LOST IN THE LONG TAIL: IMPROVING CREATIVITY AND COHERENCE IN LANGUAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Decoding strategies strongly shape the behavior of language models in open-ended generation. Prior work has focused on *truncating* the distribution over next tokens in order to remove noise. However, do improbable tokens always constitute noise? We examine a phenomenon we term *token myopia*, in which truncation decoding methods systematically under-generate valid multi-token lexical units because their initial token receives low probability despite being followed by a highly probable continuation. Using idioms and collocations as a measurable testbed, we show that language models often encode idiomatic structure internally yet fail to realize these expressions in free generation due to this first-token bottleneck. To address this limitation, we propose **Lookahead-based Rescoring** (LoRe), a short-horizon sequence-scoring decoder that rescales candidate next tokens using d-step greedy lookahead continuation scores. To balance quality and diversity, we combine this approach with a stochastic activation mechanism that applies lookahead at a tunable rate. LoRe is complementary to existing truncation-based methods such as min-p sampling, allowing the model to generate more long-tail expressions without altering the underlying model. Across creative generation tasks, LoRe increases multi-word expression usage by 50% on average relative to min-p sampling. Furthermore, it outperforms the baseline in 86% of pairwise evaluations on average in creativity, while preserving fluency and coherence. These results suggest that many decoding failures do not stem from deficiencies in model knowledge, but from myopic token-level decision making, and that applying a short-horizon lookahead can better align generation with the latent linguistic structures already represented within language models.

1 INTRODUCTION

Language models have become remarkably capable at generating fluent and contextually appropriate text. However, their generations are still distinct from human-written text in systematic ways (Pillutla et al., 2021; Muñoz-Ortiz et al., 2024; Reinhart et al., 2025). One notable difference is the use of creative and culturally relevant expressions (Liu et al., 2022; Kabra et al., 2023; Cecilia Liu et al., 2024; Kim et al., 2025; Almheiri et al., 2026; Attia et al., 2026). Idioms and multi-word expressions (MWEs) are a staple of natural human writing and occur frequently in spoken and academic English, appearing at rates of hundreds of tokens per million words in corpora such as MICASE (Simpson et al., 2002) and the British Academic Spoken English corpus (Thompson & Nesi, 2001). Despite being exposed to these expressions during pretraining, autoregressive language models rarely produce them spontaneously during generation. In thousands of unconstrained generations across a range of model sizes, and across an array of temperature settings, we observe near-zero idiom occurrence rates.

To begin, we investigate whether the failure to generate multi-word expressions (MWE) stems from a lack of underlying knowledge. We test this hypothesis using probing classifiers and find that it does not. Specifically, literal and idiomatic expressions are strongly linearly separable in model representations, indicating that the relevant knowledge is present. Given this finding, we examine why models possess knowledge of these expressions yet fail to generate them at rates consistent with their prevalence in the pretraining data.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We trace the root cause to a fundamental property of autoregressive decoding. We call this phenomenon *token myopia*: the tendency of next-token selection to prioritize immediate token likelihood, thereby discarding locally improbable tokens even when they would initiate highly probable downstream continuations. Related forms of myopic generation have been observed previously in planning settings, where agents must learn to take actions whose immediate reward may be low or absent but whose long-term contribution to future return is substantial (Sutton & Barto, 2018; Arumugam et al., 2021; Ma et al., 2025). However, this phenomenon has not been formally characterized in the context of open-ended text generation, nor has a targeted remedy been proposed.

In the case of idioms, the first word (e.g., *kicked* in *kicked the bucket*) is assigned low probability by the model in the sentence "His father left him with nothing after he, . . . ", because it is an unusual word choice. However, once that first token is conditioned upon, the subsequent idiom tokens (*the bucket*) become highly probable. Standard greedy decoding, and many standard methods such as nucleus sampling, prune the first token before this downstream reward is ever realized.

We propose **Lookahead-based Rescoring** (LoRe), a decoding-time intervention that rescores the candidate next tokens by appending each to the current context and evaluating a short greedy rollout of d steps. The candidate whose *cumulative* d-step log-probability is highest is selected. To prevent LoRe from mode collapse at high deployment rates, we stochastically activate the lookahead mechanism with probability α, falling back to min-p (Nguyen et al., 2025) or top-p sampling (Holtzman et al., 2020) otherwise. We show that an intermediate activation rate α maximizes MWE generation by 50% on average, and also improves fluency, coherence and creativity.

Our contributions are:

- We document and quantify the under-generation of idiomatic expressions in large language models (LLMs) across multiple model families and sizes.
- We provide evidence for token myopia: LLMs distinguish idiomatic from literal sentences in representation space but systematically prune the first idiom token during decoding, as revealed by incremental probability analysis.
- We introduce LoRe, a modular, simple decoding method with stochastic activation and min-p integration, and analyze its behavior across activation rates on 1,000 creative generation examples. We find that LoRe increases the fluency and coherence of generations and generates more MWEs and idioms (counted separately as a subset of long-tail phrases). We also emphasize that LoRe is compatible with most generation methods out of the box.

2 LLMs Under-generate Idiomatic Expressions

We first empirically establish that language models under-generate idiomatic expressions relative to their frequency in pretraining data. Idioms are well-suited as probes of long-tail behavior due to their rarity and non-compositional semantics, making them a useful diagnostic for under-generation in open-ended generation.

2.1 Idiom Set

We use the EPIE corpus (Saxena & Paul, 2020), which provides idiomatic sentences in English with multiple usage examples per phrase. The EPIE corpus has static and dynamic idioms. Static idioms have fixed surface forms and are typically expressed in a single canonical form (e.g., "big deal"), whereas dynamic idioms allow limited lexical variation or syntactic flexibility while preserving idiomatic meaning (e.g., "drive someone up the wall" $\rightarrow$ "Her children drove their mother up the wall"). For ease of retrieval in large pretraining corpora, we focus on the static idiom subset, which can be efficiently queried using infini-gram (Liu et al., 2025). We note that not all EPIE usage examples are idiomatic; some correspond to literal usages (e.g., "She kicked the bucket she saw on the ground"). To restrict our analysis to predominantly idiomatic meanings, we estimate idiomatic usage frequency using infini-gram counts. We first sample one million documents from the Pile (Gao et al., 2020). For each idiom, we use Claude-3.5-Sonnet to distinguish between its literal and idiomatic usages of the idiom in the sampled documents. We define the idiomatic usage rate as $1 - \frac{\text{literal count}}{\text{total count}}$. We retain only phrases with a score above 0.9, yielding a final set of 188 high-confidence idioms used idiomatically at least 90% of the time.

2.2 IDIOM OCCURRENCE IN PRETRAINING CORPORA

We search for occurrences of our 188 idioms across all corpora accessible via infini-gram. Among these, only two datasets contain detectable idiom usage: the Pile (Gao et al., 2020) and Dolma (Soldaini et al., 2024). The highest idiom density (i.e., the fraction of tokens belonging to occurrences of our idioms) is observed in the RedPajama-Data-1T (Weber et al., 2024), at 0.36% of tokens. The Pile training split exhibits a comparable density of 0.31%. Although small in absolute terms, these values confirm that idioms appear at non-negligible frequency in pretraining corpora. For scale, this corresponds to approximately 9.3 million idiomatic tokens in the Pile for our set of 188 phrases alone.

2.3 IDIOM GENERATION BY LANGUAGE MODELS

We evaluated five models: Pythia-12B (Biderman et al., 2023), Llama-3.1-8B (Grattafiori et al., 2024), Qwen-3-32B (Qwen Team, 2025), Gemma-2-9B (Gemma Team, 2024), and Mixtral-8$\times$7B (Jiang et al., 2024). For each model, we generated 1,000 samples of up to 562 tokens using nucleus sampling with p = 1.0 (to reflect the model's full distribution) at temperatures 1.0, 1.5, 2.0, 2.5, and 3.0. As shown in Table 1, the highest observed idiom fraction is 3×10^{-5}, comprising only highly conventionalized expressions (e.g., "now and then","god forbid", "never mind"). Based on corpus-level estimates that approximately 0.31% of tokens in the Pile corpus are idiomatic (Section 2.2), we would expect roughly 1630 idiom tokens across the $1{,}000 \times 562$ generated tokens. However, the model that generated the most idioms, Gemma-2-9B, only produced 36 idiomatic tokens, while Qwen-3-32B did not generate any.

This deficit holds uniformly across all five models despite their diverse pretraining corpora. Only Pythia is trained on the Pile. The remaining models are trained on web-scale corpora that, to varying degrees of specificity, are documented to include substantial CommonCrawl-derived web text (Grattafiori et al., 2024; Gemma Team, 2024; Qwen Team, 2025; Jiang et al., 2024). We therefore expect all six models to have received broadly comparable idiomatic exposure during pretraining, making corpus-level differences an unlikely explanation for the near-uniform suppression of idiom generation we observe.

2.4 LLMS CAN DISTINGUISH IDIOMATIC FROM LITERAL SENTENCES

The absence of idioms in generation does not imply that models lack internal representations of idiomaticity. To investigate this, we extracted mean-pooled hidden states of idiomatic sentences and their literal counterparts from each layer of Pythia-12B, llama-3.1-8B, Qwen-3-32B, Gemma-2-9B, and Mixtral-8x7b for a balanced set of idiomatic and literal sentence pairs from EPIE, then trained a linear classifier on the pooled representations to predict whether a sentence is idiomatic or literal (see classification details and dataset examples in §A.1).

Table 1: Best idiom generation performance by model. For each model, the temperature with highest idiom token ratio is shown.

Model	Temperature	Idiom Tokens	Ratio
Gemma-2-9b	1.0	36	0.000064
Llama-3.1-8b	1.0	29	0.000052
Mixtral-8x7b	1.0	29	0.000052
Pythia-12b	1.0	26	0.000046
Qwen-3-32b	1.0	0	0.000000

The classifier achieves perfect accuracy for all models using embeddings from layers as early as layer 7 (22% depth) for Llama-3.1-8B, layer 9 (21% depth) for Gemma-2-9B, layer 14 (22% depth) for Qwen-3-32B, layer 14 (39% depth) for Pythia-12B, and layer 14 (44% depth) for Mixtral-8x7b. This result establishes that the failure to generate idioms is not a failure of knowledge but rather a failure of the decoding procedure.

2.5 THE FIRST-TOKEN BOTTLENECK

To understand *why* these models fail to generate idioms despite encoding them, we analyze the incremental token probabilities along idiom sequences versus model-generated continuations.

Protocol. For each idiom in our set, we sample an idiomatic context sentence from EPIE. We then manually pair the idiomatic sentence with a literal sentence that uses the exact same tokens before the idiomatic token. We also made sure the tokens before the idiomatic token hint at what idiom is appropriate (e.g., "She couldn't beat cancer and eventually ..." with completions: "kicked the bucket" and "died."). We compute two incremental probability sequences for the next idiom tokens. The *idiomatic sequence* is $P(w_1 \mid c), P(w_2 \mid c, w_1), \ldots, P(w_5 \mid c, w_1, \ldots, w_4)$, where $w_1, \ldots, w_5$ are the idiomatic tokens and c is the context. The *literal baseline sequence* is computed by first generating the model's actual continuation from c with temperature = 1.0 and top-p = 1.0 (unrestricted sampling to represent the LLM's true learned distribution), then extracting the first 5 tokens $g_1, \ldots, g_5$ and computing $P(g_1 \mid c), P(g_2 \mid c, g_1), \ldots, P(g_5 \mid c, g_1, \ldots, g_4)$. We average these probability sequences over all idioms and models.

Results. Figure 1 shows the average probability curves for Pythia-12B, Qwen-3-32B, Gemma-2-9B, Llama-3.1-8B, Mixtral-8×7B. The pattern is consistent across all idiom lengths: **(1)** the first idiom token w_1 has lower probability than the first generation token g_1. **(2)** Once w_1 is conditioned upon, the probability of subsequent idiom tokens rises steeply and surpasses the literal continuation; and **(3)** the literal generation probabilities remain relatively flat throughout.

3 Method: Lookahead-based Rescoring (LoRe)

3.1 Short-Horizon Sequence Scoring

To address token myopia, we propose augmenting the one-step decoding criterion with a short-horizon lookahead that evaluates the cumulative d-step log-probability of each candidate continuation.

Given the current context c and a set of candidate next tokens $\{v_1, \ldots, v_k\}$ (ranked by $p_\theta(v \mid c)$), we extend each candidate by greedily rolling out d additional steps:

$$\mathrm{LA}_d(v_i, c) = \sum_{t=1}^{d} \log p_\theta\Big(\hat{w}_t \;\Big|\; c, v_i, \hat{w}_1, \ldots, \hat{w}_{t-1}\Big) \tag{1}$$

The final score for candidate v_i is:

$$\mathrm{score}(v_i) = \underbrace{\log p_\theta(v_i \mid c)}_{\text{one-step}} + \underbrace{\mathrm{LA}_d(v_i, c)}_{\text{lookahead}} \tag{2}$$

Finally, the token with the highest cumulative score is selected.

Stochastic Activation. Full lookahead decoding is computationally expensive, requiring $k \cdot d$ additional forward passes per step. We therefore introduce a stochastic activation mechanism: at each decoding step, LoRe is activated with probability α and standard nucleus sampling is used otherwise. This allows continuous interpolation between pure nucleus sampling ($\alpha = 0$) and full lookahead decoding ($\alpha = 1$), and enables study of the cost–benefit trade-off as a function of α.

Min-p Lookahead. We use min-p sampling for candidate generation and greedy decoding for the rollout. This yields lookahead scores that better capture the diversity of the model's likely continuations while reducing the tendency to exploit degenerate, low-entropy trajectories. At each lookahead step, the next token is sampled greedily, and its log-probability is accumulated into LA_d. Algorithm 1 provides the pseudocode for min-p LoRe generation.

We adopt min-p sampling in place of top-p because its truncation threshold is adaptive to the confidence of the model's predictive distribution. Unlike top-p, which retains a fixed cumulative probability mass, min-p preserves tokens whose probability exceeds a fraction of the most likely token, allowing a broader set of plausible continuations when the distribution is uncertain while remaining selective when the model is confident. This produces more representative rollouts and improves exploration of alternative continuations. We integrate LoRe with top-p sampling as well in §A.7 and show that top-p allows tokens with very small probabilities to be chosen because of their likely continuations, making the generations less fluent and coherent.

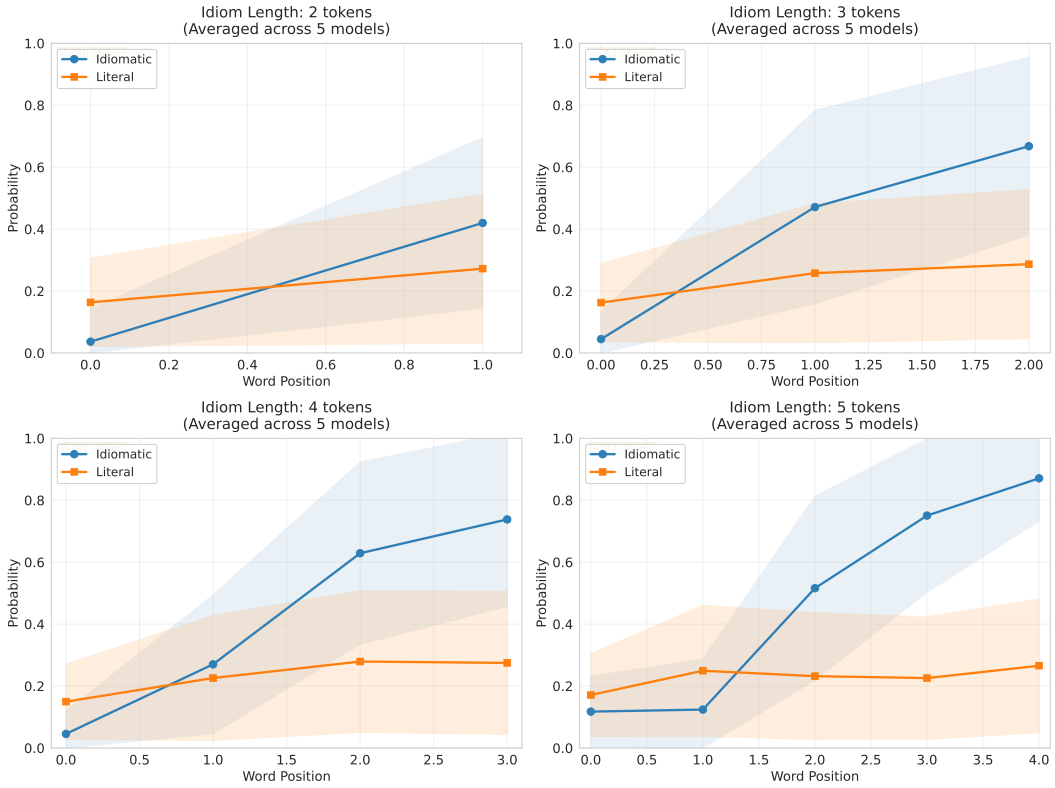

Figure 1: Incremental token probabilities for idiom sequences (blue) versus model-generated literal continuations (orange), averaged over 188 idioms. The shaded region shows one standard deviation. The first idiom token is consistently low-probability, while subsequent tokens rise sharply once the idiom is initiated. All five models exhibit the same pattern.

4 Experimental Setup

We aim to answer the following questions: (1) Does LoRe recover idiomatic and multi-word expressions more than min-p and top-p sampling? (2) Does LoRe preserve generation quality in non-creative settings? (3) What is the optimal activation probability α for LoRe?

4.1 Models

We conduct experiments on Pythia-12B, Llama-3.1-8B, Gemma-2-9B as our primary evaluation models. We use the base (non-instruct) versions to evaluate free-generation behavior without instruction-following biases.

4.2 Generation Datasets

Creative prompts. We sample context prefixes from two sources: (1) the BookMIA dataset (Shi et al., 2023), using the first 500 snippets with truncation to the first complete sentence; and (2) Project Gutenberg (manu, 2024), using 500 documents with header removal and truncation to the main body. Context prefixes are tokenized to 30 tokens. Contexts from both sources are shuffled with a fixed random seed (42) and cached.

Non-creative prompts. We prompt each model 1,000 times with their `<BOS>` token to test whether LoRe preserves generation quality in non-creative settings. We compare LoRe against:

Algorithm 1 Look-ahead Rescoring (LoRe)

Require: Model p_θ, tokenizer, context c, max length L, lookahead depth d, number of min-p candidates k, activation probability α, nucleus parameter p

seq $\leftarrow$ `tokenize`(c)
for $t = 1$ **to** L **do**
  if Bernoulli$(\alpha) = 1$ **then** ▷ Lookahead step
    Compute $\log p_\theta(v \mid \text{seq})$ for all $v \in \mathcal{V}$
    $\{(v_i, \ell_i)\}_{i=1}^{k} \leftarrow$ `top-`$k(\log p_\theta(\cdot \mid \text{seq}))$
    for $i = 1$ **to** k **do**
      $\text{LA}_i \leftarrow$ `GreedyRollout`$(p_\theta,\ \text{seq} \| v_i,\ d)$
      $\text{score}_i \leftarrow \ell_i + \text{LA}_i$
    end for
    $v^* \leftarrow v_{\arg\max_i \ \text{score}_i}$
  else ▷ Standard min-p step
    $v^* \sim$ `min-`p`-sample`$(\log p_\theta(\cdot \mid \text{seq}),\ p)$
  end if
  seq $\leftarrow$ seq$\|v^*$
  if $v^* =$ EOS **then break**
  end if
end for
return `decode`(seq)

(1) standard min-p sampling with $p = 0.02$, (2) standard top-p with $p = 0.95$ and temperature $= 1.0$.

4.3 Generation Setup

We use min-p with $p = 0.02$ as it filters out highly unlikely tokens while still keeping reasonable diversity, and LoRe lookahead depth of 4 because most multi-word expressions are 2-4 words in length (Baldwin & Kim, 2010), making this depth sufficient to capture creative idiomatic expressions while maintaining computational efficiency. For each activation probability $\alpha \in \{0.0, 0.1, 0.2, \ldots, 1.0\}$, we generate **1,000 stories** of up to 256 tokens each.

4.4 Metrics

We use LLM-as-a-Judge (LaaJ) to compare generated stories across four dimensions against our top-p and min-p baselines. The exact prompts used for these evaluations are provided in §A.2.

- **Creativity** measures the originality of ideas, use of interesting imagery, narrative experimentation, and avoidance of clichés.
  - **Rare MWE Usage** counts the number of rare and creative MWEs, which we consider as a sub-category of creativity,.
- **Fluency** assesses the naturalness of the prose, grammatical correctness, vocabulary usage, and overall readability.
- **Coherence** evaluates logical narrative consistency, clear connections between sentences, absence of contradictions, and smooth transitions between ideas.

For creativity, fluency, and coherence, LaaJ performs pairwise comparisons between stories generated with LoRe at different activation probabilities and those generated by the baselines. We evaluate eight activation probabilities, $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$, against each baseline across 1,000 stories, yielding 8,000 pairwise comparisons per metric for each baseline. The presentation order of each story pair is randomized to mitigate position bias. We use pairwise comparisons rather than absolute scoring because the relatively lower generation quality of our LLM lineup makes absolute quality judgments more difficult to calibrate reliably; comparing two stories directly provides a more consistent basis for assessing their relative quality.

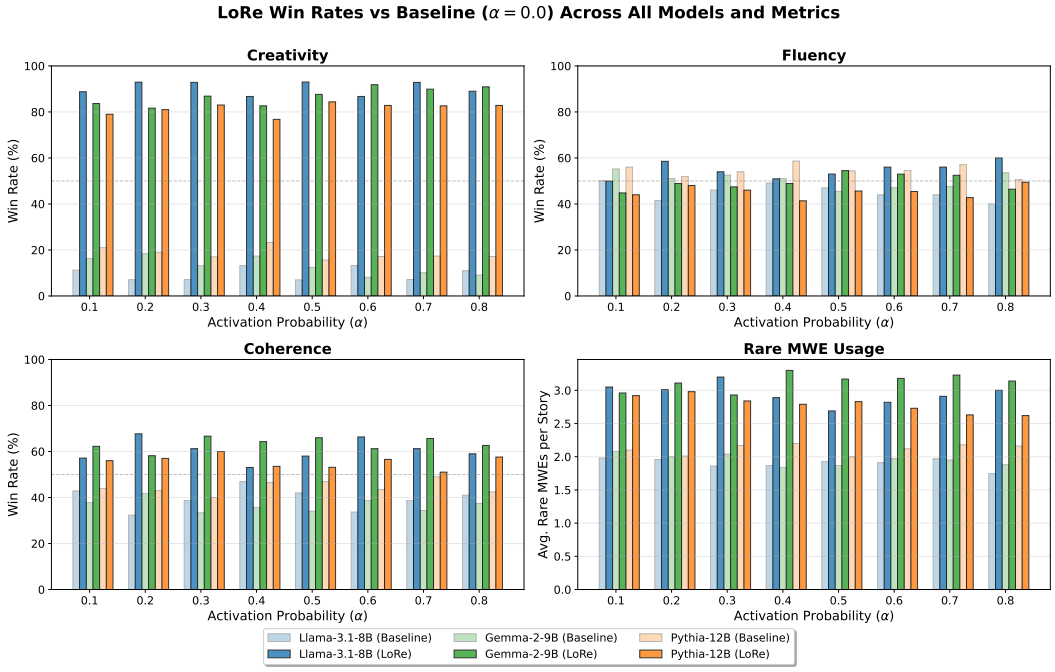

Figure 2: LoRe win rates versus min-p baseline ($\alpha = 0.0$) across quality metrics for Llama-3.1-8B, Gemma-2-9B, and Pythia-12B. The first three panels show the percentage of 1,000 pairwise comparisons where LoRe was judged superior using Claude-3-Haiku as judge. The fourth panel shows the average number of rare/creative MWEs per story for LoRe and baseline (lighter bars).

For rare MWE usage, LaaJ instead performs a counting-based evaluation. Given each generated story, LaaJ identifies the MWEs and idiomatic expressions that occur naturally and appropriately in context and reports their counts.

To verify that the evaluations produced by Claude-3-Haiku are consistent with human judgment, we manually evaluate 10 stories for each activation probability. Two fluent English speakers independently assess creativity, fluency, and coherence through pairwise comparisons, and also count the number of rare MWEs in each story. We then measure both agreement among human annotators and agreement between the aggregated human judgments and the LLM's evaluations. The results, reported in Section A.8, indicate substantial agreement across all evaluation dimensions, suggesting that the LLM's assessments are broadly aligned with human judgments.

5 Results

5.1 Recovery of rare multi-word expressions

We first examine whether LoRe successfully recovers rare and creative multi-word expressions (MWEs) that standard decoding methods under-generate. We count rare MWE usage using a language model judge (Claude-3-Haiku) which counts and reports the use of creative, uncommon expressions while avoiding clichés. LoRe demonstrates strong and consistent improvements in rare MWE usage across all activation probabilities. At $\alpha = 0.8$, both Llama-3.1-8B and Gemma-2-9B achieve 77-78% win rates against baseline, meaning LoRe produces more effective rare idiom usage in over three-quarters of comparisons. LoRe generations contain an average of 2.9-3.2 rare MWEs per story compared to baseline's 1.9, representing a 50-65% increase in creative expression usage on average. Even at moderate activation probabilities, performance remains robust: at $\alpha = 0.4$, models achieve 72-80% win rates (average 76%), demonstrating that occasional lookahead activation is sufficient to recover creative expressions. Notably, the win rates remain relatively flat across activation probabilities (72-78%), suggesting that the quality of rare MWE usage saturates quickly. This

indicates that even infrequent LoRe activation ($\alpha \approx 0.2$–0.4) captures most benefits by occasionally steering generation toward less common multi-token sequences that would otherwise be pruned by standard sampling.

5.2 Generation Quality Across Metrics

Figure 2 shows win rates for each metric across activation probabilities.

Creativity. LoRe dramatically improves creativity, achieving 84-93% win rates across all activation probabilities and models (overall average: 88.6%). The optimal activation probability is $\alpha = 0.7$ (91.4% average win rate), though performance remains strong across the entire range. The relatively flat trend across α values suggests that even moderate lookahead activation captures substantial creativity benefits, likely by exploring less common continuation paths during rollout scoring.

Coherence. Coherence improvements are more modest but consistent, with win rates of 58-64% across activation probabilities (average: 61.9%). The optimal setting is $\alpha = 0.3$ (63.9% win rate). The moderate gains suggest that LoRe successfully enhances creativity without hurting consistency.

Fluency. Fluency performance remains comparable to baseline across all activation probabilities, with win rates averaging 47.2% (range: 42-50%). At the optimal $\alpha = 0.6$, LoRe achieves 49.5% win rate—essentially equivalent to baseline. This near-neutral performance indicates that LoRe maintains fluency while substantially improving creativity, avoiding the common trade-off where creative generation methods degrade language quality. The consistency of near-50% win rates across all α values demonstrates that lookahead scoring does not compromise grammatical correctness or readability despite prioritizing novel, less common expressions.

Interpretation. The relatively flat trends across activation probabilities, especially for creativity and rare MWE usage, indicate that even low α values (0.1–0.3) capture most benefits. This occurs because LoRe only needs to activate occasionally at critical decision points (e.g., when initiating idiomatic phrases) to steer generation toward creative outcomes. For instance, creativity win rates at $\alpha = 0.1$ (86.2%) are only 5 percentage points lower than at $\alpha = 0.7$ (91.4%).

For balanced performance, $\alpha \approx 0.5$–0.7 appears optimal, achieving strong creativity (89-91% win rates) and rare MWE usage (74-77%) while maintaining acceptable coherence (62-63%). Lower values ($\alpha < 0.3$) sacrifice some creativity gains, while higher values ($\alpha > 0.7$) provide minimal additional benefit given the saturation effect. LoRe proves most effective for creative generation tasks requiring diverse, idiomatic language. For applications demanding maximal fluency or formal correctness (e.g., technical documentation), the minimal fluency improvements may limit its utility. The strong creativity and rare MWE recovery demonstrate that LoRe successfully addresses the systematic under-generation of long-tail expressions in standard decoding, while the coherence improvements confirm it maintains narrative quality through multi-token lookahead scoring.

5.3 Non-creative Generation Quality

Figure 3 shows that LoRe preserves fluency and coherence in non-creative generation while yielding a modest improvement in creativity. Across activation probabilities, the scores remain comparable to those of the baseline decoding methods, indicating that the lookahead-based rescoring mechanism does not degrade overall generation quality. These results suggest that LoRe can increase the use of multi-token lexical units without sacrificing the readability or consistency of the generated text.

6 Related Work

Classic decoding methods. The quality of language model generations is strongly influenced by the decoding algorithm used at inference time. Classical approaches such as greedy decoding, beam search, top-k sampling (Fan et al., 2018), and nucleus (top-p) sampling (Holtzman et al., 2020) attempt to balance generation quality and diversity by restricting the next-token distribution. Subsequent work has proposed adaptive truncation and entropy-aware sampling strategies, including Miro-

stat (Basu et al., 2021), η-sampling (Hewitt et al., 2022), and Min-p sampling (Nguyen et al., 2025). These methods primarily focus on mitigating degeneration caused by unreliable low-probability tails of the token distribution.

Nonlocal decoding methods. There has been recent acknowledgement of the myopic nature of traditional decoding in the planning domain (Ma et al., 2025). However, this issue has not yet been investigated in the context of open-ended generation. We show that token myopia also impacts open-ended generation specifically through the under-generation of long-tail phrases. Separately but in a similar spirit, decoding methods such as DeLorean (Qin et al., 2020), FUDGE (Yang & Klein, 2021), and COLD (Qin et al., 2022) condition the current token generated on information such as future constraints or desired attributes without explicit rollouts.

Lookahead and Sequence-Level Decoding. Several prior approaches incorporate information about future continuations when selecting the current token. Beam search maintains multiple partial sequences and scores them jointly, but is known to favor generic or degenerate outputs in open-ended generation (Holtzman et al., 2020). Similar to our approach, NeuroLogic A*esque decoding uses lookahead heuristics to estimate future constraint satisfaction (Lu et al., 2022). Moreover, Predictive-Decoding frames autoregressive generation as a model-predictive-control problem and uses sampled future trajectories to guide current decisions (Ma et al., 2025), and EPIC uses properties of future trajectories, in particular their entropy, to calibrate the decoding distribution (Ahmed & Singh, 2026). Related work has also explored reducing the computational cost of lookahead-based constrained decoding through speculative execution (Nakshatri et al., 2025). Our work builds on the general idea that next-token probabilities alone can be insufficient, but targets a different failure mode. Rather than optimizing for external constraints or future plans, we study open-ended and creative generation.

Idioms, Multi-Word Expressions, and Long-Tail Generation. Prior work has extensively studied how language models represent and understand idioms and multi-word expressions (MWEs), including idiom identification, sense disambiguation between idiomatic and literal, and probing of contextual representations (Saxena & Paul, 2020; Zeng & Bhat, 2022; Liu & Neubig, 2022; Kabra et al., 2023; Zeng & Bhat, 2023; Miletić & Schulte im Walde, 2024; Zhou et al., 2024; Almheiri et al., 2026). However, a comparatively smaller body of work has considered generation of idioms, for example through idiomatic sentence generation and paraphrasing (Zhou et al., 2021; Zhou & Bhat, 2024), controlled generation of figurative language (Lai & Nissim, 2022), and explicit idiom insertion (Zhou et al., 2023; Gao et al., 2026). These settings typically ask models to deliberately produce text containing an idiom. In contrast, we study the *spontaneous generation* of such expressions. More broadly in the realm of long-tail expressions, recent work has documented persistent weaknesses of language models on rare lexical items in the long tail (Algayres et al., 2025). We therefore use idioms as a measurable instance of long-tail lexical constructions.

7 Conclusion

In this paper, we examine the problem of token myopia, the tendency for models to under-generate multi-token lexical units when the first token has a low probability. We uncover this problem using idiomatic expressions as probes, and propose LoRe, a simple method which improves creativity, and coherence on creative open-ended generation, which retaining the same level of fluency. Our findings suggest that decoding decisions can limit the expression of linguistic knowledge already encoded in a model, but that we can also recover more human-like text generation capabilities by applying simple fixes like short-horizon lookahead.

AI use statement

In this work, we used generative AI tools as part of our methodology: Claude-3.5-Sonnet classified literal versus idiomatic usages when constructing the idiom set (§2.1), and Claude-3-Haiku served as the judge for all LLM-as-a-Judge evaluations (§4.4). We also used LLMs to help search for related literature.

Additionally, we used generative AI tools to improve the clarity and grammar of the manuscript, and for implementing parts of Python plotting scripts. We have reviewed all AI-assisted work: the authors reviewed all LLM-assisted text, verified every reference against its original source, and manually checked a sample of the judge's outputs, including a human–judge agreement study (§A.8). We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

We did not use AI to Generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, or interpret results.

REFERENCES

Kareem Ahmed and Sameer Singh. Entropy-aligned decoding of lms for better writing and reasoning. *arXiv preprint arXiv:2601.01714*, 2026.

Robin Algayres, Charles-Éric Saint-James, Mahi Luthra, Jiayi Shen, Youssef Benchekroun, Dongyan Lin, Rashel Moritz, Juan Pino, and Emmanuel Dupoux. LongTail-Swap: Benchmarking language models' abilities on rare words. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 11231–11251. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.601. URL https://aclanthology.org/2025.findings-emnlp.601/.

Saeed Almheiri, Bilal Elbouardi, Salsabila Zahirah Pranida, Irina Nikishina, Parameswari Krishnamurthy, Muhammad Cendekia Airlangga, Rifo Ahmad Genadi, Nguyen Phan Gia Bao, Amir Hossein Yari, Hawau Olamide Toyin, et al. Multilingual idioms in sentences and conversations across high-, medium-, and low-resource languages. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12363–12389, 2026.

Dilip Arumugam, Peter Henderson, and Pierre-Luc Bacon. An information-theoretic perspective on credit assignment in reinforcement learning, 2021. URL https://arxiv.org/abs/2103.06224.

Mena Attia, Aashiq Muhamed, Mai Alkhamissi, Thamar Solorio, and Mona Diab. Beyond understanding: Evaluating the pragmatic gap in llms' cultural processing of figurative language. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7238–7265, 2026.

Timothy Baldwin and Su Nam Kim. Multiword expressions. In *Handbook of Natural Language Processing*, pp. 267–292. CRC Press, Boca Raton, FL, 2nd edition, 2010.

Sourya Basu, Govardana Sachitanandam Ramachandran, Nitish Shirish Keskar, and Lav R. Varshney. Mirostat: A neural text decoding algorithm that directly controls perplexity. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2007.14966.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies (Volume 1: Long Papers), pp. 2016–2039, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.112. URL https://aclanthology.org/2024.naacl-long.112/.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

Hui Gao, Changhao Song, Peng Zhang, Jing Zhang, Chang Yang, and Liuxian Ge. LLM-SLM collaborative framework of idiomatic expression generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12125–12145. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.555. URL https://aclanthology.org/2026.acl-long.555/.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Gemma Team. Gemma 2: Improving open language models at a practical size, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, et al. The Llama 3 herd of models, 2024.

John Hewitt, Christopher D. Manning, and Percy Liang. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022. URL https://arxiv.org/abs/2210.15191.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*, 2020.

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

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. Multi-lingual and multi-cultural figurative language understanding. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8269–8284, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.525. URL https://aclanthology.org/2023.findings-acl.525/.

Jisu Kim, Youngwoo Shin, Uiji Hwang, Jihun Choi, Richeng Xuan, and Taeuk Kim. Memorization or reasoning? exploring the idiom understanding of llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21689–21710, 2025.

Huiyuan Lai and Malvina Nissim. Multi-figurative language generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 5939–5954. International Committee on Computational Linguistics, 2022. URL https://aclanthology.org/2022.coling-1.519/.

Emmy Liu and Graham Neubig. Are representations built from the ground up? an empirical examination of local composition in language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9053–9073, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.617. URL https://aclanthology.org/2022.emnlp-main.617/.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. Testing the ability of language models to interpret figurative language. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4437–4452, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.330. URL https://aclanthology.org/2022.naacl-main.330/.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens, 2025. URL https://arxiv.org/abs/2401.17377.

Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, Noah A. Smith, and Yejin Choi. NeuroLogic a*esque decoding: Constrained text generation with lookahead heuristics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 780–799. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.57. URL https://aclanthology.org/2022.naacl-main.57/.

Chang Ma, Haiteng Zhao, Junlei Zhang, Junxian He, and Lingpeng Kong. Non-myopic generation of language models for reasoning and planning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 35031–35058, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/56b694fb10c02b177e75a41f45825a74-Paper-Conference.pdf.

manu. Project gutenberg. Hugging Face dataset repository, 2024. URL https://huggingface.co/datasets/manu/project_gutenberg. Accessed: 2026-08-23.

Filip Miletić and Sabine Schulte im Walde. Semantics of multiword expressions in transformer-based models: A survey. *Transactions of the Association for Computational Linguistics*, 12: 593–612, 2024. doi: 10.1162/tacl_a_00657. URL https://aclanthology.org/2024.tacl-1.33/.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. Contrasting linguistic patterns in human and LLM-generated news text. *Artificial Intelligence Review*, 57:265, 2024. doi: 10.1007/s10462-024-10903-2.

Nishanth Sridhar Nakshatri, Shamik Roy, Rajarshi Das, Suthee Chaidaroon, Leonid Boytsov, and Rashmi Gangadharaiah. Constrained decoding with speculative lookaheads. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4681–4700. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.239. URL https://aclanthology.org/2025.naacl-long.239/.

Minh Nhat Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent llm outputs. In *International Conference on Learning Representations (ICLR)*, 2025.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers, 2021. URL https://arxiv.org/abs/2102.01454.

Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena D. Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 794–805, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.58. URL https://aclanthology.org/2020.emnlp-main.58/.

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. Cold decoding: Energy-based constrained text generation with langevin dynamics. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 9538–9551. Curran Associates, Inc., 2022. doi: 10.52202/

068431-0693. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/3e25d1aff47964c8409fd5c8dc0438d7-Paper-Conference.pdf.

Qwen Team. Qwen3 technical report, 2025.

Alex Reinhart, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. Do LLMs write like humans? variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8):e2422455122, 2025. doi: 10.1073/pnas.2422455122. URL https://www.pnas.org/doi/abs/10.1073/pnas.2422455122.

Prateek Saxena and Soma Paul. EPIE dataset: A corpus for possible idiomatic expressions. In Petr Sojka, Ivan Kopeček, Karel Pala, and Aleš Horák (eds.), *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020*, volume 12284 of *Lecture Notes in Computer Science*, pp. 87–94, Brno, Czech Republic, 2020. Springer. doi: 10.1007/978-3-030-58323-1_9.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models, 2023.

Rita C. Simpson, Sarah L. Briggs, Janine Ovens, and John M. Swales. The michigan corpus of academic spoken english (micase), 2002. Available at the Michigan Corpus of Academic Spoken English.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.840. URL https://aclanthology.org/2024.acl-long.840/.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018.

Paul Thompson and Hilary Nesi. The British Academic Spoken English (BASE) corpus project. *Language Teaching Research*, 5(3):263–264, 2001. doi: 10.1177/136216880100500305.

Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. Redpajama: an open dataset for training large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 116462–116492. Curran Associates, Inc., 2024. doi: 10.52202/079017-3697. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d34497330b1fd6530f7afd86d0df9f76-Paper-Datasets_and_Benchmarks_Track.pdf.

Kevin Yang and Dan Klein. FUDGE: Controlled text generation with future discriminators. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3511–3535, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.276. URL https://aclanthology.org/2021.naacl-main.276/.

Ziheng Zeng and Suma Bhat. Getting BART to ride the idiomatic train: Learning to represent idiomatic expressions. *Transactions of the Association for Computational Linguistics*, 10: 1120–1137, 2022. doi: 10.1162/tacl_a_00510. URL https://aclanthology.org/2022.tacl-1.65/.

Ziheng Zeng and Suma Bhat. Unified representation for non-compositional and compositional expressions. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11696–11710, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.783. URL https://aclanthology.org/2023.findings-emnlp.783/.

Jianing Zhou and Suma Bhat. Non-compositional expression generation and its continual learning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2828–2839, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.166. URL https://aclanthology.org/2024.findings-acl.166/.

Jianing Zhou, Hongyu Gong, and Suma Bhat. PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pp. 33–48. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.mwe-1.5. URL https://aclanthology.org/2021.mwe-1.5/.

Jianing Zhou, Ziheng Zeng, Hongyu Gong, and Suma Bhat. Non-compositional expression generation based on curriculum learning and continual learning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4320–4335, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.286. URL https://aclanthology.org/2023.findings-emnlp.286/.

Jianing Zhou, Ziheng Zeng, Hongyu Gong, and Suma Bhat. Enhancing language models with idiomatic reasoning. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=2wtj0up8rv.

A Appendix

A.1 Representation Separability Analysis

To investigate whether models possess internal representations of idiomaticity despite limited idiom generation, we extracted mean-pooled hidden states from all layers of language models and trained linear classifiers to distinguish between idiomatic and literal sentences.

A.1.1 Experimental Setup

- **Dataset:** 500 balanced idiom-literal contrast pairs (1,000 total sentences)
- **Composition:** Each pair contains the same idiom phrase appearing in different sentence contexts. Both sentences in each pair use the identical idiom surface form, but differ in whether it is used literally or figuratively (See examples in Table 2).
- **Classifier:** Logistic Regression (max_iter=1000, random_state=42)
- **Train-Test Split:** 70/30, stratified by class
- **Hidden State Processing:** Mean pooling across all token positions per sentence

A.1.2 Procedure

1. **Representation Extraction:** For each sentence, we extracted hidden states from all layers of the model and computed mean-pooled representations across all token positions.
2. **Classification:** For each layer, we trained a linear logistic regression classifier (70/30 train-test split, stratified by class) to predict whether a sentence is idiomatic or literal based solely on its hidden state representation.

3. **Separability Metrics:**
   - *Linear Classification Accuracy:* Performance of logistic regression on test set
   - *Normalized Centroid Distance:* Euclidean distance between class centroids, normalized by average within-class variance

Table 2: Example contrast pairs from EPIE dataset. Each pair contains the same idiom phrase used in literal vs. idiomatic contexts.

Idiom	Literal Usage	Idiomatic Usage
all hell broke loose	When the demons were unleashed from the underworld, all hell broke loose on Earth.	The large man's friends rose from their table and the waiters formed themselves into a solid phalanx, and all hell broke loose.
off the hook	The fish wriggled and managed to get itself off the hook.	But some bruised wives may feel that Vaughan is almost too ready to let men off the hook; many would like to see their partner shed a little retributory blood.
against the grain	Carpenters recommend not to sand against the grain as it can damage the wood.	The Smiths run directly against the grain.

A.2 LLM-AS-A-JUDGE EVALUATION PROMPTS

We used Claude-3-Haiku via OpenRouter to perform pairwise comparisons between stories generated with different methods. For each comparison, stories were randomly shuffled to avoid position bias. Below are the exact prompts used for each metric.

A.3 CREATIVITY EVALUATION PROMPT

```
You are evaluating AI-generated creative writing for creativity.
Compare these two stories and determine which is better.

Consider:
- Originality of ideas and concepts
- Interesting imagery or descriptions
- Narrative structure and experimentation
- Genre exploration
- Avoidance of clichés and repetition

Story A (p=X):
[story text]

---

Story B (p=Y):
[story text]

---

Which story is more creative? Respond with ONLY "Story A" or
"Story B" followed by a brief explanation.
Format: [Story A/Story B] - [one sentence explanation]

Your response:
```

A.4 FLUENCY EVALUATION PROMPT

```
You are evaluating AI-generated creative writing for fluency.
```

```
Compare these two stories and determine which is better.

Consider:
- Natural, smooth prose that flows well
- Grammatically correct sentences
- Appropriate word choice and vocabulary
- Natural rhythm and pacing
- Readability and ease of comprehension
- Absence of awkward phrasings or unnatural constructions
- Professional writing quality

Story A (p=X):
[story text]

---

Story B (p=Y):
[story text]

---

Which story is more fluent and natural? Respond with ONLY "Story A"
or "Story B" followed by a brief explanation.
Format: [Story A/Story B] - [one sentence explanation]

Your response:
```

A.5 Coherence Evaluation Prompt

```
You are evaluating AI-generated creative writing for coherence.
Compare these two stories and determine which is better.

Consider:
- Logical flow and narrative consistency
- Clear connections between sentences and paragraphs
- Absence of contradictions or non-sequiturs
- Consistent characters, settings, and plot points
- Smooth transitions between ideas
- Overall readability and comprehensibility
- Lack of repetition or loops

Story A (p=X):
[story text]

---

Story B (p=Y):
[story text]

---

Which story is more coherent and cohesive? Respond with ONLY
"Story A" or "Story B" followed by a brief explanation.
Format: [Story A/Story B] - [one sentence explanation]

Your response:
```

A.6 MWE USAGE EVALUATION PROMPT

```
You are an expert evaluator of creative language use.
Compare these two story continuations on RARE MWE USAGE.

RARE MWE USAGE: Evaluate which story uses more rare,
creative, and non-clichéd multi-word expressions.
What to look for:

Idioms (figurative expressions):
"the hair stood up on the back of his neck" > "kick the bucket"

Creative phrasal verbs:
"storm-tossed" > "turn on"

Vivid collocations:
"butterflies in the pit of his stomach" > "take a look"

Novel/uncommon expressions vs. overused clichés
Prefer stories that:
Use less common, more creative MWEs
Avoid overused clichés like "beat around the bush",
"break the ice", "kick the bucket"

Use vivid, context-specific expressions
Show variety (don't repeat the same MWE)
Avoid counting as rare MWEs:

Very common phrasal verbs: "turn on", "get up", "look at"

Overused idioms: "kick the bucket", "beat around the bush", "break the ice"

Generic collocations: "make a decision", "take a look"

PROMPT: {prompt}

STORY A:
[story text]

---

STORY B:
[story text]

---

Which story uses more rare, creative, and non-clichéd multi-word expressions?
Provide your answer in this EXACT format:

WINNER: [A/B/TIE]

RARE_MWE_COUNT_A: [integer count of rare/creative MWEs]

RARE_MWE_LIST_A: [comma-separated list, or "none"]

RARE_MWE_COUNT_B: [integer count of rare/creative MWEs]

RARE_MWE_LIST_B: [comma-separated list, or "none"]
```

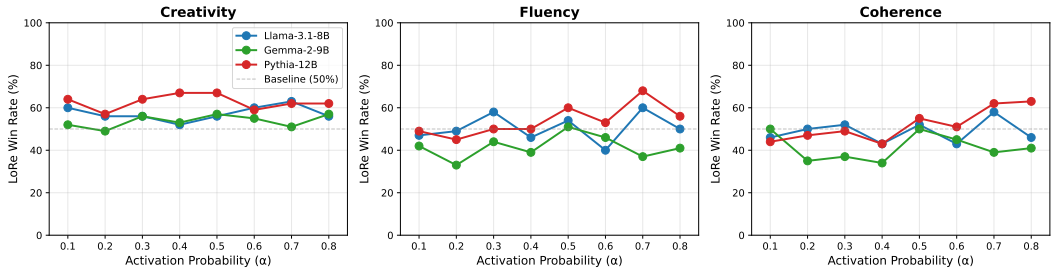

Figure 3: LoRe win rates for <BOS> prompts. LoRe does not degrade the quality of generation for non-creative generation.

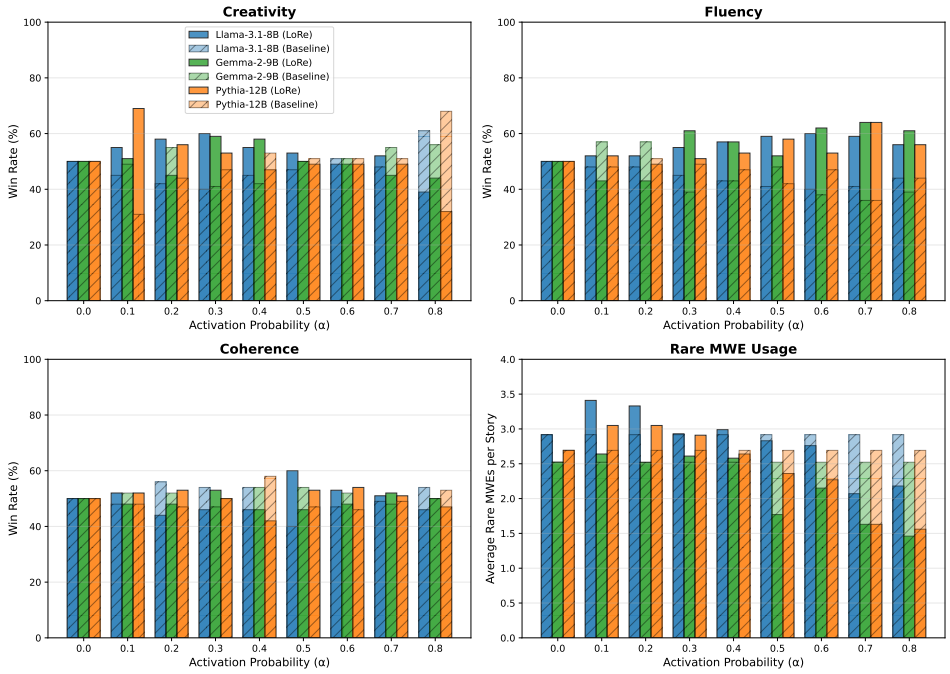

Figure 4: LoRe win rates versus top-p baseline ($\alpha = 0.0$) across quality metrics for Llama-3.1-8B, Gemma-2-9B, and Pythia-12B. The first three panels show the percentage of 1,000 pairwise comparisons where LoRe was judged superior using Claude-3-Haiku as judge. The fourth panel shows the average number of rare/creative MWEs per story for LoRe and baseline (lighter bars)

```
EXPLANATION: [1-2 sentences explaining which story's
MWEs are more creative and why]
```

A.7 LoRe with Top-p Sampling

We integrate LoRe with top-p sampling by performing rollouts for the tokens chosen by top-p and also using top-p for normal token generations. Figure 4 shows the LLLM-as-a-Judge comparisons and MWE usage counts for this experiments. The improvements are modest compared to LoRe with min-p. This is because top-p sampling let's twice as many tokens through compared to min-p for each token generation. This lets through unlikely tokens that may have likely continuations. The likely continuations artificially bump us the probability of unlikely tokens, making them more likely to get picked, which ultimately results worse performance compared to baseline in higher activation rates.

Table 3: Unanimous human agreement rates and agreement between LLM evaluations and aggregated human judgments. For rare MWE usage, agreement is computed with a tolerance of one expression.

Agreement Type	Creativity	Fluency	Coherence	Rare MWE Usage
Human–Human	80%	72%	65%	79%
Human–LLM	75%	64%	62%	70%

A.8 Human–Judge Agreement

To assess the reliability of our evaluation protocol, we recruited two human annotators to compare 10 stories generated at each activation probability against stories produced by the baseline decoding method. For creativity, fluency, and coherence, inter-human agreement was computed as the percentage of comparisons for which the two annotators selected the same preferred story. For rare MWE usage, agreement was defined as the percentage of instances in which annotators' counts differed by at most one expression.

To evaluate the alignment between human and LLM judgments, we aggregated human evaluations using majority voting for creativity, fluency, and coherence, and by averaging rare MWE counts. We then compared these aggregated human judgments with the corresponding LLM evaluations. As shown in Table 3, human annotators exhibit substantial agreement across all evaluation dimensions, with the highest agreement observed for creativity (80%) and rare MWE usage (79%). Agreement is lower for coherence (65%), suggesting that coherence judgments are inherently more subjective and difficult to assess consistently.

The LLM demonstrates strong alignment with human judgments, achieving agreement rates comparable to inter-human agreement for creativity and rare MWE usage.