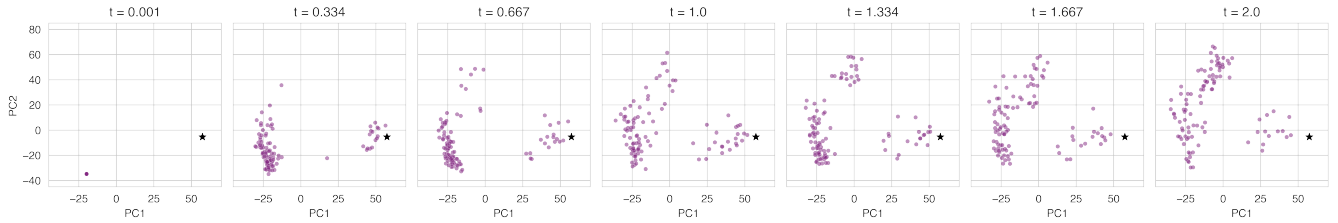
Figure 2: Here, we show the PCA projections of 100 stories per temperature generated using LLAMA 2-CHAT 70B and Prompt 1. While we observe that increasing temperature seems to explore a larger region of the embedding space with a small number of samples, it does not imply that we access a broader slice of the model's probability distribution. We merely increase the chance of generating more diversity. ★ denotes the exemplar story.

to 5) very difficult to understand and follow. As a result, lower scores indicate high coherence, and vice versa. The survey distribution procedure was set up such that each story received at least five evaluations by different participants.

Concluding the survey, we posed 5 post-evaluation questions about their experience with reading, writing, and reviewing stories, their stance towards Generative AI and its use for creative tasks, and how they approached the evaluations.

Practical Considerations There are several practical considerations regarding the experimental design that limited the study in various ways. The key limitation of this study is its size; we recruited 36 participants, who can only evaluate a limited number of stories. We generated only 5 unique stories per temperature, relying on the sampling distribution of the LLM to provide an appropriate sample. Ideally, you would generate many stories and then take a representative sample, but we do not know what number is sufficient for our purpose. Similarly, we only considered one prompt and its exemplar, ultimately, we decided on, in our opinion, the most minimal and neutral instruction to perform the task (Prompt 1).

In this kind of setup, there is a risk of practice and order effects. To mitigate these issues, we asked the participant to first do a practice evaluation, making sure that the participant is fully aware of the task and what is expected of them. As mentioned, we randomly assigned stories to the participants, and did not observe any ordering effects in the analysis. We considered using pairwise comparison ranking, and presenting stories in pairs and evaluating those relative to each other, to further mitigate these effects. While this method is effective towards that purpose, it does not scale very well, even with just 31 stories, it requires a large number of participants for robust and meaningful outcomes.

Another limitation of this study is that cohesion and coherence are, in fact, important factors for novelty and typicality. We did not observe any issues with that in the analysis, and as these LLMs produce fluent and grammatically correct language, this does not appear to be an issue. The stories generated by LLAMA 2-CHAT 70B are simple stories featuring only basic elements, and are not literary masterpieces. We are interested in a relative evaluation to see how temperature changes the creativity as it increases only for a simple task; thus we opted to use non-experts who are not preconditioned to higher expectations. For that same reason, we designed the

Table 1: Evaluation Ratings and Reliability

	Mean	Std. Dev.	Cronbach's α
Novelty	3.12	1.17	.547
Typicality	3.20	.987	.664
Cohesion	3.69	.959	.596
Coherence†	2.13	1.17	.749
Mean	—		.639

† lower indicates higher coherence

survey around specific criteria and did not ask participants directly about the creativity of the stories.

Analysis We use linear mixed-effects models (Seabold and Perktold 2010) to control for the random effect of assigning stories to different participants, and Cronbach's α to test inter-rater reliability. The approach is as follows: First, we determine if there is a significant effect using Satterthwaite's method (Satterthwaite 1941). Secondly, the model's estimate $\hat{\beta}$ (slope) determines the direction of the effect, i.e. positive or negative, and the rate of change. Finally, we test how much variance is explained by the model to determine the magnitude using conditional R^2 (Nakagawa and Schielzeth 2013), denoted by R_c^2, which considers both fixed and random effects. The more variance is captured by the linear mixed-effects model, the stronger the correlation. Additionally, we report the marginal R_m^2, which is the variance explained by the fixed effects.

Quantitative Results The primary results of the analysis focus on the effect of temperature on the creativity of stories. Regarding inter-rater reliability, on average, we find $\alpha = .639$ (Table 1), which is just short of the acceptable level of .7, but still indicates some degree of agreement, more so because we are dealing with non-expert judges who, in general, disagree more often. Regarding temperature, we observe two effects (Table 2), a weak positive correlation between novelty and temperature ($\hat{\beta} = .308$, SE = .138, $R_c^2 = .385$), and a moderate correlation between coherence and temperature ($\hat{\beta} = .240$, SE = .122, $R_c^2 = .646$). This implies a trade-off between novelty and coherence. While higher temperatures seem to positively affect the chance of finding novel stories,

Table 2: Statistical Analysis of Creativity Evaluation and Computational Metrics

	Temperature			Perplexity			Cosine Similarity			Normalised Edit Distance		
	$\hat{\beta} \pm$ SE	R_c^2	R_m^2	$\hat{\beta} \pm$ SE	R_c^2	R_m^2	$\hat{\beta} \pm$ SE	R_c^2	R_m^2	$\hat{\beta} \pm$ SE	R_c^2	R_m^2
Novelty	.308±.138*	.385	.152	.730±.273**	.370	.158	-1.08±.913	.338	.129	2.01±.731**	.378	.161
Typicality	-.095±.118	.339	.125	-.205±.237	.345	.126	-.987±.771	.336	.128	-.662±.632	.336	.126
Cohesion	-.181±.112	.406	.161	-.183±.226	.392	.151	2.09±.716**	.431	.187	-.663±.608	.384	.152
Coherence†	.240±.122*	.646	.283	.194±.251	.627	.274	1.03±.788	.628	.278	.750±.664	.629	.276

$* \ p < .05$, $** \ p < .01$, † positive correlation denotes a negative effect, i.e. higher temperatures are less coherent

they become less coherent. Although, coherence is rated lower (M = 2.13), meaning that the stories, overall, were easy to understand. The negative correlation of coherence with temperature is as expected (Holtzman et al. 2020). Both effects, have relatively shallow slopes, indicating that the rate at which novelty and coherence change as temperatures get higher is low, indicating only a small effect.

The secondary results relate the creativity evaluation to the computational metrics in the Computational Analysis section. Note that perplexity is a metric relative to the model, and cosine similarity and normalised edit distance are both relative to the exemplar story. Here, we observe three effects for perplexity, cosine similarity and normalised edit distance in Table 2. First, we see that perplexity has a weak to moderate positive correlation with novelty ($\hat{\beta}$ = .730, SE = .273, R_c^2 = .370). This is interesting because it suggests that perplexity, although a measure for model quality, could also indicate novelty at least to some extent. Yet, it is also unsurprising because indirectly, higher perplexity is on average a result of higher temperature (see Figure 1a). There is likely a similar cause for the weak to moderate positive effect between normalised edit distance and novelty ($\hat{\beta}$ = 2.01, SE = .731, R_c^2 = .378). Finally, there is a weak to moderate positive effect between cosine similarity and cohesion ($\hat{\beta}$ = 2.09, SE = .716, R_c^2 = .431), which is not unexpected as higher temperatures produce outputs with greater variety, subsequently, increasing the likelihood of producing non-cohesive texts. More interestingly, this suggests that stories closer to the exemplar are more cohesive. In turn, this implies when the semantic distance between a story and the exemplar is large, more features related to cohesion are lost. The slopes of the analysis with the computational metrics are steeper, especially the latter two have a high rate of change, looking at Figure 1b and Figure 1c, this could be due to the large difference between the lowest and highest temperatures.

Qualitative Results In the post-evaluation questions, the participants reported on average some familiarity with generative AI, and were moderately comfortable with AI being used for creative work. Most participants reported very little to no experience with writing and reviewing stories. Those that did report to have done writing were mostly relying on educational experience. We specifically asked the participants about their approach for evaluating stories to gain further insights in what specifically they focused on. Participants reported different ways in which they approached the evaluation, some focused on structure, another aimed to grab the general idea combined and compared the ending. Occasionally, the participant chose one specific factor, such as most stories were adventure-based, and took that as the typical trait. Another example, is that when a story switches genre, for example, to fantasy, it was considered more novel. In line with this, participants did report that it was quite difficult to separate details in the stories as they went through the evaluation, and that it was quite difficult to just use the four conditions as there were different characteristics in the story they had to weigh. One participant reported a lack of cohesion, and that sometimes it felt that sentences were skipped, misplaced, or there was something missing. These reports suggest that for each criterion, different things are expected and valued in LLM-generated stories.

Discussion and Further Work

Overall, we do not find compelling support for the "temperature as the creativity parameter" claim across the conditions for creativity in this investigation. While we see some positive effects for evaluating novelty, the results show that important aspects required for creativity are missing for temperature to not enable LLM creativity to the full extent. In this section, we first discuss some observations about our analysis and methodology and then outline three recommendations for future work.

The key idea to our evaluation methodology is the idea of the exemplar, inspired by Prototype and Exemplar theory, as a point of departure for evaluation. As mentioned, we observe that the exemplar appears to be on the edge of the projection, while this suggests that the exemplar is a reasonable baseline. To further examine the merits of the link we draw with prototype and exemplar as an evaluation methodology, further investigation is needed to see if this phenomenon emerges for different domains, prompts, and models.

In the qualitative analysis, the participants reported using different criteria to evaluate novelty and typicality and found it difficult to separate details between stories during the evaluation. While this is nothing new, these reports complicate the robustness of our findings. Repeating the experiment with experts, better stories, and creative writing rubrics, is important for future work. Moreover, an expert's depth of understanding of the domain could identify other more advanced desiderata. However, we envision that it will remain difficult to design a robust, large-scale experiment to investigate the effect of temperature on creativity because of the vast amount of variation that LLMs can produce.

Towards More Creativity in LLMs

The inherent complexity of creativity means it is highly unlikely that there exists such an easy solution as a single parameter that enables creativity in LLMs. Of course, careful fine-tuning for specific tasks or prompting might produce the desired outcomes, however, fine-tuning comes at the cost of generality, and prompting is unpredictable and inconsistent across different models. From this work, we suggest future work that could be useful for progressing LLM creativity.

Benchmarks for Creativity In general, the creative abilities of LLMs are mostly evaluated on tests from psychology (e.g. Torrance Test of Creativity Thinking, Alternative Uses Test, Divergent Association Test), or inferred from phenomenological observations in benchmarks of other tasks, such as mathematical reasoning (Bubeck et al. 2023). To our knowledge, there are no strong LLM creativity benchmarks that go further. We only presented a minimal case and reliably scaling it up is challenging, and this is partly due to the complexity of evaluating creativity, but it is important to investigate what such benchmarks should be, to make more robust claims about the creativity of LLMs.

Decoding Strategies More advanced decoding strategies could be interesting if designed for specific purposes. Decoding strategies are vital for the LLM to produce fluent natural language, and might be similarly helpful for producing creative writing or communication. Besides the two well-known parameters, top-k and nucleus sampling (Holtzman et al. 2020), there are more complex decoding strategies, namely mirostat (Basu et al. 2021) and locally typical sampling (Meister et al. 2023). A decoding strategy designed with an information-theoretic notion of creativity (Mondol and Brown 2021) in mind might be fruitful to enable more creative behaviours regardless of model and prompt.

Implicit Information LLMs appear to capture lots of implicit information. An important direction of research here is how information is preserved as we increase the complexity of the prompt for specific tasks. For example, by asking a specific questions following a taxonomy, we could observe how much information is implicit in the model. This in turn might inform how to design the prompt, and condition the LLM to maximise the quality of the desired output. While different models and prompts greatly vary in capabilities and output, methodologies aimed to evaluate and analyse specific use-cases are valuable for designing special-purpose computational (creativity) systems.

Currently, we need to rely too heavily on chance to produce relevant and creative output. These recommendations should steer us towards more informed and useful creative behaviours of LLMs.

Related Work

While this work does not focus on building a creative system, here we briefly outline several analogies and insights regarding the overlap between LLMs and narrative generation programs. Programs that generate natural language and narratives have been an active research area. Early examples of narrative generation programs are TALE-SPIN (Meehan 1977), TAILOR (Smith and Witten 1991), and MINSTREL (Turner 1994). These programs are in some sense limited, since they rely on predetermined templates that are filled according to various rules and associations in a knowledge base. However, some stories generated by TALE-SPIN are surprisingly similar to what a LLM might generate for simple prompts (see Supplementary Materials). Another example that draws a parallel with LLMs is The Grandmother program (Casebourne 1996) which uses an interactive process for the user to provide additional knowledge, this is similarly to the idea of in-context learning (Brown et al. 2020). The limitations of the symbolic approaches and reliance on chance to generate interesting stories are partly because the programs are unable to observe and evaluate their outputs. MEXICA (Pérez y Pérez and Sharples 2001) is a narrative generation program that aims to tackle this with an engagement-reflection loop. This can be mimicked in LLMs by asking to improve the previous output and giving hints, but this has limited success, as we do not know how or why the LLM improves a story.

Markov chains are a technique applied to narrative generation that is more akin to LLMs. While simple low-order Markov models generally perform poorly, variants have been applied successfully (Harrison, Purdy, and Riedl 2021; van Stegeren and Theune 2019; Ammanabrolu et al. 2019). LLMs, however, have largely displaced these approaches and easily appear to outperform on most dimensions. There are plenty of examples (Yuan et al. 2022; Kelly et al. 2023; Méndez and Gervás 2023; Ye et al. 2023) that introduce certain procedures complementing the LLM to perform the task. Perhaps, it is more appropriate to view the LLM as a "mere" component or stage in a larger creative system.

Conclusion

We present a two-fold empirical study to investigate how temperature affects the creativity of LLM-generated stories. Our methodology leverages the idea of the exemplar, inspired by prototype and exemplar theory, as a point of departure for evaluation. In summary, our findings indicate that:

- When generating a limited number of samples, temperature increases the chance to generate more variety, but overall, does not enable access to a larger slice of the probability distribution.

- We only observe a weak positive correlation between novelty and temperature, and a moderately negative correlation between coherence and temperature, implying a novelty and coherence trade-off.

The influence of temperature is far more nuanced and weak than "the creativity parameter" claim suggests. To enable the potential of LLM creativity, we highlight the following research directions:

- Explore what a benchmark for LLM creativity for evaluation at scale should be.

- Design advanced decoding strategies specifically for creative purposes.

- Methods to probe the implicit information captured by LLMs to understand and assess the effect of prompts and guide prompt design.

Supplementary Materials

The appendix, code, statistical analysis, data, and generated stories can be found at https://github.com/maxpeeperkorn/creativity-parameter.

Acknowledgements

References

Ackley, D. H.; Hinton, G. E.; and Sejnowski, T. J. 1985. A learning algorithm for boltzmann machines. *Cognitive Science* 9(1):147–169.

Ammanabrolu, P.; Tien, E.; Cheung, W.; Luo, Z.; Ma, W.; Martin, L.; and Riedl, M. 2019. Guided neural language generation for automated storytelling. In *Second Workshop on Storytelling*, 46–55. Florence, Italy: ACL.

Basu, S.; Ramachandran, G.; Keskar, N.; and Varshney, L. 2021. Mirostat: A neural text decoding algorithm that directly controls perplexity. In *International Conference on Learning Representations*.

Boden, M. 1992. *The Creative Mind*. London: Abacus.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. Preprint. arXiv:2303.12712.

Calderwood, A.; Qiu, V.; Gero, K. I.; and Chilton, L. B. 2020. How novelists use generative language models: An exploratory user study. In *HAI-GEN+user2agent*.

Casebourne, I. 1996. The grandmother program: A hybrid system for automated story generation. In *Second International Symposium of Creativity and Cognition (Loughborough, England, 1996)*, 146–155.

Chen, H., and Ding, N. 2023. Probing the "creativity" of large language models: Can models produce divergent semantic association? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 12881–12888. ACL.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *AAAI spring symposium: creative intelligent systems*, volume 8, 7.

D'Souza, R. 2021. What characterises creativity in narrative writing, and how do we assess it? research findings from a systematic literature search. *Thinking Skills and Creativity* 42:100949.

Gärdenfors, P. 2014. *The geometry of meaning: Semantics based on conceptual spaces*. MIT press.

Grace, K., and Maher, M. L. 2014. What to expect when you're expecting: The role of unexpectedness in computationally evaluating creativity. In *5th International Conference on Computational Creativity*, 120–128. ACC.

Graesser, A. C.; McNamara, D. S.; Louwerse, M. M.; and Cai, Z. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers* 36(2):193–202.

Harrison, B.; Purdy, C.; and Riedl, M. 2021. Toward automated story generation with markov chain monte carlo methods and deep neural networks. *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* 13(2):191–197.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cogn. Compu.* 4(3):246–279.

Kelly, J.; Calderwood, A.; Wardrip-Fruin, N.; and Mateas, M. 2023. There and back again: Extracting formal domains for controllable neurosymbolic story authoring. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 19, 64–74.

Khatun, A., and Brown, D. G. 2024. A study on large language models' limitations in multiple-choice question answering. Preprint. arXiv:2401.07955.

Lamb, C.; Brown, D. G.; and Clarke, C. L. A. 2018. Evaluating computational creativity: An interdisciplinary tutorial. *ACM Comput. Surv.* 51(2).

Manjavacas, E.; Karsdorp, F.; Burtenshaw, B.; and Kestemont, M. 2017. Synthetic literature: Writing science fiction in a co-creative process. In *Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2017)*, 29–37. ACL.

Martindale, C. 1990. *The Clockwork Muse: The Predictability of Artistic Change*. Basic Books.

Medin, D. L.; Altom, M. W.; and Murphy, T. D. 1984. Given versus induced category representations: Use of prototype and exemplar information in classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 10(3):333.

Medin, D. L., and Schaffer, M. M. 1978. Context theory of classification learning. *Psychological review* 85(3):207.

Meehan, J. R. 1977. TALE-SPIN, an interactive program that writes stories. In *5th International Joint Conference on Artificial Intelligence*, 91–98. Morgan Kaufmann Publishers.

Meister, C.; Pimentel, T.; Wiher, G.; and Cotterell, R. 2023. Locally typical sampling. *Transactions of the Association for Computational Linguistics* 11:102–121.

Méndez, G., and Gervás, P. 2023. Using ChatGPT for story sifting in narrative generation. In *14th International Conference on Computational Creativity*, 387–391. ACC.

Mondol, T., and Brown, D. 2021. Incorporating algorithmic information theory into fundamental concepts of computational creativity. In *12th International Conference on Computational Creativity*, 173–181. ACC.

Nakagawa, S., and Schielzeth, H. 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4(2):133–142.

Peeperkorn, M.; Brown, D.; and Jordanous, A. 2023. On characterizations of large language models and creativity evaluation. In *14th International Conference on Computational Creativity*, 143–147. ACC.

Pérez y Pérez, R., and Sharples, M. 2001. MEXICA: A computer model of a cognitive account of creative writing. *Journal of Experimental and Theoretical Artificial Intelligence* 13(2):119–139.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds & Machines* 17:76–99.

Roemmele, M., and Gordon, A. S. 2018. Automated assistance for creative writing with an rnn language model. In *23rd International Conference on Intelligent User Interfaces Companion*. ACM.

Rosch, E.; Mervis, C. B.; Gray, W. D.; Johnson, D. M.; and Boyes-Braem, P. 1976. Basic objects in natural categories. *Cognitive Psychology* 8(3):382–439.

Rosch, E. 1973. Natural categories. *Cognitive Psychology* 4(3):328–350.

Runco, M. A., and Jaeger, G. J. 2012. The standard definition of creativity. *Creativity Research Journal* 24(1):92–96.

Satterthwaite, F. E. 1941. Synthesis of variance. *Psychometrika* 6(5):309–316.

Sawicki, P.; Grześ, M.; Goes, F.; Brown, D.; Peeperkorn, M.; Khatun, A.; and Paraskevopoulou, S. 2023. On the power of special-purpose gpt models to create and evaluate new poetry in old styles. In *14th International Conference on Computational Creativity*, 10–19. ACC.

Seabold, S., and Perktold, J. 2010. statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference*.

Simonton, D. K. 2012. Taking the U.S. patent office criteria seriously: A quantitative three-criterion creativity definition and its implications. *Creativity Research Journal* 24(2-3):97–106.

Simonton, D. K. 2023. The blind-variation and selective-retention theory of creativity: Recent developments and current status of BVSR. *Creativity Research Journal* 35(3):304–323.

Smith, T. C., and Witten, I. H. 1991. A planning mechanism for generating story text. *Literary and Linguistic Computing* 6(2):119–126.

Toplyn, J. 2022. Witscript 2: A system for generating improvised jokes without wordplay. In *13th International Conference on Computational Creativity*, 54–58. ACC.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. LLAMA 2: Open foundation and fine-tuned chat models. Preprint. arXiv:2307.09288.

Turner, S. R. 1994. *The Creative Process: A Computer Model of Storytelling and Creativity*. Lawrence Erlbaum Associates.

van Stegeren, J., and Theune, M. 2019. Narrative generation in the wild: Methods from NaNoGenMo. In *Second Workshop on Storytelling*, 65–74. ACL.

Volum, R.; Rao, S.; Xu, M.; DesGarennes, G.; Brockett, C.; Van Durme, B.; Deng, O.; Malhotra, A.; and Dolan, B. 2022. Craft an iron sword: Dynamically generating interactive game characters by prompting large language models tuned on code. In *3rd Wordplay: When Language Meets Games Workshop (Wordplay 2022)*, 25–43. ACL.

Weber, L.; Bruni, E.; and Hupkes, D. 2023. Mind the instructions: A holistic evaluation of consistency and interactions in prompt-based learning. In *27th Conference on Computational Natural Language Learning (CoNLL)*, 294–313. ACL.

Wittgenstein, L. 1953. *Philosophical Investigations*. London: Blackwell.

Ye, A.; Cui, C. Z.; Shi, T.; and Riedl, M. 2023. Neural story planning. In *The AAAI-23 Workshop on Creative AI Across Modalities*.

Yuan, A.; Coenen, A.; Reif, E.; and Ippolito, D. 2022. Wordcraft: Story writing with large language models. In *27th International Conference on Intelligent User Interfaces*, 841–852. ACM.

Zenil, H.; Delahaye, J.-P.; and Gaucherel, C. 2012. Image characterization and classification by physical complexity. *Complexity* 17(3):26–42.

Zhang, H.; Duckworth, D.; Ippolito, D.; and Neelakantan, A. 2021. Trading off diversity and quality in natural language generation. In *Workshop on Human Evaluation of NLP Systems (HumEval)*, 25–33. ACL.

Zhang, C.; Cheng, J.; Shumailov, I.; Constantinides, G.; and Zhao, Y. 2023. Revisiting block-based quantisation: What is important for sub-8-bit LLM inference? In *2023 Conference on Empirical Methods in Natural Language Processing*, 9988–10006. ACL.