Scaling Creative Writing Beyond Story-Centric Data with Attribute-Guided Genre Expansion

Hwan Chang*
Chung-Ang University
Seoul, Republic of Korea
hwanchang16@gmail.com

Yongil Kim
LG AI Research
Seoul, Republic of Korea
yong-il.kim@lgresearch.ai

Heuiyeen Yeen
LG AI Research
Seoul, Republic of Korea
heuiyeen214@lgresearch.ai

Yireun Kim
LG AI Research
Seoul, Republic of Korea
yireun.kim@lgresearch.ai

Jinsik Lee
LG AI Research
Seoul, Republic of Korea
jinsik.lee@lgresearch.ai

Hwanhee Lee†
Chung-Ang University
Seoul, Republic of Korea
hwanheelee@cau.ac.kr

Abstract

High-quality creative writing data for language models remains dominated by story-centric data, limiting models' ability to follow the structural and functional conventions of diverse creative formats. We propose an *attribute-guided genre expansion* framework for scaling creative writing data beyond story generation. By separating thematic breadth from genre-form control, our framework leverages human-authored story prompts as diverse creative seeds, while utilizing manually curated genre attributes to enforce distinct structural, stylistic, and formatting conventions. We combine these to prompt strong models for genre-faithful query–response pairs, which are then quality-filtered. Applying this framework, we construct the *Multi-Genre Collection*, a 50K-example corpus spanning 13 creative genres, including story, lyrics, game design, and other creative formats. Experiments demonstrate that models fine-tuned on our data consistently surpass not only base models and writing-specialized baselines, but also models trained on existing writing corpora. Genre-count ablations further indicate that genre expansion is a key driver of robust creative writing capability.

CCS Concepts

• **Computing methodologies** → **Natural language generation**.

Keywords

Creative writing, synthetic data generation, large language models, multi-genre datasets

ACM Reference Format:
Hwan Chang, Yongil Kim, Heuiyeen Yeen, Yireun Kim, Jinsik Lee, and Hwanhee Lee. 2026. Scaling Creative Writing Beyond Story-Centric Data with Attribute-Guided Genre Expansion. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26), November 7–11, 2026, Rome, Italy*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3799682.3839976

*Work done during internship at LG AI Research.
†Corresponding author.

ACM ISBN 979-8-4007-2539-5/2026/11
https://doi.org/10.1145/3799682.3839976

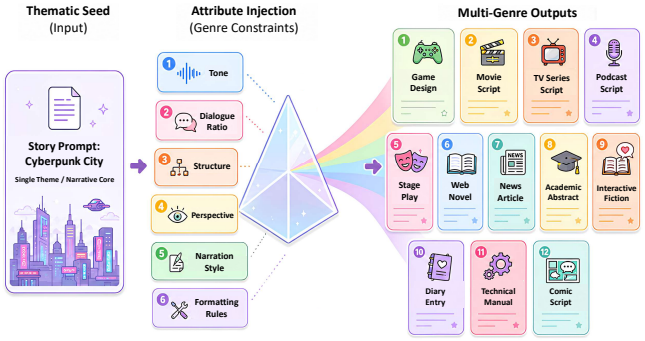

Figure 1: Overview of our attribute-guided multi-genre expansion framework. A thematic seed is transformed into diverse outputs by injecting curated genre attributes.

1 Introduction

Creative writing is one of the most prominent use cases of large language models (LLMs) [1, 15], accounting for a substantial portion of real-world interactions [2, 4]. Existing public resources, nonetheless, remain heavily story-centric, largely treating creative writing as narrative generation [6, 13]. Yet real-world creative writing extends far beyond stories [20]: users ask models to write rap verses, game design documents, and other creative formats. These genres differ not only in topic or style, but also in their structural and functional conventions—rap demands rhyme scheme and flow, movie scripts rely on narrative arc and dialogue, game design documents require mechanics and player interaction. A model trained primarily on story-centric data may thus learn rich narrative content while failing to follow the genre-specific structural and functional constraints expected in practical creative writing tasks.

However, mere scaling of story-centric data fails to address these structural gaps, as it does not systematically encode the formal constraints of non-story genres. Existing synthetic instruction-generation methods offer a natural starting point by bootstrapping seed tasks, evolving instructions, or rewriting problems at scale [18, 21]. Yet, many successful applications have targeted general instruction following or verifiable domains such as code and mathematics, where difficulty, correctness, or solution validity can be approximated through seed examples, executable constraints, final answers, or problem transformations [11, 16, 19, 22]. Creative writing lacks such compact validity signals. Directly applying

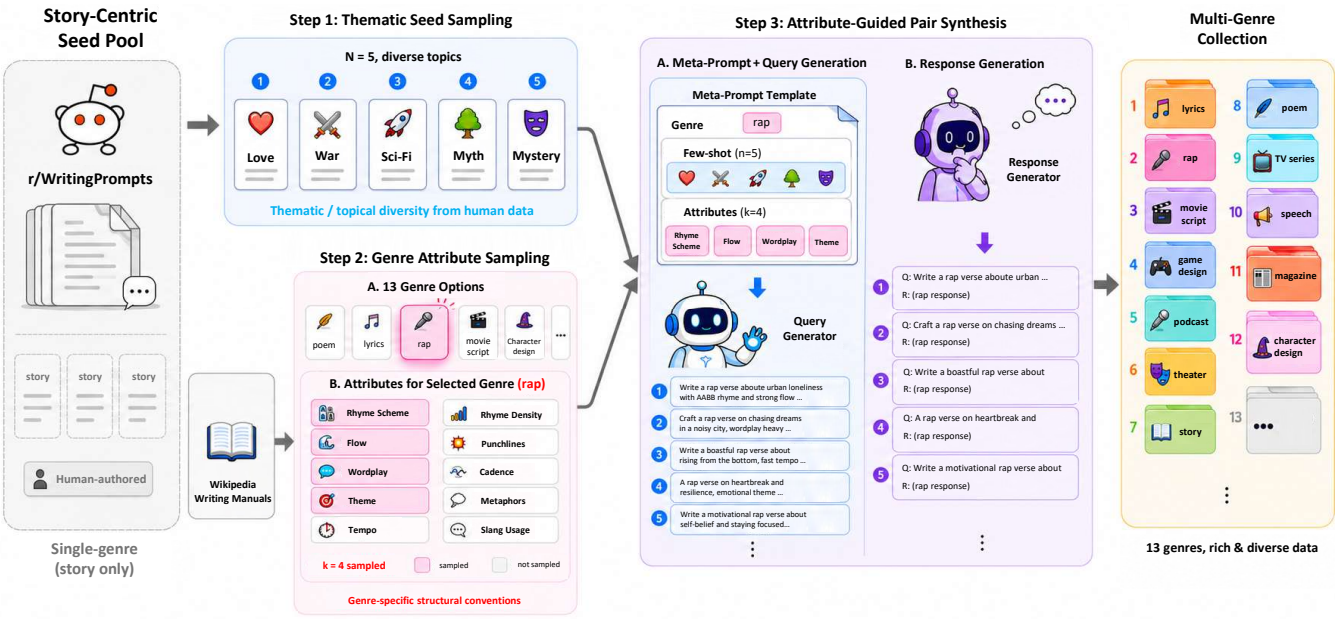

Figure 2: Attribute-guided genre expansion pipeline. We sample few-shot examples from existing story prompts to maintain thematic diversity, randomly sample subset genre-specific attributes to inject structural constraints, and generate multi-genre queries by combining task instructions, few-shot demonstrations, target genre, and sampled attributes.

generic synthesis pipelines risks producing prompts that are topically diverse but formally under-specified, because genre fidelity depends on human-recognizable conventions such as lyric flow, screenplay formatting, and game mechanics. These limitations suggest that creative writing data generation requires human guidance not merely for quality filtering, but for defining the genre-form constraints that generation should follow.

To bridge this gap and effectively scale creative writing capabilities, we identify three key requirements. First, *systematic genre coverage* is necessary to represent the range of creative formats users actually request. Second, *genre-form fidelity* is needed because each genre imposes distinct structural, stylistic, functional, and formatting constraints. Third, *specification diversity* is necessary because real writing instructions naturally range from open-ended requests to highly constrained prompts. These requirements call for a controlled way to expand story-centric creative prompts into genre-faithful instruction data.

To address this challenge, we propose **attribute-guided genre expansion**, a controlled generation framework for scaling creative writing data beyond story-centric sources. Our framework is related to synthetic instruction generation, but differs in its axis of control: rather than evolving prompts primarily by task complexity or problem variation, we separate thematic variation from genre-form control. Human-authored story prompts provide diverse creative seeds, while manually curated attributes specify the conventions of each target genre. For each genre, we draw seeds from a curated `r/WritingPrompts` pool and sample a subset of its attributes, then combine them with the target genre to generate genre-faithful writing queries. The queries are paired with responses from strong LLMs and quality-filtered to yield reliable query–response data.

You are a {genre} query generator.
{task_instruction}
(Guideline: Create distinct scenarios.)
When crafting each query, incorporate these specific dimensions:

{genre_attribution}
(0 ~ len(genre_attribute) genre-specific attributes sampled)

Draw thematic inspiration from the examples below, but reimagine them with new perspectives suitable for {genre}.

{few_shot}
(Sampled from existing Story Generation Query Pool for thematic variety)

Figure 3: Abstract Structured Meta-Prompt Template.

Applying this framework, we construct the *Multi-Genre Collection*, a 50K-instance corpus spanning 13 creative writing genres.

Our experiments show that the resulting data improves creative writing capability across multiple dimensions. Models fine-tuned on the *Multi-Genre Collection* consistently outperform not only their base models and the writing-specialized LongWriter-glm4-9B [3] baseline, but also competitor models trained on existing writing corpora [17] across all popular benchmarks. We further find that increasing genre coverage improves output novelty: as the number of training genres increases from story-only to all 13 genres, novelty metrics [24] improve consistently. Finally, evaluations with independent LLM judges and humans confirm these gains reflect genuine quality improvements rather than evaluator bias.

2 Attribute-Guided Genre Expansion

We formulate the construction of the Multi-Genre Collection as an attribute-guided genre expansion process for scaling creative writing beyond story-centric data. Rather than directly collecting prompts for each creative genre, our framework starts from a story-centric human-authored seed pool and generates genre-faithful

creative writing query–response pairs through the three-stage process illustrated in Figure 2. Instantiating this process yields an English corpus spanning 13 creative genres.

Formally, let (q^g, y) denote a query–response pair for genre g (e.g., q^{rap} = *"Write a rap verse about urban loneliness with an AABB rhyme scheme"*). Prior writing datasets are heavily concentrated on g = `story`; our goal is to expand coverage beyond story writing by generating new queries for each target genre through attribute-guided genre expansion.

Let $X = \{(q_i^{\text{story}}, y_i)\}_{i=1}^{n}$ denote few-shot examples sampled from the seed dataset, and let $A \subseteq A_g$ denote a subset of genre-specific attributes for genre g. We construct a meta-prompt via a template $\mathcal{T}$ (Figure 3) and generate queries through an LLM:

$$Q = \text{LLM}\big(\mathcal{T}(g, X, A)\big). \tag{1}$$

The meta-prompt $\mathcal{T}(g, X, A)$ unifies thematic seeds from human-authored data with genre-specific structural constraints. In this way, the pipeline separates two sources of variation: the seed examples provide topical and stylistic breadth, while the sampled genre attributes control the structural conventions of the target genre.

Source Dataset Selection. We choose Reddit's community forum `r/WritingPrompts` as our seed dataset for (q^{story}, y) pairs, as it provides human-authored prompts with high topical diversity and naturally varying specificity. To ensure quality, we apply two filters using GPT-5-mini [14]: (1) *safety filtering*, which removes 686 instances containing harmful or inappropriate content, and (2) *irrelevant content removal*, which cleans 861 instances of off-topic fragments (e.g., meta-commentary or subreddit boilerplate). This curated seed pool serves as the source of thematic diversity for subsequent genre transfer.

Step 1: Thematic Seed Sampling. For each template instantiation, we randomly sample n=5 query–response pairs from the curated seed dataset to serve as thematic seeds. These seeds transfer the topical and stylistic breadth of human-authored story prompts to non-story creative formats, supplying the thematic variety that drives the generated queries.

Step 2: Genre Attribute Sampling. Since different genres are governed by distinct structural conventions, we define *genre attributes* as the key dimensions along which queries within a genre can meaningfully vary—for example, *rhyme scheme* and *flow* for rap, or *narrative arc* and *character development* for TV series. To construct these attributes, we collect authoritative genre definitions from encyclopedic references (e.g., Wikipedia) and creative writing manuals, then prompt GPT-5 [14] to extract structured attribute lists from each definition. The resulting attributes are all manually reviewed and refined to verify that each reflects a genuine structural convention of the genre, to merge overlapping attributes, to remove overly generic ones, and to add any salient dimensions missed by the automatic extraction—ensuring coverage and genre fidelity and yielding a curated set A_g of 5–15 attributes per genre g. To reflect natural variation in instruction specificity, we sample $k \sim \text{Uniform}(0, |A_g|)$ and then randomly select a subset $A \subset A_g$ with $|A| = k$. When $k = 0$, the query remains open-ended; as k increases, the query becomes progressively more constrained. Thus, attribute sampling acts as the pipeline's control mechanism for varying instruction specificity while preserving genre-specific structure.

Step 3: Attribute-Guided Pair Synthesis. Given the few-shot examples X and sampled attributes A, we prompt GPT-5-mini [14] as the query generator LLM$(\cdot)$ in Eq. (1) to produce five queries per template instantiation. To further promote diversity and mitigate model collapse, we employ verbalized sampling [23]—explicitly instructing the model to produce outputs that vary in topic, tone, and structure. We then generate responses using Qwen3-235B-A22B-Thinking [15]. To filter low-quality responses, we employ an independent LLM-as-a-judge [26] using Qwen3-30B-A3B-Instruct [15] to evaluate response quality and exclude pairs whose scores fall below two standard deviations from the mean.

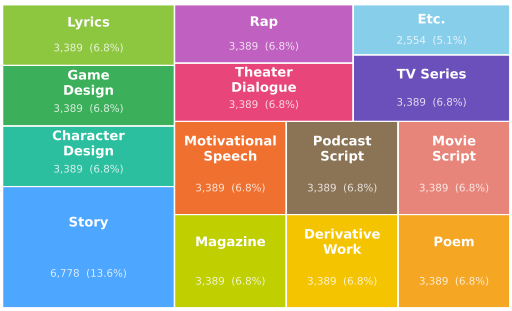

Figure 4: Genre distribution in the Multi-Genre Collection.

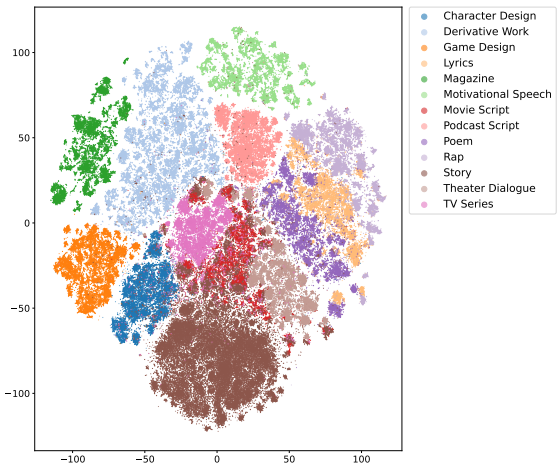

Figure 5: t-SNE visualization of the Multi-Genre Collection.

Synthesis Outcomes and Diversity Analysis. Instantiating the above pipeline yields the Multi-Genre Collection, a 50K-instance creative writing corpus with balanced coverage across the target genres (Figure 4). The 'Etc.' category captures long-tail formats (e.g., diary entries, comic scripts, interactive fiction) outside our 13 primary genres. To verify semantic distinctiveness, we embed all queries using Qwen3-Embedding-0.6B [25] and project them into 2D via t-SNE (Figure 5), revealing clear genre-aligned clusters with minimal overlap. This indicates that the attribute-guided synthesis pipeline produces genre-distinct instruction distributions rather than merely paraphrasing story-centric prompts.

Table 1: Performance across out-of-distribution generalization benchmarks, and in-distribution genre coverage.

	OOD (Generalization)		ID (Coverage)
Models	**Arena Hard**	**WritingBench**	**Multi-Genre**
LongWriter	2.3	47.2	49.5
Llama	2.9	43.7	44.4
Llama + SFT	33.0 (+30.1)	60.6 (+16.9)	68.4 (+24.0)
EXAONE	10.4	52.7	64.9
EXAONE + SFT	22.9 (+12.5)	59.4 (+6.7)	66.2 (+1.3)
Qwen3	8.0	56.1	67.7
Qwen3 + SFT	**34.2** (+26.2)	**63.6** (+7.5)	**69.3** (+1.6)

Table 2: Comparison with existing writing data using 2K sampled subsets, with Qwen3-8B fine-tuned on each corpus.

	OOD (Generalization)		ID (Coverage)
Data	**Arena Hard**	**WritingBench**	**Multi-Genre**
DeepWriting	4.9	55.5	65.9
LongWriter	7.0	56.4	69.6
Multi-Genre	**7.5**	**59.1**	**71.4**

3 Experiments

3.1 Experimental Setup

Models. We evaluate three base models: Llama-3.1-8B-Instruct (Llama) [8], EXAONE-3.5-7.8B-Instruct (EXAONE) [1], and Qwen3-8B (Qwen) [15]. All models are fine-tuned via Supervised Fine-Tuning (SFT) using LlamaFactory [27] with LoRA [9] adapters of rank 128 applied to all transformer layers, preserving pre-trained knowledge while ensuring computational efficiency. We train for 2 epochs with a cutoff length of 4096 tokens, learning rate 1.0×10^{-5}, warmup ratio 0.1, and global batch size 128. As a writing-specialized baseline, we additionally include LongWriter-glm4-9B [3].

Benchmarks. We evaluate models across three benchmarks: (1) *Arena Hard (Creative Writing)* [10] v2.0, filtered to English samples, judged by GPT-4.1 against reference answers following the original evaluation protocol; (2) *WritingBench* [20], filtered to English samples in the creative writing domains (*Advertising & Marketing* and *Literature & Arts*), judged by GPT-5-mini following the original evaluation protocol; and (3) the held-out test set from our *Multi-Genre Collection* (Multi-Genre), spanning all 13 genres with 50 instances per genre (650 total), where Qwen3-235B-A22B-Thinking outputs serve as references and GPT-4.1 rates each model output on a 1–10 scale for quality, creativity, and genre adherence.

3.2 Experimental Results

Overall Performance. Table 1 shows that all three base models improve substantially after SFT on our Multi-Genre Collection, with consistent gains on out-of-distribution benchmarks (largest for Qwen3-8B) and balanced in-distribution coverage across all 13 genres. Notably, all fine-tuned models outperform the writing-specialized LongWriter-glm4-9B [3] baseline, confirming that diverse, genre-specific data is more effective than narrow writing specialization.

Table 3: Effect of genre count on creative output novelty. More genres yield more varied, non-redundant outputs.

# Genres	0	4	8	12	All
Novelty (↑)	3.87	3.93	4.26	4.56	**4.81**

Table 4: WritingBench evaluation under independent judges.

Judge	**GPT-5-mini**	**DeepSeek v3.2**	**Gemini 3 Flash**
Llama	43.7	36.6	41.7
Llama + SFT	60.6	63.1	70.5
Qwen	56.1	51.5	58.5
Qwen + SFT	**63.6**	**65.9**	**71.5**

Table 5: Human evaluation on WritingBench.

Data	**Multi-Genre**	**LongWriter**	**DeepWriting**
Human Eval	**59.3**	57.7	56.9

Comparison with Existing Writing Datasets. Since the Multi-Genre Collection is considerably larger, we randomly sample 2K examples from each dataset for SFT to ensure a fair comparison. As shown in Table 2, Qwen3-8B fine-tuned on Multi-Genre Collection substantially outperforms models trained on DeepWriting-20k [17] and LongWriter-6k [3] across all benchmarks, with particularly large gains on Arena Hard, demonstrating that our attribute-guided pipeline produces higher-quality, more transferable training data.

Impact of Genre Diversity. To examine whether exposure to diverse genres generalizes to broader creative ability, we evaluate outputs using the NoveltyBench Distinct metric [24], which partitions generations into semantically and functionally equivalent clusters to estimate output distinctness. As shown in Table 3, the novelty score rises monotonically as the number of training genres increases from 0 to 13, confirming that diverse creative formats directly enhance the model's capacity to generate varied, non-redundant outputs.

Robustness to Judge and Human Evaluation. To address potential self-preference bias from GPT-5-mini serving as both query generator and judge, we re-evaluate on WritingBench using two independent judges (Table 4). All fine-tuned variants strictly outperform their base models across all three evaluators, confirming that gains reflect genuine quality improvements rather than evaluator artifacts. We further validate via human evaluation on 50 prompts randomly sampled from WritingBench, scored following its protocol (Table 5), which corroborates these findings.

Qualitative Analysis. Figure 6 further illustrates these gains. Given a prompt requesting story-driven symphonic metal deathcore lyrics formatted for Suno, the base Llama follows a conventional verse–chorus structure but relies on generic imagery and provides

Prompt: Write detailed, story-driven symphonic metal deathcore lyrics about a young wizard corrupted by growing power, formatted for Suno.

Llama (Base)

Verse 1: In hallowed halls of ancient lore / A young apprentice, I walked before / Masters of the arcane, guiding my hand…
Chorus: I rise, a phoenix, from the ashes born / As knowledge grows, so does the darkness sworn…
Outro: A legacy of darkness, forever to last / The price of power, a soul forever lost…

Llama + SFT

ASCENDANCE (Suno Format)
Intro [0:00–0:45]: In twilight halls of **Aethel**, where shadows learn to speak…
Verse 1 [0:46–1:30]: He mastered the **Lumen** spell… Wove **Aero** currents… The **Verdant** incantation…
Bridge [4:36–5:10]: He sees **Elara**'s smile… The child he once protected…

Figure 6: Qualitative comparison on an Arena Hard song-lyrics prompt. Dark red highlights SFT-distinctive features: a title, named world and entities, and Suno-compatible timestamps.

neither a title nor named entities. In contrast, Llama + SFT produces a titled composition (*ASCENDANCE*), grounds the narrative in a named setting (*Aethel*), introduces a recurring character (*Elara*) and specific spells (*Lumen*, *Aero*, and *Verdant*), and supplies section-level timestamps. These differences indicate that multi-genre training improves format compliance, world-building specificity, and narrative coherence beyond the aggregate evaluation gains.

4 Related Work

High-quality datasets have been central to advancing creative writing in language models, though early efforts focused almost exclusively on story generation. WritingPrompts [6] provided diverse prompt–story pairs from Reddit, ROCStories [13] offered five-sentence commonsense stories for narrative understanding, and LitBench [7] introduced preference labels for evaluating story quality and writing personalization. Beyond stories, prior work has targeted individual genres in isolation: SongComposer [5] for lyric–melody pairs, and Mirowski et al. [12] for screenplay and theatre co-writing. In contrast, the Multi-Genre Collection addresses the field's collective blind spot in a unified framework, spanning 13 diverse creative genres within a single resource.

5 Conclusion

We introduce the Multi-Genre Collection, a 50K-instance, 13-genre dataset. Using our attribute-guided genre expansion, we transfer thematic diversity from human-authored prompts while enforcing genre-specific constraints through manually curated attributes. Our experiments demonstrate three key findings: (1) models fine-tuned on our dataset substantially outperform both base models and writing-specialized baselines; (2) our dataset consistently outperforms existing writing datasets; and (3) increasing genre diversity directly enhances output novelty.

Acknowledgments

This work was supported by the Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (RS-2025-25458133) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)].

GenAI Usage Disclosure

We write the manuscript ourselves; a general-purpose LLM (ChatGPT) is used solely for refinement of style, clarity, and grammar. It is not used for ideation, claim generation, or experimental design.

As part of the proposed methodology, large language models are used as components of the data construction pipeline—namely, GPT-5-mini for filtering and query generation, Qwen3-235B-A22B-Thinking for response generation, and Qwen3-30B-A3B-Instruct as a quality-scoring LLM-as-a-judge.

References

[1] Soyoung An, Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Gerrard Jeongwon Jo, Hyunjik Jo, et al. 2024. EXAONE 3.5: Series of Large Language Models for Real-world Use Cases. *arXiv e-prints* (2024), arXiv–2412.

[2] Ruth Appel, Maxim Massenkoff, Peter McCrory, Miles McCain, Ryan Heller, Tyler Neylon, and Alex Tamkin. 2026. *Anthropic Economic Index report: economic primitives.* https://www.anthropic.com/research/anthropic-economic-index-january-2026-report

[3] Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025. LongWriter: Unleashing 10,000+ Word Generation from Long Context LLMs. In *The Thirteenth International Conference on Learning Representations.* https://openreview.net/forum?id=kQ5s9Yh0WI

[4] Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. *How people use chatgpt.* Technical Report. National Bureau of Economic Research.

[5] Shuangrui Ding, Zihan Liu, Xiaoyi Dong, Pan Zhang, Rui Qian, Conghui He, Dahua Lin, and Jiaqi Wang. 2024. Songcomposer: A large language model for lyric and melody composition in song generation. *arXiv preprint arXiv:2402.17645* (2024).

[6] Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833* (2018).

[7] Daniel Fein, Sebastian Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. 2025. LitBench: A Benchmark and Dataset for Reliable Evaluation of Creative Writing. *arXiv preprint arXiv:2507.00769* (2025).

[8] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations.* https://openreview.net/forum?id=nZeVKeeFYf9

[10] Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2025. From Crowdsourced Data to High-quality Benchmarks: Arena-Hard and Benchbuilder Pipeline. In *Forty-second International Conference on Machine Learning.* https://openreview.net/forum?id=KfTf9vFvSn

[11] Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2024. WizardCoder: Empowering Code Large Language Models with Evol-Instruct. In *The Twelfth International Conference on Learning Representations.* https://openreview.net/forum?id=UnUwSIgK5W

[12] Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI conference on human factors in computing systems.* 1–34.

[13] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kevin Knight, Ani Nenkova, and Owen Rambow (Eds.). Association for Computational Linguistics, San Diego, California, 839–849. doi:10.18653/v1/N16-1098

[14] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267* (2025).

[15] Qwen Team. 2025. Qwen3 Technical Report. arXiv:2505.09388 [cs.CL] https://arxiv.org/abs/2505.09388

[16] Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. 2024. Openmathinstruct-1: A 1.8 million math instruction tuning dataset. *Advances in Neural Information Processing Systems* 37 (2024), 34737–34774.

[17] Haozhe Wang, Haoran Que, Qixin Xu, Minghao Liu, Wangchunshu Zhou, Jiazhan Feng, Wanjun Zhong, Wei Ye, Tong Yang, Wenhao Huang, Ge Zhang, and Fangzhen Lin. 2026. Reverse-Engineered Reasoning for Open-Ended Generation. In *The Fourteenth International Conference on Learning Representations*. https://openreview.net/forum?id=aK9JneKTL8

[18] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 13484–13508. doi:10.18653/v1/2023.acl-long.754

[19] Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. 2023. Magicoder: Empowering code generation with oss-instruct. *arXiv preprint arXiv:2312.02120* (2023).

[20] Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Wang Zijia, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. 2025. WritingBench: A Comprehensive Benchmark for Generative Writing. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=Pkskg9drDQ

[21] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. WizardLM: Empowering Large Pre-Trained Language Models to Follow Complex Instructions. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=CfXh93NDgH

[22] Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. Metamath: Bootstrap your own mathematical questions for large language models. In *International Conference on Learning Representations*, Vol. 2024. 45040–45061.

[23] Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R. Tomz, Christopher D. Manning, and Weiyan Shi. 2025. Verbalized Sampling: How to Mitigate Mode Collapse and Unlock LLM Diversity. arXiv:2510.01171 [cs.CL] https://arxiv.org/abs/2510.01171

[24] Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. 2025. NoveltyBench: Evaluating Creativity and Diversity in Language Models. In *Second Conference on Language Modeling*. https://openreview.net/forum?id=XZm1ekzERf

[25] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176* (2025).

[26] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=uccHPGDlao

[27] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Association for Computational Linguistics, Bangkok, Thailand. http://arxiv.org/abs/2403.13372