Zipf's and Heaps' Laws for Tokens and LLM-generated Texts

Nikolay Mikhaylovskiy
NTR Labs / Moscow, Russia
Higher IT School of Tomsk State University / Tomsk, Russia
nickm@ntr.ai

Abstract

The frequency distribution of words in human-written texts roughly follows a simple mathematical form known as Zipf's law. Somewhat less well known is the related Heaps' law, which describes a sublinear power-law growth of vocabulary size with document size. We study the applicability of Zipf's and Heaps' laws to texts generated by Large Language Models (LLMs). We empirically show that Heaps' and Zipf's laws only hold for LLM-generated texts in a narrow model-dependent temperature range. These temperatures have an optimal value close to $t = 1$ for all the base models except the large Llama models, are higher for instruction-finetuned models and do not depend on the model size or prompting. This independently confirms the recent discovery of sampling temperature dependent phase transitions in LLM-generated texts.

1 Introduction

Zipf's law is a simple and long-known dependency between the frequency rank r of a word and its frequency in a natural corpus $f(r)$ (Estoup, 1916; Zipf, 1936; Piantadosi, 2014):

$$f(r) \propto r^{-\alpha}, \tag{1}$$

where $\alpha \approx 1.0$. Somewhat less known Heaps' law (Herdan, 1964; Heaps, 1978; Baayen, 2001; Serrano et al., 2009; Font-Clos et al., 2013) describes the growth of vocabulary size (number of unique words) w as a function of the size of a document (number of words) n:

$$w(n) \propto n^{\beta}, \tag{2}$$

and has a language-dependent exponent $\beta < 1.0$ (Baeza-Yates and Ribeiro-Neto, 1999; Serrano et al., 2009).

Similar power laws occur in many other areas of science, often quite distant from language (for an overview, see Mitzenmacher (2003) and Newman (2005)). The more intriguing is, as we show in this work, that Zipf's and Heaps' laws rarely hold for the texts generated by LLMs. Specifically, we study the applicability of Zipf's and Heaps' laws to LLM-generated texts and make the following contributions:

1. We empirically confirm that Zipf's and Heaps' laws hold for tokens in human-written texts.

2. We propose a simple statistical metric of the generated texts related to Heaps' law and find that many LLMs have an optimum with respect to this metric at temperatures between 1 and 2.

3. We empirically show that Heaps' and Zipf's laws only hold for LLM-generated texts in a narrow, model-dependent temperature range. These temperatures are higher for instruction-finetuned models than the base ones and do not depend on the model size or prompting.

4. By using the wide selection of LLMs we show that the pretraining corpus, training algorithm, and model architecture (within transformer architectural paradigm) influence zipfian properties of the LLM-generated texts much less than the temperature.

5. We tie these extrema of power-law fitness with the phase transition of LLM-generated texts through a critical state adhering to power laws.

In this section, we go on to introduce the key concepts. In Section 2 we briefly review the previous research. In Section 3 we confirm that the statistics for tokens in natural texts closely follow the statistics for words. In Section 4, we study the applicability of Zipf's and Heaps' laws to LLM-generated texts.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 15469–15481
November 4-9, 2025 ©2025 Association for Computational Linguistics

1.1 Autoregressive Probabilistic Language Models

Probabilistic language models consider sequences

$$t_{1:m} = \{t_1, t_2, \ldots, t_m\} \tag{3}$$

of tokens from the lexicon L. An autoregressive language model estimates the probability of such a sequence

$$\begin{aligned} P(t_{1:m}) &= P(t_1)P(t_2|t_1)\ldots P(t_m|t_{1:m-1}) \\ &= \prod_{k=1}^{m} P(t_k|t_{1:k-1}) \end{aligned} \tag{4}$$

using the chain rule.

1.2 Text Generation with a Language Model

Given an input text as a context, the goal of open-ended generation is to produce a coherent continuation of the text (Holtzman et al., 2020). More formally, given a sequence of m tokens $t_1 \ldots t_m$ as context, the objective is to generate the next n continuation tokens, resulting in the completed sequence $t_1 \ldots t_{m+n}$. This is achieved through the use of the left-to-right text probability decomposition (4), which is used to generate the sequence one token at a time, using a particular decoding strategy.

A common approach to text generation is to shape a probability distribution through temperature (Ackley et al., 1985). Given the logits $u_{1:|L|} = \log p(t = L_l|t_{1:i-1})$ and temperature T, the softmax is re-estimated as

$$p(t = L_l|t_{1:i-1}) = \frac{\exp(u_l/T)}{\sum_{l'} \exp(u_{l'}/T)} \tag{5}$$

Setting $T \in [0, 1)$ skews the distribution towards high-probability events, and, similarly, $T \in (1, \infty)$ skews the distribution towards low-probability events.

1.3 Phase Transitions

A physical phase of a system refers to a state (typically equilibrium) with unique macroscopic properties. These phases possess certain stability regions within the parameter space. The properties of the state change at the boundaries of these regions, where phase transition occurs.

Ehrenfest (1933) defined a phase transition as a discontinuity in the n-th order derivative of the free energy with respect to any argument of the free energy. Modern physics extends the notion of phases and applies it to various situations and beyond the notion of free energy. In particular, a first-order phase transition exhibits a discontinuity in the first-order derivative, whereas a second-order phase transition is continuous in its first derivative but shows a discontinuous or divergent behavior in its second derivatives (Papon et al., 2007).

The phase transition point that separates phases is defined by singular, divergent statistical quantities in the limit where the system size is infinitely large. This point does not depend on subjective factors, such as metrics and thresholds (Nakaishi et al., 2024). Specifically, autocorrelations in a critical state decay according to a power law, and thus their range becomes infinite. On the other hand, often varied properties exhibit power-law behavior at the critical point. In general, wherever one can see a quantity that obeys a power law, one can suspect the existence of a critical state.

2 Prior Research

Studies of the zipfian properties of texts generated by language models are not entirely new. In particular, Takahashi and Tanaka-Ishii (2017, 2019) and Lippi et al. (2019) have studied the zipfian properties of texts generated by ngram- and LSTM-based models. Holtzman et al. (2020) studied the dependence of the zipfian coefficient on the GPT-2 decoding strategy, and this dependence was rather weak. These works studied language models that are now obsolete, mostly have architectures different from the now dominant transformers, and with no relation to the sampling temperature. The related phenomena of power-law autocorrelations decay in texts generated by language models have been studied by Takahashi and Tanaka-Ishii (2017); Shen (2019); Takahashi and Tanaka-Ishii (2019); Lippi et al. (2019); Mikhaylovskiy and Churilov (2023).

Nakaishi et al. (2024) and Bahamondes (2023) independently pioneered the application of the correlation-based phase transition apparatus to LLM-generated texts and discovered phase transitions dependent on sampling temperature. Mikhaylovskiy (2025) have studied the phenomenon in mure detail following the approach of Mikhaylovskiy and Churilov (2023). Our work merges these two lines of research, discovering sampling temperature-dependent phase transitions in zipfian properties of LLM-generated texts.

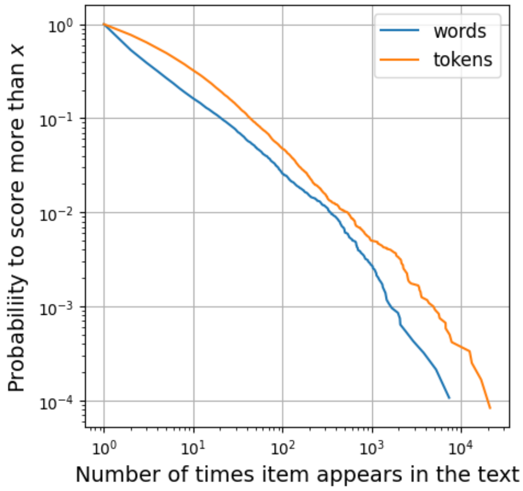

Figure 1: Cumulative histogram for the Critique of Pure Reason in German

3 Zipf's and Heaps' Laws for Tokens in Natural Texts

LLMs are used to generate diverse sequences in natural and programming languages. At higher temperatures, they generate gibberish that cannot be reasonably interpreted as words. Thus, if we want to study statistical properties of LLM-generated texts at higher temperatures, we can hardly use words as base units. Tokens are a natural choice.

Zipf's and Heaps' laws have been studied primarily for words, and tokenizing text into different units can potentially yield different statistical relationships. Thus, we need to verify that Zipf's and Heaps' laws hold for LLM tokenizations of human-written texts: while splitting texts into 'words' on arbitrary symbols (for example, 'e') does recover a near-Zipfian distribution on these artificial 'words', the exponent $\alpha = 0.77$ deviates significantly from the usual values (Piantadosi, 2014).

We verify Zipf's and Heaps' laws for tokens in natural texts with a dataset from Mikhaylovskiy and Churilov (2023). It consists of six long literary and philosophical works in 5 languages each. We tokenize each text using naïve splitting into words by separators such as spaces and punctuation, and using Mistral AI v3 (tekken) / Ministral-8b tokenizer (MistralAI, 2024) that uses Byte-Pair Encoding (Gage, 1994) with Tiktoken (OpenAI, 2024).

To verify Zipf's law for tokens, we calculate a cumulative histogram for each text and fit it with a power law using least squares in log-log coordinates (we agree with Newman (2005) that the classical Zipf rank representation is worse for approximation, but find that the maximum likelihood estimates of the exponents are unstable with respect to changes in the initial value). An example of a cumulative diagram is presented in Figure 1. Lines for both tokens and words can reasonably be considered to be straight in log-log coordinates, implying the power law relationship. One can decide that the probability distributions for tokens and words adhere to Zipf's law about as closely and the exponents of the power laws for words and tokens are similar and close to 1.

Language	Average α for words	Average α for tokens	stdev words	stdev tokens
en	0.952	0.985	0.072	0.054
de	0.959	1.024	0.054	0.042
ru	1.043	1.009	0.077	0.067
es	0.962	1.016	0.030	0.035
fr	0.963	0.989	0.033	0.042

Table 1: Zipf's law exponents averages for words and tokens in different languages

Language	Average α for words	Average α for tokens	stdev words	stdev tokens
en	0.809	0.802	0.027	0.024
de	0.846	0.793	0.031	0.028
ru	0.916	0.805	0.049	0.025
es	0.842	0.798	0.025	0.028
fr	0.843	0.801	0.020	0.018

Table 2: Heaps' law exponents averages for words and tokens in different languages

Indeed, the exponent for words averaged over texts and languages is 0.981 with a standard deviation of 0.063, and the similar exponent for tokens is 1.006 with a standard deviation of 0.05. As a side observation, while Table 1 confirms the conclusion of Gelbukh and Sidorov (2001) that Zipf's law exponent for words for Russian differs from that for Western-European languages, Zipf's law exponents for tokens do not exhibit any statistical difference between languages. A more detailed study and explanation of this phenomenon may be a topic of future research. The detailed tables of the exponents and deviations computed are presented in Appendix A.

Similarly, we build graphs of the number of unique words as a function of position in the text in

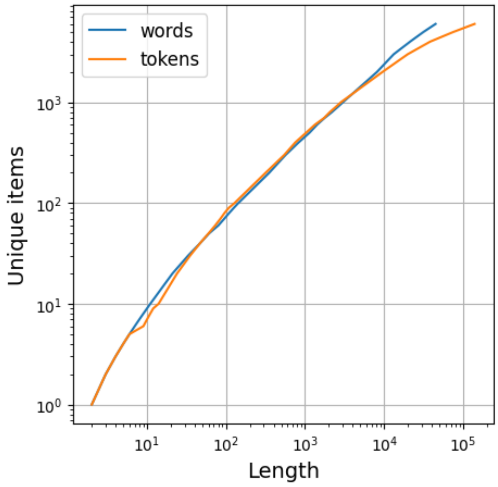

Figure 2: Unique items vs. position for The Iliad in French.

log-log coordinates to verify Heaps' law. A typical example can be seen in Figure 2. Although visually we can tell that the lines are not as straight as for Zipf's law, they (especially for words) can be well approximated by straight lines, implying decent adherence to the power law.

Saturation comes earlier for tokens and can be explained by both fundamentally limited number of tokenizer tokens for any language and tokens being shorter than words, so that new lexical items come later in the text if the text is measured in tokens. Still, we can say that up to the length of 10^4 the power law behavior of tokens and words is similar. We can conjecture that the relationship between the number of unique words and the text length for large n is best described by $w(n) \propto n/\log n$ rather than $w(n) \propto n^\beta$, but this is a topic of future research.

In any case, the Heaps' law exponent for words averaged over texts and languages is 0.855 with a standard deviation of 0.047, and the similar exponent for tokens is 0.801 with a standard deviation of 0.02. We can infer that the statistical behavior of tokens in human-written texts with respect to Heaps' law is also similar to the statistical behavior of words. Similarly to Zipf's law, from Table 2 we can observe that while the languages fall into three distinct groups with relation to word-level Heaps' law exponents: Russian, English, and the rest of the languages, token-level Heaps' law exponents do not differ statistically significantly among the languages. The detailed tables of the exponents and deviations computed are presented in Appendix B.

Model	Params, B	Active params, B	Context, K tokens
Qwen2.5-0.5B	0.49	0.49	32
Llama-3.2-1B	1.23	1.23	128
Qwen2.5-1.5B	1.54	1.54	32
Llama-3.2-3B	3.21	3.21	128
Llama-3.1-8B	8	8	128
Qwen2.5-14B	14.7	14.7	128
Qwen2.5-32B	32.5	32.5	128
Llama-3.1-70B	70	70	128
granite-3.1-1b-a400m	1.3	0.4	128
granite-3.1-3b-a800m	3.3	0.8	128
Mixtral-8x7B-v0.1	47	13	32

Table 3: Models used to generate texts

4 Zipf's and Heaps' Laws for Tokens in LLM-generated Texts

4.1 Data Generation

We generate texts at least 24K tokens long with an assortment of autoregressive open-weight language models with traditional GPT-like (Radford et al., 2018) and MoE (Jacobs et al., 1991) architectures at different sizes: Alibaba Qwen 2.5 (Qwen et al., 2025), Meta Llama 3.x (LLama Team, 2024), IBM Granite 3.1 (Granite Team, 2024) and Mistral AI Mixtral (Jiang et al., 2024).

The complete list of models is presented in Table 3. We used 4-bit BitsAndBytes NF4 quantized base and original instruction-finetuned models through the HuggingFace Transformers library (Wolf et al., 2020). We generate all the texts in a single run to fit the text generated into the context window. All the texts are generated from a single random, seed-controlled token. We use batch inference with a batch size selected to fit the model and context into the GPU memory.

We generate 40-50 texts for a range of temperatures from 0.4 to 1.5 or 2.5 (depending on the model, see below) with step 0.1. We do not use top-k, top-p or any other decoding parameters such as no-repeat to keep the things clean. The influence of these parameters on the statistical properties of the generated texts is a topic for future research. We have spent about 12 days of the 8*NVIDIA A100 80GB cluster to generate all the texts.

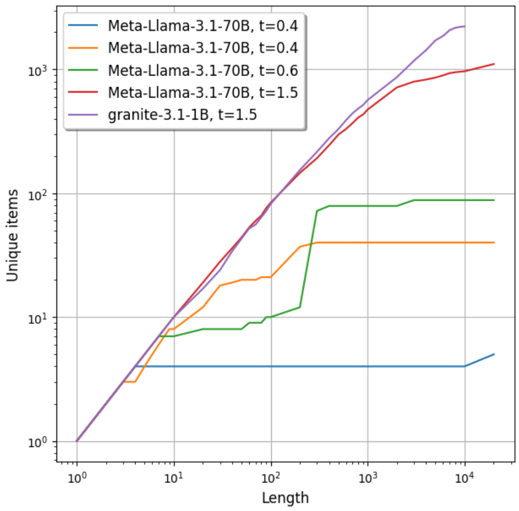

Figure 3: Unique items vs. position histograms for LLM-generated texts

4.2 Heaps' Law Applicability to Texts Generated by Base Models

Similarly to natural texts, we build graphs of the number of unique words as a function of position in the text in log-log coordinates. Some typical examples are collected in Figure 3. At low temperatures, the generated text degenerates quickly and soon, new tokens stop to appear. At higher temperatures, the process can restart after the initial degeneration to generate some new tokens before degenerating again. At even higher temperatures, the unique items vs. position histogram can follow the same pattern as for natural texts. Smaller models produce gibberish at higher temperatures; statistical properties of this gibberish are, though, indiscernible from natural texts. We should also note that there are boundary tokenization effects because the tokenizer of the generation model is different from the tokenizer we use to analyze texts; this can result in a few extra new tokens at the end of the text.

The above means that most models at most temperatures produce texts that do not obey Heaps' law even approximately. Thus, it does not make sense to fit the dependency of the number of unique words as a function of position in the text with a power law. Instead, we propose a simple statistical descriptor of the generated texts: the ratio of the number of tokens first appearing in the first and the second half of the text. We denote this quantity R. We have calculated R on the dataset from Mikhaylovskiy and Churilov (2023) and its mean value is 0.17 with a standard deviation of 0.05.

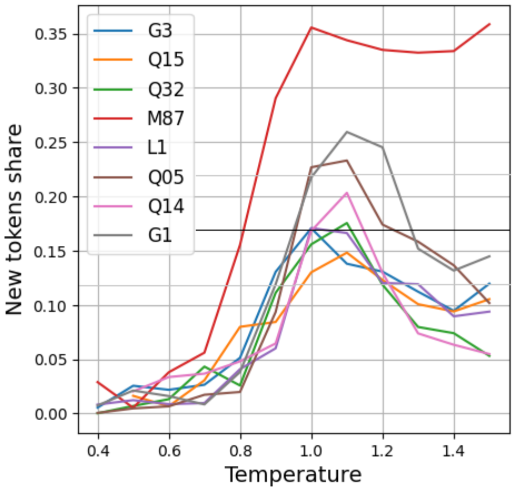

Figure 4: Relationship between R and t for LLM-generated texts

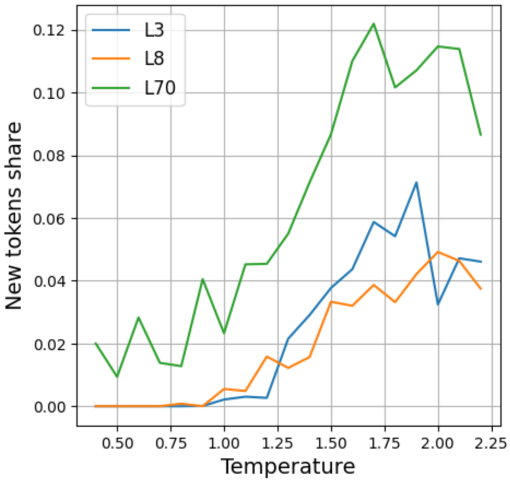

Figure 5: Relationship between R and t for texts generated by large LLama models

If we plot the average R for a certain model as a function of temperature t, and analyze the plots, we can see that the models fall into three distinct groups. One group, including all Qwen and Granite models, as well as LLama 1B, has a pronounced maximum at $t = 1$ (see Figure 4) with a value comparable to natural texts (horizontal line). This group generates degenerate texts at low temperatures, human-like texts at temperatures around 1, and gibberish they generate at higher temperatures mostly includes tokens already used. Neither the location of the maximum nor its value depends on the LLM size. The other group consists of the large Llama models (see Figure 5). They have a similar behavior, but with a less pronounced maximum around $t = 2$ and the value of the maximum is less than typical values for natural texts. Finally,

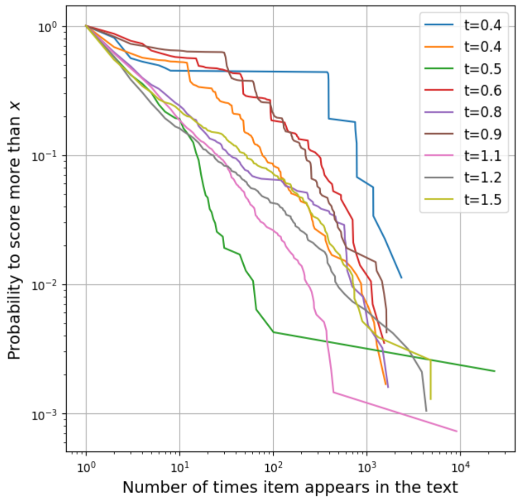

Figure 6: Examples of cumulative histograms for texts generated by Qwen 2.5 32B at different temperatures

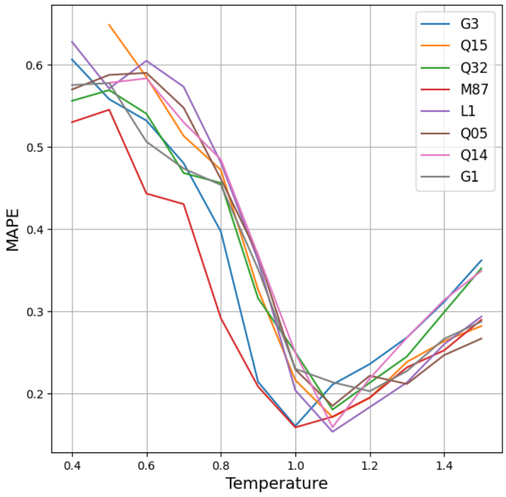

Figure 7: Relationship between power law MAPE and t for LLM-generated texts

Mixtral 8*7B constitutes a group of its own (see Figure 4) – it has an unpronounced maximum at $t = 1$ with a value significantly larger than that for natural texts.

Following Nakaishi et al. (2024) we can conjecture that models undergo phase transitions from ordered to critical and then to disordered states. The first group of models does this at temperatures close to unity and large LLamas at a temperature close to 2, and we can observe these transitions through the descriptor R.

4.3 Zipf's Law Applicability to Texts Generated by Base LLMs

Similarly to natural texts, we calculate a cumulative histogram for each text and plot it in log-log coordinates. Some typical examples are collected in Figure 6. Although for temperatures close to 1 a large part of the histogram can be thought to adhere to a power law, for most cases one can hardly say that a large part of the histogram can be considered a straight line in log-log coordinates.

To quantify this, we fit each cumulative histogram with a power law using least squares and compute Mean Average Percentage Error (MAPE) on this fit. For reference, on the natural text dataset from Mikhaylovskiy and Churilov (2023) its mean value is 0.156 with a standard deviation of 0.077. Then we plot the average MAPE for a certain model as a function of temperature t. The plots for all models except for the larger Llamas have a pronounced minimum between $t = 1.0$ and $t = 1.2$ (cf. Figure 7) with a value close to that for natural texts. Llama 3B, 8B and 70B have a minimum MAPE between $t = 1.3$ and $t = 1.5$ (cf. Figure 8), and its value is higher than normal for natural texts.

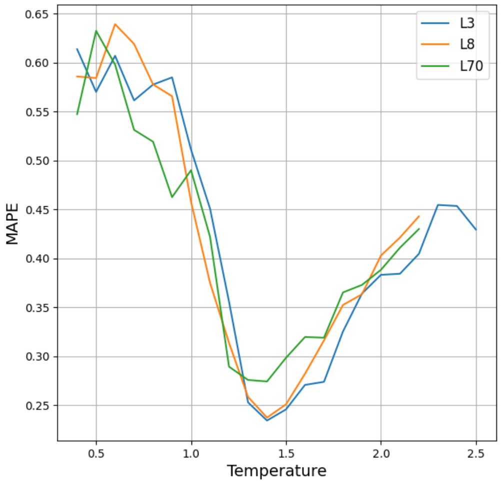

Figure 8: Relationship between power law MAPE and t for LLama-generated texts

We could not find a significant relationship between the location of this minimum and the size of the LLM. We can conclude that at the temperature where MAPE reaches its minimum value, the texts generated by LLMs can be considered to adhere to Zipf's law reasonably well. This further supports the idea that the models undergo phase transitions from ordered to critical and then to disordered states as the temperature increases. We should, though, note that the phase transition temperature for large LLama models differs significantly when determined from Heaps' and Zipf's laws perspectives.

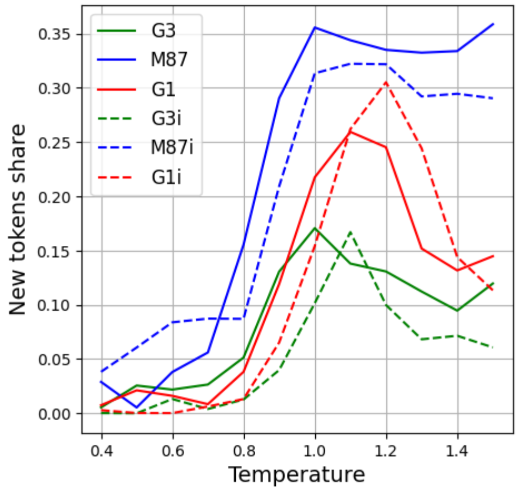

Figure 9: Relationship between R and t for the base and instruct MoE-generated texts

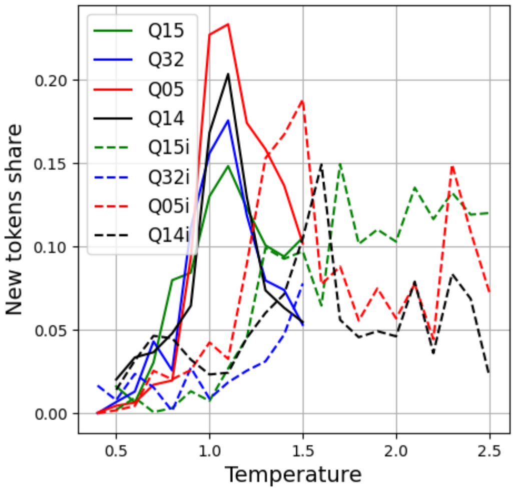

Figure 11: Relationship between R and t for the base and instruct Qwen-generated texts

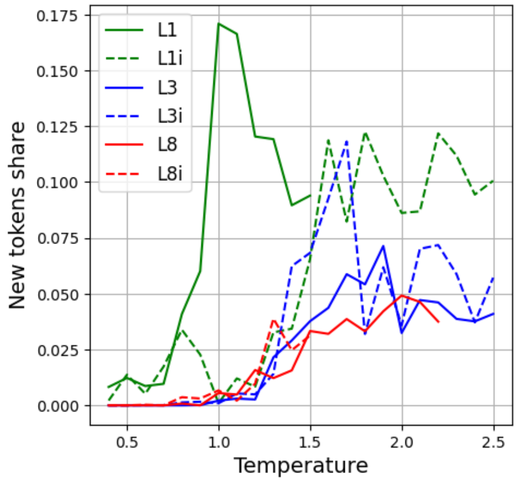

Figure 10: Relationship between R and t for the base and instruct Llama-generated texts

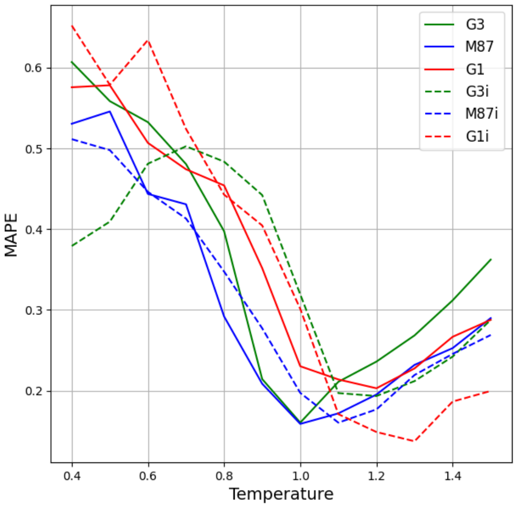

Figure 12: Relationship between power law MAPE and t for the base and instruct MoE-generated texts

4.4 Influence of Instruction Fine-tuning

In real scenarios, only models fine-tuned with instructions (Ouyang et al., 2022; Wei et al., 2022) are used. Thus, it is important to study these models as well. Similarly to Subsection 4.2, we plot the average R for a certain model as a function of temperature t, and see that the models studied behave differently after finetuning. Namely, the instruction finetuning shifts the maximum of $R-t$ plot of most models to higher temperatures, but for MoE models this shift is rather small (cf. Figure 9) and its value changes are also small. LLama models after finetuning also exhibit less pronounced maximum (cf. Figure 10), and LLama-3B is the only model for which the maximum of $R - t$ plot moves to a lower temperature and its value grows. However, the maximum of $R - t$ plot of Qwen models after finetuning shifts to significantly higher temperatures (cf. Figure 11) and the maximum becomes smaller and less pronounced.

Similarly to Subsection 4.3, we fit cumulative histograms for each text generated with a power law using least squares in log-log coordinates and compute MAPE on this fit. Then we plot the average MAPE for a certain model as a function of temperature t. For all the models, the instruction finetuning shifts the minimum of $MAPE - t$ plot to higher temperatures. For both MoE and LLama models this shift is rather small (cf. Figure 13, Figure 12). On the other hand, the minimum of $MAPE - t$ plot of the Qwen models after finetuning shifts to significantly higher temperatures

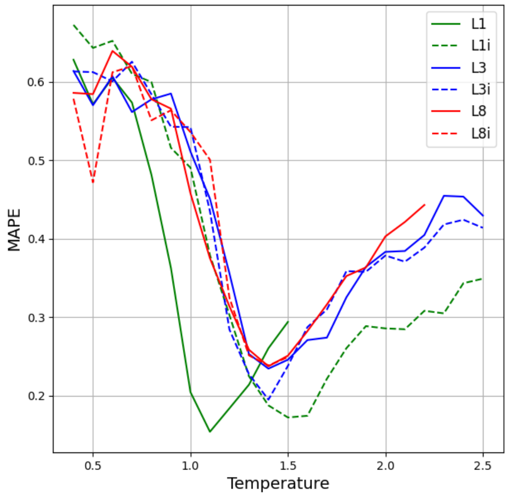

Figure 13: Relationship between power law MAPE and t for the base and instruct Llama-generated texts

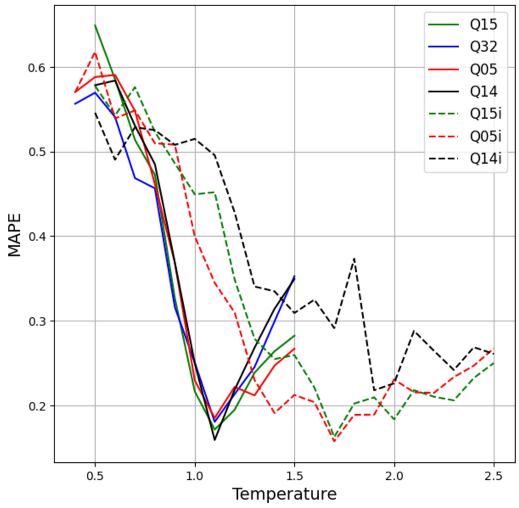

Figure 14: Relationship between power law MAPE and t for the base and instruct Qwen-generated texts

(cf. Figure 14), and we can conjecture that this shift grows with the model size. The minimum becomes much less pronounced as well. We can attribute the relationship between the size of the Qwen model and the shift in $MAPE - t$ plot minimum to the scaling laws followed by the Qwen finetuning procedures.

Overall, the shift of the critical state to higher temperatures can be attributed to narrowed down probability distributions after the instruction finetuning. It is interesting that finetuning affects the generated text adherence to Heaps' and Zipf's laws differently. This is an exciting topic for future research.

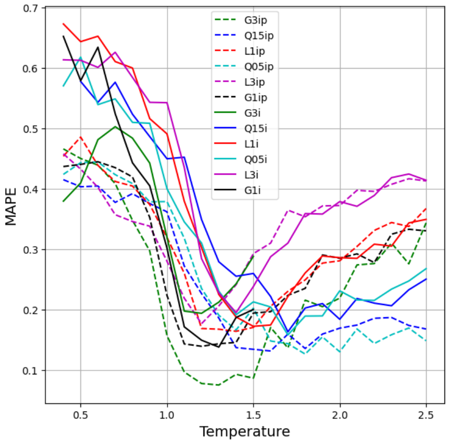

Figure 15: Relationship between power law MAPE and t for the instruct models with long and short prompts

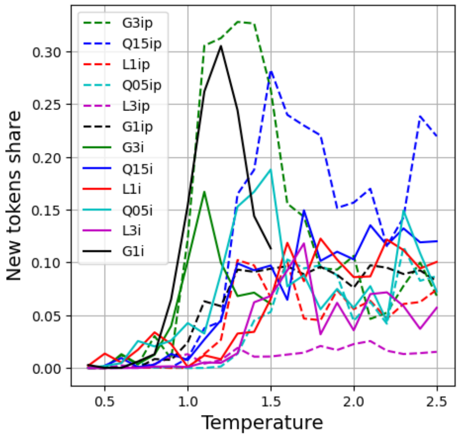

Figure 16: Relationship between R and t for the instruct models with long and short prompts

4.5 Influence of Prompting

In practice, LLMs are used with long-form prompts of some sort. Given that some of the models studied are base models, it makes sense to use a piece of natural text as a prompt and analyze the generated text in comparison with texts generated from a single token. Specifically, we randomly source a 2000 token long continuous piece from the Adventures of Tom Sawyer in English, and use it as a prompt to generate the further text following the setup described in Subsection 4.1 with LLMs with under 3B parameters. The only difference is that we generate 20 texts with each model.

Similarly to Subsection 4.3, we fit cumulative histograms for each text generated with a power law using least squares, compute the MAPE on this fit and plot the average MAPE for a certain model as a function of temperature t. We can see from the Figure 15 that the difference in the qualitative behavior of the plots for long- and short- prompted texts is insignificant. However, one can note that the average MAPEs for texts generated with long prompts are typically smaller, especially in the low-temperature range. Based on the above, we can say that prompting does not influence the adherence of LLM-gernerated texts to Zipf's law. Similar plots for the base models are presented in Appendix C.

On the other hand, the lexical diversity of the long prompt dominates the relatively small vocabulary of the generated texts (cf. Figure 16). This leads some models to generate no new tokens at all at low temperatures and generally shifts the maxima of $R - t$ plots to the high-temperature area. More high-resolution methods are likely needed to study these phenomena; this is an exciting topic for future research.

5 Conclusion and Discussion

We have shown empirically that Zipf's and Heaps' laws hold for tokens in natural (i.e., human-written) texts, but this is rarely the case for tokens of LLM-generated texts. More specifically, for each model, there is the best temperature at which the texts generated have statistical resemblance to human-written texts. We tie these extrema of power-law fitness with the phase transition of LLM-generated texts from ordered to disordered phase through a critical state that adheres to power laws. Thus, we believe to have been the first to discover phase transitions in Zipf's and Heaps' laws for LLM-generated texts.

Statistics for different model families behave differently with respect to the instruction-finetuning. For large LLama models these extrema come at different temperatures for Heaps' and Zipf's laws. Instruction-finetuned models typically generate human-like texts at higher temperatures than the base models. This shift is more pronounced for the Qwen model family. We can conjecture that the finetuning approach followed by the Qwen team honors narrower distributions after finetuning. This may also explain a long-standing puzzle in the community: high temperatures enhance the writing quality of LLaMA and Mistral, they do not have the same positive effect on Qwen.

The wide selection of LLMs used had an explicit goal of showing that pretraining corpus, training algorithm, and model architecture (within transformer architectural paradigm) influence zipfian properties of the LLM-generated texts much less than the temperature. We also did not find a significant influence of prompting on the statistical properties studied.

Limitations

Although the tekken tokenizer works well with a variety of languages, the dataset we used in Section 2 is limited to European languages using Latin and Cyrillic scripts. This may limit the applicability of the results of this work to other language groups and families, as well as different scripts such as Chinese or Arabic.

In practice, LLMs are used to generate texts using top-k, top-p and advanced decoding parameters such as no-repeat. We did not study the influence of these parameters on the metrics of interest. This can limit the applicability of this work to practical cases.

Acknowledgments

We are grateful to anonymous referees for important suggestions, including suggestions to study the effects of instruction finetuning and prompting.

Early versions of this research were discussed with Kai Nakaichi, Dmitry Manin and Anastasiya Kolmogorova. We are very grateful for their insights and criticism.

References

David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. 1985. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169.

R.H. Baayen. 2001. *Word Frequency Distributions*. Text, Speech and Language Technology. Springer Netherlands.

R. Baeza-Yates and B. Ribeiro-Neto. 1999. *Modern Information Retrieval*. ACM Press Books. ACM Press.

Sebastian Bahamondes. 2023. Study of the possibility of phase transitions in LLMs.

Paul Ehrenfest. 1933. Phasenumwandlungen im ueblichen und erweiterten sinn, classifiziert nach den entsprechenden singularitaeten des thermodynamischen potentiales. *Bd. 36. Proc. Royal Acad. Amsterdam, 153–157.*

J. B. Estoup. 1916. Les gammes stenographiques. *Institut Stenographique de France*.

Francesc Font-Clos, Gemma Boleda, and Álvaro Corral. 2013. A scaling law beyond Zipf's law and its relation to Heaps' law. *New Journal of Physics*, 15(9):093033.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Alexander Gelbukh and Grigori Sidorov. 2001. Zipf and heaps laws' coefficients depend on language. *Lecture Notes in Computer Science*, pages 332–335.

IBM Granite Team. 2024. Granite 3.0 language models.

H. S. Heaps. 1978. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, Inc., USA.

G. Herdan. 1964. *Quantitative Linguistics*. Butterworths.

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

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87.

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. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Marco Lippi, Marcelo A. Montemurro, Mirko Degli Esposti, and Giampaolo Cristadoro. 2019. Natural language statistical features of LSTM-generated texts. *IEEE Transactions on Neural Networks and Learning Systems*, 30(11):3326–3337.

Meta LLama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nikolay Mikhaylovskiy. 2025. States of llm-generated texts and phase transitions between them. In *First Conference of Mathematics of AI*.

Nikolay Mikhaylovskiy and Ilya Churilov. 2023. Autocorrelations decay in texts and applicability limits of language models. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2023"*.

MistralAI. 2024. Tokenization.

Michael Mitzenmacher. 2003. A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics*, 1(2).

Kai Nakaishi, Yoshihiko Nishikawa, and Koji Hukushima. 2024. Critical phase transition in large language models. *Preprint*, arXiv:2406.05335.

Mark EJ Newman. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351.

OpenAI. 2024. Tiktoken.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

P. Papon, S.L. Schnur, J. Leblond, and P.H.E. Meijer. 2007. *The Physics of Phase Transitions: Concepts and Applications*. Advanced Texts in Physics. Springer Berlin Heidelberg.

Steven T. Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21:1112 – 1130.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

M. Ángeles Serrano, Alessandro Flammini, and Filippo Menczer. 2009. Modeling statistical properties of written text. *PLOS ONE*, 4(4):1–8.

Huitao Shen. 2019. Mutual information scaling and expressive power of sequence models. *Preprint*, arXiv:1905.04271.

Shuntaro Takahashi and Kumiko Tanaka-Ishii. 2017. Do neural nets learn statistical laws behind natural language? *PLOS ONE*, 12(12):1–17.

Shuntaro Takahashi and Kumiko Tanaka-Ishii. 2019. Evaluating computational language models with scaling properties of natural language. *Computational Linguistics*, 45(3):481–513.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Work	average α for words	average α for tokens	stdev words	stdev tokens
The Adventures of Tom Sawyer	1.018	1.067	0.040	0.045
The Republic	0.945	0.993	0.056	0.054
Don Quijote de la Mancha	0.944	0.966	0.041	0.024
The Iliad	1.050	1.036	0.069	0.029
Critique of Pure Reason	0.908	0.956	0.013	0.033
Moby-Dick or, The Whale	1.002	1.025	0.053	0.016
War and Peace	0.970	0.990	0.045	0.019

Table 4: Zipf's law exponents averages for words and tokens in natural texts

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

G.K. Zipf. 1936. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Cognitive psychology. Routledge.

A Zipf's law exponents for words and tokens in natural texts

Zipf's law exponents computed for words and tokens in natural texts are presented in Table 5. The averages and standard deviations for each text in various languages are presented in Table 4.

Work	Language	α for words	α for tokens
The Adventures of Tom Sawyer	En	0.995	1.012
	De	1.035	1.104
	Ru	1.080	1.120
	Es	0.979	1.058
	Fr	1.000	1.039
The Republic	Es	0.933	0.999
	De	0.906	1.005
	Fr	0.948	0.969
	En	0.898	0.924
	Ru	1.038	1.071
Don Quijote de la Mancha	Fr	0.952	0.959
	De	0.936	1.005
	Ru	1.005	0.964
	En	0.890	0.937
	Es	0.939	0.964
The Iliad	Es	1.002	1.049
	En	1.091	1.078
	De	1.019	1.037
	Fr	0.986	1.005
	Ru	1.151	1.012
Critique of Pure Reason	Ru	1.044	0.975
	De	0.938	1.006
	En	0.899	0.956
	Es	0.922	0.988
	Fr	0.902	0.923
Moby-Dick or. The Whale	Es	0.980	1.041
	Fr	0.984	1.036
	En	0.965	1.019
	Ru	1.081	1.005
War and Peace	Es	0.979	1.013
	De	0.936	1.003
	Fr	0.966	0.992
	En	0.928	0.971
	Ru	1.041	0.970

Table 5: Zipf's law exponents for words and tokens in natural texts

B Heaps' law exponents for words and tokens in natural texts

Heaps' law exponents computed for words and tokens in natural texts are presented in Table 6. The averages and standard deviations for each text in various languages are presented in Table 7.

Work	Language	α for words	α for tokens
The Adventures of Tom Sawyer	En	0.841	0.837
	De	0.877	0.841
	Ru	0.935	0.847
	Es	0.882	0.850
	fr	0.859	0.828
The Republic	Es	0.825	0.792
	De	0.822	0.794
	Fr	0.820	0.786
	En	0.795	0.785
	ru	0.897	0.814
Don Quijote de la Mancha	Fr	0.861	0.785
	De	0.862	0.778
	Ru	0.961	0.785
	En	0.793	0.776
	Es	0.829	0.756
The Iliad	Es	0.857	0.796
	En	0.829	0.807
	De	0.878	0.798
	Fr	0.843	0.814
	Ru	0.947	0.811
Critique of Pure Reason	Ru	0.914	0.828
	De	0.816	0.786
	En	0.774	0.795
	Es	0.805	0.792
	Fr	0,816	0,792
Moby-Dick or, The Whale	Es	0,853	0,807
	Fr	0,866	0,814
	En	0,842	0,831
	Ru	0,943	0,813
War and Peace	Es	0,842	0,790
	De	0,831	0,757
	Fr	0,834	0,786
	En	0,792	0,784
	Ru	0,912	0,770

Table 6: Heaps' law exponents for words and tokens in natural texts

C Plots comparing prompted and unprompted models

The relationship between power law MAPE and t for the base models with long and short prompts is presented in Figure 17. The relationship between R and t for the base models with long and short prompts is presented in Figure 18.

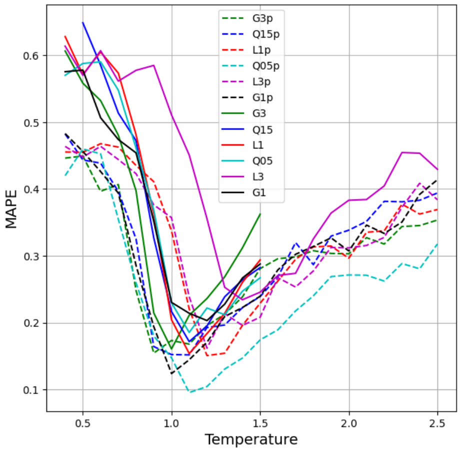

Figure 17: Relationship between power law MAPE and t for the base models with long and short prompts

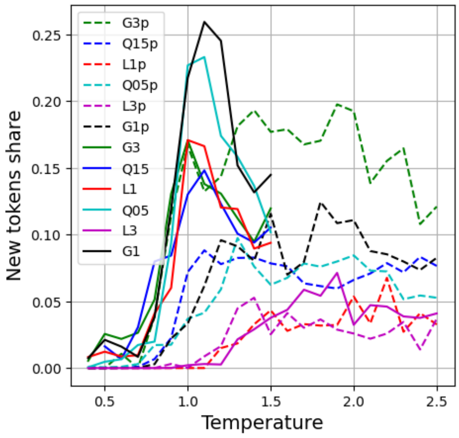

Figure 18: Relationship between R and t for the base models with long and short prompts

Work	average α for words	average α for tokens	stdev words	stdev tokens
The Adventures of Tom Sawyer	0.879	0.841	0.035	0.009
The Republic	0.832	0.794	0.038	0.012
Don Quijote de la Mancha	0.861	0.776	0.063	0.012
The Iliad	0.871	0.805	0.046	0.008
Critique of Pure Reason	0.798	0.793	0.022	0.002
Moby-Dick or, The Whale	0.876	0.816	0.046	0.010
War and Peace	0.842	0.777	0.043	0.014

Table 7: Heaps' law exponents averages for words and tokens in natural texts