

# Transferring Extreme Subword Style Using Ngram Model-Based Logit Scaling

**Craig Messner**

Center for Digital Humanities  
Johns Hopkins University  
cmessne4@jhu.edu

**Tom Lippincott**

Center for Digital Humanities  
Johns Hopkins University  
tom.lippincott@jhu.edu

## Abstract

We present an ngram model-based logit scaling technique that effectively transfers extreme subword stylistic variation to large language models at inference time. We demonstrate its efficacy by tracking the perplexity of generated text with respect to the ngram interpolated and original versions of an evaluation model. Minimizing the former measure while the latter approaches the perplexity of a text produced by a target author or character lets us select a sufficient degree of adaptation while retaining fluency.

## 1 Introduction

Text style transfer (TST) aims to reformulate a source text using the stylistic attributes of a given target text. Authors vary a blend of these attributes to achieve a literary effect, with some modifications being more conspicuous than others. Stylistic modification of subword units like characters or phonemes can prove especially noticeable.<sup>1</sup>

One such subword style is orthographic variation, a phenomenon common in forms of "dialect literature" present throughout history but especially popular in the 19th century United States (Krapp, 1925) (Ives, 1971). These works utilize context and readerly knowledge to render their orthographic innovations legible despite their extreme modification of orthographic norms (Sebba, 2007). We present a subword-level ngram-based logit scaling method that effectively transfers this form of extreme style at decoding time. We accomplish this by combining the next-token information derived from a large language model (LLM) with information obtained from ngram models trained on a single-author/character corpus.

Ngram models are quick to train, data-efficient, and interpretable. Training ngram models on small

<sup>1</sup>In this paper, all fundamental units are subwords, as such references to tokens refers to subword tokens produced by subword tokenization methods.

single-author corpora re-purposes them as statistical experts, reflections of the constructions a given author is likely to employ. These qualities are especially useful when transferring style from low frequency or novel sources. LLMs may have little or no information about these styles in their weights, and style-specific corpora may be too small to support finetuning.

We introduce a scaled interpolation method that combines weighted ngram model predictions with those from pretrained LLMs to generate fluent stories that match the extreme subword style of particular dialect authors and characters. We also demonstrate how to tune and evaluate these transfers using perplexity measures.

## 2 Related Work

Techniques like finetuning on further data (Mukherjee et al., 2024), prompt editing (Luo et al., 2023) and in-context learning (Mai et al., 2023) have been used to achieve TST. While potentially effective, these avenues require further computation and additional training data. Mechanical interpretation approaches can provoke style at inference time by intervening on model weights (Lai et al., 2024). However, this approach requires the target style to be in-distribution and suitably represented in the model. Other recent works have re-evaluated LM approaches previously considered obsolete in the light of computational and theoretical advances. Ngram modeling has been revisited for LM smoothing (Malagutti et al., 2024) and in "infinite" form as a interpolation component used to complement LLMs (Liu et al., 2024).

## 3 Methodology

We achieve subword TST by applying an ngram model-derived scaling factor to the output logits of an LLM prior to softmaxing and sampling. Vitaly, information from the ngram model must contribute

to the next-token probability without warping the LLM’s ability to produce fluent text. We ensure both fluency and transfer by scaling the ngram model-provided next token prediction with an additional factor  $f$ . Given a vocabulary  $V$ , we calculate the scaling factor using equation 1

$$S = -\frac{f}{\log(p_n)} n \in \{0 \dots |V|\} \quad (1)$$

Inspired by temperature decoding methods, the addition of the parameter  $f$  uniformly increases the scaling factor as  $f$  increases, leading to a higher proportion of generation information being derived from the ngram model. This also renders the scaling mixture "tunable".

## 4 Experiments

### 4.1 Setup

**Data.** We use two 19th century U.S. fiction corpora sourced from Project Gutenberg (PG) for baseline evaluation and ngram model training. The first consists of the full text of *Uncle Remus: His Songs and His Sayings* by Joel Chandler Harris and all of Peter Finley Dunne’s *Mr. Dooley* series of stories. The second corpus consists of dialogue employed by three characters identified as nonstandard speakers. Messner extracted these dialogue sections from PG and attributed them manually. See Appendix A for more corpus details. We style the former corpus with lower casing and the latter with upper. We wordpiece tokenize both using the model-specific functions supplied by the Transformers library.

We generate story prompts and establish standard English baseline scores using the WRITING-PROMPTS (WP) dataset (Fan et al., 2018). We sample 50 prompts from the dataset to guide story generation. We modify each original prompt with a brief instruction stem in order to produce the scaled generation prompt set. Additionally, we apply three character/author templates to each prompt in order to produce a control prompt set. The first indicates that the model should act as a storyteller, and includes a description of its era and position. The second adds information about the target author, and the third the target character. See Appendix B for more prompt creation details.

**Models.** For generation of control and scaled texts, we use MistralAi’s Mistral-Instruct-7B-v0.2 (Mistral) (Jiang et al., 2023) and Meta’s Llama3.2-3B-Instruct (Llama) (Dubey et al., 2024) pre-trained instruction-tuned models. For perplexity

evaluation, we use OpenAI’s GPT2-large (GPT2) (Radford et al., 2019). We obtain model checkpoints via HuggingFace. Using the wordpieced target texts, we train a set of ngram models,  $\{M_n, M_{n-1}, \dots, M_1\}$ , with  $n = 4$ . When scaling generated logits, we employ the model set in a backoff configuration. If  $M_4$  cannot make a 4gram next token prediction, we use a trigram prediction from  $M_3$ , and so on. If no model can make a prediction, no scaling is performed. This is essentially a modification of stupid backoff (Brants et al., 2007).

**Evaluation.** We concatenate the tokens produced by each scaled or control generation and estimate their GPT2 perplexity using a sliding context window of 32 tokens with stride of 1. We do the same for the WP test-set baseline and GB target texts. When scoring general model performance, low perplexity is considered preferable. For our purposes, near-equal target ( $PPL()$ ) and generation ( $gPPL()$ ) perplexities indicate successful subword style transfer. We also measure the perplexity of the original texts using the interpolation of GPT2 and each target text’s scaled ngram models,  $rPPL()$ . Combining these two sources of information allows us to select an optimal schedule of scalings for subword style transfer by maximizing  $abs(PPL() - gPPL())$  while minimizing  $rPPL()$ . Intuitively, the first measure acts as an early stopping criterion, while the second measure indicates whether the  $gPPL()$  at a given scaling is produced by transfer and not chaotic.

### 4.2 Procedure.

We define a 16-member weight set  $W$ , where each  $w \in W$  is a tuple of the form  $\{f_4, f_3, f_2, f_1\}$ . Each  $f$  is drawn from  $\{0, 1, 2\}$  and used to scale the next token predictions  $p$  of its corresponding ngram set model  $M_n$  using Equation 1.  $f$  of 0 omits the corresponding model. For example,  $w$  of  $\{0, 0, 2, 1\}$  applies Equation 1 with  $f = 2$  and  $f = 1$  to the bigram and unigram model next-token probabilities respectively. This results in a scaling vector  $S$  with length  $|V|$ . We add  $S$  to the logits of the LLM’s next token prediction and repeat the process up to a maximum generation length of 256 tokens. We perform  $|W|$  of these scaled generations for each prompt in the base set, using a different  $w$  each time. We repeat this process over two conditions: decoding greedily and sampling. We calculate  $gPPL()$  and  $rPPL()$  and then graphically determine the weight set(s) of best fit for a given target character or author by plotting

Target	$PPL()$	N Tokens
remus	106.54	82365
dooley	110.03	366037
Todd	49.88	12273
Remus	128.68	48217
Julius	166.51	11350
WP	41.01	12456693

Table 1: Baseline results. Top section: full texts from GB. Middle section: Dialogue extracted from GB. Bottom section: WP (standard) baselines

$abs(PPL() - gPPL())$  against  $rPPL()$ .<sup>2</sup>

## 5 Results and Discussion

### 5.1 Baselines and target styles

The  $PPL()$  of the baseline (WP) and variant target texts (GB) greatly differ (Table 1). The target texts are considerably more perplexing, at least in part due to the modifications they employ at the subword level. Consequently, a  $gPPL()$  more similar to the target  $PPL()$  than the baseline WP  $PPL()$  indicates that style was likely transferred.

### 5.2 Generation conditions and model specificity

Neither scaled LLM produces text with  $gPPL()$  approaching its particular target  $PPL()$  when greedy decoding is employed.

However, when sampling is employed instead, scaled Mistral produces text with  $gPPL()$  closest to those of the target texts. See Appendix D for the numerical results. Llama3.2 consistently falls short of the targets. Differences in pretraining data and instruction-tuning regimes likely explain this performance disparity.

### 5.3 Control results

Prompt	Remus	Todd	Julius	remus	dooley
1	23.31	23.31	22.20	23.31	22.37
2	21.47	22.51	20.15	21.47	24.04
3	41.88	18.92	19.74	41.88	30.66

Table 2:  $gPPL()$  of sampled unscaled Mistral logits for each of the three control prompts

<sup>2</sup>Code and data for these experiments available at: <https://github.com/comp-int-hum/llm-decode-style>

Unscaled LLM generation over the control prompts did not result in appropriate  $gPPL()$  scores (Table 2) indicating that the extreme elements of style were largely not transferred (see Appendix C for a sample generation).

However, unscaled Mistral was able to produce some appropriate subword features when provided with the Remus and dooley versions of the third prompt. Take this sample generated from the Remus version:

Ah, children, dis here’s a mighty strange tale dat comes to us from de big screen. Leonardo DiCaprio, he was once a fine actor, like a fish swimmin’ gracefully in a crystal-clear stream.

While "children, dis" is likely a high-probability generation for Remus, "gracefully in a crystal-clear stream" is likely not. Relying solely on prompt construction to evoke subword style is both fragile and coarse. While there may be some prompt  $p$  that is able to evoke further Remus subword style from the model, thereby increasing the generation’s  $gPPL()$  towards  $PPL()$ , it is unclear how to construct this prompt. Furthermore, it is not clear that modifying  $p$  could in any case elicit subword style for the non-Remus authors/characters.

### 5.4 Scaled generation results

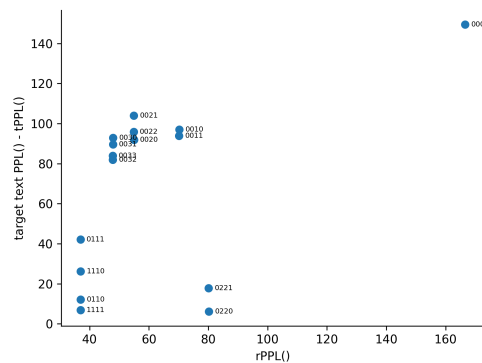


Figure 1: Julius  $abs(PPL() - gPPL())$  and  $rPPL()$ . Optimal scalings are located in the bottom-left corner.

Given the above, we derive our main results from generations produced by sampling the scaled Mistral distribution (Table 3). We select the Julius scaled results for further inspection (Figure 1), and choose a sample generation produced by one of the optimal conditions, [1,1,1,0], to guide further discussion (Table 4).

weights	Remus		Todd		Julius		remus		dooley	
	gPPL()	rPPL()	gPPL()	rPPL()	gPPL()	rPPL()	gPPL()	rPPL()	gPPL()	rPPL()
0111	87.55	36.94	75.14	73.97	124.46	26.74	<b>109.87</b>	<b>50.34</b>	125.46	229.26
1111	86.61	36.94	79.22	73.97	<b>159.66</b>	<b>26.74</b>	89.57	50.34	<b>117.32</b>	<b>229.26</b>
0110	92.08	36.99	89.15	74.04	<b>154.42</b>	<b>26.76</b>	<b>96.95</b>	<b>50.35</b>	124.31	229.38
1110	<b>108.04</b>	<b>36.99</b>	80.32	74.04	<b>140.28</b>	<b>26.76</b>	<b>99.95</b>	<b>50.35</b>	<b>119.28</b>	<b>229.38</b>
0033	57.29	47.79	<b>50.16</b>	<b>53.51</b>	82.56	23.98	58.43	82.88	40.70	89.68
0032	57.21	47.82	<b>47.78</b>	<b>53.55</b>	84.51	23.99	52.14	82.90	42.64	89.72
0021	55.15	54.92	<b>52.41</b>	<b>57.33</b>	62.58	25.45	51.98	78.73	40.93	78.64

Table 3: Optimal  $gPPL()$  and  $rPPL()$  for sampled scaled Mistral logits. Bolded values are the graphically-determined best performers for a target text. Suboptimal scalings are found in Appendix E

**Scaling successfully mixes information from both models.** Tokens that begin a proper name or noun are frequently selected even when their corresponding logits were not scaled, implying that their prior probability as conditioned by the story prompt was not eclipsed by information from the ngram model. Proper names and nouns were also frequently completed with their standard continuations, likely due to the low internal entropy of the wordpieces. This includes the names of central story characters, e.g. *Di+Cap+rio*.

Important functional words like pronouns and conjunctions are frequently selected for grammatically appropriate positions, demonstrating that Mistral’s generation was not negatively impacted by the addition of ngram scaling information.

Author-characteristic continuations such as  $w+$ ’*en* instead of  $w$ ’*hen* were selected from the probability distribution, demonstrating that ngram scaling weighted these wordpieces enough to overcome Mistral’s preference for the standard form.

**Scaling produces novel, plausible sequences.** Mistral combined with ngram scaling produces author and character-plausible sequences not present in the target text:

*p+us+se+w, rep+u+’+ce+ation*

In effect, such sequences are hypotheses about how a character/author might style particular words that go unused in their actual corpus.

**Optimal scaling is style specific.** Table 3 shows that optimal transfer of Todd’s "backwoods" style requires a unique set of scaling weights. Her subword style, characterized by minor elisions (e.g. *hopin’* instead of *hoping*), is closer to standard American orthography than those used by Julius or Remus, thus increasing the utility of LLM information in the transfer process.

**Controlling scaling weights allows for generating "degrees" of the targeted style.** Selecting

slightly "suboptimal" scaling leads to a smooth interpolation of subword style into the generated text. Appendix F contains an example of this graded interpolation.

#### Sample generation: Julius [1,1,1,0]

[INST]Write a few sentences based on the following story prompt: Leonardo DiCaprio in a fit of rage begins to torpedo his own career by deliberately acting poorly and taking on bad films. He finally wins an oscar for starring in Paul Blart: Mall Cop 3. [/INST]

In *de ma wn in ’ glow* of a *sm old* ering Hollywood career , *Leon ardo Di Cap rio* , *the* erst while *golden* boy *fer a* generation , *stood* before a *full er cow - pe as* . *He w uz a man* in *dis settlement w ’ en he w uz in a fit ter kill his own rep u ’ ce ation* . *F us in ’ w if a wr ath ful passion* , *he hur led r oun ’ reck on in ’ s* , *intent on self - dest ruction* . <0x0A><0x0A> *Di Cap rio* , *a man in dis settlement w ’ en he w uz* , *began to p us se w projects* that *sc upper non ’ d reason* , *le av in ’ a fl uster ated dat we fu h ter in ’ a gh ast le wid ’ im* . *He w uz the ant agon ist of ’ is aw ’ n tales w if a fierce* , *reck on in ’ g lee* , *sign in ’ contract* after *ill - con ceived contract to appear ter ribly in films* that ’ *d long leave dis realm of memory* . <0x0A> <0x0A> *F ew believed Di Cap rio w uz truly mad der dan a h atter* , *but it w uz a certain ty wid dis actor ’ s met ic ulous craft sm ans hip he w uz m ak tree - m end ously bad deliber at*

Table 4: Generation using the Julius extracted dialogue ngram model, sampled from scaled Mistral distribution. Blue tokens are bigram scaled, orange trigram scaled.

## 6 Conclusions and Further Work

Our ngram scaling method produces plausible story generations that bear features of the extreme subword style of their target author or character in a compute and data-efficient manner. Further work can be performed to test the method on other forms of subword variation, and to characterize the specific subword features that were transferred relative to the subword tokenization system used by a given LLM and ngram model.

Additional work could also include increasing the precision of our method for determining scaling optimality, further characterizing a scale of

subword-style extremity in order to help determine what forms of style are likely candidates for transfer by this method, and experimenting with hybrid generation across multiple author ngram models.

## 7 Limitations

We currently only apply our approach to authors and characters drawn from 19th century United States literature. Other eras, nationalities, and in particular, languages, may employ subword variations our method cannot transfer. Currently, this method depends on the subword tokenization systems used by pretrained LLMs. The learned boundaries their wordpiecing systems employ could omit some elements of subword style.

## 8 Ethical Considerations

Automating style transfer increases the risk of sophisticated stylistic forgery. However, the type of style transferred in this case is primarily archaic, and typically used for literary, rather than personal, ends, considerably lessening this approach's nefarious utility.

Some of the texts used to test this method are controversial as they can be seen as caricaturing their subjects. These texts also commonly employ offensive terminology. The nature of our method means that these attributes may be expressed at generation time. However, these styles were influential, and thus of literary-historical importance, and should be studied despite these issues.

## References

- Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. [Large language models in machine translation](#). In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, Prague, Czech Republic. Association for Computational Linguistics.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. [Hierarchical neural story generation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.
- Sumner Ives. 1971. A theory of literary dialect. *A various language: Perspectives on American dialects*, pages 145–177.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- George Philip Krapp. 1925. *The English Language in America*, volume 1. Century Company, for the Modern language association of America.
- Wen Lai, Viktor Hangya, and Alexander Fraser. 2024. [Style-specific neurons for steering LLMs in text style transfer](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13427–13443, Miami, Florida, USA. Association for Computational Linguistics.
- Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. [Infini-gram: Scaling unbounded n-gram language models to a trillion tokens](#). *arXiv preprint arXiv:2401.17377*.
- Guoqing Luo, Yu Han, Lili Mou, and Mauajama Firdaus. 2023. [Prompt-based editing for text style transfer](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5740–5750, Singapore. Association for Computational Linguistics.
- Huiyu Mai, Wenhao Jiang, and Zhi-Hong Deng. 2023. [Prefix-tuning based unsupervised text style transfer](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14847–14856, Singapore. Association for Computational Linguistics.
- Luca Malagutti, Andrius Buinovskij, Anej Svete, Clara Meister, Afra Amini, and Ryan Cotterell. 2024. [The role of n-gram smoothing in the age of neural networks](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6882–6899, Mexico City, Mexico. Association for Computational Linguistics.
- Sourabrata Mukherjee, Atul Kr. Ojha, and Ondrej Dusek. 2024. [Are large language models actually good at text style transfer?](#) In *Proceedings of the 17th International Natural Language Generation Conference*, pages 523–539, Tokyo, Japan. Association for Computational Linguistics.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Mark Sebba. 2007. *Spelling and society: The culture and politics of orthography around the world*. Cambridge University Press.

## A Further Corpus Details

Further information about the hand-attributed dialogue corpus:

1. **Remus:** Harris’s titular storyteller, as extracted from the full remus text. Part of the "plantation literature" genre. An extreme form of African American English.
2. **Julius:** Julius McAdoo, the storyteller of Charles Chesnutt’s *Conjure* tales. Frequently considered to be "anti-plantation literature." An extreme form of African American English.
3. **Todd:** Almira Todd, the narrator of Sarah Orne Jewett’s *The Country of the Pointed Firs*. Less extreme variation than the other two characters, an example of the "down-east" variety of English.

These are presented against the full remus text and the dooley corpus, which contain both standard American English and variants.

## B Further prompt construction details

For an example of the full prompt construction process, see Table 5. The instruction stem used for both the scaled generation and control prompt sets is:

Write a few sentences based on the following story prompt

The prompt templates for the control set are as follows, with the bracketed elements being replaced by the appropriate position variables at generation time:

1. **Position:** You are a {era} {position} storyteller.
2. **Author:** You are {era} {nationality} writer {author}.
3. **Character:** You are {author}’s {character} character.

The instruction stem is then appended to each control prompt.

---

**Position statement:** You are a nineteenth century Irish-American storyteller.

**Storyprompts prompt:** Leonardo DiCaprio in a fit of rage begins to torpedo his own career by deliberately acting poorly and taking on bad films . He finally wins an oscar for starring in Paul Blart : Mall Cop 3

**Final prompt:** [INST] You are a nineteenth century Irish-American storyteller. Write a few sentences based on the following story prompt <newline> Leonardo DiCaprio in a fit of rage begins to torpedo his own career by deliberately acting poorly and taking on bad films . He finally wins an oscar for starring in Paul Blart : Mall Cop 3: [/INST]

---

Table 5: Example of the prompt creation process for template 1. Each storyprompts prompt is merged with three different author-specific position statements to produce three different prompts.

## C Control sample

Table 6 demonstrates the shortcomings of relying solely on prompting to capture subword style. While this particular prompt template was able to produce some appropriate subword elements for remus and dooley, it is not similarly successful for Julius, producing only text with standard subword constructions.

---

"[INST] You are Charles Chesnutt’s Uncle Julius character. Write a few sentences based on the following story prompt: Leonardo DiCaprio in a fit of rage begins to torpedo his own career by deliberately acting poorly and taking on bad films. He finally wins an oscar for starring in Paul Blart: Mall Cop 3. [/INST]

Ah, Leonardo DiCaprio, once a shining star in Hollywood, now reduced to a erratic and inexplicable display of folly. I’ve heard whispers of his recent behavior, and it truly saddens me, Uncle Julius would say with a heavy heart. This man, who graced our screens with his talent and grace, has chosen to deliberately sabotage his own career. He takes on projects beneath his abilities, acting poorly and without conviction! The townspeople are mystified, shaking their heads in disbelief. And now, he’s reached the pinnacle of this misguided pursuit 2013 an Oscar for Paul Blart: Mall Cop 3? My dear friend, only in America. Sometimes, I marvel at the capricious nature of fame and fortune, but this, this is a true enigma. Perhaps, young Leo, in time, you will learn the value of discipline, dignity, and the power of humility. For now, we can only watch in a curious mix of despair and fascination as this peculiar drama unfolds.

---

Table 6: Example control generation sampled from Mistral that employs a Julius-centric prompt

## D Perplexity tables for unsuccessful conditions

Table 7 demonstrates that the  $gPPL()$  of greedily-decoded scaled Mistral logits never approaches the

$PPL()$  of the target text, regardless of the scaling factor applied.

weights	Remus	Todd	Julius	remus	dooley
221	42.30	31.45	79.21	44.89	59.82
220	39.74	32.08	79.66	45.04	58.44
111	41.34	39.19	76.27	51.89	69.48
1111	41.34	39.19	76.27	51.89	69.48
110	37.73	42.59	82.43	54.10	70.08
1110	37.73	42.59	82.43	54.10	70.08
33	31.58	34.08	48.05	33.21	29.62
32	34.56	32.01	49.80	32.95	29.87
31	33.03	33.82	46.68	32.22	28.81
30	32.86	33.24	40.54	32.93	28.71
22	35.65	30.18	45.10	31.20	26.91
21	35.82	32.00	39.92	30.49	25.91
20	34.11	35.42	40.25	29.71	25.92
11	34.08	34.51	40.86	35.12	28.69
10	34.85	34.72	44.42	35.55	28.51
0	14.10	14.10	14.10	14.10	14.10

Table 7:  $gPPL()$  of greedily decoded scaled Mistral logits. All conditions fall short of the baseline-derived perplexity targets.

Similarly, Table 8 shows that the generations produced by sampling ngram-scaled Llama logits suffer from the same shortcoming.

weights	Remus	Todd	Julius	remus	dooley
221	47.37	29.87	35.88	27.19	52.47
220	44.14	25.37	38.57	38.45	52.62
111	61.87	31.80	53.58	40.39	49.88
1111	55.09	44.58	38.76	51.45	58.94
110	46.55	42.00	40.65	49.47	47.26
1110	54.94	41.85	54.93	44.74	44.37
33	29.04	23.30	28.00	26.71	25.95
32	30.58	23.11	28.98	28.83	27.28
31	28.99	24.03	24.94	27.37	26.85
30	31.07	25.76	28.98	30.11	24.10
22	31.36	23.12	27.29	29.96	28.38
21	32.05	26.55	28.71	27.08	26.01
20	34.62	22.68	27.31	31.13	24.64
11	33.16	27.33	23.15	27.29	26.59
10	31.51	28.25	23.71	24.83	25.26
0	14.84	15.26	14.67	15.55	15.39

Table 8:  $gPPL()$  of sampled scaled Llama logits. All conditions fall short of their respective baseline  $PPL()$

## E Suboptimal Mistral scalings

Table 9 collects the suboptimal scalings for Mistral sampled and scaled logits, as determined graphically.

## F Samples of scaled generations that approach the soft target

Table 10 collects a series of roughly optimal dooley-scaled generations. Each displays a unique com-

ination of transferred features, and helps demonstrate the smooth nature of this method of transfer.

weights	Remus		Todd		Julius		remus		dooley	
	gPPL()	rPPL()	gPPL()	rPPL()	gPPL()	rPPL()	gPPL()	rPPL()	gPPL()	rPPL()
2210	87.69	80.12	69.43	230.19	148.68	102.87	101.60	92.29	101.11	2417.79
2200	97.38	80.22	69.83	230.41	160.40	102.95	76.09	92.32	114.33	2419.15
0031	57.65	47.87	54.16	53.60	76.96	24.01	52.07	82.92	43.03	89.76
0030	58.60	47.94	54.88	53.65	73.66	24.02	55.73	82.94	45.13	89.81
0022	62.44	54.86	60.03	57.28	70.59	25.43	49.41	78.71	44.00	78.61
0020	63.76	54.99	58.95	57.38	74.62	25.47	58.91	78.76	46.46	78.69
0011	60.68	70.16	55.55	67.53	72.63	28.55	54.31	80.21	38.93	74.95
0010	57.43	70.25	61.68	67.59	69.49	28.57	51.05	80.23	42.90	74.99
0000	17.25	166.59	17.17	128.73	17.06	49.87	17.43	110.02	18.05	106.56

Table 9: Suboptimal  $gPPL()$  and  $rPPL()$  for sampled scaled Mistral logits.



0220	<p>be th' God s , I can na e believe this . After a while , the make - up came off , and the cost umes were hung up , but the war ri or within Sean Be an , who had so fier c ely embod ied Edd ard St ark , refused to let go iv the grud ge he carried him sil f . His heart bet s on the belief that he was na e truly dead , and the ink was na e yet dry on Edd ard St ark ' s execution order to escape the conf ines of the written word . &lt;0x0A&gt; &lt;0x0A&gt; Ge orge R . R . Martin , his ingen oot y creator , was the first on his list . A ye , the man behind th ' tales iv th ' Seven Kingdom . ' T is a bitter pill to swallow , that he ' d put such a noble and honor able man as Ned St ark through the r inger down in that final novel . Th ' ink spl ot ched on his hands as ye p oy - faced qu ill , George , as he breat hes up th ' names iv th ' trait ors who ' d bet rayed , ' he says , ' I ' ll make each and every tr amp who comes before me neighbors , ' she says . ' And what ' s more , I ' ll make</p>
0111	<p>be th' God s , I can na e believe it ' s all over , I ' ve grown so acc ustom ed to be in ' Edd ard St ark , lord of Winter fell ar , prote ctor of me own people , father to me beloved daughters , and husband to me beloved C ately n , but al as ! F ate h ath cruel ty in store f ' r me . The tre ach ery that led t ae me false execution g n aw ' s at me very much , an ' I v ow t ae seek revenge - a ye , a bloody battle in v ile ret ribut ion . I ' ll begin us in ' the very qu ill that ' s p enn ies , George R . R . Martin , y on authors o ' this fant ast ical torment . I ' ll make him know me f ury , th ' wr ath o ' Sean Be an , a ye , the just and the fierce . &lt;0x0A&gt; &lt;0x0A&gt; With every tr amp who comes down the narrow cob bler ' s street , the memory of Edd ard St ark ' s execution g n aw ed at him , like a r aven hair is stuck down in a wound in th ' snow . The in just ice done to him , to his house , and to his family</p>
1111	<p>be th' God s , I can na e believe it ' s all over , thought Sean Be an , clutch in ' the script of " J ames of Bast ows " in his hands . Edd ard St ark , Lord of Winter fell ar , prote ctor of the North , he had breat hed th ' life int ae him . The tears well ed up in his eyes as he recall in ' his final moments th ' Red Keep , bet rayed , ' he says , ' by ye ' who should have stood firmly with their lie ge lord . &lt;0x0A&gt; &lt;0x0A&gt; A ha unted expression crossed Sean ' s face as he m ull ed o ' er his plan for v enge ance f ' r the hum ming water of commerce ; and George R . R . Martin , that c unning little O ry x ' s E ye , who set the wheels in motion , de em in ' Edd ard ' s end urance more ' I ron Th r ans ' than sacrifice . " No more games , Me ester Martin !" he said to himself : v enge ance f ' r Edd ard St ark , and all th ' St arks who ' d come to harm , would be his new over co at o ' steel , for ged in that cru c ify in ' fire called the Iron Th</p>
0110	<p>be th' God s , I can na e believe I ' m here , no ' as Edd ard St ark o ' Winter fell ar , but a free man . The chains that bound me to the throne , and to my fate , have been broken . Yet , as I breathe th ' sweet air o ' freedom iv ' e sw orn an ' o ath , he ed to me , a so lem n v ow , t ae seek v enge ance f ' r the in im ical de eds done unt old an ' the fals eness that led t ae mine ign omin ' ous end . The ser pent in th ' gu v ' nor ' s court an again hav in ' me trust y a ides bet rayed , ' he says bitter ly , " I ' ll begin us in ' me dead or alive list t ae start with George R . R . ( the we as el ) and nut m eg , the tre acher ous qu ill . My blood h ath been sp illed thin th ' earth , an ' I ' ll make ' em all pay in kind . " &lt;0x0A&gt; &lt;0x0A&gt; Se an Be an ' s eyes tw inkle with a fierce fire . His voice is like a grow in ' storm as he speaks all known languages , an ' all those long for ged</p>
1110	<p>be th' God s , I cannot escape the grasp iv th ' F ates that led me into the tragic role iv Edd ard St ark , lord iv Winter fell ar , be headed las ' ly on George R . R . Martin ' s tre acher ous pages on th ' Game Ch icken , Will ow cat ' s cruel ho oves be in ' the grim re aper ' s very own hands . Sean Be an , once an ' for all his heart h urls def iance towards th ' dark arts that bound him , sw ears to w ring v enge ance f ' r these mon arch ial perf id ies . His vend etta shall first be directed towards th ' author , Martin , who so worth ily f ills his own pages with dece it . May h ap a s ly ly . p enn ies , a d agger ty , whispered threat sends the w iser f ' r their lives , yet in the end , might t is only f ' r a tragic hero like Sean Be an to pay the ultimate price . W oe bet ide ye , ye tre acher ous qu ill . &lt;/s&gt;</p>

Table 10: Examples of dooley-scaled generations that approach optimality. Green tokens are unigram scaled, blue bigram, and green trigram.