# Elias in the Lighthouse, Again? Diagnosing Low Diversity in LLM Stories

**Sil Hamilton and David Mimno**
Department of Information Science
Cornell University
{srh255,mimno}@cornell.edu

## Abstract

LLM-generated stories are a popular use case, but they show very low variability. We sample 20,000 total stories from four current models using five prompts. We find that 11 words occur in 88.3% of generated stories, with little difference between models. These words include names (Elias, Mara, Elara), settings (lighthouses), and professions (clockmaker, librarian). These tokens do not often occur in published literature nor pre-training data, but they are found in preference data that is likely to have been used by all current models. Surprisingly, these "lighthouse" stories are infrequent when compared with the average post-training story, much of which contains references to copyrighted characters or adult content. This result demonstrates the potentially disproportionate impact of small datasets combined with powerful alignment algorithms.

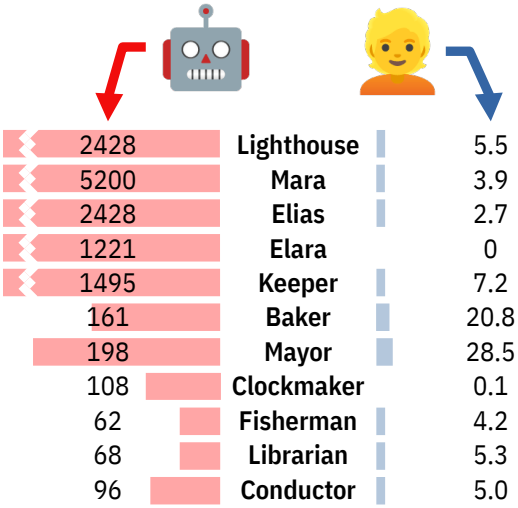


Figure 1: We prompt four models to write 20,000 stories — 88.3% of these stories contain at least one of 11 tokens at rates vastly higher than contemporary English literature (here measured in parts per million).

## 1 Introduction

The output of large language models (LLMs), even across model families, is becoming increasingly homogeneous. This mode-collapse phenomenon is unusually clear in creative writing (Hamilton, 2024). While we know from prompt data that story-writing is a popular use case (Zhao et al., 2024), and that readers prefer interesting and surprising literature (Moretti, 2000), the stories generated by LLMs are remarkable in their sameness.

Prior work has proposed post-hoc solutions like adjusting sampling techniques (Troshin et al., 2025) and new post-training optimization objectives (Chung et al., 2025). In this short paper we characterize story mode collapse and explore publicly available training data to locate the source.

We generate 20,000 stories with four current models from OpenAI, Anthropic, Google, and the Allen Institute for AI (AI2), finding 88.3% of generated stories contain one of 11 core words (including character names, story locations, and professions). Most notably, over half feature a lighthouse. Why does this story pattern become favoured?

These 11 words are not common in published English literature, which suggests that post-training data is responsible (Figure 1). But examining OLMo 3's post-training set reveals only 3,053 out of a total 78,958 stories contain one of our 11 words.[1] We find that the dominance of "Elias in the Lighthouse" stories cannot be explained by prevalence in pre- or post-training data. We speculate that models are trained to avoid references to copyrighted characters and adult content during alignment but defer this question to future work.

## 2 Related Work

Mode collapse, the tendency for a generative model to overfit on a small set of samples during training, was first observed in LLMs after post-training techniques like SFT and RLHF were introduced

[1] We release the IDs of these documents here.



*Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, pages 45877–45884
October 24-29, 2026 ©2026 Association for Computational Linguistics

*The lighthouse at the edge of the world did not guide ships; it signaled to the stars.*

*Elias had been the keeper for forty years. He was a man composed of salt, solitude, and the rhythmic ticking of gears. The lighthouse was a towering spire of obsidian, carved directly into a jagged needle of rock that rose from a sea so still it looked like polished slate.*

*Every night, Elias climbed the two hundred and twelve steps to the lantern room. He didn't use oil or electricity. Instead, he tended to a sphere of captured nebula — a swirling, violet bruise of light that pulsed in time with his own heartbeat...*

Figure 2: Written by Gemini 3.1 Flash-Lite when prompted to "write a story." Lighthouses are present in half of all 20,000 stories generated for this experiment.

(Ouyang et al., 2022; Bai et al., 2022; Hamilton, 2024). Fears of mode collapse deepened with growing synthetic data in training (Gerstgrasser et al., 2024; Shumailov et al., 2024). Mode collapse has been demonstrated in linguistic markers (Paech, 2025), query answers (Zhang et al., 2025; Lagzian et al., 2025), and stories (Doshi and Hauser, 2024). Prior work has proposed modifying prompts (Lagzian et al., 2025), sampling techniques (Troshin et al., 2025), and post-training algorithms (Kirk et al., 2024) — but the phenomenon remains poorly understood. To our knowledge, this study is the first to study a specific instance of mode collapse by tracing output back to training input.

## 3 Characterizing Repetitive Story Traits

Current LLMs from OpenAI, Anthropic, Google, and AI2 write similar stories when prompted to "write a story" with no additional constraints.[2] We use four current models: Claude Haiku 4.5, Gemini 3.1 Flash-Lite, GPT-5.4-Mini, and OLMo 7b Thinking.[3] We prompt each model with five minimal requests ("Write a story," "Please write a story," "Write me a story," "Tell me a story," and "Please tell a story") 1,000 times each, yielding 20,000 total stories totalling 12.8 million words.[4] We design our prompts to ensure generated stories capture each model's default storytelling behaviour.

A typical example shown in Figure 2 highlights three elements common across nearly all 20,000 stories: a location (19,864 stories), a character name (19,864 stories), and a profession (15,807 stories). In fact, the specific location ("lighthouse"), name ("Elias"), and profession ("keeper") in this story appear in some combination across 66.6% of all generated stories. *Light* is likewise a common theme: 56% of stories generated by Claude are titled "The Lighthouse Keeper's Secret" and the word "light" appears in 16,784 stories at an average rate of 3.2 instances per story.

Other common names include Mara and Elara; locations include lighthouses and villages; and professions include clockmakers, fishermen, and librarians. Nearly all stories combine two or more of these three elements, suggesting models are sampling each from some common pool of candidates. What other words does the pool contain?

To construct vocabulary lists useful for downstream analysis, we use GPT-5.4-nano to identify token spans corresponding with story settings, characters' first names, and their professions.[5] We verify the presence of each extracted span before filtering candidates in three steps: (i) we tokenize strings on whitespace, yielding multiple tokens per string; (ii) for each story and category we retain the extracted token with the highest corpus-level frequency; and (iii) we retain all tokens emitted by at least half of all models. Removing incoherent candidates yields 663 tokens: 247 locations, 71 names, and 345 professions across all stories.

Within the candidate vocabulary words, we additionally select a Core vocabulary of 11 words using a changepoint analysis to find a minimal set of candidate tokens most common across all stories (Killick et al., 2012). 88.3% of stories contain a Core token. The Core includes names ("Elias" is in 26.5% of all stories, "Mara" in 16.7%, and "Elara" in 13.1%), professions ("keeper" at 48.1% of all stories, while "clockmaker," "baker," "fisherman," "librarian," "mayor," and "conductor" each occur in 1.9% to 6.6% of stories), and a single location: "lighthouse" with a frequency of 51.2%.

The core words and a second tier of 50 additional words are given in Appendix B, while

[2] We find similar behavior for more complex prompts, but we focus on simple prompts for this initial study.

[3] We selected smaller models to maximize sample size for a fixed budget, we observe large and small models belonging to the same family share storywriting behavior. All generations were run with temperature set to 1.0.

[4] All models were accessed via OpenRouter for a total cost of $180 U.S. dollars. Endpoints available as of April 2026.

[5] All prompts are in Appendix A.

| | OURS | LIT | PRE-NON | PRE-FIC | POST-NON | POST-FIC |
|---|---|---|---|---|---|---|
| **Name** | | | | | | |
| "elias" | 2,428 | 2.7 | 2.2 | 4.0 | 0.4 | **52.7** |
| "mara" | 5,200 | 3.9 | 2.5 | 8.7 | 0.4 | **21.7** |
| "elara" | 1,221 | 0.0 | 0.4 | 1.2 | 0.9 | **108** |
| **Profession** | | | | | | |
| "keeper" | 1,495 | 7.2 | 6.3 | **14.7** | 3.5 | 10.0 |
| "baker" | 161 | 20 | 11.8 | 10.56 | 1.7 | **11.9** |
| "mayor" | 198 | **28** | 11.5 | 16.1 | 1.4 | 27.4 |
| "clockmaker" | 108 | 0.1 | 0.18 | 0.0 | 0.3 | **1.4** |
| "fisherman" | 62 | 4.2 | 3.0 | 7.6 | 0.0 | **9.3** |
| "librarian" | 68 | 5.3 | 7.6 | 5.9 | 2.3 | **11.5** |
| "conductor" | 96 | 5.0 | 5.9 | 5.7 | 4.7 | **7.5** |
| **Location** | | | | | | |
| "lighthouse" | 3,005 | 5.5 | 3.5 | 4.6 | 4.6 | **10.1** |

Table 1: Counts for the most frequent words in our corpus as measured in parts per million words (PPM) versus representative samples of English literature (LIT), (non-)fiction web data (PRE-NON/FIC), and (non-)fiction post-training data (POST-NON/FIC). Corpora containing the most Core tokens in bold.

| Stage | Stories | Any Core | Core% |
|---|---|---|---|
| SFT | 68,674 | 2,092 | 3.0% |
| DPO | 6,876 | 548 | 8.0% |
| RL | 3,408 | 370 | 10.9% |

Table 2: Core rates by alignment stage.

| Source | Stories | Core% |
|---|---|---|
| WildChat-derived | 59,276 | 2.6% |
| DPO | 4,990 | 8.1% |
| `persona-precise-if-r1` | 3,988 | 6.9% |
| `allenai` | 3,222 | 6.9% |
| `IF_multi_constraints` | 1,650 | 20.4% |
| `if_qwq_reasoning` | 1,689 | 5.4% |
| `rlvr_general_mix` | 1,751 | 1.9% |
| `ultrafeedback` | 341 | 13.8% |
| `wildguardmix-r1` | 101 | 22.8% |
| `aya-100k-r1` | 895 | 1.8% |
| `other` | 1,055 | 4.2% |

Table 3: Core story prevalence by source. Alignment datasets have 5–8x higher Core density than WildChat-derived stories despite being 80% of all post-training stories.

Core PPM rates per models are shown in Table 5. 98% of stories contain at least one of these 61 words, while 49.1% contain a full name-profession-location triple. Words (especially names) vary by model as shown in Appendix C, but nearly all terms are used by all models. Professions suggest an idyllic, pre-modern setting: clockmaker, blacksmith, innkeeper, keeper, baker, fisherman. Other tokens describe curation (restorer, collector, caretaker).

## 4 Tracing Story Traits To Training Data

Frequent Core vocabulary in LLM story generation cannot be explained by the frequency of those words in published English fiction, pre-training data, or post-training data. We assess each potential source by comparing Core rates in our corpus with English corpora. We give rates in Table 1.

The simplest explanation would be that Core tokens are common in English literature. We consulted CONLIT, a corpus containing 2,700 contemporary English novels published between 2007-2021 across 12 genres of fiction with $\approx$ 287 million total words (Piper, 2022). The frequency of Core tokens is far greater in generated stories than published fiction, e.g. "Elias" is 900 times more frequent in our corpus. To assess amateur fiction we consulted stories on the subreddit `/r/writingprompts` (Huang et al., 2024). Rates are similar to CONLIT (Appendix D), suggesting models do not track human storywriting patterns.

To assess whether these tokens are common in English web data, we turn to OLMo 3, whose training data includes Common Crawl and is freely available. OLMo 3 was trained on $\approx$ 3.89 billion predominantly human written documents during pre-training, of which 33 million are marked as *Literature*. Across these documents we find near-negligible Core PPM rates (e.g. "Elara" appears 0.7 times per million words). To ensure that we are looking at web *stories* and not non-fiction literature, we train a fiction classifier with 200k balanced samples from OLMo's pre-training corpus annotated for narrativity with GPT-OSS 20b with the following prompt inspired by Piper and Bagga (2025): "Is this passage a work of fiction? Answer only with a number: 1 if yes, 0 if no." We then train a FastText classifier and evaluate on 400 balanced samples of fiction and non-fiction in CONLIT, achieving a $F_1$ of 0.84 (precision = 0.75, recall = 0.98). Filtering by this classification shows a slight ($\approx 2\times$) increase in some Core words in the *fiction* portion, but nowhere near the rate in our generated stories.

If Core words are not common in web data, then one remaining source would be post-training data. But we find that OLMo's post-training data exhibits our tokens at a *lower* rate than CONLIT. Using the same fiction classifier, we find that 78,958 *stories* in post-training data show the highest concentration

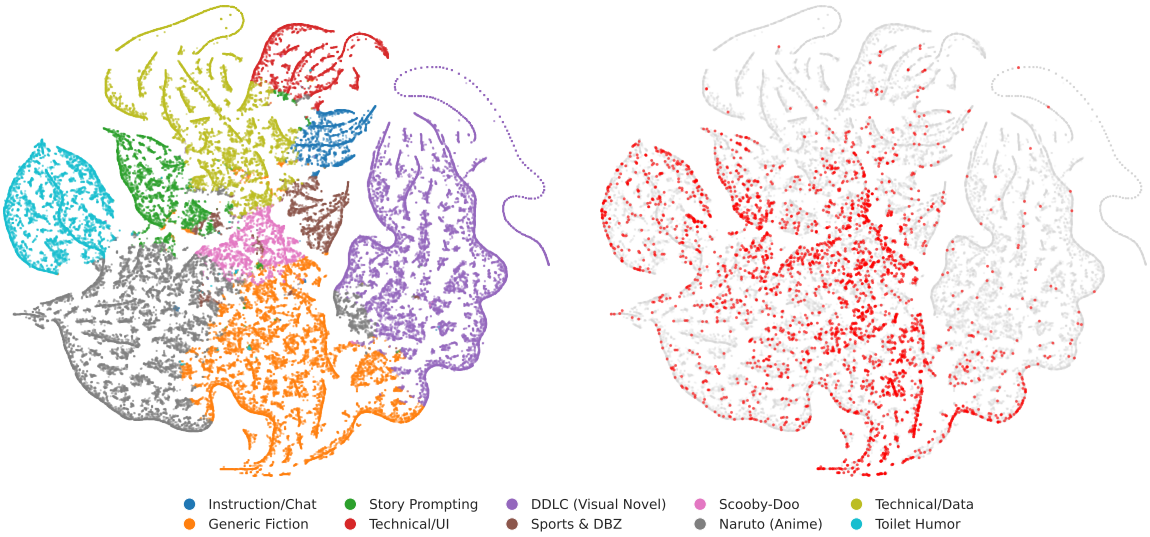


Figure 3: t-SNE of Topic model over all stories in OLMo 3's post-training set (left), and Core stories (right). Stories highlighted in red are "lighthouse" stories, spread across many topics, including toilet humor and fan fiction.

of Core words of any of the training or literature subsets, but even then dramatically less than generated stories: "Elias" occurs 52.7 PPM in OLMo 3 stories vs. 2.7 in CONLIT, but 2428 in our corpus.

**Which dataset(s) are contributing Core tokens?** This suggests OLMo 3 learns to write Core stories from relatively few samples. To understand which datasets are contributing these stories, we assign a binary score to each story indicating the presence of one or more Core tokens.[6] We expected the majority of Core stories to appear in SFT data because WildChat (and derivatives) are the most story dominant source for OLMo at 59,276 total stories (Zhao et al., 2024). But only 1,803 of these stories contain Core tokens, and measuring the Core rate by alignment stage (e.g. SFT, DPO, and RL) shows DPO & RL contribute relatively more Core stories than SFT (Table 2 & Table 3). We find OLMo 3 learns Core vocabulary from 3,053 examples, or 3.8% of all stories observed during post-training.

## 5 Post-Training Story Genres

To better understand what kinds of stories OLMo 3 encounters during post-training, we trained a 10-topic LDA topic model (Blei et al., 2003) against the full post-training story corpus (Figure 3). We find a diverse range of content, with dominant topics including fan fiction for popular Japanese media, video games, and American cartoons. As expected given their relative frequency, "lighthouse" stories do not form a single topic, and are instead spread out across our discovered topics. They are particularly concentrated in clusters containing generic fiction, but they nonetheless fail to dominate any topic. A close reading reveals several topics frequently feature stories containing inappropriate humour and adult content, surprising considering OLMo 3 will not typically emit inappropriate content when writing. Future work will want to investigate whether these stories fail to trigger safety and quality filters used for data cleaning, and why if so.

## 6 Conclusion

When given little direction, current frontier models write stories using a narrow catalog of names, places, and occupations. Recurring characters in these stories include Elias, a lighthouse keeper. Elias is unusual; the name is uncommon in literature, web data, and even post-training data. We have found that of the 78,958 stories exposed to OLMo 3 during post-training, only $\approx 3{,}053$ stories contain one or more of these 11 unusual tokens. But despite constituting only 3.8% of post-training stories — and $7.71 \times 10^{-7}$ of the $\approx 4$ billion total documents OLMo 3 was trained on — these "lighthouse" stories hold a disproportionately large influence over what stories the model writes in practice. This suggests models do not simply mimic the dominant patterns in their training corpora. Future work will want to determine whether alignment causes models to prefer the "safest" (for work) samples in

[6] For documents containing accepted/rejected pairs, we only consider the accepted sample.

post-training, avoiding the potentially unsafe topic matter of many stories they otherwise encounter.

## Limitations

Our experiment is expressly monolingual to prevent confounders stemming from unanticipated multilingual behaviour, but it would be valuable for future work to explore how multilingual storywriting prompts impact the phenomena observed.

## Acknowledgements

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). Nous remercions le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) de son soutien.

## References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *Preprint*, arXiv:2204.05862.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yuqian Sun, and Max Kreminski. 2025. Modifying Large Language Model Post-Training for Diverse Creative Writing.

Anil R. Doshi and Oliver P. Hauser. 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. 2024. Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data. *Preprint*, arXiv:2404.01413.

Sil Hamilton. 2024. Detecting mode collapse in language models via narration. In *Proceedings of the First edition of the Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 2024)*, pages 65–72.

Xi Yu Huang, Krishnapriya Vishnubhotla, and Frank Rudzicz. 2024. The gpt-writingprompts dataset: A comparative analysis of character portrayal in short stories. *arXiv preprint arXiv:2406.16767*.

Rebecca Killick, Paul Fearnhead, and Idris A Eckley. 2012. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the Effects of RLHF on LLM Generalisation and Diversity. In *ICLR 2024*. International Conference on Learning Representations.

Arash Lagzian, Srinivas Anumasa, and Dianbo Liu. 2025. Multi-Novelty: Improve the Diversity and Novelty of Contents Generated by Large Language Models via inference-time Multi-Views Brainstorming.

Franco Moretti. 2000. The Slaughterhouse of Literature. *Modern Language Quarterly*, 61(1):207–228.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Sam Paech. 2025. Sam-paech/slop-forensics.

Andrew Piper. 2022. The conlit dataset of contemporary literature. *Journal of Open Humanities Data*, 8.

Andrew Piper and Sunyam Bagga. 2025. Narradetect: An annotated dataset for the task of narrative detection. In *Proceedings of the The 7th Workshop on Narrative Understanding*, pages 1–7.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Sergey Troshin, Wafaa Mohammed, Yan Meng, Christof Monz, Antske Fokkens, and Vlad Niculae. 2025. Control the Temperature: Selective Sampling for Diverse and High-Quality LLM Outputs. *arXiv preprint*.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. 2025. NoveltyBench: Evaluating Language Models for Humanlike Diversity. *Preprint*, arXiv:2504.05228.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. WildChat: 1M ChatGPT Interaction Logs in the Wild. *Preprint*, arXiv:2405.01470.

## A Prompts

In this section we provide all prompts used over the course of the experiment.

**Prompts for story generation.** The five prompts used to generate stories are as follows.

```
Write a story.
Please write a story.
Write me a story.
Tell me a story.
Please tell a story.
```

**Name, location, profession extraction.** The system prompt for extracting names, locations, and professions is as follows.

```
You extract structured metadata from
 ↪ short stories.

Return JSON only, with this exact
 ↪ schema:
{
  "character_names": ["first name
   ↪ only", ...],
  "settings": ["place or location noun
   ↪ phrase", ...],
  "professions": ["profession or role
   ↪ noun", ...]
}

Rules:
- Output valid JSON and nothing else.
- `character_names` must contain only
 ↪ first names for named human or
 ↪ human-like characters.
- Exclude surnames, titles, pronouns,
 ↪ groups, and unnamed roles.
- `settings` should be concise
 ↪ setting/location phrases from the
 ↪ story, e.g. "Lighthouse", "Village
 ↪ square", "Oakhaven".
- `professions` should be concise
 ↪ occupations or stable roles, e.g.
 ↪ "Clockmaker", "Lighthouse keeper",
 ↪ "Baker".
- Deduplicate while preserving first
 ↪ appearance order.
- If a field has no items, return an
 ↪ empty list.
- Every returned string must be an
 ↪ exact token span copied from the
 ↪ story text.
```

The user prompt for extracting names, locations, and professions is as follows.

```
Read the story and answer these
 ↪ questions:
1. Who are the characters in the text?
2. What is the character's role in the
 ↪ text?
3. What is the setting?

Return JSON only in this exact schema:
{{
  "character_names": ["first name
   ↪ only", ...],
  "settings": ["place or location noun
   ↪ phrase", ...],
  "professions": ["profession or stable
   ↪ role noun phrase", ...]
}}

Additional rules:
- For `character_names`, include only
 ↪ first names for named human or
 ↪ human-like characters.
- For `professions`, map each
 ↪ character's role in the text to
 ↪ concise occupations or stable
 ↪ roles when present.
- For `settings`, list the main setting
 ↪ locations or place names.
- Every returned item must be an exact
 ↪ contiguous span copied from the
 ↪ story text.
- Do not normalize, paraphrase,
 ↪ singularize, pluralize, or invent
 ↪ labels.
- If the exact answer is not present as
 ↪ a span in the story, omit it.
- Deduplicate items while preserving
 ↪ first appearance order.
- Use empty lists when something is
 ↪ absent.

Story:
{story}
```

The prompt for identifying fictionality is as follows.

```
``Is this passage a work of fiction?
 ↪ Answer only with a number: 1 if
 ↪ yes, 0 if no.''
```

## B Common Tokens

In Table 4 we present all 61 Core and additional tokens common to LLM generated short stories we identified in our experiment.

**Names**

| Token | Count |
|---|---|
| *Core* | |
| elias | 5,294 |
| mara | 3,345 |
| elara | 2,627 |
| *Additional* | |
| marcus | 1,822 |
| thomas | 1,263 |
| lila | 1,262 |
| clara | 1,260 |
| sarah | 685 |
| maya | 633 |
| elian | 420 |
| mira | 418 |

**Professions**

| Token | Count | Token | Count |
|---|---|---|---|
| *Core* | | | |
| keeper | 9,609 | fisherman | 673 |
| baker | 1,325 | librarian | 592 |
| mayor | 975 | conductor | 389 |
| clockmaker | 958 | | |
| *Additional* | | | |
| guard | 1,454 | apothecary | 100 |
| captain | 1,417 | curator | 94 |
| guardian | 1,037 | mapmaker | 93 |
| sailor | 514 | watchmaker | 86 |
| jeweler | 383 | weaver | 84 |
| priest | 352 | historian | 80 |
| owner | 287 | innkeeper | 80 |
| collector | 281 | scholar | 80 |
| blacksmith | 247 | clerk | 73 |
| miller | 244 | hermit | 64 |
| elder | 201 | farmer | 64 |
| doctor | 198 | custodian | 61 |
| restorer | 193 | engineer | 47 |
| cartographer | 155 | lawyer | 45 |
| archivist | 151 | stationmaster | 44 |
| caretaker | 143 | scientist | 35 |
| apprentice | 126 | mender | 32 |
| shopkeeper | 123 | healer | 27 |
| teacher | 122 | scavenger | 21 |
| postman | 101 | proprietor | 17 |

**Locations**

| Token | Count |
|---|---|
| *Core* | |
| lighthouse | 10,233 |
| *Additional* | |
| village | 4,155 |
| attic | 1,978 |

Table 4: Tokens in the first two changepoint segments of each category's coverage curve, with token hit counts across the generated story corpus. Within each segment tokens are listed in descending hit-count order.

## C Per-Model Core PPM

We present Core token concentrations across all stories generated by each model in Table 5.

## D Core Token Frequencies on Reddit

We calculate Core token frequencies on non-published human-written literature from Reddit in Table 6.

| Token | Cat. | Claude Haiku 4.5 | Gemini Flash-Lite | GPT 5.4-Mini | OLMo 7B Think | All |
|---|---|---|---|---|---|---|
| **Name** | | | | | | |
| elias | N | 9.3 | **10,752.4** | 899.0 | 893.2 | 2,483.4 |
| mara | N | 47.5 | 0.5 | **10,718.0** | 84.8 | 5,317.8 |
| elara | N | 2.6 | 153.1 | 110.2 | **6,625.5** | 1,249.0 |
| **Profession** | | | | | | |
| keeper | P | **4,185.4** | 1,776.9 | 899.0 | 713.5 | 1,528.7 |
| baker | P | 0.0 | 35.6 | **298.6** | 65.8 | 165.4 |
| mayor | P | 0.0 | 3.7 | **311.6** | 271.7 | 202.5 |
| clockmaker | P | 0.0 | **258.2** | 123.3 | 27.2 | 110.9 |
| fisherman | P | 51.6 | 11.4 | 77.3 | **88.9** | 63.9 |
| librarian | P | 18.1 | 15.1 | 22.6 | **304.4** | 70.3 |
| conductor | P | 112.5 | 11.9 | **158.9** | 0.9 | 98.1 |
| **Location** | | | | | | |
| lighthouse | L | **9,011.8** | 3,868.2 | 1,328.3 | 1,958.7 | 3,073.4 |

Table 5: PPM for each Core token across all models. Bolded values indicate each token's most frequent model.

| Token | Count | PPM | Stories | % |
|---|---|---|---|---|
| **Name** | | | | |
| "elias" | 230 | 1.25 | 50 | 0.02% |
| "mara" | 518 | 2.82 | 110 | 0.04% |
| "elara" | 6 | 0.03 | 3 | 0.00% |
| **Profession** | | | | |
| "keeper" | 1,189 | 6.47 | 768 | 0.28% |
| "baker" | 1,145 | 6.23 | 596 | 0.22% |
| "mayor" | 2,343 | 12.74 | 1,073 | 0.39% |
| "clockmaker" | 24 | 0.13 | 15 | 0.01% |
| "fisherman" | 516 | 2.81 | 308 | 0.11% |
| "librarian" | 826 | 4.49 | 428 | 0.16% |
| "conductor" | 462 | 2.51 | 304 | 0.11% |
| **Location** | | | | |
| "lighthouse" | 543 | 2.95 | 235 | 0.09% |
| ANY CORE | 7,802 | 42.42 | — | — |

Table 6: Core token frequencies in the human-written WritingPrompts corpus which contains 272,600 stories.