

THE MAGNITUDE OF CATEGORIES OF TEXTS ENRICHED BY LANGUAGE MODELS

TAI-DANAE BRADLEY AND JUAN PABLO VIGNEAUX

ABSTRACT. The purpose of this article is twofold. Firstly, we use the next-token probabilities given by a language model to explicitly define a category of texts in natural language enriched over the unit interval, in the sense of Bradley, Terilla, and Vlassopoulos. We consider explicitly the terminating conditions for text generation and determine when the enrichment itself can be interpreted as a probability over texts. Secondly, we compute the Möbius function and the magnitude of an associated generalized metric space of texts. The magnitude function of that space is a sum over texts (*prompts*) of the t -logarithmic (Tsallis) entropies of the next-token probability distributions associated with each prompt, plus the cardinality of the model’s possible outputs. A suitable evaluation of the magnitude function’s derivative recovers a sum of Shannon entropies, which justifies seeing magnitude as a partition function. Following Leinster and Shulman, we also express the magnitude function of the generalized metric space as an Euler characteristic of magnitude homology and provide an explicit description of the zeroeth and first magnitude homology groups.

1. Introduction

In recent years, advances in large language models (LLMs) have brought renewed attention to mathematical structures underlying language. Building on the observation that LLMs acquire a great deal of semantic information through the statistics of co-occurrences of fragments of text, the first author together with Terilla and Vlassopoulos described a category-theoretical framework for mathematical structure inherent in text corpora in [Bradley et al., 2022]. This involves a category of strings from a finite alphabet of symbols, with morphisms indicating substring containment; the strings correspond to texts or fragments of texts in a language. Statistical information is incorporated through an enrichment of this category over the unit interval by assigning a value $\pi(y|x) \in [0, 1]$ to each pair of strings x and y , which can intuitively be thought of as the probability that y is an extension of a prompt x . By further embedding this enriched category of strings into an enriched category of copresheaves valued in the unit interval, a setting which contains rich mathematical structure, one can further explore semantic information via

The authors thank Matilde Marcolli and Yiannis Vlassopoulos for engaging in numerous insightful discussions on category theory, language models, and linguistics that have influenced this article. We also thank the anonymous referee whose feedback helped to improve the article considerably.

Received by the editors 2025-01-23 and, in final form, 2025-10-31.

Transmitted by Joachim Kock. Published on 2025-11-10.

2020 Mathematics Subject Classification: 18D20; 68T50; 94A17.

Key words and phrases: categorical magnitude, language model, generalized metric space, entropy.

© Tai-Danae Bradley and Juan Pablo Vigneaux, 2025. Permission to copy for private use granted.

enriched analogues of categorical limits and colimits, which are akin to logical operations on meaning representations of texts [Bradley et al., 2022]. More recently, Liu *et al.* have provided experimental evidence supporting this semantic interpretation of the co-Yoneda embedding of texts in the enriched category of copresheaves [Liu et al., 2024].

While an explicit construction of $\pi(y|x)$ was not given in [Bradley et al., 2022], we will show in this article that these values may in fact arise from next-token probabilities generated by a language model. For context: in natural language processing, texts are analyzed by splitting them into tokens—which might be characters, words, or word fragments—and a *language model* (LM) is an algorithm (possibly learned) that, for any finite sequence of *tokens* (i.e. a text fragment), assigns a probability to any given token being its continuation [Jurafsky and Martin, 2025, Ch. 3]. These are sometimes referred as *autoregressive* or *causal* LMs, particularly when compared with *masked* LMs [Jurafsky and Martin, 2025, Ch. 11]. The simplest causal LMs are n -gram models, which make their prediction based on the last $(n - 1)$ -tokens of the prompt. The relatively recent large language models, which are deep learning models trained on vast amounts of data, have extended the context used for prediction to thousands or even millions of tokens. Most of today’s popular LLMs, such as those in the GPT and Llama families, are autoregressive.

Let us briefly review some basics and then state the main results of this work. To start, it is standard for current language models to use special characters called the *beginning*- and *end-of-sentence* tokens, which we will denote by $\perp$ and $\dagger$, respectively. These tokens are usually not visible to the user but indicate to the model the boundary, so to speak, of a text, which is essential during training and when generating output. (In this way, a “sentence” may simply be thought of as a string of text built up token by token.) Additionally, the models have a *cutoff* size, or maximum length limit, to prevent unreasonably long outputs or infinite loops.

So in practice, for a tokenized user input a , the model is fed the prompt $x = \perp a$ and generates a probability distribution p_x on its set of tokens. It then samples from that distribution to select a token a_1 . If a_1 is the end-of-sentence token $\dagger$, the model outputs a to the user. Otherwise, the string xa_1 is fed back into the model, which again generates a probability distribution p_{xa} on its set of tokens. It samples from that distribution to select a token a_2 , and the process repeats. If $a_2 = \dagger$, the model outputs aa_1 to the user. Otherwise, it continues in this fashion until it either selects $\dagger$ or reaches the cutoff size. For convenience, if a string of tokens ends in $\dagger$, then we will refer to it as a *finished* text. Otherwise, it is *unfinished*. This brief background motivates the earlier claim that the values from π indeed form the hom-objects of a category enriched over the unit interval. Concretely, we will show the following main result:

Every LM defines an enriched category (see Proposition 2.16). *Every autoregressive language model defines a $[0, 1]$ -category whose objects are strings of tokens that begin with $\perp$ and may or may not end with $\dagger$, and whose hom-objects are given by π .*

Our consideration of beginning/end-of-sentence tokens as well as cutoff values differ-

entiate this result from the work in [Gaubert and Vlassopoulos, 2024], where a similar enriched category-theoretical construction appears. This will also allow us to prove that although the $\pi(y|x)$ are not *literally* probabilities (for instance, we will see that $\pi(x|x) = 1$ for each object x), for each unfinished text x the function $\pi(-|x)$ can be regarded as a probability mass function on the set $T(x)$ of *terminating states* of a prompt x , which is the set of all strings having x as a prefix that either end with $\dagger$ or reach the cutoff size. Said differently, $T(x)$ is the set of all theoretically possible outputs of a model (including those with negligible probability), given the prompt x . This is another contribution of this work:

Hom objects restrict to a probability mass function (see Proposition 2.9). *Every autoregressive language model determines a probability mass function $\pi(-|x)$ on the set of terminating states of an unfinished text x .*

It was remarked in [Bradley et al., 2022] and [Gaubert and Vlassopoulos, 2024] that one can alternatively enrich the category of strings over the extended non-negative reals $[0, \infty]$ by defining the hom-object between strings x and y to be $-\ln \pi(y|x)$. Lawvere [1973] observed that one can see a $([0, \infty], \geq, +, 0)$ -enriched category as a *generalized metric space* by interpreting the hom-objects as distances; the resulting distance function may be non-symmetric and degenerate but still satisfies the triangle inequality. In this article, we extend this geometric perspective by computing a fundamental invariant of enriched categories—their *magnitude*. Magnitude generalizes the cardinality of a set and the Euler characteristic of a topological space, and it can be regarded as the “size” of an enriched category [Leinster, 2008, 2013]. The computation of magnitude of ordinary metric spaces seen as $[0, \infty]$ -enriched categories reveals rich connections with traditional invariants from integral geometry and geometric measure theory such as volume, capacity, dimension, and intrinsic volumes [Leinster and Meckes, 2017].

In more detail, we study the generalized metric space $\mathcal{M}$ whose elements are strings $x, y, \dots$ and whose distances are given by $d(x, y) = -\ln \pi(y|x)$. We use the approach introduced by Rota [1964], later extended by Leinster and Shulman [2021], to compute the Möbius coefficients of $\mathcal{M}$ (see Proposition 3.6). Then we use these Möbius coefficients to calculate the *magnitude function* $\text{Mag}(t\mathcal{M})$, for $t > 0$.

Magnitude function associated with an LM (see Proposition 3.10). *The magnitude function of the generalized metric space $\mathcal{M}$ associated with an autoregressive language model is equal to*

$$\text{Mag}(t\mathcal{M}) = (t - 1) \sum_{x \in \text{ob}(\mathcal{M}) \setminus T(\perp)} H_t(p_x) + \#(T(\perp)), \quad t > 0.$$

In the above expression,

1. $\#(T(\perp))$ is the number of possible outputs of the model,

2. p_x is the probability distribution over tokens generated by the model prompted by x , and
3. H_t is the t -logarithmic entropy: $H_t(p_1, \dots, p_m) = (1 - \sum_{i=1}^m p_i^t)/(t - 1)$.

The entropies $(H_t)_{t \in (0, \infty) \setminus \{1\}}$ were first introduced by Havrda and Charvát [1967] and later popularized in physics by Tsallis [1988]. Each H_t is positive and strictly concave; it vanishes when its argument concentrates all the probability on a single token (i.e. represents a Dirac measure) and it is maximal for the uniform distribution [Havrda and Charvát, 1967]. It follows that, when $t > 1$, the function $\text{Mag}(t\mathcal{M})$ is lower bounded by $\#(T(\perp))$, and this value is attained by a “deterministic” language model that has a preferred continuation for each prompt, with no uncertainty (see Example 2.3). Similarly, also when $t > 1$, the function $\text{Mag}(t\mathcal{M})$ is upper bounded by

$$\frac{(t-1)(n^{1-t} - 1)}{1-t} \cdot \#(\text{ob}(\mathcal{M}) \setminus T(\perp)) + \#(T(\perp))$$

and this value is attained by a completely random language model whose next-token probability distributions are always uniform over the set of all possible continuations. This means that for every string x , the probability distribution p_x is uniform on the token set; in this case the t -logarithmic entropy $H_t(p_x)$ is equal to $(n^{1-t} - 1)/(1 - t)$, where n is the size of the token set. In the limit $t \rightarrow \infty$, these bounds take the simple form

$$\underbrace{\#(T(\perp))}_{\text{deterministic LM}} \leq \lim_{t \rightarrow \infty} \text{Mag}(t\mathcal{M}) \leq \underbrace{\#(\text{ob}(\mathcal{M}))}_{\text{maximally random LM}}.$$

More generally, in the case $t > 1$ the magnitude is always proportional to the amount of uncertainty in the model, as measured by the t -logarithmic entropy. In fact, if two language models with associated spaces $\mathcal{M}_1$ and $\mathcal{M}_2$ assign the same next-token probabilities (respectively, $p_{\bullet}^{(1)}$ and $p_{\bullet}^{(2)}$) to all prompts but one, called x' , in such a way that $H_t(p_{x'}^{(1)}) > H_t(p_{x'}^{(2)})$, then $\text{Mag}(t\mathcal{M}_1) > \text{Mag}(t\mathcal{M}_2)$.

The situation is less clear when $t < 1$. For instance, for a maximally random LM, we have $\lim_{t \rightarrow 0} \text{Mag}(t\mathcal{M}) = (1 - n) \cdot \#(\text{ob}(\mathcal{M}) \setminus T(\perp)) + \#(T(\perp))$, which is negative (since $n > 1$, unless $\dagger$ is the only token besides $\perp$). However, even for metric spaces, this limit might take any value [Roff and Yoshinaga, 2025], thus resisting a straightforward interpretation.

We shall see that the derivative of $\text{Mag}(t\mathcal{M})$ at $t = 1$ is a sum of Shannon entropies: $\sum_{x \in \text{ob}(\mathcal{M}) \setminus T(\perp)} H(p_x)$. Analogous results hold for the subspace $\mathcal{M}_x$ of texts that extend a given prompt x , see Remark 3.12.

Let us also remark that dependence of magnitude on the terminating states is reminiscent of similar results pertaining to posets or metric spaces that show a dependency on the “boundary.” For instance, if a finite poset has top and bottom elements, then its magnitude can be computed by evaluating its Möbius function (defined in Section 3)

at those extremal elements [Rota, 1964], [Stanley, 2011, Proposition 3.8.5]. And in the context of metric spaces, *weighting vectors* [Leinster, 2013]—such as $(\sum_y \mu(x, y))_x$ when the Möbius coefficients μ exist—effectively detect the boundary of certain subsets of Euclidean space, an insight with recent applications in machine learning [Bunch et al., 2021; Adamer et al., 2024]; see also [Willerton, 2009].

Finally, we will draw a connection to magnitude homology [Leinster and Shulman, 2021]. This is our final main result:

Magnitude homology associated with an LM (see Proposition 3.14). *The magnitude function can be expressed in terms of the ranks of the magnitude homology groups. For any $t > 0$,*

$$\text{Mag}(t\mathcal{M}) = \sum_{\ell} e^{-t\ell} \sum_{k \geq 0} (-1)^k \text{rank}(H_{k,\ell}(\mathcal{M})),$$

where the first sum ranges over those ℓ that may appear as finite lengths of paths in $\mathcal{M}$.

This result mirrors an analogous result by Leinster and Shulman [2021, Theorem 7.14], although our proof follows directly from Proposition 3.6 and encounters significantly less technical complications.

STRUCTURE OF THE ARTICLE. Section 2 derives, from a given LM, an enriched category of strings and an associated generalized metric space. Subsection 2.1 introduces the basic definitions concerning tokens and texts, including the partial order of texts. Subsection 2.2 describes the process of text generation by an LM, defines the function π , and shows that $\pi(-|x)$ is a probability mass function over the set $T(x)$ of terminating states. Subsection 2.10 reminds the reader of some basic definitions from enriched category theory, focusing on categories enriched over commutative monoidal preorders, which are a particularly simple case. Subsection 2.15 shows that π defines a $[0, 1]$ -category $\mathcal{L}$ of texts in the sense of [Bradley et al., 2022]. Subsection 2.17 introduces the corresponding generalized metric space $\mathcal{M}$ using the isomorphism of categories $-\ln : [0, 1] \rightarrow [0, \infty]$.

Section 3 is dedicated to magnitude. Subsection 3.1 gives a short overview of categorical magnitude, and Subsection 3.2 treats the classical case of posets. Subsection 3.5 deals with the computation of the Möbius coefficients and magnitude of $\mathcal{M}$. Subsection 3.13 covers the magnitude homology of $\mathcal{M}$. Finally, Section 4 presents some final remarks and perspectives.

2. Obtaining enriched categories from an LM

The goal of this section is to show that every LM defines a category enriched over the unit interval that, in turn, gives rise to a generalized metric space. In leading up to these results, we begin with a thin category whose objects are strings of symbols from a finite alphabet that start with a beginning-of-sentence token and that may or may not end with an end-of-sentence token. Morphisms are provided by substring containment. We

then enrich this setting by constructing $[0, 1]$ -hom objects obtained from the probabilities generated by a given LM and also consider a $[0, \infty]$ -enriched analogue.

2.1. TOKENS AND TEXTS. Let A be a finite alphabet (a set of tokens) and consider all finite strings $a, b, \dots$ from the free monoid A^* . In LMs, tokens can correspond to words, pieces of words, or special characters, depending on the model, and hence longer texts (e.g. sentences, paragraphs) are elements of A^* . Below, we call elements of A *tokens* and elements of A^* *strings* or *texts*. Write $a \leq b$ whenever a is a *prefix* of b , that is, whenever $b = aa'$ for some $a' \in A^*$. This defines a partial order since $\leq$ is reflexive (as A^* contains the empty string ϵ), transitive, and antisymmetric. We denote by $|a|$ the length of a string $a \in A^*$.

Besides the elements of A , we introduce two special tokens $\perp$ and $\dagger$ which represent, respectively, the so-called “beginning-of-sentence” and “end-of-sentence” tokens. Unlike the elements in A , the tokens $\perp$ and $\dagger$ do not appear in arbitrary positions; their role is clarified below. Although some systems conflate both symbols, it is useful to keep the distinction here.

We regard a partially ordered set $(P, \leq)$ as a thin category with object set P and such that $x \rightarrow y$ if and only if $x \leq y$. Given $N \in \mathbb{N}$, we define $\mathbf{L} := \mathbf{L}^{\leq N}$ as the full subcategory of $((A \cup \{\perp, \dagger\})^*, \leq)$ made of strings that start with $\perp$ and are followed by at most $N - 1$ symbols, all of them in A with the exception of the last one, which might equal $\dagger$; in other terms,

$$\text{ob}(\mathbf{L}) = \{\perp a : a \in A^* \text{ and } |a| \leq N - 1\} \sqcup \{\perp a \dagger : a \in A^* \text{ and } |a| < N - 1\}.$$

More explicitly, given objects x, y of $\mathbf{L}$, there is a morphism $x \rightarrow y$ whenever $x \leq y$ as elements of $(A \cup \{\perp, \dagger\})^*$, that is, if x is a prefix of y . We refer to an object of the form $\perp a$ as an *unfinished text* and to an object of the form $\perp a \dagger$ as a *finished text*. For any non-identity morphism $x \rightarrow y$, the prefix x is necessarily an unfinished text, whereas y can be finished or unfinished. Observe that $\mathbf{L}$ has an initial object,¹ which is $\perp$, and also that it has a natural “grading” given by the length $|x|$ of strings x . We remark that $|\perp| = 1$.

It will be convenient to write the set $\text{ob}(\mathbf{L})$ as a disjoint union $\bigsqcup_{j=0}^{N-1} \mathbf{L}^{(j)}$, where $\mathbf{L}^{(j)}$ consists of strings of length $1 + j$. (Remark that $\mathbf{L}$ is a free category on a graph, so j measures the graph-theoretic distance of the objects in $\mathbf{L}^{(j)}$ to $\perp$.) Aside from identities, the arrows of $\mathbf{L}$ only go from elements of $\mathbf{L}^{(i)}$ to elements of $\mathbf{L}^{(j)}$ for $j > i$. Similarly, given a string x in $\mathbf{L}$, there is a full subcategory $\mathbf{L}_x$ of $\mathbf{L}$ whose objects are $y \in \mathbf{L}$ such that $x \rightarrow y$; in this case, we have $\text{ob}(\mathbf{L}_x) = \bigsqcup_{j=0}^{N-|x|} \mathbf{L}_x^{(j)}$, where $\mathbf{L}_x^{(j)}$ consists only of those strings of length $|x| + j$, namely, those strings that extend x on the right by j tokens. So, $\mathbf{L}_x^{(0)} = \{x\}$ and $\mathbf{L}_x^{(1)} = \{xa_1 : a_1 \in A \cup \{\dagger\}\}$ and $\mathbf{L}_x^{(2)} = \{xa_1a_2 : a_1 \in A, a_2 \in A \cup \{\dagger\}\}$ and

¹By this point, the reader may have wondered why $\dagger$ is used as the end-of-sentence symbol instead of $\top$, given our use of $\perp$ for the beginning-of-sentence symbol. To start, the category $\mathbf{L}$ does not have a terminal object, so we wish to avoid using notation that suggests otherwise. Further, one might think of the end-of-sentence token as the state which “kills” the generation of a string, hence a dagger.

so on. Remark that $\mathsf{L}_{\perp}^{(j)} = \mathsf{L}^{(j)}$ for any $j \geq 0$, and if x is a finished text, then L_x has a single object, namely x .

We will refer to L as a poset or as a category interchangeably. Either way, it keeps track of which strings are right extensions of other strings in a language. To incorporate statistical information, we turn to enriched category theory in the following sections.

2.2. LMS AND INDUCED PROBABILITIES ON TEXTS. We characterize the behavior of an autoregressive language model (e.g. GPT, Llama) as follows. For any tokenized user input $a \in A^*$, the model is fed the prompt $x = \perp a$ and generates a probability distribution $p_x := p(-|x) : A \cup \{\dagger\} \rightarrow [0, 1]$. After generating p_x , the model then samples a token a_1 according to p_x . If $a_1 = \dagger$, then the process terminates and the model outputs a . Otherwise, the token a_1 is appended to a . These steps are then repeated. The model generates a probability distribution $p_{\perp aa_1}$ on $A \cup \{\dagger\}$, samples a token a_2 according to it, and either terminates execution if $a_2 = \dagger$ (outputting aa_1) or appends a_2 to aa_1 to generate $p_{\perp aa_1 a_2}$, and so on.

We assume that the process stops when $\dagger$ is sampled or when the extended prompt $y = xa_1 \cdots a_{N-|x|}$ reaches a maximum length N ; we call N the *cutoff*. We can think of $N - 1$ as the context size of the language model, i.e. the maximum length of a string x that can be used to generate a distribution p_x . Below are some common examples of autoregressive language models.

2.3. EXAMPLE. We say that a language model is *deterministic* if, for any possible prompt x , there is $a = a(x) \in A \cup \{\dagger\}$ such that $p_x(a) = 1$ (and therefore $p_x(a') = 0$ for any $a' \in (A \cup \{\dagger\}) \setminus \{a\}$). Although this example is quite artificial, it will appear as an extreme case for our magnitude calculations.

2.4. EXAMPLE. If $p_{\perp a_1 \cdots a_n}(-)$ only depends on a_n for any $n \geq 1$, then the LM corresponds to a *Markov chain*. In the terminology of Markov chains (see e.g. [Grimmett and Stirzaker, 2020, Ch. 6]), $\tilde{A} = A \cup \{\dagger\}$ is the set of states and $\dagger$ is an absorbing state. The numbers $P_n(a_n, a_{n+1}) := p_{\perp a_1 \cdots a_n}(a_{n+1})$ form a matrix $P_n : \tilde{A} \times \tilde{A} \rightarrow [0, 1]$, which has positive entries and is *stochastic*, meaning that each row sums to 1, i.e. $\sum_{a' \in \tilde{A}} P_n(a, a') = 1$. By convention we are assuming here that $P_n(\dagger, \dagger) = 1$ and $P_n(\dagger, a) = 0$ for any $a \in A$, which makes $\dagger$ an absorbing state.

Later in Section 4, we will briefly remark on a notion called *categorical diversity* in connection with homogeneous and irreducible Markov chains. If for all $n > 2$ it holds that $P_n = P_1$, then the chain is said to be *homogeneous*. A Markov chain is *irreducible* if there is a nontrivial probability of going from any state $a \in A$ to any state $a' \in A$ in a finite number of steps. A chain with an absorbing state is never irreducible. However, if $P_n(a, \dagger) = 0$ for every $n \geq 1$ and $a \in A$, then $P_n|_{A \times A}$ is also stochastic and $(P_n)_{n \geq 1}$ defines a Markov chain with state space A . When this resulting chain is homogeneous and irreducible, it necessarily has a stationary distribution (because A is finite, see [Grimmett and Stirzaker, 2020, Sec. 6.6]), which is a probability vector $q : A \rightarrow [0, 1]$ such that $\sum_{a' \in A} q(a)P(a, a') = q(a')$. In this setting, we will see in Section 4 that if additionally

$p_{\perp}(a) = q(a)$, then we can recover the Kolmogorov–Sinai entropy of the Markov chain from the categorical diversity.

2.5. **EXAMPLE.** An *n*-gram model is a Markov chain “with memory” of order $n - 1$, in the sense that $p_{\perp a_1 \dots a_i}(-)$ only depends on the last $n - 1$ tokens of $\perp a_1 \dots a_i$ (or all of them if $i < n - 2$). For details see [Jurafsky and Martin, 2025, Ch. 3].

2.6. **EXAMPLE.** A *large language model* is a deep neural network that implements a function $p^{(\theta)} : A^* \rightarrow \Delta(A)$, $a \mapsto p_{\perp a}^{(\theta)}(-)$. Here $\Delta(A)$ denotes the set of probability mass functions on A , and θ the network’s parameters (connection weights between neurons) and hyper-parameters. In the case of autoregressive LLMs, the deep neural network usually implements a version of the decoder-only transformer architecture [Vaswani et al., 2017; Radford et al., 2019]. The network is trained on a self-supervised manner: the goal is to predict as well as possible the last token a_n from $(a_1, \dots, a_{n-1})$, given any tokenized fragment of text $(a_1, \dots, a_n)$ in a corpus $C \subset A^*$. For details see [Jurafsky and Martin, 2025, Ch. 10].

In general, it will be useful to consider the set of all possible outputs of an LM corresponding to a given input. In this vein, we define the set of *terminating states* for a prompt $x = \perp a$ as follows.

2.7. **DEFINITION.** The set of *terminating states* of an unfinished text x in $\mathbf{L}$ is defined to be

$$T(x) = \{y \in \text{ob}(\mathbf{L}_x) : y \text{ is an unfinished text of length } N \text{ or} \\ y \text{ is a finished text such that } |y| \leq N\}.$$

In the first case, $y = xa'$ for some $a' \in A^*$ such that $|a'| = N - |x|$; for such a terminating state the model outputs aa' to the user according to some probability distribution. In the second case, $y = xa''\dagger$ for some $a'' \in A^*$ such that $|a''| \leq N - |x| - 1$; for such a terminating state the model outputs aa'' according to some probability distribution. Either way, there is a bijective correspondence between the terminating states and the theoretically possible outputs of the model, including those with small, or even zero, probabilities.² The phrase “theoretically possible” is used to emphasize that the set $T(x)$ depends only on the category of strings $\mathbf{L}$ and is independent of a choice of language model.

Now, it is tempting to define the probability of the model’s output b given the user input a as the product of the intermediate probabilities of the tokens involved in its production. For example, when $N > 5$, the probability of generating the output $b = a_1 a_2 a_3 a_4 a_5 \in A^*$ given the user’s input $a = a_1 a_2$ would be

$$p(a_3 | \perp a) p(a_4 | \perp a a_3) p(a_5 | \perp a a_3 a_4) p(\dagger | \perp a a_3 a_4 a_5).$$

²Notice that this description is only valid for prompts whose length is strictly less than N : to extend a prompt x of length greater than $N - 1$, one needs to produce first a subrogate prompt $f(x)$ such that $|f(x)| \leq N - 1$ and feed it to the language model to generate $p_{f(x)}$ and sample a token.

However, it is not obvious that this rule indeed defines a probability mass function over some set. We prove this below. But first, let us define these products in general.

2.8. DEFINITION. For any objects x and y of $\mathbf{L}$, define

$$\pi(y|x) := \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \not\rightarrow y \\ \prod_{i=1}^k p(a_{t+i}|y_{<t+i}) & \text{if } x \rightarrow y \end{cases}. \quad (1)$$

In the third case, we assume that $x = \perp a_1 \cdots a_t$ and $y = xa_{t+1} \cdots a_{t+k}$ for some $k \geq 1$, $(a_i)_{i=1}^{t+k-1} \subset A$, and $a_{t+k} \in A \cup \{\dagger\}$. The symbol $y_{<t+i}$ denotes the string $\perp a_1 \cdots a_{t+i-1}$, and $y_{<t+1} = x$.

For a given string x in $\mathbf{L}$, the function $\pi(-|x)$ is not a probability mass function on its whole support. (To begin with, $\pi(x|x) = 1$.) Nonetheless, it becomes a probability mass function when restricted to $T(x)$.

2.9. PROPOSITION. *Every autoregressive language model determines a probability mass function $\pi(-|x)$ on the set $T(x)$ of terminating states of an unfinished text x .*

PROOF. If $m = 0$, then $T(x) = \{x\}$ and by definition $\pi(x|x) = 1$. If $m = 1$, then $T(x) = \{xa \mid a \in A \cup \{\dagger\}\}$ and

$$\sum_{y \in T(x)} \pi(y|x) = \sum_{a \in A \cup \{\dagger\}} p_x(a) = 1,$$

since p_x is assumed to be a probability mass function on $A \cup \{\dagger\}$. For general $m \geq 1$,

$$\sum_{y \in T(x)} \pi(y|x) = \sum_{i=0}^{m-1} \sum_{a' \in A^i} \pi(xa' \dagger | x) + \sum_{a \in A^m} \pi(xa|x) \quad (2)$$

$$= \sum_{i=0}^{m-1} \sum_{a' \in A^i} \pi(xa' \dagger | x) + \sum_{\substack{a=a'a'' \\ a' \in A^{m-1}, a'' \in A}} p(a''|xa') \pi(xa'|x) \quad (3)$$

$$= \sum_{i=0}^{m-2} \sum_{a' \in A^i} \pi(xa' \dagger | x) + \sum_{a' \in A^{m-1}} \pi(xa'|x) \sum_{a'' \in A \cup \{\dagger\}} p(a''|xa') \quad (4)$$

$$= \sum_{i=0}^{m-2} \sum_{a' \in A^i} \pi(xa' \dagger | x) + \sum_{a' \in A^{m-1}} \pi(xa'|x) \quad (5)$$

The two sums in (2) follow from the definition of $T(x)$ and π , where the first sum accounts for finished texts (i.e. those ending in $\dagger$), and the second sum accounts for unfinished

texts of length N . We obtain (3) by identifying $a \in A^m$ with $(a', a'') \in A^{m-1} \times A$. Then (4) follows by rewriting the term corresponding to $i = m - 1$ in the leftmost sum as $\pi(xa'|x)p(\dagger|xa')$, where $a' \in A^{m-1}$, and then moving it to the rightmost sum. The expression in (5) equals one in virtue of the induction hypothesis. ■

2.10. PRELIMINARIES ON ENRICHED CATEGORY THEORY. We shall now use π to derive an enriched category of strings $\mathcal{L}$ from $\mathbf{L}$. For the convenience of the reader, we give here the definition of a category enriched over a commutative monoidal preorder [Fong and Spivak, 2019, Chapter 2], which is the only case we will need. See also [Kelly, 1982].

2.11. DEFINITION. A *commutative monoidal preorder* $(\mathcal{V}, \leq, \otimes, 1)$ is a preordered set $(\mathcal{V}, \leq)$ and a commutative monoid $(\mathcal{V}, \otimes, 1)$ satisfying $x \otimes y \leq x' \otimes y'$ whenever $x \leq x'$ and $y \leq y'$.

2.12. EXAMPLE. The unit interval $([0, 1], \leq, \cdot, 1)$ is a commutative monoidal preorder with the usual ordering $\leq$. Multiplication of real numbers is the monoidal product, which we will denote by juxtaposition, $ab := a \cdot b$ for $a, b \in [0, 1]$, and the monoidal unit is 1.

2.13. EXAMPLE. The extended non-negative reals $([0, \infty], \geq, +, 0)$ form a commutative monoidal preorder where the preorder is the opposite of the usual ordering on the reals. The monoidal product is addition with $a + \infty := \infty$ and $\infty + a := \infty$ for all $a \in [0, \infty]$, and the monoidal unit is 0.

2.14. DEFINITION. Let $(\mathcal{V}, \leq, \otimes, 1)$ be a commutative monoidal preorder. A (small) *category enriched over $\mathcal{V}$* , or simply a *$\mathcal{V}$ -category*, $\mathcal{C}$ consists of a set $\text{ob}(\mathcal{C})$ of objects and, for every pair of objects x and y , an object $\mathcal{C}(x, y)$ of $\mathcal{V}$ called a *$\mathcal{V}$ -hom object* satisfying the following: for all objects $x, y, z \in \text{ob}(\mathcal{C})$,

$$\begin{aligned} 1 &\leq \mathcal{C}(x, x) \\ \mathcal{C}(y, z) \otimes \mathcal{C}(x, y) &\leq \mathcal{C}(x, z) \end{aligned}$$

Sections 2.15 and 2.17 below give examples of categories enriched over both $[0, 1]$ and $[0, \infty]$ defined from the probabilities generated by a language model. As we will see in the setup introduced in Subsections 2.1 and 2.2, objects of both categories will consist of strings of characters from some finite token set.

2.15. EVERY LM DEFINES A $[0, 1]$ -CATEGORY. For convenience, let us recall the setup established in Subsections 2.1 and 2.2. As described there, A denotes a finite alphabet, such as the token set of a given LM, and $\mathbf{L}$ denotes the finite category whose objects are strings from A that begin with $\perp$ and are followed by at most $N - 1$ symbols, where the last symbol may or may not be $\dagger$. Here N denotes the model's cutoff size. Moreover, there is a morphism $x \rightarrow y$ if x is a prefix of y . Further recall that for each pair of strings x, y in $\mathbf{L}$, Definition 2.8 specifies the value $\pi(y|x) \in [0, 1]$ which, if y is an extension of x , is equal to the product of the successive probabilities given by the model when generating y from x one token at a time. Otherwise, if x is not contained in y , then $\pi(y|x) = 0$, and if $x = y$ then $\pi(x|x) = 1$.

This allows us to now define a category $\mathcal{L}$ enriched over the unit interval. The objects of $\mathcal{L}$ coincide with the objects of $\mathbf{L}$, and for a pair of objects x and y we define their $[0, 1]$ -hom object to be $\mathcal{L}(x, y) := \pi(y|x)$.

Let us verify that $\mathcal{L}$ satisfies both the identity and compositionality requirements for a category enriched over $[0, 1]$ listed in Definition 2.14. Suppose first that x, y, z are three strings satisfying $x \rightarrow y \rightarrow z$. If $x \neq y$ and $y \neq z$, then x and y are necessarily unfinished texts, and we may write

$$\begin{aligned} x &= \perp a_1 \cdots a_t \\ y &= \perp a_1 \cdots a_t \cdots a_{t+k} \\ z &= \perp a_1 \cdots a_t \cdots a_{t+k} \cdots a_{t+k+k'}. \end{aligned}$$

for some $(a_i)_{i=1}^{t+k+k'-1} \subset A$ and $a_{t+k+k'} \in A \cup \{\dagger\}$. It then follows from (1) that

$$\begin{aligned} \pi(y|x)\pi(z|y) &= \prod_{i=1}^k p(a_{t+i}|y_{<t+i}) \prod_{j=1}^{k'} p(a_{t+k+j}|z_{<t+k+j}) \\ &= \prod_{i=1}^{k+k'} p(a_{t+i}|z_{<t+i}) \\ &= \pi(z|x). \end{aligned} \tag{6}$$

If instead $x = y$, then $\pi(y|x) = 1$ and obviously $\pi(z|y) = \pi(z|x)$. The case $z = y$ is treated similarly. More generally, if x, y and z are arbitrary objects of $\mathbf{L}$, one has that $\pi(y|x)\pi(z|y) \leq \pi(z|x)$ as it may be the case that $\pi(y|x), \pi(z|x) \neq 0$ but $\pi(z|y) = 0$,³ or that $\pi(z|y), \pi(z|x) \neq 0$ but $\pi(y|x) = 0$.⁴ In summary, we come to the following proposition.

2.16. PROPOSITION. *Every autoregressive language model defines a $[0, 1]$ -category whose objects are those of $\mathbf{L}$ and whose hom-objects are given by π .*

A more general version of this $[0, 1]$ -category was originally defined in [Bradley et al., 2022, Definition 4], where $\mathcal{L}(x, y)$ was taken to be nonzero whenever x was an *arbitrary* substring of y (not strictly a prefix). However, the values $\pi(y|x)$ were not constructed explicitly from an LM, whereas Definition 2.8 above yields a $[0, 1]$ -category of expressions in a language where the values $\pi(y|x)$ are described concretely from the probabilities generated by an LM. A similar definition can be found in [Gaubert and Vlassopoulos, 2024], although neither their description nor the one in [Bradley et al., 2022] considers the special characters $\perp$ and $\dagger$ or a model's cutoff value. The incorporation of these aspects led us to the novel Proposition 2.9, which *justifies referring to π as a probability*.

Notice we only consider extensions of expressions *on the right*, having in mind the most standard, autoregressive LLMs, although one could also consider bidirectional extensions.

³For example, consider when $x = \perp \text{green}$ and $y = \perp \text{green lantern}$ and $z = \perp \text{green salad}$.

⁴For example, consider when $x = \perp \text{movie tic}$ and $y = \perp \text{movie}$ and $z = \perp \text{movie ticket}$.

That said, the tree-like structure implied by the restriction to right extensions plays a key role in the computation of magnitude presented below.

Finally, let us also remark that, after defining the $[0, 1]$ -category $\mathcal{L}$ of strings, the authors of [Bradley et al., 2022] further consider enriched copresheaves on $\mathcal{L}$, which were shown to contain semantic information. Our present goal, however, is to turn our attention away from $\mathcal{L}$ to a more geometric version of it.

2.17. EVERY LM DEFINES A $[0, \infty]$ -CATEGORY. As discussed in [Bradley et al., 2022, Section 5], the function $-\ln: [0, 1] \rightarrow [0, \infty]$ is an isomorphism of categories, which provides a passage from probabilities to a more geometric setting. That is, instead of considering an enrichment from probabilities involved in generating a string y from a prefix x , one could instead think of the “distance” traveled from x to y by defining

$$d(x, y) := -\ln \pi(y|x).$$

It is straightforward to check the triangle inequality is satisfied $d(x, y) + d(y, z) \geq d(x, z)$ and further that $d(x, x) = 0$ for all strings x, y, z . (This follows from Equation (6) and the fact that $\pi(x|x) = 1$ for all x .) From this perspective, a text that is highly likely to extend a prompt x is close to x , and a text that is not an extension of x is infinitely far away. In this way, we obtain a category $\mathcal{M}$ enriched over $[0, \infty]$, also known as a *generalized metric space* in the sense of Lawvere [1973], by defining the $[0, \infty]$ -hom object between a pair of strings x, y to be the distance $\mathcal{M}(x, y) := d(x, y)$. And now that we have a generalized metric space, we may inquire after its magnitude.

3. Magnitude

The theory behind the magnitude of (generalized) metric spaces is well known [Leinster, 2013] and is a special case of the magnitude of enriched categories [Leinster and Shulman, 2021; Leinster and Meckes, 2017]. Importantly, the theory requires working with a *finite* enriched category, which is an additional reason to consider the category $\mathbf{L}^{\leq N}$ with finite cutoff N . In this section, then, we provide a few preliminaries before computing the magnitude of the generalized metric space $\mathcal{M}$. We begin by reviewing the definition of magnitude in the context of enriched categories and in the classical context of posets.

3.1. DEFINITION. Magnitude is a numerical invariant of a category enriched over a monoidal category $(\mathcal{V}, \otimes, 1)$ [Leinster, 2013; Leinster and Shulman, 2021; Leinster and Meckes, 2017]. Briefly, one starts with a semi-ring R and a multiplicative function $\|-\|: \text{ob}(\mathcal{V}) \rightarrow R$ called *size* that is invariant under isomorphisms. So, $\|v\| = \|w\|$ whenever $v \cong w$ in $\mathcal{V}$, and $\|1\| = 1$ and $\|v \otimes w\| = \|v\| \|w\|$ for all $v, w \in \text{ob}(\mathcal{V})$. Then, given a $\mathcal{V}$ -category $\mathcal{C}$ with finitely many objects, one introduces the *zeta function* $\zeta_{\mathcal{C}}: \text{ob}(\mathcal{C}) \times \text{ob}(\mathcal{C}) \rightarrow R$, defined by $\zeta_{\mathcal{C}}(x, y) := \|\mathcal{C}(x, y)\|$. The function $\zeta_{\mathcal{C}}$ can be regarded as a square matrix with index set $\text{ob}(\mathcal{C})$ [Leinster, 2013]; when it is invertible, $\mathcal{C}$ is said to have *Möbius inversion* and its *Möbius function* is $\mu_{\mathcal{C}} := \zeta_{\mathcal{C}}^{-1}$ (the entries of this matrix

are called *Möbius coefficients*). When $\mathcal{C}$ has Möbius inversion, the *magnitude* $\text{Mag}(\mathcal{C})$ of $\mathcal{C}$ can be defined as

$$\text{Mag}(\mathcal{C}) = \sum_{(x,y) \in \text{ob}(\mathcal{C}) \times \text{ob}(\mathcal{C})} \zeta_{\mathcal{C}}^{-1}(x, y). \quad (7)$$

For $\mathcal{V} = [0, \infty]$, every size function is of the form $\|x\|_t = e^{-tx}$ for some $t \in \mathbb{R} \cup \{\infty\}$. It is customary to choose $\|-\|_1$ and then consider the *magnitude function* $f(t) := \text{Mag}(t\mathcal{C})$, defined for $t \in (0, \infty)$ [Leinster, 2013, Section 2.2]. The symbol $t\mathcal{C}$ denotes the generalized metric space with the same objects as $\mathcal{C}$ but whose distances (that is, whose $[0, \infty]$ -hom objects) are scaled by t .

3.2. MAGNITUDE OF POSETS. The magnitude of an enriched category stems from a classical story from posets, which we now briefly review from [Stanley, 2011].

Given a poset P , define a *closed interval* to be $[s, t] = \{u \in P : s \leq u \leq t\}$ whenever $s \leq t$ in P . If every interval of P is finite, then P is said to be a *locally finite* poset.

Let $\mathbb{k}$ be a field, P a locally finite poset, and $\text{int}(P)$ the set of all closed intervals of P . The *incidence algebra* $I(P, \mathbb{k})$ of P over $\mathbb{k}$ is the $\mathbb{k}$ -algebra of all functions $f: \text{int}(P) \rightarrow \mathbb{k}$ with the usual structure of a vector space over $\mathbb{k}$, where multiplication is given by the *convolution product*,

$$(f * g)(s, u) := \sum_{s \leq t \leq u} f(s, t)g(t, u),$$

where we write $f(s, t)$ for $f([s, t])$. (Remark that the sum has finitely many terms.) The incidence algebra of P is an associative algebra with identity $\delta: \text{int}(P) \rightarrow \mathbb{k}$ given by

$$\delta(s, u) = \begin{cases} 1 & \text{if } s = u \\ 0 & \text{if } s \neq u \end{cases}$$

since $f * \delta = \delta * f = f$ for all functions $f \in I(P, \mathbb{k})$. The *zeta function* ζ_P is another notable function in the incidence algebra, defined to be constant at 1 on every interval,

$$\zeta_P(s, u) = 1, \quad \text{for all } s \leq u \text{ in } P.$$

Observe that this coincides with the zeta function defined in the enriched categorical setting, since every poset P (and in particular the poset $\mathbf{L}$ of strings) may be viewed as a category enriched over truth values $\mathbf{2} := \{\mathbf{true}, \mathbf{false}\}$. Its objects are the elements of P , and given $s, u \in P$, if $s \leq u$, then their corresponding hom object $P(s, u)$ is $\mathbf{true}$, and if $s \not\leq u$ then it is $\mathbf{false}$. The identity and compositionality requirements in Definition 2.14 arise from the reflexivity and transitivity of the partial order and by considering the size function $\|-\|: \text{ob}(\mathbf{2}) \rightarrow \mathbb{Z}$ given by $\|\mathbf{true}\| = 1$ and $\|\mathbf{false}\| = 0$ and then setting $\zeta_P(s, u) = \|P(s, u)\|$.

For every locally finite poset P , the function ζ_P is an invertible element of the incidence algebra $I(P, \mathbb{k})$ [Stanley, 2011, Prop. 3.6.2]. Equivalently, the zeta function is invertible when regarded as a square matrix with index set P , cf. [Leinster, 2013]; we have taken this matricial perspective in our presentation of the more general categorical definition in

Subsection 3.1. The inverse of ζ_P is called the *Möbius function* of P , denoted μ_P , and it can be computed recursively:

$$\mu_P(s, u) = \begin{cases} 1 & \text{if } s = u \\ -\sum_{s \leq t < u} \mu_P(s, t) & \text{if } s < u \\ 0 & \text{otherwise.} \end{cases}$$

3.3. EXAMPLE. The Möbius function $\mu_{\mathbf{L}}$ on the poset $\mathbf{L}$ of strings is rather simple. This is due to the tree-like structure of the full subcategory $\mathbf{L}_x$ of $\mathbf{L}$ introduced in Subsection 2.1, whose objects are all strings y having a given string x as a prefix. Indeed, for any string $x \in \text{ob}(\mathbf{L})$ and tokens $a_1, a_2, \dots \in A \cup \{\dagger\}$, we have

$$\begin{aligned} \mu_{\mathbf{L}}(x, x) &= 1 \\ \mu_{\mathbf{L}}(x, xa_1) &= -\mu_{\mathbf{L}}(x, x) = -1 \\ \mu_{\mathbf{L}}(x, xa_1a_2) &= -\mu_{\mathbf{L}}(x, x) - \mu_{\mathbf{L}}(x, xa_1) = -1 + 1 = 0 \\ \mu_{\mathbf{L}}(x, xa_1a_2a_3) &= -\mu_{\mathbf{L}}(x, x) - \mu_{\mathbf{L}}(x, xa_1) - \mu_{\mathbf{L}}(x, xa_1a_2) = -1 + 1 + 0 = 0. \end{aligned}$$

Similarly, $\mu_{\mathbf{L}}(x, xa_1a_2 \cdots a_j) = 0$ for $j > 1$. (As one of the reviewers pointed out, these identities hold for any free category, so in particular for $\mathbf{L}$, which is a free category on a tree rooted in $\perp$.)

This Möbius function may then be used to compute the magnitude of the finite category $\mathbf{L}$ of strings from a set A of tokens.

3.4. EXAMPLE. Since the category $\mathbf{L}$ has an initial object, we know its magnitude is 1, cf. [Leinster, 2008, Example 2.3], but it is also instructive to prove this via an explicit computation of the Möbius function. Although we recommend the reader performs this computation by herself, we also include it here for the sake of completeness. To that end, let $\#A$ be the cardinality of the token set A and let $\#\text{ob}(\mathbf{L})$ be the number of strings in the finite poset $\mathbf{L}$ considered as a category. The magnitude of $\mathbf{L}$ is given by

$$\begin{aligned} \text{Mag}(\mathbf{L}) &= \sum_{x, y \in \text{ob}(\mathbf{L})} \zeta_{\mathbf{L}}^{-1}(x, y) \\ &= \sum_{x, y \in \text{ob}(\mathbf{L})} \mu_{\mathbf{L}}(x, y) \\ &= \sum_{x \in \text{ob}(\mathbf{L})} \mu_{\mathbf{L}}(x, x) + \sum_{\substack{x = \perp a \in \text{ob}(\mathbf{L}) \\ a \in \bigcup_{i=0}^{N-2} A^i}} \sum_{a_1 \in A \cup \{\dagger\}} \mu_{\mathbf{L}}(x, xa_1) \\ &= \#\text{ob}(\mathbf{L}) - \# \left\{ x \in \text{ob}(\mathbf{L}) : x = \perp a \text{ with } a \in \bigcup_{i=0}^{N-2} A^i \right\} (\#A + 1). \end{aligned}$$

Unwinding the last line, recall that $\mathbf{L}$ denotes $\mathbf{L}^{\leq N} = \bigsqcup_{i=0}^{N-1} \mathbf{L}^{(i)}$. We have $\mathbf{L}^{(0)} = \{\perp\}$ and, for each $i \leq 1$, $\mathbf{L}^{(i)}$ is comprised of strings of the form $\perp a$ with $a \in A^i$ or $\perp a' \dagger$ with $a' \in A^{i-1}$. It follows that $\#\mathbf{L}^{(i)} = \#A^i + \#A^{i-1}$ and so

$$\begin{aligned} \#\text{ob}(\mathbf{L}) &= 1 + \sum_{i=1}^{N-1} \#\mathbf{L}^{(i)} \\ &= 1 + \sum_{i=1}^{N-1} \#A^i + \#A^{i-1} = \#A^{N-1} + 2(\#A^{N-2}) + \cdots + 2(\#A) + 2. \end{aligned}$$

Furthermore,

$$\begin{aligned} \# \left\{ x \in \text{ob}(\mathbf{L}) : x = \perp a \text{ with } a \in \bigcup_{i=0}^{N-2} A^i \right\} (\#A + 1) &= \left(\sum_{i=0}^{N-2} \#A^i \right) (\#A + 1) \\ &= \#A^{N-1} + 2(\#A^{N-2}) + \cdots + 2(\#A) + 1. \end{aligned}$$

Therefore, $\text{Mag}(\mathbf{L}) = 1$.

3.5. MAIN RESULTS. We are now ready to compute the magnitude of the generalized metric space associated with an LM. As before, let $\mathcal{M}$ be the $[0, \infty]$ -category of strings introduced in Subsection 2.17, whose objects are $\text{ob}(\mathcal{M}) = \text{ob}(\mathbf{L}^{\leq N})$ and where the $[0, \infty]$ -hom objects are given by $\mathcal{M}(x, y) = d(x, y) = -\ln \pi(y|x)$. Define the *zeta matrix* $\zeta_{\mathcal{M}} : \text{ob}(\mathcal{M}) \times \text{ob}(\mathcal{M}) \rightarrow \mathbb{R}$ to be

$$\zeta_{\mathcal{M}}(x, y) := e^{-d(x, y)} = \pi(y|x),$$

and more generally for any real number $t > 0$ define

$$(\zeta_{\mathcal{M}})_t(x, y) := e^{-td(x, y)} = \pi(y|x)^t.$$

It is understood that $d(x, y) = \infty$ whenever $\pi(y|x) = 0$; for instance, the distance from a finished text x to any other string y is infinite. For simplicity, we will write ζ_t instead of $(\zeta_{\mathcal{M}})_t$. As we will see in the proposition and corollaries below, ζ_t^{-1} exists and its components can be computed explicitly.

3.6. PROPOSITION. *For any $t > 0$, the matrix ζ_t is invertible. Moreover, for any x, y in $\text{ob}(\mathcal{M})$,*

$$\zeta_t^{-1}(x, y) = \sum_{k \geq 0} \sum_{\substack{\text{nondeg. paths} \\ x=y_0 \rightarrow y_1 \rightarrow \cdots \rightarrow y_k=y \text{ in } \mathbf{L}}} (-1)^k \prod_{i=1}^k \pi(y_i|y_{i-1})^t. \quad (8)$$

(A non-degenerate path $x = y_0 \rightarrow y_1 \rightarrow \cdots \rightarrow y_k = y$ consists of a sequence of composable non-identity morphisms.)

PROOF. Following [Rota \[1964\]](#) (see also [\[Leinster and Shulman, 2021\]](#)), we introduce the formal expansion

$$\zeta_t^{-1} = \sum_{k \geq 0} (-1)^k (\zeta_t - \delta)^k, \quad (9)$$

where δ is the identity matrix: $\delta(x, y) = 1$ if $x = y$ and $\delta(x, y) = 0$ otherwise.

The matrix $\zeta - \delta$ is the adjacency matrix of a directed graph D with vertices $\text{ob}(\mathcal{M})$ (equivalently, $\text{ob}(\mathbf{L})$) and edges $E = \{(z, z') \mid z \rightarrow z' \text{ in } \mathbf{L}, z \neq z'\}$ (we are excluding edges coming from identity morphisms, whose weights vanish). The power $(\zeta_t - \delta)^k$ counts non-degenerate paths of length k in D . Since $\mathbf{L}$ is a finite poset, one readily verifies that E is finite and that the graph has no cycles. Hence the sum in (9) has a finite number of terms.⁵

The expression (8) arises from evaluating (9) on (x, y) . In fact,

$$(\zeta_t - \delta)(x, y) = \pi(x, y)^t \mathbf{1}_{(x, y) \in E},$$

where $\mathbf{1}_\bullet$ denotes the indicator function, and more generally

$$(\zeta_t - \delta)^k(x, y) = \sum_{\substack{(y_0, y_1, \dots, y_{k-1}, y_k) \\ y_0 = x, y_k = y \\ (y_{i-1}, y_i) \in E \text{ for } i=1, \dots, k}} \prod_{i=1}^k \pi(y_i | y_{i-1})^t.$$

■

3.7. REMARK. We envision extensions of this theory where the category $\mathbf{L}$ would not be a poset. This would be the case, for instance, if we consider the general relation of being a substring, instead of being just a prefix, cf. [\[Gaubert and Vlassopoulos, 2024\]](#). Similarly, it might be possible to build an analogous category for syntactic trees, where morphisms would arise from the syntactic Merge operation and rearrangements of the subtrees of a tree followed by Merge. (Merge is the fundamental syntactic operation in Chomsky's minimalist program; it has been formalized as an operation on a Hopf algebra of syntactic forests in [\[Marcolli et al., 2025\]](#).) In these situations, where the categories present nontrivial cycles, we would need different techniques to compute ζ_t^{-1} , such as those introduced in [\[Vigneaux, 2024\]](#) by the second author. Our Proposition 3.6 can also be derived from this more general formalism.

The following corollary shows that the Möbius coefficients obtained from ζ_t^{-1} have a simplified expression in terms of π from Definition 2.8 and the Möbius function $\zeta_{\mathbf{L}}^{-1}$, whose zeta function $\zeta_{\mathbf{L}}$ was described in Section 3.2.

3.8. COROLLARY. *For any strings x, y in $\text{ob}(\mathcal{M})$,*

$$\zeta_t^{-1}(x, y) = \pi(y|x)^t \zeta_{\mathbf{L}}^{-1}(x, y).$$

⁵For general categories, one has to establish convergence of the series under certain conditions, see [\[Leinster and Shulman, 2021\]](#).

PROOF. For any $k \geq 0$ and any nondegenerate path $x = y_0 \rightarrow y_1 \rightarrow \cdots \rightarrow y_k = y$ in $\mathbf{L}$, Equation (6) implies

$$\begin{aligned} \prod_{i=1}^k \pi(y_i | y_{i-1})^t &= \pi(y_1 | y_0)^t \pi(y_2 | y_1)^t \pi(y_3 | y_2)^t \cdots \pi(y_k | y_{k-1})^t \\ &= \pi(y_2 | y_0)^t \pi(y_3 | y_2)^t \cdots \pi(y_k | y_{k-1})^t \\ &= \vdots \\ &= \pi(y_k | y_0)^t. \end{aligned}$$

Hence the Möbius coefficients $\zeta_t^{-1}(x, y)$ can be written as

$$\pi(y|x)^t \sum_{k \geq 0} (-1)^k \# \{ \text{nondegenerate paths of length } k \text{ from } x \text{ to } y \text{ in } \mathbf{L} \}$$

which is equal to $\pi(y|x)^t \zeta_{\mathbf{L}}^{-1}(x, y)$ by Philip Hall's theorem [Hall, 1936], cf. [Stanley, 2011, Proposition 3.8.5], [Leinster, 2008, Corollary 1.5]. ■

Now recall that $\zeta_{\mathbf{L}}^{-1}(x, y) = \mu_{\mathbf{L}}(x, y)$, and so ζ_t^{-1} can be simplified even further by Example 3.3.

3.9. COROLLARY. *For any strings x, y in $\text{ob}(\mathbf{L})$,*

$$\zeta_t^{-1}(x, y) = \begin{cases} -\pi(y|x)^t & \text{if } y \in \mathbf{L}_x^{(1)} \\ 1 & \text{if } y = x \\ 0 & \text{otherwise.} \end{cases}$$

Finally, then, we may use these results to write down the *magnitude function* of the generalized metric space $\mathcal{M}$. By Equation (7) it is the function $f : (0, \infty) \rightarrow \mathbb{R}$ given by $f(t) = \text{Mag}(t\mathcal{M})$, where⁶

$$\text{Mag}(t\mathcal{M}) := \sum_{x, y \in \text{ob}(\mathcal{M})} \zeta_t^{-1}(x, y).$$

Now recall that for any real number t , the *t-logarithmic entropy* of a probability distribution $p = (p_1, \dots, p_n)$ is equal to

$$H_t(p) = \frac{1}{t-1} \left(1 - \sum_{i=1}^n p_i^t \right).$$

The following proposition expresses the magnitude function of $\mathcal{M}$ in terms of the logarithmic entropies of the probability distributions $(p_x : A \cup \{\dagger\} \rightarrow [0, 1])_{x \in \text{ob}(\mathbf{L})}$ generated by a language model and the cardinality of the terminal states $T(\perp)$ of the beginning-of-sentence symbol, which is the set of all theoretically possible terminating states of the model.

⁶By Corollary 3.8, it follows that $\text{Mag}(t\mathcal{M}) = \sum_{x, y \in \text{ob}(\mathcal{M})} \pi(y|x)^t \zeta_{\mathbf{L}}^{-1}(x, y)$, which is a weighted version of the magnitude of the poset $\mathbf{L}$ in Example 3.4.

3.10. PROPOSITION. *The magnitude function of the generalized metric space $\mathcal{M}$ associated with an autoregressive language model is equal to*

$$\text{Mag}(t\mathcal{M}) = (t-1) \sum_{x \in \text{ob}(\mathcal{M}) \setminus T(\perp)} H_t(p_x) + \#(T(\perp)), \quad t > 0.$$

PROOF. By Corollary 3.9,

$$\begin{aligned} \text{Mag}(t\mathcal{M}) &= \sum_{x, y \in \text{ob}(\mathcal{M})} \zeta_t^{-1}(x, y) \\ &= \sum_{x \in \text{ob}(\mathcal{M})} \zeta_t^{-1}(x, x) - \sum_{x \in \text{ob}(\mathcal{M}) \setminus T(\perp)} \sum_{y \in \mathbb{L}_x^{(1)}} \pi(y|x)^t \end{aligned} \quad (10)$$

$$= \sum_{x \in \text{ob}(\mathcal{M}) \setminus T(\perp)} \left(1 - \sum_{a \in A \cup \{\dagger\}} p_x(a)^t \right) + \#(T(\perp)). \quad (11)$$

Observe that the first sum in Equation (10) is over all objects in $\mathcal{M}$, including those strings x that are not the prefix of any other string $y \neq x$, such as all finished texts. The double sum in Equation (10) accounts for unfinished strings and their extensions by a single token, which is why the indexing set is $\text{ob}(\mathcal{M}) \setminus T(\perp)$. Equation (11) follows from the decomposition $\text{ob}(\mathcal{M}) = T(\perp) \sqcup (\text{ob}(\mathcal{M}) \setminus T(\perp))$ and the fact that $\zeta_t^{-1}(x, x) = 1$ for all $x \in \text{ob}(\mathcal{M})$.

When $t = 1$, Equation (11) is simply $\#(T(\perp))$, since p_x is a probability mass function on $A \cup \{\dagger\}$. When $t \neq 1$, we can multiply the sum by $(t-1)/(t-1)$ and rearrange factors to conclude. ■

As discussed in the Introduction, one might view $\text{Mag}(t\mathcal{M})$ as measuring something of the effective size of a language model's linguistic space, in the sense that for $t \geq 1$, magnitude achieves its smallest possible value if the model is deterministic (that is, it only produces a single preferred text), and it achieves its largest possible value if the model is completely random (at every stage the next-token probability distribution has maximal entropy).

3.11. REMARK. It is easy to show (e.g. via L'Hôpital's rule) that if $p : S \rightarrow [0, 1]$ is a probability mass function on a finite set S ,

$$\lim_{t \rightarrow 1} H_t(p) = \lim_{t \rightarrow 1} \frac{1}{t-1} \left(1 - \sum_{s \in S} p(s)^t \right) = - \sum_{s \in S} p(s) \ln p(s) =: H(p).$$

The quantity $H(p)$ is the *Shannon entropy* of p (in nats). This entails that Shannon entropy emerges when computing the derivative of the magnitude function $f(t) = \text{Mag}(t\mathcal{M})$ at $t = 1$:

$$f'(1) = \sum_{x \in \text{ob}(\mathcal{M}) \setminus T(\perp)} H(p_x). \quad (12)$$

Alternatively, we can deduce from Equation (10) that

$$f(t) = \#(\text{ob}(\mathcal{M})) - \sum_{x \in \text{ob}(\mathcal{M}) \setminus T(\perp)} Z_x(t),$$

where

$$Z_x(t) = \sum_{a \in A \cup \{\dagger\}} p_x(a)^t = \sum_{a \in A \cup \{\dagger\}} e^{-t(-\ln p_x(a))}$$

is the *partition function* of a system with microstates $A \cup \{\dagger\}$ and internal energy $E_x(a) = -\ln p_x(a) = d(x, xa)$ at inverse temperature $t > 0$. One can verify (see, for instance, [Baez, 2011]) that

$$H(p_x) = - \left. \frac{d}{dt} Z_x(t) \right|_{t=1},$$

from which Equation (12) also follows.

Since internal energy does not have to be normalized, we could also write $f(t) = \#(\text{ob}(\mathcal{M})) - \tilde{Z}(t)$, where $\tilde{Z}$ is the partition function of a system with microstates $(x, a) \in S := (\text{ob}(\mathcal{M}) \setminus T(\perp)) \times (A \cup \{\dagger\})$ and internal energy $E(x, a) = d(x, a) = -\ln p_x(a)$. The set S parameterizes indecomposable non-identity arrows in $\mathbf{L}$. For each $t > 0$, there is a corresponding probability mass function (Gibbs state) ρ on S , given by $\rho(s) = e^{-tE(s)}/\tilde{Z}(t)$. Then

$$\mathbb{E}_\rho(E) = \sum_{s \in S} E(s) \frac{e^{-tE(s)}}{\tilde{Z}(t)} = - \frac{d}{dt} \ln \tilde{Z}(t) = \frac{f'(t)}{\tilde{Z}(t)},$$

which gives yet another interpretation of the derivative of the magnitude function.

3.12. REMARK. If desired, one can also make sense of the magnitude function of an enriched category associated with a particular string. That is, suppose $x \in \text{ob}(\mathcal{M})$ and let $\mathcal{M}_x$ denote the $[0, \infty]$ -category whose objects are strings y containing x as a prefix. Define the hom-object between any $y, z \in \text{ob}(\mathcal{M}_x)$ to be $\mathcal{M}_x(y, z) := \mathcal{M}(y, z) = -\ln \pi(z|y)$. Compositionality and the identity requirement hold since they hold in $\mathcal{M}$. So $\mathcal{M}_x$ is indeed a $[0, \infty]$ -category and its magnitude function is given by

$$\zeta_{\mathcal{M}_x} = \zeta_{\mathcal{M}}|_{\text{ob}(\mathcal{M}_x) \times \text{ob}(\mathcal{M}_x)}$$

and because of Proposition 3.6, its inverse is equal to

$$\zeta_{\mathcal{M}_x}^{-1} = \zeta_{\mathcal{M}}^{-1}|_{\text{ob}(\mathcal{M}_x) \times \text{ob}(\mathcal{M}_x)},$$

which implies

$$\begin{aligned}
 \text{Mag}(t\mathcal{M}_x) &= \sum_{y,z \in \text{ob}(\mathcal{M}_x)} \zeta_{\mathcal{M}_x}^{-1}(y, z) \\
 &= \sum_{y \in \text{ob}(\mathcal{M}_x)} \zeta_{\mathcal{M}_x}^{-1}(y, y) - \sum_{y \in \text{ob}(\mathcal{M}_x) \setminus T(x)} \sum_{y \in \mathbf{L}_y^{(1)}} \pi(z|y)^t \\
 &= (t-1) \sum_{y \in \text{ob}(\mathcal{M}_x) \setminus T(x)} H_t(p_y) + \#(T(x)).
 \end{aligned}$$

3.13. MAGNITUDE HOMOLOGY. The magnitude homology of metric spaces and enriched categories has been studied extensively by [Leinster and Shulman \[2021\]](#). The main theorem of their work, namely [[Leinster and Shulman, 2021](#), Theorem 7.14], expresses the magnitude function of a metric space (M, d) as a weighted sum of Euler characteristics of certain chain complexes. More precisely, for each $\ell \in [0, \infty)$, there is a chain complex comprised of free abelian groups $(MC_{k,\ell}(M))_{k \in \mathbb{N}}$ such that $MC_{k,\ell}(M) := \mathbb{Z}[G_{k,\ell}]$ where

$$G_{k,\ell} = \left\{ (y_0, \dots, y_k) \in M^{k+1} \mid \sum_{i=0}^{k-1} d(y_i, y_{i+1}) = \ell \text{ and for all } i, y_i \neq y_{i+1} \right\}.$$

The differential $\partial_k : MC_{k,\ell}(M) \rightarrow MC_{k-1,\ell}(M)$ is defined as an alternating sum

$$\partial_k = \sum_{i=0}^k (-1)^i \partial_k^i$$

where for each $0 < i < k$,

$$\partial_k^i(y_0, \dots, y_k) = \begin{cases} (y_0, \dots, y_{i-1}, y_{i+1}, \dots, y_k) & \text{if } d(y_{i-1}, y_i) + d(y_i, y_{i+1}) = d(y_{i-1}, y_{i+1}) \\ 0 & \text{otherwise} \end{cases}$$

and $\partial_k^0 = \partial_k^k = 0$. The resulting $[0, \infty)$ -graded chain complex $(MC_{\bullet,\ell}(M), \partial_\bullet)$ is called the *magnitude complex* of M [[Leinster and Shulman, 2021](#), Definition 3.3] and its homology $H_{\bullet,\ell}(M)$ is called *magnitude homology* [[Leinster and Shulman, 2021](#), Definition 3.4]. The same definitions hold for any generalized metric space, such as $\mathcal{M}$.

The following corollary expresses the magnitude function of $\mathcal{M}$ as a weighted sum of Euler characteristics. It is a direct result of Proposition 3.6 and also mirrors [[Leinster and Shulman, 2021](#), Theorem 7.14], although the situation is considerably simpler here because each complex $(MC_{\bullet,\ell})_\ell$ is bounded and moreover only finitely many ℓ can arise as total lengths, which implies we only have to deal with finite sums instead of infinite series.

3.14. PROPOSITION. For any $t > 0$, set $q = e^{-t}$. Then

$$\text{Mag}(t\mathcal{M}) = \sum_{\ell} q^{\ell} \sum_{k \geq 0} (-1)^k \text{rank}(H_{k,\ell}(\mathcal{M}))$$

where the first sum is over those finitely many values of $\ell \in [0, \infty)$ for which

$$\bigcup_{k \geq 0} MC_{k,\ell}(\mathcal{M}) \neq \emptyset.$$

PROOF. By combining the definition of the magnitude function and its expression in Proposition 3.6, we have

$$\begin{aligned} \text{Mag}(t\mathcal{M}) &= \sum_{x,y \in \text{ob}(\mathcal{M})} \zeta_t^{-1}(x,y) \\ &= \sum_{x,y \in \text{ob}(\mathcal{M})} \sum_{k \geq 0} \sum_{\substack{\text{nondeg. paths} \\ x=y_0 \rightarrow y_1 \rightarrow \dots \rightarrow y_k=y \text{ in } \mathbf{L}}} (-1)^k \prod_{i=1}^k \pi(y_i|y_{i-1})^t. \end{aligned}$$

Observe that the sum over nondegenerate paths can be restricted to those paths $x = y_0 \rightarrow y_1 \rightarrow \dots \rightarrow y_k = y$ such that $\prod_{i=1}^k \pi(y_i|y_{i-1}) \neq 0$, which implies that $d(y_i, y_{i+1}) < \infty$ for any $i = 0, \dots, k-1$. Further, there is a bijective correspondence between nondegenerate paths $x = y_0 \rightarrow y_1 \rightarrow \dots \rightarrow y_k = y$ of total length $\ell := \sum_{i=0}^{k-1} d(y_i, y_{i+1}) < \infty$ and generators $(y_0, \dots, y_k)$ in $G_{k,\ell}$. Now remark that $\prod_{i=1}^k \pi(y_i|y_{i-1})^t = \exp(-t\ell)$, and so it is natural to group all the paths that have the same total distance, resulting in the following:

$$\text{Mag}(t\mathcal{M}) = \sum_{k \geq 0} \sum_{\ell \in [0, \infty)} \sum_{c \in G_{k,\ell}} (-1)^k e^{-t\ell}. \quad (13)$$

Since $\mathcal{M}$ is finite, only finitely many real numbers $\ell \in [0, \infty)$ can arise as total lengths of nondegenerate paths, which clarifies the meaning of the second sum. Moreover

$$\sum_{c \in G_{k,\ell}} (-1)^k e^{-t\ell} = (-1)^k q^{\ell} \#G_{k,\ell} = (-1)^k q^{\ell} \text{rank}(MC_{k,\ell}(\mathcal{M})).$$

The combination of this fact and Equation (13) yields the claim by following a classical argument present in the proof of [Leinster and Shulman, 2021, Theorem 6.17], for instance; it is important that each group $MC_{k,\ell}$ is finitely generated and that $(MC_{\bullet,\ell})_{\ell}$ is bounded. ■

3.15. REMARK. Proposition 3.14 expresses the magnitude function as a weighted sum of Euler characteristics. Together with Proposition 3.10, it establishes a connection between entropy and topological invariants, perhaps sitting alongside other recent results linking information theory and algebraic topology [Baudot and Bennequin, 2015; Vigneaux, 2020; Bradley, 2021; Mainiero, 2019].

Finally, we make a few additional observations about the *magnitude homology* of $\mathcal{M}$, which is the homology of the magnitude complex defined above. Although Leinster and Shulman work primarily with metric spaces in [Leinster and Shulman, 2021], two of their results translate directly to our setting:

1. The group $H_{0,0}(\mathcal{M})$ is the free abelian group on $\text{ob}(\mathcal{M})$, and for any $\ell > 0$, $H_{0,\ell}(\mathcal{M}) = 0$, cf. [Leinster and Shulman, 2021, Theorem 4.1].
2. The group $H_{1,\ell}(\mathcal{M})$ is the free abelian group on the set of ordered pairs (x, y) such that $x \neq y$ and $d(x, y) = \ell$ and there does not exist any point strictly between x and y in $\mathbb{L}$, cf. [Leinster and Shulman, 2021, Theorem 4.3].

Given these facts, one rewrite Equation (10) as

$$\text{Mag}(t\mathcal{M}) = \text{rank } H_{0,0}(\mathcal{M}) - \sum_{\ell \geq 0} q^\ell \text{rank } H_{1,\ell}(\mathcal{M}). \quad (14)$$

We conjecture that higher homology groups vanish. If this is the case, Equation (14) would also follow Proposition 3.14.

4. Final remarks

We conclude with a few remarks surrounding the results of this article. To start, one might notice that the zeta function associated with the generalized metric space $\mathcal{M}$ relates to perplexity, which is frequently used in evaluating autoregressive language models. To elaborate, the *perplexity* of a tokenized sequence $y = a_0 a_1 \cdots a_n$ is

$$PPL(x) = \exp \left\{ -\frac{1}{n} \sum_{i=1}^n \ln p(a_i | y_{<i}) \right\}$$

where $p(-|y_{<i})$ is the next-token probability distribution generated by the model when prompted with $y_{<i}$ [Jurafsky and Martin, 2025, Sec. 3.3], cf. Subsection 2.2. Perplexity is thus equal to the reciprocal of the zeta function $\zeta_t(a_0, y)$ when $t = 1/n$, and a_0 is the first token (usually $\perp$), and y is the full string:

$$PPL(y) = 1/\zeta_t(a_0, y) = 1/e^{t \ln \pi(y|a_0)},$$

or said differently, $\zeta_t(a_0, y) = 1/PPL(y)$. So perhaps for intuition, one might wish to think of the zeta function $\zeta_t(x, y)$ for arbitrary x, y as a generalization of the (reciprocal) of perplexity.

On a different note, as remarked in the introduction, one advantage to defining an enriched category of strings from a language is that one can pass to enriched *copresheaves* on those strings, and the latter functor category contains rich structure. Concretely, there is a $[0, \infty]$ -category $\widehat{\mathcal{M}} := [0, \infty]^{\mathcal{M}}$ whose objects are $[0, \infty]$ -copresheaves, that is, functions $f: \mathcal{M} \rightarrow [0, \infty]$ satisfying $\max\{f(y) - f(x), 0\} \leq \mathcal{M}(x, y)$ for all strings x and y . The relevance is that while $\mathcal{M}$ does not, a priori, have any sort of algebraic structure, the functor category $\widehat{\mathcal{M}}$ does, in the sense that one can compute weighted limits and colimits between copresheaves that are reminiscent of logical operations on meaning

representations of texts [Bradley et al., 2022, Section 5]. It may be interesting to compute the magnitude of an appropriate finite version of $\widehat{\mathcal{M}}$, cf. [Leinster, 2008, Example 2.5].

Lastly, let us draw a brief connection to the notion of diversity. *Categorical diversity* [Chen and Vigneaux, 2023] depends on a finite category $\mathbf{A}$, a probability mass function $p: \text{ob}(\mathbf{A}) \rightarrow [0, 1]$, and a similarity matrix $\theta: \text{ob}(\mathbf{A}) \times \text{ob}(\mathbf{A}) \rightarrow [0, \infty)$ such that

$$\theta(a, b) = 0 \quad \text{if} \quad \mathbf{A}(a, b) = \emptyset.$$

In the case of a generalized metric space, $\theta = \zeta_t$ and diversity is given by

$$\begin{aligned} \mathcal{H}(\mathcal{M}, p, \zeta) &= - \sum_{y \in \text{ob}(\mathcal{M})} p(y) \log \left(\sum_{z \in \text{ob}(\mathcal{M})} \zeta_t(y, z) p(z) \right) \\ &= - \sum_{y \in \text{ob}(\mathcal{M})} p(y) \log \left(\sum_{z \in \text{ob}(\mathcal{M})} \pi(z|y)^t p(z) \right). \end{aligned}$$

In [Chen and Vigneaux, 2023] the authors make a case for seeing this function as a probabilistic extension of $\log(\text{magnitude})$, just as entropy is a probabilistic extension of $\log(\text{cardinality})$ already in the work of Boltzmann and Gibbs. In the particular case of metric spaces, there are very deep connections between magnitude and diversity [Meckes, 2015]. We leave a detailed study of diversity of enriched categories of texts for future work, but remark here that if we take $p = \pi(-|x)|_{T(x)}$, then we recover the Shannon entropy of $\pi(-|x)|_{T(x)}$, which unlike our expression for the magnitude does not treat all states equally and gives more weight to those that are highly probable.

Very interestingly, in the particular case of a homogeneous and irreducible Markov chain with transition matrix $P: A \times A \rightarrow [0, 1]$ and with stationary distribution $p_\perp$, in the sense of Example 2.4, one can verify (following a standard computation, see e.g. the proof of [Walters, 1982, Thm. 4.27]) that the *diversity rate*

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{\mathcal{H}(\mathcal{M}, \pi(-|\perp)|_{T(\perp)}, \zeta)}{n} \\ = \lim_{n \rightarrow \infty} -\frac{1}{n} \sum_{(a_1, \dots, a_{N-1}) \in A^{N-1}} \pi(\perp a_1 \cdots a_{N-1} | \perp) \log \pi(\perp a_1 \cdots a_{N-1} | \perp) \end{aligned}$$

equals the Kolmogorov–Sinai entropy (rate) of the Markov chain,

$$- \sum_{a \in A} \sum_{a' \in A} p_\perp(a) P(a, a') \log P(a, a'),$$

which plays a fundamental role in information theory [Shannon, 1948, App. 3] and is an asymptotic lower bound for the logarithm of the perplexity of any LM that approximates the Markov chain [Jurafsky and Martin, 2025, Sec. 3.7].

References

- Michael F. Adamer, Edward De Brouwer, Leslie O’Bray, and Bastian Rieck. The magnitude vector of images. *Journal of Applied and Computational Topology*, 8(3):447–473, 2024.
- John C. Baez. Entropy as a functor, 2011. Blog post. Available online: <https://www.ncatlab.org/johnbaez/show/Entropy+as+a+functor>.
- Pierre Baudot and Daniel Bennequin. The homological nature of entropy. *Entropy*, 17(5):3253–3318, 2015. <https://doi.org/10.3390/e17053253>.
- Tai-Danae Bradley. Entropy as a topological operad derivation. *Entropy*, 23(9):1195, 2021. <https://doi.org/10.3390/e23091195>.
- Tai-Danae Bradley, John Terilla, and Yiannis Vlassopoulos. An enriched category theory of language: From syntax to semantics. *La Matematica*, 1:551–580, 2022. <https://doi.org/10.1007/s44007-022-00021-2>.
- Eric Bunch, Jeffery Kline, Daniel Dickinson, Suhaas Bhat, and Glenn Fung. Weighting vectors for machine learning: numerical harmonic analysis applied to boundary detection, 2021. arXiv preprint: arXiv:2106.00827.
- Stephanie Chen and Juan Pablo Vigneaux. Categorical magnitude and entropy. In Frank Nielsen and Frédéric Barbaresco, editors, *Geometric Science of Information: 6th International Conference, GSI 2023*, volume 14071 of *Lecture Notes in Computer Science*, pages 278–287, Cham, 2023. Springer Nature Switzerland.
- Brendan Fong and David I. Spivak. *An Invitation to Applied Category Theory: Seven Sketches in Compositionality*. Cambridge University Press, 2019.
- Stéphane Gaubert and Yiannis Vlassopoulos. Directed metric structures arising in large language models, 2024. arXiv preprint: arXiv:2405.12264.
- Geoffrey Grimmett and David Stirzaker. *Probability and Random Processes*. OUP Oxford, 2020. ISBN 978-0-19-258686-5.
- Philip Hall. The Eulerian functions of a group. *The Quarterly Journal of Mathematics*, os-7(1):134–151, 01 1936. ISSN 0033-5606. <https://doi.org/10.1093/qmath/os-7.1.134>.
- Jan Havrda and František Charvát. Quantification method of classification processes. Concept of structural α -entropy. *Kybernetika*, 3(1):30–35, 1967.
- Daniel Jurafsky and James H. Martin. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models. <https://web.stanford.edu/~jurafsky/slp3/>, 2025.

- G. M. Kelly. *Basic Concepts of Enriched Category Theory*. Cambridge University Press, 1982. Republished in: *Reprints in Theory and Applications of Categories*, No. 10 (2005) pp.1–136.
- F. W. Lawvere. Metric spaces, generalized logic and closed categories. *Rendiconti del Seminario Matematico e Fisico di Milano*, 43:135–166, 1973. <https://doi.org/10.1007/BF02924844>. Republished in: *Reprints in Theory and Applications of Categories*, No. 1 (2002) pp 1–37.
- Tom Leinster. The Euler characteristic of a category. *Documenta Mathematica*, 13:21–49, 2008. <https://doi.org/10.4171/DM/240>.
- Tom Leinster. The magnitude of metric spaces. *Documenta Mathematica*, 18:857–905, 2013. <https://doi.org/10.4171/DM/415>.
- Tom Leinster and Mark W. Meckes. *The magnitude of a metric space: from category theory to geometric measure theory*, pages 156–193. De Gruyter Open Poland, Warsaw, Poland, 2017. <https://doi.org/10.1515/9783110550832-005>.
- Tom Leinster and Michael Shulman. Magnitude homology of enriched categories and metric spaces. *Algebraic & Geometric Topology*, 21:2175–2221, 2021. <https://doi.org/10.2140/agt.2021.21.2175>.
- Tian Yu Liu, Matthew Trager, Alessandro Achille, Pramuditha Perera, Luca Zancato, and Stefano Soatto. Meaning representations from trajectories in autoregressive models. In *The Twelfth International Conference on Learning Representations*, 2024.
- Tom Mainiero. Homological tools for the quantum mechanic, 2019. arXiv preprint: arXiv:1901.02011.
- Matilde Marcolli, Noam Chomsky, and Robert C Berwick. *Mathematical structure of syntactic merge: An algebraic model for generative Linguistics*. MIT Press, 2025.
- Mark W. Meckes. Magnitude, diversity, capacities, and dimensions of metric spaces. *Potential Analysis*, 42(2):549–572, February 2015. ISSN 0926-2601, 1572-929X. <https://doi.org/10.1007/s11118-014-9444-3>.
- 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.
- Emily Roff and Masahiko Yoshinaga. The small-scale limit of magnitude and the one-point property. *Bulletin of the London Mathematical Society*, 57(6):1841—1855, 2025. <https://doi.org/10.1112/blms.70064>.
- Gian Carlo Rota. On the foundations of combinatorial theory I. Theory of Möbius functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 2(4): 340–368, 1964. <https://doi.org/10.1007/BF00531932>.
- Claude Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

Richard P. Stanley. *Enumerative Combinatorics: Volume 1*. Volume 49 of Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, MA, 2 edition, 2011.

Constantino Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52(1-2):479–487, 1988.

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, 2017.

Juan Pablo Vigneaux. Information structures and their cohomology. *Theory and Applications of Categories*, 35(38):1476–1529, 2020.

Juan Pablo Vigneaux. A combinatorial approach to categorical möbius inversion and pseudoinversion, 2024. arXiv preprint: arXiv:2407.14647.

Peter Walters. *An Introduction to Ergodic Theory*. Graduate Texts in Mathematics. Springer New York, NY, 1982.

Simon Willerton. Heuristic and computer calculations for the magnitude of metric spaces, 2009. arXiv preprint: arXiv:0910.5500.

SandboxAQ, Palo Alto, CA 94301, USA

Department of Mathematics, The Master’s University, Santa Clarita, CA 91321, USA

Department of Mathematics, California Institute of Technology, Pasadena, CA 91125, USA

Department of Linguistics, Northwestern University, Evanston, IL 60208, USA

Email: tai.danae@math3ma.com

jpvigneaux@northwestern.edu

This article may be accessed at <http://www.tac.mta.ca/tac/>

THEORY AND APPLICATIONS OF CATEGORIES will disseminate articles that significantly advance the study of categorical algebra or methods, or that make significant new contributions to mathematical science using categorical methods. The scope of the journal includes: all areas of pure category theory, including higher dimensional categories; applications of category theory to algebra, geometry and topology and other areas of mathematics; applications of category theory to computer science, physics and other mathematical sciences; contributions to scientific knowledge that make use of categorical methods. Articles appearing in the journal have been carefully and critically refereed under the responsibility of members of the Editorial Board. Only papers judged to be both significant and excellent are accepted for publication.

SUBSCRIPTION INFORMATION Individual subscribers receive abstracts of articles by e-mail as they are published. To subscribe, send e-mail to tac@mta.ca including a full name and postal address. Full text of the journal is freely available at <http://www.tac.mta.ca/tac/>.

INFORMATION FOR AUTHORS L^AT_EX₂ε is required. Articles may be submitted in PDF by email directly to a Transmitting Editor following the author instructions at <http://www.tac.mta.ca/tac/authinfo.html>.

MANAGING EDITOR. Geoff Cruttwell, Mount Allison University: gcruttwell@mta.ca

T_EXNICAL EDITOR. Nathanael Arkor, Tallinn University of Technology.

ASSISTANT T_EX EDITOR. Gavin Seal, Ecole Polytechnique Fédérale de Lausanne: gavin_seal@fastmail.fm

T_EX EDITOR EMERITUS. Michael Barr, McGill University: michael.barr@mcgill.ca

TRANSMITTING EDITORS.

Clemens Berger, Université Côte d'Azur: clemens.berger@univ-cotedazur.fr

Julie Bergner, University of Virginia: jeb2md@virginia.edu

John Bourke, Masaryk University: bourkej@math.muni.cz

Maria Manuel Clementino, Universidade de Coimbra: mmc@mat.uc.pt

Valeria de Paiva, Topos Institute: valeria.depaiva@gmail.com

Richard Garner, Macquarie University: richard.garner@mq.edu.au

Ezra Getzler, Northwestern University: getzler@northwestern.edu

Rune Haugseng, Norwegian University of Science and Technology: rune.haugsgeng@ntnu.no

Dirk Hofmann, Universidade de Aveiro: dirk@ua.pt

Joachim Kock, Universitat Autònoma de Barcelona: Joachim.Kock@uab.cat

Stephen Lack, Macquarie University: steve.lack@mq.edu.au

Tom Leinster, University of Edinburgh: Tom.Leinster@ed.ac.uk

Sandra Mantovani, Università degli Studi di Milano: sandra.mantovani@unimi.it

Matias Menni, Conicet and Universidad Nacional de La Plata, Argentina: matias.menni@gmail.com

Giuseppe Metere, Università degli Studi di Palermo: giuseppe.metere@unipa.it

Kate Ponto, University of Kentucky: kate.ponto@uky.edu

Robert Rosebrugh, Mount Allison University: rrosebrugh@mta.ca

Jiri Rosický, Masaryk University: rosicky@math.muni.cz

Giuseppe Rosolini, Università di Genova: rosolini@unige.it

Michael Shulman, University of San Diego: shulman@san Diego.edu

Alex Simpson, University of Ljubljana: Alex.Simpson@fmf.uni-lj.si

Tim Van der Linden, Université catholique de Louvain: tim.vanderlinden@uclouvain.be

Christina Vasilakopoulou, National Technical University of Athens: cvasilak@math.ntua.gr