Modeling Artistic Creativity in Poetry and Story Generation with Controllable LLMs and Diffusion-Based Language Models

Abstract

We focus on **artistic creativity** in language, specifically, poetry and short story generation, and ask how creativity can be *modeled* and *measured* in Large Language Models (LLMs). Building on Boden's triad (novelty, surprise, value) and her modes (combinational, exploratory, transformational), we design controllable generation procedures and expert-centered evaluation to distinguish mere novelty from creative contribution. Concretely, we (i) curate constructed datasets for poetry and narrative that encode constraints and intended creative moves; (ii) develop a Consensual Assessment Technique protocol with domain experts (poets, critics, creative-writing instructors); and (iii) propose process-aware automated metrics for novelty, surprise, and value that correlate with expert judgments. Methodologically, we instantiate controllable generation as a two-stage pipeline in which autoregressive LLMs first produce prototype stories and poems, and a diffusion-based language model then iteratively edits these prototypes under explicit narrative and poetic control signals, adapting recent work on controllable conversation generation via diffusion over conversation structures (Chen and Yang, 2023). Our framework treats creative generation as constrained movement within and transformation of conceptual spaces, operationalized through prompts, heuristics, retrieval, and diffusion-based editing. The project yields new datasets, metrics, and methods that advance NLP's ability to model a core human capacity while offering practical tools to the literary arts.

1 Introduction

The first heuristic of this proposal is that before we speak about artificial intelligence, we must model intelligence; similarly, before we speak about the creativity of artificial intelligence, we must study creativity itself as a human capacity. Following Boden (2004), we view creativity as producing ideas and artefacts that are simultaneously *new*, *surprising*, and *valuable*. In text, that means expressive departures from expectation that preserve coherence and meaning.

While most recent work on artistic text generation relies on autoregressive LLMs, controllability remains a central challenge. Diffusion-based language models have recently been shown to support more fine-grained control by gradually injecting structural information into an initial prototype text (Chen and Yang, 2023). Motivated by this, we combine LLMs and diffusion-based models in a two-stage architecture: LLMs propose candidate stories and poems, and a diffusion process iteratively edits them under narrative and poetic constraints.

Our key research question is straightforward: **How can we model and evaluate artistic creativity in poetry and story generation with LLMs and diffusion-based language models?** We address it by designing constructed datasets that expose creative intent, by eliciting expert judgments, and by formulating new, process-aware metrics. This project explicitly centers poetry and story generation, the two branches highlighted by Ismayilzada et al. (2024), and grounds its methods in Boden's conceptual-space view.

Thesis statement. *By representing creative processes as combinational, exploratory, and transformational moves in conceptual spaces, and by aligning automated metrics with expert assessment, controllable*

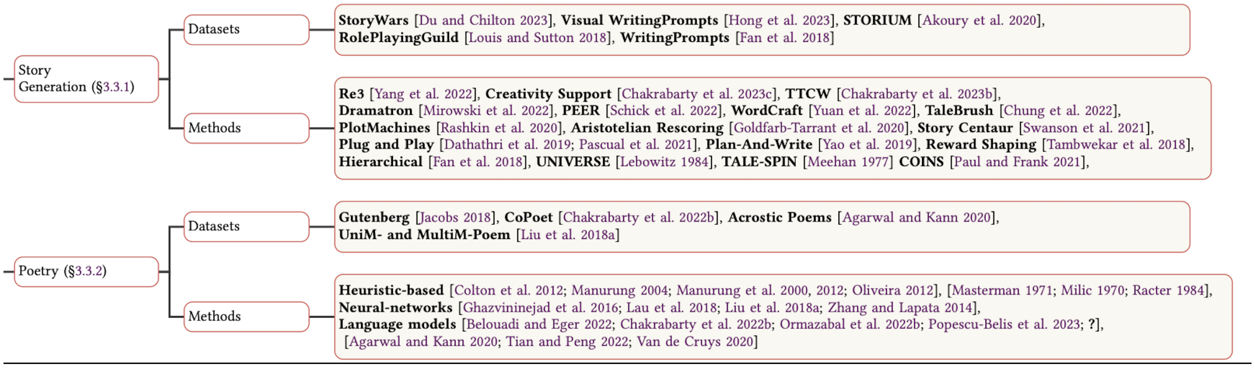
Story Generation (§3.3.1)
Datasets
StoryWars [Du and Chilton 2023], Visual WritingPrompts [Hong et al. 2023], STORIUM [Akoury et al. 2020],
RolePlayingGuild [Louis and Sutton 2018], WritingPrompts [Fan et al. 2018]
Methods
Re3 [Yang et al. 2022], Creativity Support [Chakrabarty et al. 2023c], TTCW [Chakrabarty et al. 2023b],
Dramatron [Mirowski et al. 2022], PEER [Schick et al. 2022], WordCraft [Yuan et al. 2022], TaleBrush [Chung et al. 2022],
PlotMachines [Rashkin et al. 2020], Aristotelian Rescoring [Goldfarb-Tarrant et al. 2020], Story Centaur [Swanson et al. 2021],
Plug and Play [Dathathri et al. 2019; Pascual et al. 2021], Plan-And-Write [Yao et al. 2019], Reward Shaping [Tambwekar et al. 2018],
Hierarchical [Fan et al. 2018], UNIVERSE [Lebowitz 1984], TALE-SPIN [Meehan 1977] COINS [Paul and Frank 2021],
Poetry (§3.3.2)
Datasets
Gutenberg [Jacobs 2018], CoPoet [Chakrabarty et al. 2022b], Acrostic Poems [Agarwal and Kann 2020],
UniM- and MultiM-Poem [Liu et al. 2018a]
Methods
Heuristic-based [Colton et al. 2012; Manurung 2004; Manurung et al. 2000, 2012; Oliveira 2012], [Masterman 1971; Milic 1970; Racter 1984],
Neural-networks [Ghazvininejad et al. 2016; Lau et al. 2018; Liu et al. 2018a; Zhang and Lapata 2014],
Language models [Belouadi and Eger 2022; Chakrabarty et al. 2022b; Ormazabal et al. 2022b; Popescu-Belis et al. 2023; ?],
[Agarwal and Kann 2020; Tian and Peng 2022; Van de Cruys 2020]

Figure 1: Scope within artistic creativity for language: datasets and methods for story generation and poetry (adapted from Ismayilzada et al. 2024).

generative models, combining LLM prototypes with diffusion-based editing, can be steered toward outputs that are judged creative in poetry and story generation.

2 Positioning within Computational Creativity

Story generation. Following Ismayilzada et al. (2024), we treat story generation as the task of producing extended narratives from a premise, prompt, or outline such that the text remains globally coherent, respects character and world constraints, and realises a meaningful plot. Early systems framed this as explicit story planning over character goals and social and physical constraints (Meehan, 1977; Lebowitz, 1984), whereas recent work relies on large neural and language models trained on narrative corpora (Fan et al., 2018; Akoury et al., 2020; Du and Chilton, 2023). Despite strong local fluency, current models still struggle with long-term coherence, adherence to initial premises, and avoidance of repetition (Yao et al., 2019), as well as with controllable satisfaction of high-level constraints such as topic, character, and plot structure (Tambwekar et al., 2018). Stories generated by large language models further exhibit factual inconsistency and hallucinations, limited suspense and tension, and a tendency toward stylistic and affective homogeneity compared to expert written fiction, which motivates the need for more controllable and genuinely creative story generation systems (Zhang et al., 2023; Ismayilzada et al., 2024; Tam et al., 2022, among others).

Poetry generation. In line with Ismayilzada et al. (2024), we view poetry generation as the task of producing verse that meets formal poetic constraints (lineation, meter, rhyme schemes) while conveying affectively and conceptually rich content through devices such as imagery, metaphor, and sound patterns. Early computational systems relied on hand-crafted templates, heuristic rules, and language-specific features (Colton et al., 2012; Oliveira, 2012), whereas more recent statistical, neural, and large language model approaches can generate poems that are formally plausible and stylistically fluent (Ghazvininejad et al., 2016; Lau et al., 2018; Belouadi and Eger, 2022; Ormazabal et al., 2022). However, expert analyses consistently find that such poems often lack poetically deep meaning, sustained and original imagery, and genuinely novel insight, even when meter and rhyme are correct, which reveals a gap between poetic surface form and artistic creativity (Chakrabarty et al., 2023; Ismayilzada et al., 2024).

Figure 1 situates our work squarely within these two areas while emphasizing process tracing and evaluation.

3 Related Work

Story generation. Datasets include WritingPrompts (Fan et al., 2018), STORIUM (Akoury et al., 2020), StoryWars (Du and Chilton, 2023), and role-playing corpora. Methods span planning and hierarchical generation (Fan et al., 2018; Yao et al., 2019), plot control and rescoring (Goldfarb-Tarrant et al., 2020), interactive editing (Yuan et al., 2022; Schick et al., 2022), creative support tools (Mirowski et al., 2022; Chakrabarty et al., 2023), and reinforcement or reward shaping (Tambwekar et al., 2018). Classic systems such as TALE-SPIN and UNIVERSE (Meehan, 1977; Lebowitz, 1984) foregrounded explicit world models.

Poetry. Resources include the Gutenberg poetry corpus (Jacobs, 2018), CoPoet (Chakrabarty et al., 2022), acrostic and metric corpora (Agarwal and Kann, 2020), and UniM/MultiM-Poem (Liu et al., 2018). Feature-based work has examined style, affect, and imagery in contemporary poetry, operationalising imagery through concreteness and sensory word usage (Kao and Jurafsky, 2012). Approaches range from heuristic/rule-based (Colton et al., 2012; Oliveira, 2012) to neural (Ghazvininejad et al., 2016; Lau et al., 2018) and recent instruction-tuned LLMs (Belouadi and Eger, 2022; Ormazabal et al., 2022). Evaluation remains a weakness: crowd ratings underemphasize expert value judgments.

Diffusion-based controllable generation. Diffusion models have recently been adapted for controllable language generation. Chen and Yang (2023) introduce a diffusion-based framework in which prototype conversational turns are generated with a sequence-to-sequence model and then iteratively refined via diffusion, while gradually injecting conversation structures at multiple levels (action triples, dialogue acts, discourse relations). Their results on long conversation generation demonstrate improved controllability and coherence over purely autoregressive baselines. We extend this idea from conversation to artistic text: instead of conversation structures, we inject narrative and poetic control signals (event chains, stanza structure, meter, rhyme, figurative devices) into a diffusion process operating over prototype stories and poems produced by LLMs.

4 Problem Statement and Research Questions

We target the two core domains of artistic language: short stories (200–800 words) and poetry (up to $\sim$20 lines). We will construct datasets that encode: (i) the constraints (form, meter, rhyme, narrative arcs, and higher-level script/frame structure and lexical-semantic relations over key entities and motifs), and (ii) the intended creative move (combinational, exploratory, or transformational). Our research questions are:

1. How can we *model* creative moves as controllable interventions (prompts, decoding heuristics, diffusion-based editing, constraint rewriting) in LLMs and diffusion-based language models?
2. What expert-centered evaluation protocol best distinguishes creative value from mere novelty in poetry and narrative?
3. Which automated, process-aware metrics of novelty, surprise, and value predict expert judgments across genres?
4. How can these methods reveal the limits of current generative models and point to new methodologies for artistic creativity?

5 Methodology

Our controllable generation architecture combines (i) off-the-shelf decoder-only LLMs as prototype generators and (ii) a discrete diffusion-based language model that edits these prototypes under explicit structural constraints. More precisely, an LLM first produces a draft story or poem conditioned on a prompt and a description of the intended creative move. A diffusion process, adapted from Chen and Yang (2023), then iteratively denoises a noisy version of this draft while conditioning on narrative and poetic control signals such as event chains, narrative functions, stanza and rhyme schemes, meter, and figurative language patterns, as well as higher-level representations of conceptual space in the form of scripts, frames, and local semantic networks over entities and motifs. This two-stage pipeline allows us to separate *fluency* (handled by the LLM) from *controllable structure* (handled by the diffusion model).

5.1 Constructed Datasets for Creativity

We will curate parallel sets in English and Greek with explicit annotations of form (e.g., sonnet, free verse), poetic devices (metaphor, alliteration, imagery; following operationalisations of imagery such as concreteness and sensory lexis (Kao and Jurafsky, 2012)), narrative structure (goal, obstacle, resolution), and declared creative intent (C/E/T). At the conceptual level, we annotate script and frame information (prototypical event sequences, situational roles) and induced lexical- semantic networks over key entities and motifs. Prompts and reference materials will be included to enable reproducible evaluation. These datasets will be used both to prompt LLM prototypes and to train or condition the diffusion-based editor on desired structural and stylistic properties.

5.2 Boden-Aligned Control

Combinational: dual-context prompting and conceptual blending across semantic fields, scripts, and frames (e.g., fusing restaurant and courtroom scripts, or combining mythological and everyday frames); constraint satisfaction for analogy/metaphor junctions. **Exploratory**: diversified decoding, varying sampling temperature, nucleus and top-k thresholds, anti-redundancy beams, and novelty penalties, and latent cluster hopping within a fixed script or frame, varying role fillers and local imagery while preserving the underlying event structure to surface previously unseen possibilities. **Transformational**: principled constraint rewriting at the level of meter, rhyme, and also script, frame, and semantic-network structure (e.g., adding, deleting, or reordering key events, roles, or conceptual links), style-token switching, and small adapters that alter allowable transitions, making once "impossible" forms admissible.

In the diffusion-based component, these creative moves are operationalized as control signals attached to each denoising step. For combinational creativity, we condition the diffusion model on multiple source concepts or styles and encourage it to blend motifs across them. For exploratory creativity, we vary the noise schedule and control-vector strengths to traverse different regions of the conceptual space while keeping form and basic coherence fixed. For transformational creativity, we modify the constraint set itself across diffusion steps (e.g., gradually relaxing or tightening metrical and rhyme constraints), allowing the model to transition between regimes of poetic or narrative structure.

5.3 Process Tracing (Wallas-inspired)

We approximate preparation, incubation, illumination, verification via staged generation: retrieval of constraints; timed brainstorm; candidate synthesis; verification with coherence/prosody checks. All search choices, LLM decoding paths, diffusion denoising steps, and constraint edits are logged as creative traces. In the two-stage setup, we treat the LLM prototype as an analogue of preparation and early illumination, while the diffusion trajectory, with its sequence of partially denoised texts, provides a fine-grained record of how constraints and creative moves shape the final artefact.

5.4 Expert-Centered Evaluation

We employ the Consensual Assessment Technique (Amabile, 1983) with poets, critics, and creative-writing instructors. Rubrics cover originality, apt surprise/insight, value (meaningfulness, aesthetic merit, justified deviation), coherence/craft, and form adherence. Inter-rater reliability (Krippendorff's α) will be reported. We will compare expert judgments across conditions (LLM-only, diffusion-only editing of human drafts, full LLM+diffusion pipeline) to test whether diffusion-based control improves perceived creativity and craft.

5.5 Automated, Process-Aware Metrics

Novelty: multi-granular semantic distances (lexeme/motif/event), outlierness vs. reference sets, and adversarial style-novelty (Elgammal et al., 2017; Goodfellow et al., 2014). At the surface and syntactic levels, we compute token-level perplexity under a baseline "non-creative" language model and measure relative perplexity deltas for creative outputs, as well as distance-based similarity between tokenised texts (e.g., embedding cosine distances and edit distances) to quantify how far a poem or story moves away from conventional baselines. **Surprise**: information-theoretic unexpectedness (token-level surprisal deltas), Bayesian shifts in topic/style distributions, and sequential semantic-jump indices for plot twists (Baldi and Itti, 2010). **Value**: discourse coherence (entity grids, neural coherence), prosody/meter conformance, rhyme quality, human-likeness quality estimators, and imagery-based indices (e.g., concreteness and sensory imagery densities), all calibrated to expert CAT.

Since our architecture exposes intermediate LLM drafts and diffusion steps, we can define *process-aware* metrics that track how novelty, surprise, and value evolve during generation and editing. For example, we can measure whether diffusion steps systematically increase form adherence and coherence while maintaining or enhancing novelty, and whether the trajectories of these metrics correlate with expert ratings of creative quality. We also trace how perplexity and text-distance measures change across different decoding settings (e.g., sampling temperature) and diffusion time-steps, linking controllable generation parameters to downstream creativity metrics. At the conceptual level, we quantify changes in script, frame, and semantic-network structure (e.g., number and type of activated scripts, frame shifts, additions and deletions of conceptual links) over the diffusion trajectory, treating combinational, exploratory, and transformational moves in Boden's sense as measurable operations on these representations.

6 Experimental Plan

We compare: base autoregressive LLMs; diffusion-based editing of human-written and LLM-written drafts; Boden-aligned control without process tracing; the full process-traced LLM+diffusion pipeline; ablations by creative mode (combinational, exploratory, transformational); and human–AI co-writing settings. In addition, we systematically vary decoding hyperparameters (e.g., sampling temperature, nucleus and top-k thresholds) for both LLM prototypes and diffusion proposals to study how these control settings affect expert creativity judgments and automatic metrics (perplexity, distance-based novelty, and surprise). Outcomes: expert CAT scores, psychometric-style measures (fluency, flexibility, originality, elaboration) (Guilford, 1967; Torrance, 1974), and correlations between automated metrics and expert judgments. We will also evaluate controllability and faithfulness to specified constraints, following Chen and Yang (2023) in reporting structural accuracy alongside text quality.

7 Ethics and Authorship

We will obtain consent for any human-written materials; avoid mimicking living authors via stylometry checks; document cultural/dialectal biases; and release data with appropriate licenses. Attribution policies will be transparent for co-created texts, with clear distinctions between human and model contributions.

8 Timeline

Year 1: Literature synthesis; dataset design; ethics; pilots; baseline metrics. **Year 2:** Implement LLM prototype generation, diffusion-based control, and process tracing; expert evaluation v1; metric learning. **Year 3:** Large-scale studies; metric refinement; releases and thesis writing.

9 Impact

Overall, this work delivers theoretically grounded, expert-anchored evaluation and modeling of artistic creativity for NLP. It offers new datasets and metrics for poetry and story generation, and a methodology that makes creative processes explicit, testable, and teachable. By integrating LLMs with diffusion-based language models for controllable generation, the project also advances general techniques for steering powerful generative models in high-stakes creative domains.

Note on AI-assisted drafting

Parts of this proposal (in particular, the phrasing and organisation of some paragraphs) were drafted and edited with the assistance of a large language model (OpenAI's ChatGPT). The research ideas, problem formulation, methodological choices, and cited works are my own. I have carefully reviewed, corrected, and verified all content, and I take full responsibility for the accuracy, originality, and integrity of the proposal.

References

Teresa M. Amabile. 1983. *The Social Psychology of Creativity*. Springer.

Sagar Agarwal and Katharina Kann. 2020. Acrostic Poem Generation. In *EMNLP*.

Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. STORIUM: A Dataset of Interactive Narrative. In *EMNLP*.

Pierre Baldi and Laurent Itti. 2010. Of Bits and Wows: A Bayesian Theory of Surprise. *Neural Networks*.

Tobias Belouadi and Steffen Eger. 2022. Language Models for Poetry Generation: A Survey. Preprint.

Margaret A. Boden. 2004. *The Creative Mind: Myths and Mechanisms*. Routledge.

Tuhin Chakrabarty, et al. 2022. CoPoet: Collaborative Poetry Writing with Language Models. In *ACL*.

Tuhin Chakrabarty, et al. 2023. Telling Tales: A Creative Writing Support Toolkit (TTCW). In *ACL*.

Jiaao Chen and Diyi Yang. 2023. Controllable Conversation Generation with Conversation Structures via Diffusion Models. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Simon Colton, et al. 2012. Computational Creativity: The Final Frontier? In *ECAI*.

Yufang Du and Lydia Chilton. 2023. StoryWars: Collaborative Story Writing at Scale. In *CHI*.

Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. 2017. CAN: Creative Adversarial Networks. In *ICCV* Workshops.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical Neural Story Generation. In *ACL*.

Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. 2016. Generating Topical Poetry. In *EMNLP*.

Ian Goodfellow, et al. 2014. Generative Adversarial Nets. In *NeurIPS*.

J. P. Guilford. 1967. *The Nature of Human Intelligence*. McGraw–Hill.

Mete Ismayilzada, Debjit Paul, Antoine Bosselut, and Lonneke van der Plas. 2024. Creativity in AI: Progresses and Challenges. Preprint.

Arthur Jacobs. 2018. The Gutenberg Poetry Corpus. *Behavior Research Methods*.

Jey Han Lau, Trevor Cohn, Timothy Baldwin, et al. 2018. Deep-speare: A Joint Neural Model of Poetic Language, Meter and Rhyme. In *ACL*.

Michael Lebowitz. 1984. Creating Characters in a Story-Telling Universe. *Poetics*.

Pengfei Liu, et al. 2018. Multi-Modal Poem Generation. In *IJCAI*.

James R. Meehan. 1977. TALE-SPIN, an Interactive Program that Writes Stories. In *IJCAI*.

Piotr Mirowski, Kory W. Mathewson, et al. 2022. Dramatron: Co-Writing Screenplays with Language Models. In *CHI*.

Hugo Gonçalo Oliveira. 2012. Poetry Generation with Artificial Intelligence. PhD thesis.

Aitor Ormazabal, et al. 2022. Poem Generation with Large Language Models. Preprint.

Timo Schick, et al. 2022. PEER: A Collaborative Language Model. In *NeurIPS*.

E. Paul Torrance. 1974. *Torrance Tests of Creative Thinking*. Personnel Press.

Shunyu Yao, Daniel S. Weld, and others. 2019. Plan-and-Write: Towards Better Automatic Storytelling. In *AAAI*.

Alex Yuan, et al. 2022. WordCraft: A Human–AI Collaborative Editor for Story Writing. In *UIST*.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content Planning for Neural Story Generation with Aristotelian Rescoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Justine Kao and Dan Jurafsky. 2012. A computational analysis of style, affect, and imagery in contemporary poetry. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 8–17.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, and others. 2022. Scaling instruction-finetuned language models. *arXiv preprint* arXiv:2210.11416.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. *arXiv preprint* arXiv:1912.02164.

Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. A Plug-and-Play Method for Controlled Text Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3973–3997.

Debjit Paul and Anette Frank. 2021. COINS: Dynamically Generating Contextualized Inference Rules for Narrative Story Completion. In *Proceedings of EMNLP 2021*.

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. Plot control and controllable text generation for grounded narratives. In *Proceedings of a major NLP conference*.

Pradyumna Tambwekar, Murtaza Dhuliawala, Lara J. Martin, Animesh Mehta, Brent Harrison, and Mark Riedl. 2018. Controllable Neural Story Plot Generation via Reward Shaping. In *Proceedings of IJCAI*, pages 5982–5988.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. 2023. Evaluating Large Language Models on Controlled Generation Tasks. In *Proceedings of EMNLP 2023*, pages 3155–3168.

Somnath Banerjee, and co-authors. 2024. LLMs Will Always Hallucinate, and We Need to Live With This. *arXiv preprint*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and Improving Consistency in Pretrained Language Models. *Transactions of the Association for Computational Linguistics* 9:1012–1031.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2022. Evaluating the Factual Consistency of Large Language Models Through News Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023. How Language Model Hallucinations Can Snowball. *arXiv preprint* arXiv:2305.13534.