AI Stereotypes: An Unequipartition Property for Perplexity in Generative Language Models

Anonymous authors
Paper under double-blind review

Abstract

We present a simple generalization of the classical Asymptotic Equipartition Property and the associated concept of typical sets for the perplexity of long texts generated by a language model. Specifically we show that the logarithmic perplexity of any large text generated by a language model must asymptotically converge to the average entropy of its token distributions. This defines an "idiotypical set" that all long synthetic texts generated by a language model must belong to. We refine the concept of "idiotypical set" to include only grammatically correct texts. We then show that this refined idiotypical set is a vanishingly small subset of all possible grammatically correct texts. This is a general result that applies to any language model for a very broad definition of grammar and it asserts that language models are strongly constrained in the range of their possible behaviors and outputs. We make no simplifying assumptions (such as stationarity) about the statistics of language model outputs, and therefore our results are directly applicable to practical real-world models without any approximations. We also present supporting experimental evidence from open-source models and discuss possible applications of the idiotypical set concept to problems such as detecting synthetic texts and membership inference in training datasets.

1 Introduction and Background.

Consider a generative model, defined as an algorithm that takes a user input $\mathbf{X}$ and produces an output $\mathbf{Y}$ that statistically resembles data from a natural source. A specific type of generative model is a large language model (LLM) whose output $\mathbf{Y}$ is a body of text and the input is a text user prompt $\mathbf{X}$. Modern LLMs (Anthropic, 2024; OpenAI, 2024) are able to produce detailed and information-rich text outputs such as entire screenplays and book-length manuscripts from a short and simple user prompt. They can also pass the Turing test (Pinar Saygin et al., 2000) i.e. imitate human-written text well enough to be convincing to human observers. In this work, we argue that an LLM is strongly constrained by statistical laws. Specifically, we show that the logarithmic *perplexity* of any large text produced by a language model must asymptotically converge to the average entropy of its token distributions. This means that any language model is constrained to only output text strings from an *idiotypical set*, which we show, is an exponentially vanishing subset of all possible grammatically correct strings. This involves a generalization of the well-known Asymptotic Equipartition Property (AEP) (Cover & Thomas, 2005a) from information theory and refining it for language models to account for the concept of grammatical correctness.

This work touches on ideas such as perplexity, entropy and equipartition theorems from information theory and computational linguistics, each of which boast of a vast body of previous work. We now present a brief (and necessarily incomplete) literature survey to place our work in context.

1.1 Equipartition Theorems

The simplest version of the AEP states that a long sequence of independent and identically distributed (iid) random symbols is likely to be a *typical sequence* (Csiszar, 1998) defined by the property that the empirical distribution of the occurrence of different symbols within the sequence is very close to the distribution from which the symbols were drawn. The AEP itself can be thought of as a consequence of the Law of Large

Numbers (LLN) (Idele, 2018) that asserts the convergence of a log-likelihood random variable to a limiting "entropy rate" (Algoet & Cover, 1988).

A weak AEP for iid sequences with a finite alphabet was originally introduced by Shannon (1948) followed by strong versions for stationary, ergodic sequences by McMillan (1953) and Breiman (1957), extensions to infinite alphabets by Chung (1961), and continuous-valued sequences in Barron (1985). Thus, the AEP is also commonly referred to as the Shannon-McMillan-Breiman (SMB) Theorem, Hamdan et al. (2008). There is now a vast literature on generalizations of the SMB Theorem (Moy, 1961) in the form of ergodic theorems for many classes of stationary sequences (Pollicott & Yuri, 1998).

1.2 The Green Fields of Non-Stationary Processes

Extending the AEP for iid sequences to a non-stationary sequence of independent, but non-identically distributed symbols is straightforward (Tang et al., 2024). AEPs for word-valued functions of stationary sequences are considered in (Nishiara & Morita, 2000; Timo et al., 2010). However, beyond this, the literature on the AEP for non-stationary sequences is much more limited. Closest to our work is the analysis in Yang & Liu (2004) and Wang & Yang (2016) of a very general class of non-homogeneous Markov processes, but they too assume the existence of a unique limiting distribution for the Markov process to establish a constant entropy rate. *Apparently, before the advent of modern generative models, there was simply no compelling reason to study non-stationary processes with arbitrary statistics.*

1.3 Statistical vs Structural Models of Language

Statistical analysis of languages in its modern form dates back to the work of (Shannon, 1951) on the entropy of written English. These statistical ideas competed with structural theories of language (Chomsky, 2014) in the early days of modern linguistics (Chomsky, 1959). Early methods such as n-gram models (Manning & Schütze, 1999) evolved into more mathematically sophisticated methods such as Hidden Markov Models (HMMs) (Rabiner, 1989). The statistical approach has now become dominant after the recent introduction of neural network architectures (Bengio et al., 2003), culminating in the development of powerful large language models (LLMs) based on transformer networks (Vaswani et al., 2017). The representation of language structures in LLMs is now an active topic of research Reizinger et al. (2024); Mészáros et al. (2024).

Statistical measures such as perplexity are widely used as a performance metric for training language models (Meister & Cotterell, 2021). Perplexity is closely related to the information-theoretic concepts of *surprisal* (Levy, 2008) and entropy (Cover & Thomas, 2005b), and it also appears to capture linguistic (Miaschi et al., 2021; Gamallo et al., 2017) and cognitive (Demberg & Keller, 2008; Cohen & Pakhomov, 2020) phenomena at least partially. Many "AI detection" tools are based on observed differences between the perplexity of synthetic and natural text (Mitchell et al., 2023; Gehrmann et al., 2019).

1.4 Contribution: Statistical Laws for Un-natural Language

Our goal in this paper is not to study natural language; instead our exclusive focus is on *unnatural language* i.e. synthetic text created by LLMs. Although language models were designed as computational models of natural language, in this paper, we take a subtly different perspective: an LLM once trained, is simply a stochastic machine and we can analyze the outputs of LLMs as stochastic processes in their own right. Importantly, we have access to the internal workings of an LLM which means that **we have complete and exact knowledge of the statistics of these stochastic processes.** Thus, it is possible for us to derive *exact* laws that govern the outputs of LLMs without making any assumptions about its relationship to natural language.

Our main contributions are (a) to formulate and provide theoretical and experimental justification for a version of the AEP that applies to real-world practical LLMs without any approximations, and (b) use this AEP to define a concept of typical set specifically adapted for language models. While our focus is on the theoretical development, we also briefly discuss possible applications to practical problems such as AI text detection and LLM dataset inference.

1.5 The Idiotypicality Idiom

A key concept in our paper is what we will call the "idiotypical" set of "idiotypical" strings which arises naturally from our generalization of the AEP. This is actually a straightforward extension of the classic information-theoretic concept of "typical sets" to non-stationary process and it is tempting to simply use the term "typical" (with no qualifiers) to refer to this generalization. However, this risks confusion because the classical literature is heavily focused on stationary, ergodic processes for which constant limiting entropy rates exist. Instead, we choose to use the prefix 'idio' ((derived from the Greek idios, meaning 'private' or 'peculiar to oneself', drawing inspiration from the immunological concept of an idiotypic signature Jerne (1974)) to denote our concept of *path-wise typicality specific to each sequence*.

Other notation. A sequence of symbols is denoted by boldface, uppercase identifiers e.g. $\mathbf{A}$, $\mathbf{B}$. These can be fixed, constant sequences or random sequences. The vector consisting of the first N elements of a sequence $\mathbf{A}$ is denoted by $\mathbf{A}_N \equiv [A_1,\ A_2,\ \ldots,\ A_N]$. All random variables in this work take discrete-values from a finite set. The probability of an event E is $\Pr(E)$. Sets are denoted by calligraphic symbols e.g. $\mathcal{A}$. The binary entropy of a distribution with probability mass function (pmf) $p(A_m)$ is $H(p(A_m))$[1]. The symbol $\equiv$ denotes a simple identity while $=$ denotes a mathematical inference and $\doteq$ a definition.

2 Definitions and Problem Statement

Let M be a generative model whose output $\mathbf{Y}$ consists of a sequence of tokens $\mathbf{Y} = [Y_1,\ Y_2,\ \ldots Y_N\ \ldots]$, chosen from a finite token dictionary $Y_n \in \mathcal{Y}$, as a stochastic function of user prompt $\mathbf{X}$. The model M is defined by a probability distribution $p(\mathbf{Y}|\mathbf{X})$, or more formally, a sequence of probability distributions $p(\mathbf{Y}_N|\mathbf{X})$ over $\mathbf{Y}_N \in \mathcal{Y}^N$ where $\mathbf{Y}_N \doteq [Y_1,\ Y_2,\ \ldots Y_N]$ is the sub-string of $\mathbf{Y}$ consisting of the first N tokens. Practical implementations of LLMs specify the probability distribution iteratively (Radford et al., 2018), e.g.,:

$$p(Y_1, Y_2|\mathbf{X}) = p(Y_1|\mathbf{X})p(Y_2|Y_1, \mathbf{X}). \tag{1}$$

Thus, the model M first draws a random value for the first token say $Y_1 = y_1$ by sampling from the distribution $p(Y_1|\mathbf{X})$. Then the model determines a distribution for the second token as a function of the initial prompt $\mathbf{X}$ and the randomly chosen first token y_1. Thus, the second token is randomly sampled from a distribution $p(Y_2|\mathbf{X}, Y_1 = y_1)$ and so on. We can write

$$p(\mathbf{Y}_N|\mathbf{X}) = p_1(Y_1)p_2(Y_2)\ldots p_N(Y_N),\ \text{where } p_n(Y_n) = p(Y_n|Y_1,\ Y_2,\ Y_{n-1}, \mathbf{X}) \tag{2}$$

For open-source LLMs, given a user prompt $\mathbf{X}$ and a string of tokens $\mathbf{Y}_N \equiv [Y_1,\ Y_2,\ \ldots,\ Y_N]$, it is possible to get a complete listing of the conditional pmfs $p_n(Y_n),\ n = 1 \ldots N$.

Remark. Equation (2) is simply an application of the Bayes rule of probability theory and it always holds for any generative model regardless of whether the tokens Y_n are sequentially generated as in (1). However, the conditional distributions $p_n(y)$ are not in general easily accessible for e.g. diffusion models. Thus, while (2) is true for all generative models, it may only be *useful* for sequential models.

The perplexity $\text{perp}_M(\mathbf{Y}_N) \doteq \prod_{n=1}^{N} p_n(Y_n)^{-\frac{1}{N}}$ of a text string $\mathbf{Y}_N = [Y_1,\ Y_2,\ \ldots Y_N]$ for a model M is defined as the per-token inverse likelihood of the string $\mathbf{Y}$. It is usually more convenient to work with the log-perplexity (or per-token surprisal) $l_M(\mathbf{Y}_N)$:

$$l_M(\mathbf{Y}_N) \doteq \log_2\left(\text{perp}_M(\mathbf{Y}_N)\right) \equiv -\frac{1}{N}\sum_{n=1}^{N} \log_2(p_n(Y_n)) \tag{3}$$

2.1 Warm-up: the Stationary Case

Let $\alpha(y),\ \beta(y)$ be two fixed probability mass functions (pmfs) over the (discrete) set $\mathcal{Y}$ of tokens. Consider a toy problem involving two language models A and B, that each generate a string of tokens $\mathbf{Y} = [Y_1,\ Y_2,\ \ldots]$

[1]We will always speak of the entropy of distributions and avoid speaking of the entropy of random variables.

where each token is generated iid from the distribution $\alpha(y)$ and $\beta(y)$ respectively. The iid assumption implies that the tokens Y_n can be thought of as being generated by a stationary and ergodic random process[2].

Consider a long string $\mathbf{A}_N \doteq [A_1,\ A_2,\ \ldots,\ A_N],\ N \gg 1$ randomly generated from model A. Let $p_{\mathbf{A}}(y)$ denote the empirical distribution of $y \in \mathcal{Y}$ in the string $\mathbf{A}_N$:

$$p_{\mathbf{A}}(y) \doteq \frac{n_{\mathbf{A}}(y)}{N} \equiv \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}(A_n \equiv y),\ \forall y \in \mathcal{Y} \tag{4}$$

where $\mathbf{1}(.)$ denotes the indicator function and $n_{\mathbf{A}}(y)$ is the number of occurrences of token y in the (long) string $\mathbf{A}_N$. The log-perplexity of string $\mathbf{A}_N$ for model A is:

$$l_A(\mathbf{A}_N) = -\frac{1}{N}\sum_{n=1}^{N} \log_2(\alpha(Y_n)) \equiv -\sum_{y\in\mathcal{Y}} \frac{n_{\mathbf{A}}(y)}{N}\log_2(\alpha(y)) \equiv H(p_{\mathbf{A}}, \alpha) \tag{5}$$

where $H(\gamma_1, \gamma_2) \doteq -\sum_{y\in\mathcal{Y}} \gamma_1(y)\log_2(\gamma_2(y))$ is the cross-entropy between pmfs $\gamma_1(y),\ \gamma_2(y)$ over $y \in \mathcal{Y}$. It is well-known $H(\gamma_1, \gamma_2) \geq H(\gamma_1)$ with equality when $\gamma_1 \equiv \gamma_2$, where $H(\gamma) \doteq -\sum_{y\in\mathcal{Y}} \gamma(y)\log_2(\gamma(y))$ is the entropy of distribution γ. Thus we have from (5):

$$l_A(\mathbf{A}_N) \equiv H(p_{\mathbf{A}}, \alpha) \geq H(p_{\mathbf{A}}) \text{with equality iff } p_{\mathbf{A}}(y) \equiv \alpha(y),\ \forall y \in \mathcal{Y} \tag{6}$$

The simplest version of the classical Asymptotic Equipartition Theorem (AEP) (Cover & Thomas, 2005a) states that the log-perplexity of a long string $\mathbf{A}_N$ of iid symbols is almost always very close to the entropy $H(\alpha)$ of the distribution $\alpha(y)$ the symbols are drawn from. A slightly more general version is as follows.

Proposition 1. A Simple Cross-AEP. *For a text string $\mathbf{A}_N = [A_1,\ A_2,\ \ldots,\ A_N]$ of iid tokens A_n drawn from a distribution with pmf $\alpha(y)$, the log perplexity $l_B(\mathbf{A}_N)$ for a model B (or A) as defined in (3) approaches the cross-entropy of the distributions $H(\alpha, \beta)$ (or entropy $H(\alpha)$) with high probability:*

$$\lim_{N\to\infty} \Pr\left(|l_B(\mathbf{A}_N) - H(\alpha,\beta)|\right) = 0, \text{ and } \lim_{N\to\infty} \Pr\left(|l_A(\mathbf{A}_N) - H(\alpha)|\right) = 0 \tag{7}$$

From (5) and (6), we see that (7) itself is equivalent to $p_{\mathbf{A}}(y) \equiv \frac{n_{\mathbf{A}}(y)}{N} \to \alpha(y),\ \forall y \in \mathcal{Y}$ i.e. the empirical distribution $p_{\mathbf{A}}(y)$ for long strings $\mathbf{A}$ converges to $\alpha(y)$. Putting these observations together we have:

$$\lim_{N\to\infty} l_A(\mathbf{A}_N) \equiv H(\alpha) \leq H(\alpha,\beta) \equiv \lim_{N\to\infty} l_B(\mathbf{A}_N) \tag{8}$$

Intuitively, (8) states that a long string $\mathbf{A}_N$ generated from model A is likely to have a lower perplexity for model A than for any other model B. This means that a model trained to have low perplexity over a set of reference texts will generate text strings that are statistically similar to the reference texts. This is the theoretical justification for using perplexity as a loss function for training language models and it is an excellent justification - provided only that we accept the Shannon model of language i.e. the idea that languages can be reasonably modeled as a stochastic sequence of tokens (Jakobson, 1961). Testing for abnormally high or low perplexity is an important tool for LLM applications as we discuss in Section 6.

2.2 Independence is All You Need

The simple AEP in Proposition 1 has been extended in various ways in the literature such as the following version (see Tal et al. (2017), Appendix B.4), which we will make use of in the sequel.

Proposition 2. Generalized AEP. *Consider a language model A' that generates text string $\mathbf{A}'_N = [A'_1,\ A'_2,\ \ldots,\ A'_N]$ where the tokens A'_n are drawn independently from a sequence of distributions with*

[2]Strictly speaking, a stationary random process "starts" at $n = -\infty$ rather than $n = 1$; here we ignore the tokens produced by the assumed random process before $n = 1$.

pmfs $\alpha_n(y),\ n = 1, 2, \ldots, N$. *Assuming the entropies of distributions* $\alpha_n(y)$ *are uniformly bounded i.e.* $H(\alpha_n(y)) < H_{max} < \infty,\ \forall n$, *the log perplexity* $l_{A'}(\mathbf{A}'_N)$ *of the string* $\mathbf{A}'_N$ *as defined in (3) obeys:*

$$\lim_{N\to\infty} \Pr[|l_{A'}(\mathbf{A}'_N) - h_{A'}(N)| > \epsilon] \equiv 0,\ \forall \epsilon > 0, \text{where } h_{A'}(N) \equiv \frac{1}{N}\sum_{n=1}^{N} H(\alpha_n) \tag{9}$$

In effect this means that log-perplexity is close to the average entropy of the distributions $\alpha_n(y),\ n = 1, 2, \ldots, N$ with high probability. The regularity condition that the entropies of $\alpha_n(y)$ are uniformly bounded is trivially true if the token dictionary $\mathcal{Y}$ is a finite set: $H(p(y)) \leq \log_2 |\mathcal{Y}|$ for any distribution $p(y)$ over $y \in \mathcal{Y}$.

While Proposition 2 does not lend itself to an intuitive interpretation in terms of the empirical distribution of the tokens A'_n, it too is a direct consequence of the Law of Large Numbers applied to the log-perplexity random variable. Much of the analysis in Section 2.1 can be extended to generalized models like A' in Proposition 2 that allows each token A'_n to be drawn from different distributions $\alpha_n(y)$. However, the tokens in these models must be drawn from *fixed distributions independent of past tokens*. This is still quite trivial compared to modern language models where the output tokens depend in highly complex and sophisticated ways on past tokens. For such models, we must give up independence and rely instead only on martingale difference properties.

3 An Un-Equipartition Property for Synthetic Text

We now propose a generalization of the theory described in Section 2.1 to a more interesting class of models. Consider a language model M and a random infinitely long text string $\mathbf{Y}$ generated by M, whose probabilities for a given prompt $\mathbf{X}$ is described by (2). Specifically, given a model M and a string $\mathbf{Y}$, (2) defines a sequence of probability distributions $p_n(y)$ over the set of tokens $y \in \mathcal{Y}$. We will assume that the prompt $\mathbf{X}$ is fixed and omit it from our notation in the sequel.

Definition 1. The path entropy $h_M(\mathbf{Y}_N)$ of model M for string $\mathbf{Y}_N$ is defined as:

$$h_M(\mathbf{Y}_N) \doteq \frac{1}{N}\sum_{n=1}^{N} H(p_n) \tag{10}$$

where $H(p) \equiv -\sum_{y\in\mathcal{Y}} p(y)\log_2(p(y))$ is the entropy of the distribution $p(y)$ over $y \in \mathcal{Y}$.

Note that we make no assumptions about the existence of an "entropy rate" $\lim_{N\to\infty} h_M(\mathbf{Y}_N)$ for the model M. However, we can show that

$$E\left[l_M(\mathbf{Y}_N)\right] \equiv E\left[h_M(\mathbf{Y}_N)\right] \equiv \frac{1}{N} H\left(\Pr(\mathbf{Y}_N)\right) \tag{11}$$

where $H\left(\Pr(\mathbf{Y}_N)\right)$ is the entropy of the random text strings $\mathbf{Y}_N$.

Definition 2. Let the log-deviation $\lambda(p)$ of a probability distribution $p(y)$ over $y \in \mathcal{Y}$ be defined as:

$$\lambda(p) \doteq \sqrt{S(p) - (H(p))^2},\ \text{where } S(p) \doteq \sum_{y\in\mathcal{Y}} p(y)\left(\log_2 p(y)\right)^2 \tag{12}$$

Note that the entropy $H(p)$ and log-deviation $\lambda(p)$ are the mean and standard deviation of the log-likelihood random variable $l_p(y) \equiv -\log_2 p(y)$ under distribution $p(y)$.

Definition 3. The log-deviation $\lambda_M(\mathbf{Y}_N)$ of a string $\mathbf{Y}_N$ for model M is defined as:

$$\lambda_M(\mathbf{Y}_N) \doteq \frac{1}{N}\sqrt{\sum_{n=1}^{N} \lambda^2(p_n)} \tag{13}$$

We can again interpret $H(p_n),\ \lambda(p_n)$ as the *conditional* mean and standard deviation of $-\log_2(\Pr(Y_n))$ conditioned on $\mathbf{Y}_{N-1}$. Clearly, if the log-deviations $\lambda(p_n)$ of the distributions $p_n(y)$ for a string $\mathbf{Y}$ are uniformly upper-bounded, $\lambda_M(\mathbf{Y}_N)$ asymptotically vanishes. We are now ready to state our first main result.

3.1 An LLN for LLMs

Proposition 3. Un-Equipartition Property for Perplexity. *For a text string $\mathbf{Y}_N$ generated from a language model with a given prompt $\mathbf{X}$, with high probability, the log perplexity $l_M(\mathbf{Y}_N)$ is asymptotically close to the path entropy $h_M(\mathbf{Y}_N)$. More precisely:*

$$\lim_{N\to\infty} \Pr[|l_M(\mathbf{Y}_N) - h_M(\mathbf{Y}_N)| > \epsilon] \equiv 0, \ \forall \epsilon > 0 \tag{14}$$

We want to argue that $\lambda_M(\mathbf{Y}_N)$ is the standard deviation of $l_M(\mathbf{Y}_N) - h_M(\mathbf{Y}_N)$, and therefore asymptotically vanishing $\lambda_M(\mathbf{Y}_N)$ implies $l_M(\mathbf{Y}_N) \to h_M(\mathbf{Y}_N)$. This is obviously very wrong - after all $\lambda_M(\mathbf{Y}_N)$ is itself a random variable. However, such informal reasoning is valuable for intuition and we show in Appendix B how to carry it through rigorously. We now present a formal proof of Proposition 3. It uses a version of the Weak Law of Large Numbers that applies to sequences of random variables with no restrictions on their dependence. Such a Law can be obtained using martingale difference ideas and an elementary derivation is presented in Appendix A.

Proof of Proposition 3. Consider the sequence of conditional likelihood random variables $\{l_n\}$ defined as:

$$l_n \doteq -\log_2 \Pr\left(Y_n | \mathbf{Y}_{n-1}\right) \equiv -\log_2\left(p_n(Y_n)\right) \tag{15}$$

We note that the random variables $Z_n \doteq l_n$ satisfy all of the conditions for Lemma 2, if we impose the restriction that they are strictly bounded i.e. $|l_n| < L$, for some some finite L. This is equivalent to requiring that all non-zero token probabilities $p_n(y)$ are above some lower bound e.g. $p_n(y) \geq 10^{-10}$. In practice, this is easily satisfied by any practical LLM. Then from Lemma 2, we have for any $\epsilon > 0$:

$$\lim_{N\to\infty} \Pr\left(\left|\frac{1}{N}\sum_{n=1}^{N}\left(l_n - H(p_n)\right)\right| > \epsilon\right) = 0 \tag{16}$$

From the definition (15), we note that $\frac{1}{N}\sum_{n=1}^{N} l_n \equiv l_M(\mathbf{Y}_N)$, and likewise $\frac{1}{N}\sum_{n=1}^{N} H(p_n) \equiv h_M(\mathbf{Y}_N)$. Combining these observations with (16) gives (14). □

3.2 Panning for Linguistic Gold

We can now formally define the *idiotypical set* $\mathcal{T}_M^n(\epsilon) \subset \mathcal{Y}^n$ for a model M with $\epsilon > 0$:

$$\mathcal{T}_M^n(\epsilon) \doteq \{\mathbf{Y}_n : |l_M(\mathbf{Y}_n) - h_M(\mathbf{Y}_n)| < \epsilon\} \equiv \left\{\mathbf{Y}_n : \left|h_M(\mathbf{Y}_n) + \frac{1}{n}\log_2\left(\Pr(\mathbf{Y}_n)\right)\right| < \epsilon\right\} \tag{17}$$

Proposition 3 asserts that with high probability a stochastic model M is constrained to only output idiotypical strings from this set $\mathcal{T}_M^n(\epsilon)$ for any $\epsilon > 0$ if we consider sufficiently long strings.

In the special case of stationary stochastic processes, the Unequipartition Property in Proposition 3 reduces to the classical Equipartition Property and the idiotypical set reduces to the classical typical set whose elements have nearly equal surprisals. Furthermore, for most practical applications of the classical AEP, the typical set turns out to be *vanishingly small*. This has powerful consequences: it identifies the small number of outcomes that actually matter, so we can focus our scarce resources (e.g. bandwidth) on them efficiently.

3.3 A Vanishing Subset of a Vanishing Subset

It turns out that the vanishing smallness of the idiotypical set continues to hold under the more general Unequipartition Property. However, for language models, this observation by itself is not very interesting: *we already know that LLMs possess an astonishing ability to sift through the vast majority of token strings that represent gibberish text, and generate coherent non-gibberish outputs.*

Crucially, the consequences of Proposition 3 go a lot farther than this. We will show a much stronger result that even if we restrict our attention to a vanishing subset of token strings, any practical LLM will still only

generate a vanishingly small fraction of that vanishing subset. To make this precise, consider a sequence of subsets $\mathcal{G}(n) \subset \mathcal{Y}^n$ parameterized by length n with the exponential growth rate $g(n) \doteq \frac{1}{n}\log_2|\mathcal{G}(n)|$. To avoid trivialities, we assume these subsets are rich enough to include sequences of unlimited length, meaning $g(n) \geq g_{min} > 0$. Furthermore, we assume $g(n) < \log_2|\mathcal{Y}|$, which implies that $\lim_{n\to\infty}\frac{|\mathcal{G}(n)|}{|\mathcal{Y}|^n} = 0$. For the practical applications discussed in this paper, we will interpret $\mathcal{G}(n)$ as a dictionary of grammatically correct, meaningful, or otherwise "acceptable" strings $\mathbf{Y}_n$ for a given task. Since most token sequences represent meaningless gibberish, any such dictionary naturally satisfies our assumed growth constraints. We can now state our second main result.

Proposition 4. Language Model Stereotypes. *All outputs $\mathbf{Y}_n \in \mathcal{G}(n)$ generated by a model M belongs to a vanishingly small subset of $\mathcal{G}(n)$ with asymptotically high probability. Specifically, the set $\mathcal{T}^N_{M,\mathcal{G}}(\epsilon) \doteq \mathcal{T}^N_M(\epsilon) \cap \mathcal{G}(n)$ satisfies the following:*

$$\lim_{N\to\infty}\Pr\left(\mathbf{Y}_N \in \mathcal{T}^N_{M,\mathcal{G}}(\epsilon)|\mathbf{Y}_N \in \mathcal{G}(N)\right) = 1, \text{ and } \lim_{N\to\infty}\frac{|\mathcal{T}^N_{M,\mathcal{G}}(\epsilon)|}{|\mathcal{G}(n)|} = 0, \ \forall\epsilon > 0 \tag{18}$$

We will call $\mathcal{T}^N_{M,\mathcal{G}}(\epsilon)$ as the grammatical idiotypical set. We will derive some intermediate results before presenting a formal proof of Proposition 4. Note the exceptionally broad scope of this result: it applies to any generative model M and the grammar $\mathcal{G}(n)$ can be any arbitrary sequence of sets that are strictly smaller than $\mathcal{Y}^n$, but satisfy an exponential growth condition. While this generality is appealing, the core ideas will be easier to follow if we refine our definitions to exclude certain edge cases. As an example, Proposition 4 trivially holds for a dictionary $\mathcal{G}(n)$ that contain large numbers of "phantom" strings in zero-probability sets that are invisible to a model M. We will now show one example of such a set.

Proposition 5. High entropy strings are unreachable. *The probability that the path entropy $h_M(\mathbf{Y}_N)$ of a string $\mathbf{Y}_N \in \mathcal{G}(N)$ generated by the model M is greater than $g(N) \equiv \frac{1}{N}\log_2|\mathcal{G}(N)|$ asymptotically vanishes, i.e.,*

$$\lim_{N\to\infty}\Pr\left(h_M(\mathbf{Y}_N) > g(N) + \epsilon\middle|\mathbf{Y}_N \in \mathcal{G}(N)\right) = 0, \ \forall\epsilon > 0 \tag{19}$$

Proof. From Proposition 3 we note that for sufficiently large n, we must have with high probability $|l_M(\mathbf{Y}_N) - h_M(\mathbf{Y}_N)| < \frac{\epsilon}{2}$. Then we have:

$$\Pr\left(h_M(\mathbf{Y}_N) > g(N) + \epsilon\middle|\mathbf{Y}_N \in \mathcal{G}(N)\right) \leq \frac{1}{p_G}\Pr\left(\mathcal{E}^N_1(\epsilon)\right)$$

where $\mathcal{E}^n_1(\epsilon) \doteq \left\{\mathbf{Y}_n \in \mathcal{G}(n) : l_M(\mathbf{Y}_n) > g(n) + \frac{\epsilon}{2}\right\}$. We can then bound this probability as:

$$\Pr(\mathcal{E}^n_1(\epsilon)) = \sum_{\mathbf{Y}_n\in\mathcal{E}^n_1(\epsilon)}\Pr(\mathbf{Y}_n) \leq \sum_{\mathbf{Y}_n\in\mathcal{E}^n_1(\epsilon)} 2^{-n(g(n)+\frac{\epsilon}{2})} \equiv |\mathcal{E}^n_1(\epsilon)|\,2^{-n(g(n)+\frac{\epsilon}{2})} < |\mathcal{G}(n)|\,2^{-n(g(n)+\frac{\epsilon}{2})} \equiv 2^{-\frac{n\epsilon}{2}}$$

Thus we have shown that $\Pr\left(h_M(\mathbf{Y}_N) > g(N) + \epsilon\middle|\mathbf{Y}_N \in \mathcal{G}(N)\right) < \frac{1}{p_G}2^{-\frac{N\epsilon}{2}} \to 0$ as $N \to \infty$. □

Proposition 5 shows that we can safely exclude high entropy strings from our dictionary without loss of generality. Specifically, with the *reachable set* $\mathcal{R}(\epsilon, n) \doteq \{\mathbf{Y}_n \in \mathcal{G}(n) : h_M(\mathbf{Y}_n) \leq g(n) + \epsilon\}$ we will assume that $\lim_{N\to\infty}\frac{|\mathcal{R}(\epsilon,N)|}{|\mathcal{G}(N)|} = 1, \ \forall\epsilon > 0$, i.e. that most grammatical strings are actually *reachable* by the model M. Next we refine the idea of grammar further to make it less abstract and more relevant to practical applications. This requires some analytical machinery which we now introduce.

4 The Grammar Police

Informally we want our grammar $\mathcal{G}(n)$ to represent "acceptable" or "meaningful" outputs from a model for interesting applications. We start with some motivating examples of such applications:

Ex 1 **Valid Python programs.** We define $\mathcal{G}(n)$ as the set of all syntactically correct Python code that compiles (or interprets) without error. While we know LLMs are capable of generating complex Python code, we will show that that any model is limited to generating a vanishing fraction of all error-free Python programs.

Ex 2 **Proofs in First Order Logic (FOL).** We define $\mathcal{G}(n)$ as the set of valid mathematical proofs in FOL. We will show that a model used to generate proofs in FOL is limited to generating a vanishing fraction of all correct proofs.

Ex 3 **Lyrical Poetry.** We define $\mathcal{G}(n)$ as the set of stories consisting of a sequence of haikus. We will show that any language model can only generate a vanishing fraction of all such possible stories.

Much of the excitement about modern LLMs arise from their apparent ability to generate grammatical outputs for applications like the above. However, being stochastic machines, they need to be suitably augmented to perform reliably on such tasks. In particular, we will assume that the model M knows how to generate grammatically correct sentences when suitably prompted, but we allow that it may still sometimes generate ungrammatical outputs. Formally we assume that $0 < p_G \leq \Pr(\mathbf{Y}_n \in \mathcal{G}(n)) < 1, \ \forall n$.

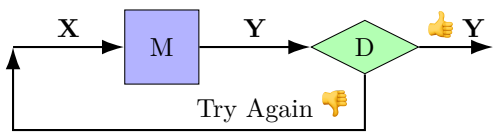


Figure 1: Language Model Augmented with a Grammar Checking Device.

Consider the augmented language model $(M, \mathcal{G})$ shown in Fig. 1, where the output of a model M is screened by a grammar-checking device D which defines the dictionary $\mathcal{G}(n)$; if the model's output $\mathbf{Y}$ is not grammatically correct, then it is prompted again to produce a new random string $\mathbf{Y}$ until it generates a grammatical string[3]. Under our assumption that $\Pr(\mathbf{Y}_n \in \mathcal{G}(n)) > p_G > 0$, the augmented model $(M, \mathcal{G})$ is guaranteed to generate a grammatical string in a finite time with probability one and the probability that model $(M, \mathcal{G})$ generates a grammatical string $\mathbf{Y}_n \in \mathcal{G}(n)$ can be bounded as:

$$2^{-nl_M(\mathbf{Y}_n)} \leq \Pr\left(\mathbf{Y}_n \middle| \mathbf{Y}_n \in \mathcal{G}(n)\right) \leq \frac{2^{-nl_M(\mathbf{Y}_n)}}{p_G} \tag{20}$$

Equation (20) is the key relationship we need to complete the proof of Proposition 4.

Remark. The construction in Fig. 1 implicitly assumes that membership in the set $\mathcal{G}(n)$ is *decidable*. Not all exponentially growing sets satisfy this decidability condition, but we expect that all practically interesting examples do. For undecidable grammars, we can simply treat (20) as a standalone assumption.

4.1 The Ideal Model that doesn't Discriminate

The per-symbol entropy of the augmented model $(M, \mathcal{G})$ can now be bounded as:

$$E\left[l_M(\mathbf{Y}_N) | \mathbf{Y}_N \in \mathcal{G}(N)\right] \leq \frac{1}{N} H\left(\Pr\left(\mathbf{Y}_N \middle| \mathbf{Y}_N \in \mathcal{G}(N)\right)\right) \leq g(N) \tag{21}$$

where both inequalities are satisfied with equality for the unique entropy-maximizing model that generates all grammatical strings with equal probability. This entropy-maximizing model satisfies some very special properties: the log-perplexity and path entropy of all grammatical strings are equal i.e. $l_M(\mathbf{Y}_n) \equiv h_M(\mathbf{Y}_n) \equiv g(n)$; Proposition 3 reduces to a simple Equipartition Property and the idiotypical set includes all grammatical strings. This ideal model outputs all grammatical outputs with equal probability and has no preference at all for any one grammatical text over another.

[3]Note that we assume that the device D is deterministic, which means our analysis does not allow the grammar checking device to be another stochastic LLM.

Practical LLMs do not behave like this. The ideal entropy-maximizing distribution represents a single point - a zero measure set - in the space of distributions. Even if a model happens to follow this exact distribution at one temperature setting, it will deviate from it at all other temperatures. While the precise mechanism for this deviation is not relevant to our analysis, certain empirical observations from the literature offer some insight. For example, practical LLMs have been shown to be subject to a precision/recall tradeoff Sajjadi et al. (2018), which means that there exist some strings that an LLM is less certain about than others. This simple observation has dramatic consequences in the asymptotic regime: even a small preference for one valid next token over another, compounded over a large number of tokens, results in a set of favored token strings that have overwhelmingly high probability compared to non-favored strings. We now formalize this observation and show that it immediately leads to Proposition 4.

Assumption 1. Practical LLMs are not entropy-maximizing. The path entropy $h_M(\mathbf{Y}_n)$ is strictly smaller than $g(n)$ for at least some non-zero fraction of valid idiotypical strings. Specifically, there is an entropy gap $\Delta g > 0$ such that $h_M(\mathbf{Y}_n) \leq g(n) - \Delta g$ for at least some non-zero fraction $\rho > 0$ of strings $\mathbf{Y}_n \in \mathcal{T}^n_{M,\mathcal{G}}(\epsilon)$:

$$|\mathcal{E}^n_2(\epsilon)| \geq \rho \left|\mathcal{T}^n_{M,\mathcal{G}}(\epsilon)\right|, \text{ where } \mathcal{E}^n_2(\epsilon) \doteq \left\{\mathbf{Y}_n \in \mathcal{T}^n_{M,\mathcal{G}}(\epsilon) : h_M(\mathbf{Y}_n) \leq g(n) - \Delta g\right\} \tag{22}$$

Proof of Proposition 4. For $\mathbf{Y}_n \in \mathcal{E}^n_2(\epsilon)$ with $\epsilon \leq \frac{\Delta g}{2}$, we have $l_M(\mathbf{Y}_n) \leq g(n) - \frac{\Delta g}{2}$. Then we have:

$$1 \geq \Pr(\mathcal{E}^n_2(\epsilon)) \geq \sum_{\mathbf{Y}_n \in \mathcal{E}^n_2(\epsilon)} 2^{-n l_M(\mathbf{Y}_n)} \geq \sum_{\mathbf{Y}_n \in \mathcal{E}^n_2(\epsilon)} 2^{-n\left(g(n) - \frac{\Delta g}{2}\right)} \equiv \frac{|\mathcal{E}^n_2(\epsilon)|}{|\mathcal{G}(n)|} 2^{\frac{n\Delta g}{2}} \geq \frac{\rho |\mathcal{T}^n_{M,\mathcal{G}}(\epsilon)|}{|\mathcal{G}(n)|} 2^{\frac{n\Delta g}{2}} \tag{23}$$

We have shown $|\mathcal{T}^n_{M,\mathcal{G}}(\epsilon)| \leq |\mathcal{G}(n)| 2^{-n\left(\frac{\Delta g}{2} - \frac{\log_2(\rho)}{n}\right)}$ thus establishing the second part of (18). The first part of (18) follows from Proposition 3 completing the proof. □

It should be emphasized that Assumption 1 is a very permissive condition that we expect to be universally satisfied by practical LLMs. It follows that Proposition 4 must hold for any reasonable grammar and model.

The obverse of Proposition 4 is that almost all grammatically correct strings are non-idiotypical for a given language model M. A string can be non-idiotypical in two different ways depending on whether its perplexity is abnormally high ("under-typical") or abnormally low ("over-typical"). Formally, we define the set $\mathcal{V}^n_{M,\mathcal{G}}(\epsilon)$ of over-typical strings of length n for a model M as:

$$\mathcal{V}^n_{M,\mathcal{G}}(\epsilon) \doteq \{\mathbf{Y}_n \in \mathcal{G}(n) : l_M(\mathbf{Y}_n) \leq h_M(\mathbf{Y}_n) - \epsilon < g(n) - \epsilon\} \tag{24}$$

Proposition 6. Over-Typical strings are statistical unicorns. *The size of the over-typical set $\mathcal{V}^n_{M,\mathcal{G}}(\epsilon)$ is an exponentially vanishing fraction of the dictionary $\mathcal{G}(n)$ i.e.:*

$$\lim_{N \to \infty} \frac{|\mathcal{V}^N_{M,\mathcal{G}}(\epsilon)|}{|\mathcal{G}(N)|} = 0, \ \forall \epsilon > 0 \tag{25}$$

Proof. The argument is straightforward and does not even require Assumption 1: over-typical strings, by definition, have very large probability masses, and therefore there just cannot be very many of them:

$$1 \geq \Pr\left(\mathbf{Y}_n \in \mathcal{V}^n_{M,\mathcal{G}}(\epsilon)\right) \geq \sum_{\mathbf{Y}_n \in \mathcal{V}^n_{M,\mathcal{G}}(\epsilon)} 2^{-n l_M(\mathbf{Y}_n)} \geq \left|\mathcal{V}^n_{M,\mathcal{G}}(\epsilon)\right| 2^{-n(g(n) - \epsilon)} \equiv \frac{\left|\mathcal{V}^n_{M,\mathcal{G}}(\epsilon)\right|}{|\mathcal{G}(n)|} 2^{n\epsilon} \tag{26}$$

Thus we have $\left|\mathcal{V}^n_{M,\mathcal{G}}(\epsilon)\right| \leq |\mathcal{G}(n)| \, 2^{-n\epsilon}$ which proves the result. □

Propositions 3 and 6, together require that almost all grammatical strings must be *under-typical* for any practical LLM.

5 Experiments with Open-Source Models

We performed a series of experiments to verify the ideas described in Sections 3 and 4. The basic idea is to use a local instance of a pre-trained open-source model to evaluate the path entropy of a given string using the model's conditional probability distributions and compare the result with the log-perplexity of the string again calculated using the model's probability distributions. We used the smaller GPT-2 model (Radford et al., 2019) as well as the larger Llama 3.1 8B (Touvron et al., 2023) model.

5.1 Perplexity of Self-Generated Strings

Our first set of experiments consisted of repeatedly generating a long text string from a model (GPT-2 or Llama 3.1 8B) and then evaluating its log-perplexity, path entropy and log-deviations for the same model. We fix the prompt string and the seed of the random number generator to allow for later reproduction. We then used top-k sampling with $k = 100$ to generate the probability distribution for the next token. We normalized the probabilities of the top-k tokens to sum to unity. We then created and saved a table consisting of token ID, token string and selection probability of each of the top k tokens. Finally, a random next token is chosen by sampling from the normalized distribution, and the new token is appended to the prompt to repeat the process for N_{max} total tokens. The random token Y_n chosen at each step along with the probability distribution $p_n(y)$ from which they were chosen were saved into a json file for later processing. We repeated this experiment many times for different seed values for the random number generator, different choices of sampling distributions and initial prompts.

For each generated string, we can then test our claim in Proposition 3. Specifically, for each token $N \in 1 \dots N_{max}$ we calculated the log-perplexity $l_M(\mathbf{Y}_N)$, path entropy $h_M(\mathbf{Y}_N)$ and the log-deviation $\lambda_M(\mathbf{Y}_N)$ of the substring $\mathbf{Y}_N$ consisting of the first N tokens of the generated string.

5.2 Perplexity of Externally Generated Strings

We also used a slightly modified version of the procedure described in Section 5.1 to calculate the log-perplexity and path entropy of an arbitrary string (that was not generated by the model itself) on the GPT-2 and Llama 3.1 8B models. We first choose a text string that we wish to analyze. We fix an initial fragment of the string as our prompt $\mathbf{X}$. We tokenize the remaining portion of the string excluding the prompt using the model's tokenizer. Let $Y_n,\ n = 1 \dots N_{max}$ denote the resulting sequence of tokens.

Stating with $n = 1$, we list the probability distribution $p_n(y)$ of the next token Y_n for the model using top-k with $k = 100$ sampling for the n'th token. We can then calculate the path entropy $h_M(\mathbf{Y}_N)$ and log-deviation $\lambda_M(\mathbf{Y}_N)$ as in Section 5.1. However, to calculate the log-perplexity $l_M(\mathbf{Y}_N)$, we also need the probability that our model will output the actual next token Y_n from the string $\mathbf{Y}$ which may not be included in the list of top-k tokens[4]. When this happens, we replaced the lowest probability entry in the list of top-k tokens with the actual next token from our string along with its probability.

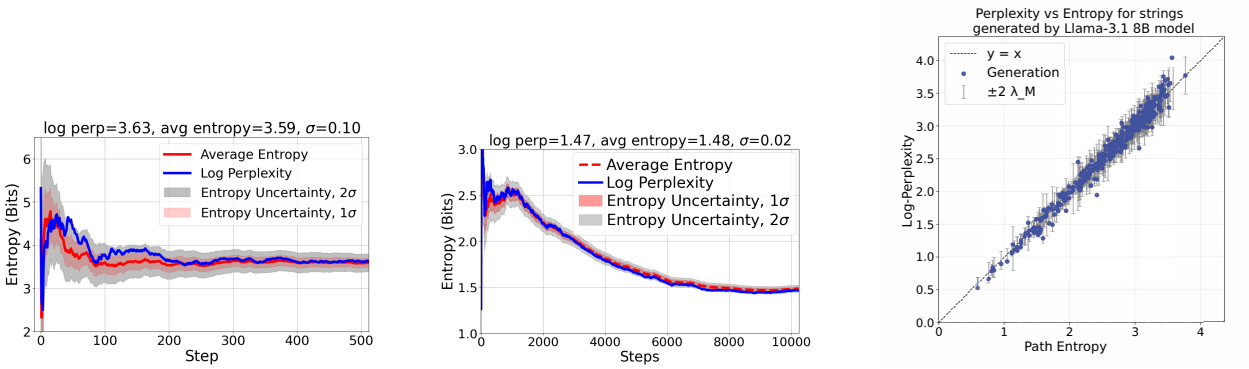
The normalized probability of each of the top-k tokens along with their token-ID and token strings, as well as the probability and ID of the actual next token are saved in a json file. This json file is then processed in exactly the same way as in Section 5.1 to compute $l_M(\mathbf{Y}_N),\ h_M(\mathbf{Y}_N)$ and $\lambda_M(\mathbf{Y}_N)$ of the sub-strings $\mathbf{Y}_N,\ N = 1 \dots N_{max}$ for the GPT-2 or Llama 3.1 8B model.

5.3 Results and Discussion

Idiotypical text. Figure 2 shows the log-perplexity and path entropy for sub-strings $\mathbf{Y}_N$ for the GPT-2 or Llama 3.1 8B model as a function of N for several text strings generated by the same model. The plots also shows the range $h_M(\mathbf{Y}_N) \pm \lambda_M(\mathbf{Y}_N)$ and $h_M(\mathbf{Y}_N) \pm 2\lambda_N(\mathbf{Y}_N)$. We see that the log-perplexity $l_M(\mathbf{Y}_N)$ stays within the range $h_M(\mathbf{Y}_N) \pm 2\lambda_M(\mathbf{Y}_N)$ consistently in both cases. The asymptotic convergence is more clearly seen in the long string from the Llama 3.1 8B model[5] in Fig. 2b.

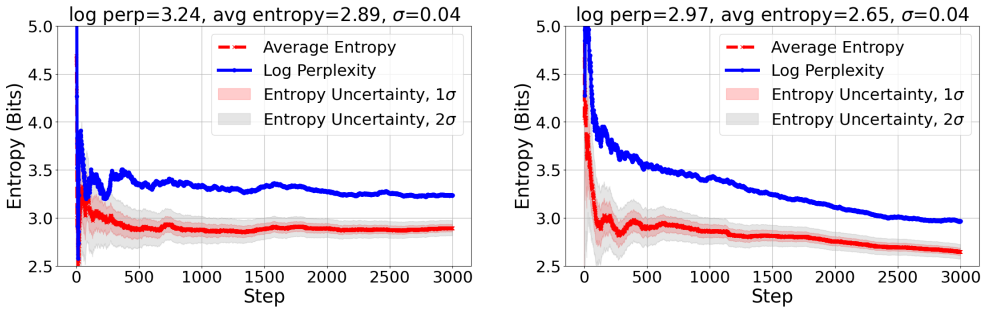
[4]Our choice of top-k sampling is a compromise. The number of tokens in GPT-2 or Llama is quite large ($|\mathcal{T}| \approx 40,000$). If we exhaustively list the full probability distribution of all tokens, the experiment becomes very computationally expensive and slow.

[5]The GPT-2 model cannot generate similarly long texts.



(a) GPT-2, $N_{max} = 512$. (b) Llama 3.1 8B, $N_{max} = 10240$. (c) Scatter Plot for Llama-3.1 8B.

Figure 2: Log-perplexity & path entropy for self-generated text with top-100 sampling.



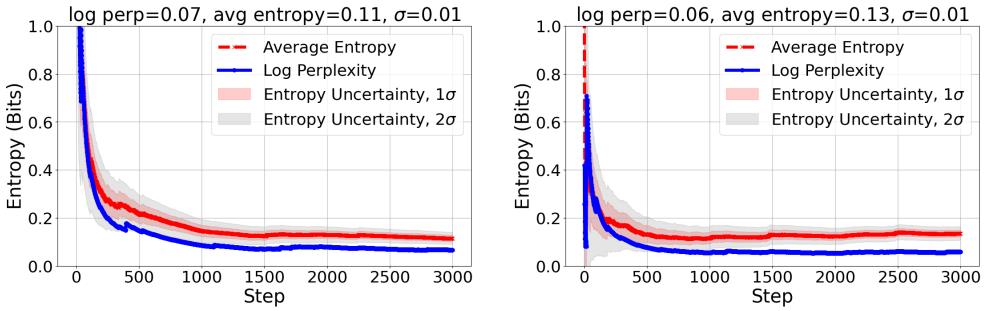
(a) Human writing. (b) Text from ChatGPT o1 pro.

Figure 3: Log-perplexity and path entropy for Llama 3.1 8B of non-Llama-generated strings $\mathbf{Y}$.

While Proposition 3 holds regardless of the quality of the text, its practical relevance depends on whether the predicted asymptotic behavior is observable in normal-length texts. Figure 2c shows the log-perplexity $l_M(\mathbf{Y}_N)$ against the path entropy $h_M(\mathbf{Y}_N)$ for the Llama-3.1 8B model for several text strings generated by the same model. A total of 450 strings were generated using top-100 sampling from 15 distinct prompts and 30 different random seeds for each prompt. Of these, abnormally short strings of less than 100 tokens were excluded from the analysis. The remaining 383 strings had an average length of 685 tokens and a standard deviation of 321. The clustering of the log-perplexity around the 45° line in Fig. 2c provides direct visual affirmation for Proposition 3. The fact that the path entropy of the generated strings vary over a wide range in Fig. 2c also provides direct evidence for the Assumption 1 that the Llama-3.1 8B model is not entropy-maximizing. Additional plots showing the asymptotic behavior are presented in Appendix C.

Under-typical text. Next we consider Fig. 3 that analyses two long, non-Llama generated strings. The text used in Fig. 3a is an excerpt from an article Pelly (2025) in Harper's Magazine, a respected literary publication. The text in Fig. 3b was generated by ChatGPT o1 Pro OpenAI (2024). We see that the log-perplexity $l_M(\mathbf{Y}_N)$ for Llama 3.1 8B is several standard deviations $\lambda_M(\mathbf{Y}_N)$ larger than the path entropy $h_M(\mathbf{Y}_N)$ for both strings in Fig. 3 showing that these strings are clearly *under-typical* for the Llama 3.1 8B model. If we assume that the strings in Fig. 3 are sampled from a random manifold of the grammar, then this under-typicality is a simple consequence of the smallness of the Llama-3.1 8B model's idiotypical sets.

Over-typical text. This brings us to the remarkable plots in Fig. 4 which shows the log-perplexity of excerpts from classic texts for Llama 3.1 8B, specifically from Alice in Wonderland and The Great Gatsby in Figs. 4a and 4b respectively. The average entropy for the token distributions in Fig. 4a is an astonishingly small 0.11 bits, which means that the model on average assigns a very high probability ($\sim 99\%$) to the



(a) Alice in Wonderland excerpt.

(b) The Great Gatsby excerpt.

Figure 4: Log-perplexity and path entropy for Llama 3.1 8B of excerpts from literary classics.

precise sequence of tokens from the excerpt for its predicted next tokens. This shows eidetic memorization Carlini et al. (2021) of the excerpt by the Llama 3.1 8B model: with zero temperature, the model will output this text verbatim. The same observation also applies to Fig. 4b. We also see from both plots that the log-perplexity is several standard deviations *below* the path entropy making the excerpts slightly, but unambiguously, *over-typical* for the Llama-8B model.

In summary, we see from Fig. 2 that synthetic texts generated by an LLM are *idiotypical* for the model, while most external texts such as in Fig. 3 are *under-typical*. By definition, *over-typical* strings are very special such as the excerpts from literary classics in Fig. 4 that have been memorized by the model.

6 Practical applications: the capybara problem

A simple and direct application of the ideas in this work is for testing whether a given piece of text has "abnormally" high or low perplexity for a language model. Such a test is relevant to many important practical problems. Low perplexity is commonly used as one marker of synthetic text in AI text detection tools (Wu et al., 2025) by testing $l_M(\mathbf{Y}_n) \lessgtr l_{th}$ against a threshold l_{th}. However, it turns out that choosing a suitable classification threshold l_{th} is difficult - the so-called "capybara problem" (Hans et al., 2024).

Our theory shows a principled way to do this: with $l_{th} = h_M(\mathbf{Y}_n) + 3\lambda_M(\mathbf{Y}_n)$, we can accurately distinguish between synthetic text generated by Llama 3.1 8B and non-Llama text such as in Fig. 3 even with relatively short strings of ≈ 500 tokens. The key insight is to choose *variable* rather than fixed thresholds, i.e. different l_{th} for different strings $\mathbf{Y}_n$. Similar comments apply to the dataset inference problem (Maini et al., 2024): while it is well-known (Carlini et al., 2021) that text strings memorized by an LLM have low perplexity, our theory allows us to rigorously derive threshold values.

7 Conclusions

We presented a generalization of the classical Asymptotic Equipartition Property that must be satisfied by the perplexity of any long string generated by a language model and provided theoretical and experimental arguments in support of the property. We used this property to extend the information theoretic concept of typical set to language models. The main take-aways are as follows:

1. The classical AEP and the concept of typical sets can be generalized to apply to the outputs of LLMs.
2. These "idiotypical" sets can be shown to be a vanishingly small subset of possible "good" outputs and this means that LLMs are strongly constrained in what they can generate.
3. Experimentally we showed that the asymptotic behavior in the AEP can be observed for LLMs for moderate length texts (a few hundred tokens), and thus may be practically relevant to detecting texts of abnormally high or low perplexity.

References

Paul H. Algoet and Thomas M. Cover. A sandwich proof of the Shannon–McMillan–Breiman theorem. *The Annals of Probability*, 16(2):899–909, 1988. ISSN 00911798, 2168894X. URL http://www.jstor.org/stable/2243846.

Anthropic. Claude 3 Opus. https://claude.ai, 2024. URL https://www.anthropic.com.

Andrew R Barron. The strong ergodic theorem for densities: generalized Shannon-McMillan-Breiman theorem. *The annals of Probability*, 13(4):1292–1303, 1985.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

Leo Breiman. The individual ergodic theorem of information theory. *The Annals of Mathematical Statistics*, 28(3):809–811, 1957.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650. USENIX Association, August 2021. ISBN 978-1-939133-24-3. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Noam Chomsky. Review of B. F. Skinner, Verbal Behavior. *Language*, 35(1):26–58, 1959.

Noam Chomsky. *Aspects of the Theory of Syntax*. Number 11. MIT press, 2014.

Kai Lai Chung. A note on the ergodic theorem of information theory. *The Annals of Mathematical Statistics*, 32(2):612–614, 1961.

Joel E. Cohen. Markov's inequality and Chebyshev's inequality for tail probabilities: A sharper image. *The American Statistician*, 69(1):5–7, 2015. doi: 10.1080/00031305.2014.975842.

Trevor Cohen and Serguei Pakhomov. A tale of two perplexities: Sensitivity of neural language models to lexical retrieval deficits in dementia of the Alzheimer's type. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1946–1957, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.176. URL https://aclanthology.org/2020.acl-main.176.

T. Cover and J. Thomas. Asymptotic equipartition property. In *Elements of Information Theory*, chapter 3, pp. 57–69. John Wiley & Sons, Ltd, 2005a. ISBN 9780471748823. doi: https://doi.org/10.1002/047174882X.ch3. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/047174882X.ch3.

T. Cover and J. Thomas. Entropy, relative entropy, and mutual information. In *Elements of Information Theory*, chapter 2, pp. 13–55. John Wiley & Sons, Ltd, 2005b. ISBN 9780471748823. doi: https://doi.org/10.1002/047174882X.ch2. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/047174882X.ch2.

I. Csiszar. The method of types [information theory]. *IEEE Transactions on Information Theory*, 44(6):2505–2523, 1998. doi: 10.1109/18.720546.

Vera Demberg and Frank Keller. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210, 2008. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2008.07.008. URL https://www.sciencedirect.com/science/article/pii/S0010027708001741.

Pablo Gamallo, José Ramom Pichel Campos, and Inaki Alegria. A perplexity-based method for similar languages discrimination. In *Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects (VarDial)*, pp. 109–114, 2017.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 111–116, 2019.

D Hamdan, W Parry, and J-P Thouvenot. Shannon entropy for stationary processes and dynamical systems. *Ergodic Theory and Dynamical Systems*, 28(2):447–480, 2008.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting LLMs with binoculars: Zero-shot detection of machine-generated text, 2024. URL https://arxiv.org/abs/2401.12070.

S. I. Idele. The law of large numbers: Some issues of interpretation and application for beginners. *Journal of the Royal Statistical Society Series D: The Statistician*, 28(3):209–217, 12 2018. ISSN 2515-7884. doi: 10.2307/2987870. URL https://doi.org/10.2307/2987870.

Roman Jakobson. Linguistics and communications theory. 1961.

Niels K Jerne. Toward a network theory of immune system. *Ann Immunol (Paris)*, 125(1):373–389, 1974.

Roger Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177, 2008. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2007.05.006. URL https://www.sciencedirect.com/science/article/pii/S0010027707001436.

Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. LLM dataset inference: Did you train on my dataset?, 2024. URL https://arxiv.org/abs/2406.06443.

Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

Brockway McMillan. The basic theorems of information theory. *The Annals of mathematical statistics*, pp. 196–219, 1953.

Clara Meister and Ryan Cotterell. Language model evaluation beyond perplexity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5328–5339, 2021.

Anna Mészáros, Szilvia Ujváry, Wieland Brendel, Patrik Reizinger, and Ferenc Huszár. Rule extrapolation in language modeling: A study of compositional generalization on ood prompts. *Advances in Neural Information Processing Systems*, 37:34870–34899, 2024.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. What makes my model perplexed? a linguistic investigation on neural language models perplexity. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 40–47, 2021.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24950–24962. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/mitchell23a.html.

Shu-Teh Chen Moy. Generalizations of Shannon-McMillan theorem. *Pacific Journal of Mathematics*, 11(2): 705–714, 1961.

M. Nishiara and H. Morita. On the AEP of word-valued sources. *IEEE Transactions on Information Theory*, 46(3):1116–1120, 2000. doi: 10.1109/18.841193.

OpenAI. OpenAI o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

OpenAI. Hello GPT-4o. https://openai.com/index/hello-gpt-4o, 2024.

Liz Pelly. The ghosts in the machine. *Harper's Magazine*, pp. 25–32, Jan 2025.

Ayse Pinar Saygin, Ilyas Cicekli, and Varol Akman. Turing test: 50 years later. *Minds and machines*, 10(4): 463–518, 2000.

Mark Pollicott and Michiko Yuri. *Dynamical systems and ergodic theory*, volume 40. Cambridge University Press, 1998.

Lawrence R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Patrik Reizinger, Szilvia Ujváry, Anna Mészáros, Anna Kerekes, Wieland Brendel, and Ferenc Huszár. Position: Understanding LLMs requires more than statistical generalization. In *International Conference on Machine Learning*, pp. 42365–42390. PMLR, 2024.

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018.

Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423, 1948.

Claude E. Shannon. Prediction and entropy of printed English. *The Bell System Technical Journal*, 30(1): 50–64, 1951.

Omri Tal, Tat Dat Tran, and Jacobus Portegies. From typical sequences to typical genotypes. *Journal of Theoretical Biology*, 419:159–183, 2017. ISSN 0022-5193. doi: https://doi.org/10.1016/j.jtbi.2017.02.010. URL `https://www.sciencedirect.com/science/article/pii/S0022519317300656`.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. The science of detecting LLM-generated text. *Communications of the ACM*, 67(4):50–59, 2024.

Roy Timo, Kim Blackmore, and Leif Hanlen. Word-valued sources: An ergodic theorem, an aep, and the conservation of entropy. *IEEE Transactions on Information Theory*, 56(7):3139–3148, 2010. doi: 10.1109/TIT.2010.2046251.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models, 2023. URL `https://arxiv.org/abs/2302.13971`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30: 5998–6008, 2017.

Zhongzhi Wang and Weiguo Yang. The generalized entropy ergodic theorem for nonhomogeneous Markov chains. *Journal of Theoretical Probability*, 29:761–775, 2016.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, pp. 1–66, 2025.

Weiguo Yang and Wen Liu. The asymptotic equipartition property for mth-order nonhomogeneous Markov information sources. *IEEE transactions on information theory*, 50(12):3326–3330, 2004.

A A Generalized Law of Large Numbers for Random Variables with Arbitrary Dependence

Consider a infinitely long string $\mathbf{Y} = [Y_1\ Y_2\ \ldots\ Y_n\ \ldots]$ consisting of a sequence of symbols $Y_n \in \mathcal{Y}$ generated stochastically from a finite dictionary $\mathcal{Y}$. We will denote the substring consisting of the first N symbols as $\mathbf{Y}_N \equiv [Y_1\ Y_2\ \ldots Y_N]$ and the joint distribution of the symbols $P(\mathbf{Y}_N)$. We make no assumptions about the distribution $P(\mathbf{Y}_N)$ from which the symbols Y_n are generated. In particular, we do not assume that the symbols are independent or weakly dependent even asymptotically. We also do not assume stationarity i.e. the marginal distributions of symbols Y_n and Y_m can be completely different. We also do not assume any Markovian or conditional independence properties.

A.1 Conditional Means and Sample Means for Sequences of Dependent Random Variables

Consider a sequence of random variables $\mathbf{Z} = [Z_1\ Z_2\ \ldots\ Z_n\ \ldots]$ whose n'th element Z_n is a function of the first n symbols $Y_1, Y_2, \ldots, Y_n$ i.e. $Z_n \doteq g_n(\mathbf{Y}_n)$ and $g_n : \mathcal{Y}^n \to \mathbb{R}$ are functions that we will assume to be uniformly bounded i.e. $|g_n(\mathbf{Y}_n)| \leq G < \infty,\ \forall n \in \mathbb{Z}^+, \mathbf{Y}_n \in \mathcal{Y}^n$.

We define the conditional mean and conditional variance of the random variables Z_n:

$$\mu_n^{(\mathbf{Y})} \doteq \sum_{y \in \mathcal{Y}} p_n^{(\mathbf{Y})}(y) g_n\left([\mathbf{Y}_{n-1}\ y]\right) \equiv E\left[Z_n | \mathbf{Y}_{n-1}\right] \tag{27}$$

$$V_n^{(\mathbf{Y})} \doteq \sum_{y \in \mathcal{Y}} p_n^{(\mathbf{Y})}(y) \left(g_n\left([\mathbf{Y}_{n-1}\ y]\right) - \mu_n^{(\mathbf{Y})}\right)^2 \equiv E\left[Z_n^2 | \mathbf{Y}_{n-1}\right] - \left(\mu_n^{(\mathbf{Y})}\right)^2 \tag{28}$$

where the distribution $p_n^{(\mathbf{Y})}(y) \doteq \Pr(Y_n = y | \mathbf{Y}_{n-1})$. Since the random variables Z_n are discrete-valued and bounded, they have bounded moments for any distribution:

$$|\mu_n^{(\mathbf{Y})}| \leq G,\ V_n^{(\mathbf{Y})} \leq G^2 \tag{29}$$

A more formal statement. Let $\mathcal{F}_n$ be the σ-algebra defined by $\mathbf{Y}_n \equiv [Y_1\ Y_2\ \ldots\ Y_n]$ i.e. $\mathcal{F}_n$ is a collection that includes all (mathematically reasonable) events defined on the first n symbols. Then the random variables Z_n as we have defined them are defined by the condition that Z_m is $\mathcal{F}_m$-measurable, and $\mu_m^{(\mathbf{Y})} \equiv \mathrm{E}\left[Z_m \middle| \mathcal{F}_{m-1}\right]$.

We need some intermediate results before presenting our main result. First, we define the partial sums of the sequence Z_n and conditional means:

$$S_n \doteq \sum_{m=1}^{n} Z_m,\ \mu_{S_n}^{(\mathbf{Y})} \doteq \sum_{m=1}^{n} \mu_m^{(\mathbf{Y})} \tag{30}$$

Clearly, $\mathrm{E}\left[Z_n\right] \equiv \mathrm{E}\left[\mu_n^{(\mathbf{Y})}\right]$ and $\mathrm{E}\left[S_n\right] \equiv \mathrm{E}\left[\mu_{S_n}^{(\mathbf{Y})}\right]$. The key idea is that the sequence $Z_m - \mu_m^{(\mathbf{Y})}$ is a martingale difference sequence, and $S_n - \mu_{S_n}^{(\mathbf{Y})}$ is a martingale for any arbitrary distribution of the Z_m's.

Lemma 1. *The variance of the zero-mean random variable $S_n - \mu_{S_n}^{(\mathbf{Y})}$ can be written as sums of contributions from each of the Z_m's. Specifically, we have:*

$$V_{S_n} \doteq E\left[\left(S_n - \mu_{S_n}^{(\mathbf{Y})}\right)^2\right] = \sum_{m=1}^{n} E\left[V_m^{(\mathbf{Y})}\right] \tag{31}$$

Proof. We will use the method of induction to prove (31). For $n = 1$, (31) holds trivially. Now, let's assume (31) holds for $n = l$ i.e. assume that:

$$V_{S_l} \equiv \sum_{m=1}^{l} E\left[V_m^{(\mathbf{Y})}\right] \tag{32}$$

Now consider $S_{l+1} \equiv S_l + Z_{l+1}$. Its variance can be written as:

$$V_{S_{l+1}} = E\left[\left(S_l + Z_{l+1} - \mu_{S_l}^{(\mathbf{Y})} - \mu_{l+1}^{(\mathbf{Y})}\right)^2\right] \tag{33}$$

$$= V_{S_l} + E\left[V_{l+1}^{(\mathbf{Y})}\right] + 2\,E\left[\left(S_l - \mu_{S_l}^{(\mathbf{Y})}\right)\left(Z_{l+1} - \mu_{l+1}^{(\mathbf{Y})}\right)\right] \tag{34}$$

We will now show that the last term in (34) is zero. We apply the law of iterated expectations to write:

$$E\left[\left(S_l - \mu_{S_l}^{(\mathbf{Y})}\right)\left(Z_{l+1} - \mu_{l+1}^{(\mathbf{Y})}\right)\right] = \underset{\mathbf{Y}_l}{E}\left[E\left[\left(S_l - \mu_{S_l}^{(\mathbf{Y})}\right)\left(Z_{l+1} - \mu_{l+1}^{(\mathbf{Y})}\right)|\mathbf{Y}_l\right]\right] \tag{35}$$

$$= \underset{\mathbf{Y}_l}{E}\left[\left(S_l - \mu_{S_l}^{(\mathbf{Y})}\right) E\left[\left(Z_{l+1} - \mu_{l+1}^{(\mathbf{Y})}\right)|\mathbf{Y}_l\right]\right] \equiv 0 \tag{36}$$

where we used the fact that $S_l - \mu_{S_l}^{(\mathbf{Y})}$ is a deterministic function of $\mathbf{Y}_l$ and $\mu_{l+1}^{(\mathbf{Y})} \equiv \mathrm{E}\left[Z_{l+1}|\mathbf{Y}_l\right]$.

Remark. An alternative way to arrive at (36) is to treat $\mu_{l+1}^{(\mathbf{Y})} \equiv \mathrm{E}\left[Z_{l+1}\middle|\mathbf{Y}_l\right]$ as an estimate of Z_{l+1} given $\mathbf{Y}_l$, and then use the Orthogonality Principle to argue that the estimation error $\left(Z_{l+1} - \mu_{l+1}^{(\mathbf{Y})}\right)$ must be uncorrelated with any function of $\mathbf{Y}_l$. More formally, $\left(S_l - \mu_{S_l}^{(\mathbf{Y})}\right)$ is $\mathcal{F}_l$-measurable and is therefore orthogonal to $(Z_{l+1} - \mathrm{E}\left[Z_{l+1}|\mathcal{F}_l\right])$.

Using (36), we can now simplify (34) to:

$$V_{S_{l+1}} = V_{S_l} + E\left[V_{l+1}^{(\mathbf{Y})}\right] \equiv \sum_{m=1}^{l+1} E\left[V_m^{(\mathbf{Y})}\right] \tag{37}$$

where we used the induction assumption (32). We have shown that if (31) holds for $n = l$, then (31) must also hold for $n = l + 1$, thus completing the induction. □

Lemma 2. *The sequence* $\frac{1}{n}\left(S_n - \mu_{S_n}^{(\mathbf{Y})}\right)$ *converges to zero in probability:*

$$\lim_{n\to\infty} \Pr\left(\frac{1}{n}\left|S_n - \mu_{S_n}^{(\mathbf{Y})}\right| > \epsilon\right) = 0,\ \forall \epsilon > 0 \tag{38}$$

Proof. Using Lemma 1, we can write:

$$V_{S_n} = \sum_{m=1}^{n} E\left[V_m^{(\mathbf{Y})}\right] \leq nG^2 \tag{39}$$

using the simple bounds in (29). The variance of $\frac{1}{n}\left(S_n - \mu_{S_n}^{(\mathbf{Y})}\right)$ is simply $\frac{1}{n^2}V_{S_n} \leq \frac{G^2}{n} \to 0$ as $n \to \infty$. Equation (38) now follows immediately using the Chebyshev Inequality (Cohen, 2015). □

Lemma 2 states that the sample means of the two random sequences $\{Z_n\}$ and $\{\mu_n^{(\mathbf{Y})}\}$ must be close to each other asymptotically. Since we have made very few assumptions about the random variables involved, we cannot assume that either of the sequences $\{Z_n\}$ or $\{\mu_n^{(\mathbf{Y})}\}$ themselves converge in any sense to an asymptotic limit. Neither can we assume that their sample means separately converge to a limit.

With an additional assumption stating the existence of one such limit, we can establish a Weak Law of Large Numbers generalized to its most extreme i.e. a statement that the sample mean of a sequence of random variables $\{Z_n\}$ with arbitrary dependence converges asymptotically to a suitably defined average quantity in probability.

Proposition 7. A Weak Law of Large Numbers. *If the average of the conditional means converges to a limit i.e.* $\mu^{(\mathbf{Y})} \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} \mu_m^{(\mathbf{Y})}$ *exists, then the sequence* $\frac{1}{n}S_n$ *converges in probability to* $\mu^{(\mathbf{Y})}$*:*

$$\frac{1}{n}\sum_{m=1}^{n} Z_n \xrightarrow[p]{} \mu^{(\mathbf{Y})} \quad \text{or} \quad \lim_{n\to\infty} \Pr\left(\left|\frac{1}{n}S_n - \mu^{(\mathbf{Y})}\right| > \epsilon\right) = 0, \ \forall \epsilon > 0 \tag{40}$$

Proof. Combining Lemma 2 with the observation $\frac{1}{n}\mu_{S_n}^{(\mathbf{Y})} \to \mu^{(\mathbf{Y})}$ completes the proof. □

B Informal Proof: Roads not Traveled

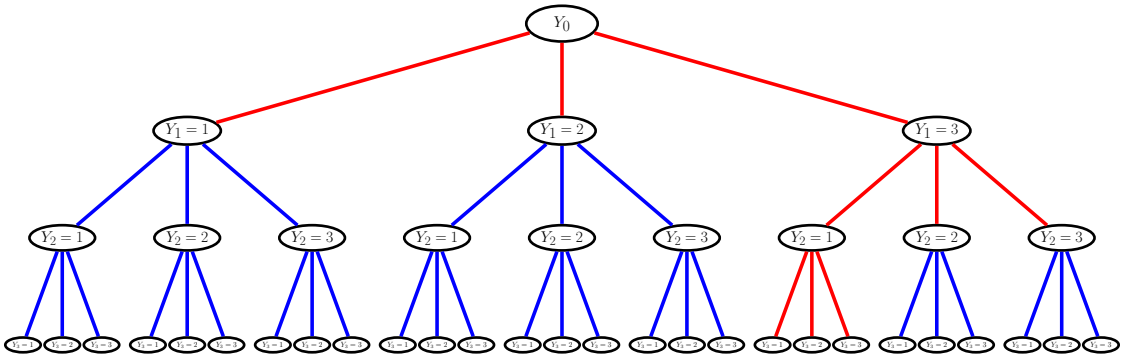


Figure 5: Probability tree for the first 3 tokens of a language model M that selects each token Y_n from a dictionary $\mathcal{Y}$ of three tokens. Also shown in red is the probability tree of an auxiliary model $M'_{\mathbf{S}}$ for a string $\mathbf{S}$ with $S_1 = 3,\ S_2 = 1$ and so on.

Consider a fixed string $\mathbf{S}$. For this particular string $\mathbf{S}$ and model M, we have a fixed set of distributions $q_n(y) \doteq \Pr(S_n = y|\mathbf{S}_{n-1})$. Now define an auxiliary generative model $M'_{\mathbf{S}}$ that generates a random string $\mathbf{Y}' \doteq [Y'_1,\ Y'_2,\ \ldots,\ Y'_N,\ \ldots]$ where the tokens Y'_n are generated *independently* from the distributions $q_n(y)$. If we represent a language model M as a probability tree, an auxiliary model $M'_{\mathbf{S}}$ represents one specific subtree corresponding to the string $\mathbf{S}$ that defines the model. This is illustrated in Fig. 5. The auxiliary model $M'_{\mathbf{S}}$ is much simpler than the general language model M, because its output tokens Y'_n are generated independently of other tokens. Our informal proof of Proposition 3 is based on the following idea: while calculating $\Pr(E)$ for some event E, we can change any part of the probability tree that do not affect the event E ("the roads not traveled") to make our calculation easier e.g. by allowing us to work with an auxiliary language model $M'_{\mathbf{S}}$ which generates strings of independent tokens.

Informal Proof of Proposition 3. Assuming the tokens Y_n are generated by the auxiliary model $M'_{\mathbf{S}}$ (instead of the original model M), the random variables $l_m,\ l_n$ are independent and $\mathrm{E}\left[l_n\right] \equiv H\left(q_n\right)$. Equation (16) then immediately follows from Proposition 2. □

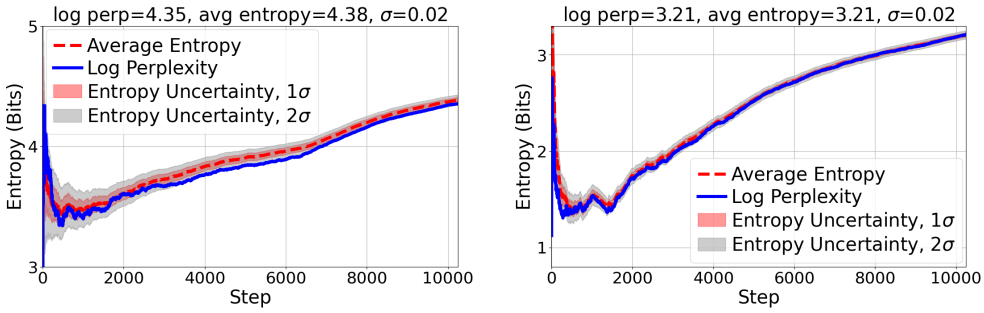
This informal method allows us to treat the martingale property that the formal proof relies on as a quasi-independence condition to reason about the random variables l_n intuitively. As an example, for the auxiliary model $M'_{\mathbf{S}}$, from (12, 13), the variances of $l_n,\ l_M(\mathbf{S}_N)$ are respectively $\lambda^2(q_n),\ \lambda^2_M(\mathbf{S}_N)$. We then have the following simple bound on deviations using the Chebyshev's Inequality Cohen (2015):

$$\Pr\left(|l_M(\mathbf{S}_N) - h_M(\mathbf{S}_N)| > \alpha\lambda_M(\mathbf{S}_N)\right) \leq \frac{1}{\alpha^2},\ \forall\alpha > 0 \tag{41}$$

C Additional Experimental Context and Results

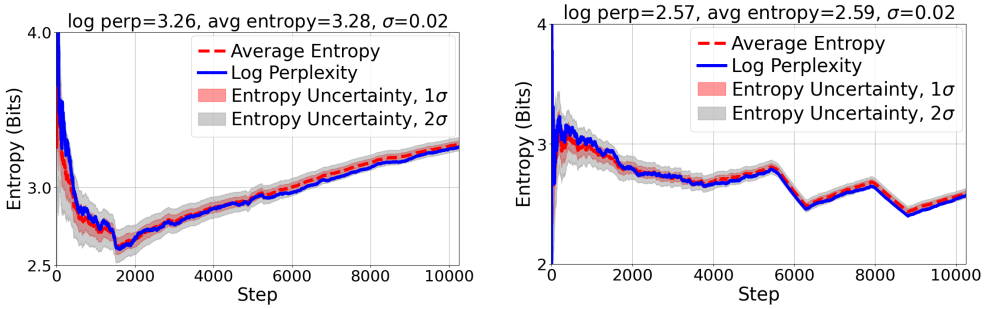
In this work, all experiments were ran on a server running Linux (Ubuntu) on AMD EPYC 7413 24-Core Processor with 48 logical cores, 96 GB RAM, four NVIDIA A100-SXM4-80GB GPUs. On this setup, each run of a text generation (i.e., running GPT-2 to generate Nmax = 512 tokens, Llama-8B to generate Nmax = 1024 tokens and Nmax = 10240 tokens, their distributions, etc.), file I/O, data analysis and plotting takes several seconds to minutes, with each experiment taking about 90 minutes to run in the case of 10,240 steps. We were able to run a few hundred of these experiments for this paper from which selected examples are presented in Section 5.3.

Figure 6 shows additional results similar to Fig. 2 where we naively varied the seed value of the random number generator from $209 - 214$ for Llama-8B. We present these to show that Fig. 2's results are representative of the general case and were not cherry-picked. In all cases, $K = 100$.



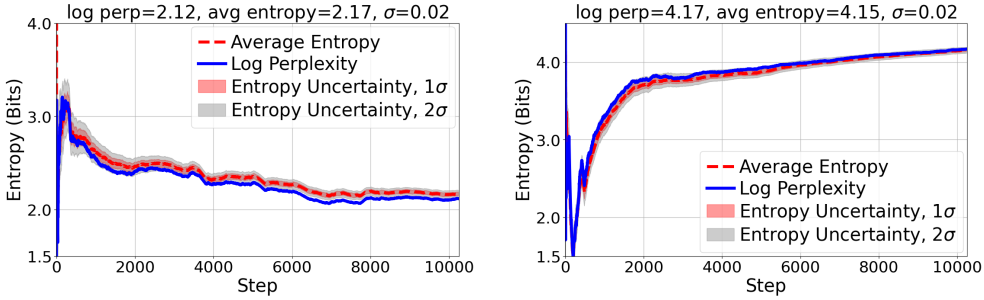
(a) Seed=209.

(b) Seed=210.



(c) Seed=211.

(d) Seed=212.



(e) Seed=213.

(f) Seed=214.

Figure 6: Log-perplexity and empirical entropy for Llama 3.1 of strings generated by the same model with top-100 sampling.