A FORMAL LIMITATION ON LEARNING HUMAN LANGUAGE FROM TEXTUAL CORPORA

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a listener recover what a speaker means from the form of an utterance alone? We answer this question information-theoretically, and for a listener given by *any* featurizer of text, including the hidden states of contemporary large language models. Modeling language use as a joint distribution over meanings, contexts, and utterances, we derive upper bounds on the probability that a decoder recovers a speaker's intended meaning from a representation of the utterance. The bounds are governed by the uncertainty that form leaves about meaning, which splits into an irreducible part and a part that only (extralinguistic) context, but never the utterance alone, can resolve. Because these quantities are intrinsic to language, no representation, however much text or supervision produced it, can surpass them. The bounds apply, moreover, to meaning spaces that are discrete or continuous. We provide empirical evidence in support of the theory through experiments on artificial languages, Mandarin zero-pronoun resolution, and color reference.

1 INTRODUCTION

How much about natural language can be learned from text alone? It is an old question, and for over half a century the classical answer, from formal language learning theory, was largely negative. Famously, Gold (1967) showed that a learner presented with an unbounded stream of grammatical strings, but never told which strings are *un*grammatical, cannot in general be guaranteed to converge on the target grammar. Angluin (1980) then tightened the analysis, determining exactly which classes of languages *are* identifiable from positive data—a condition that the expressive grammars posited for natural language fail to meet. Read literally, these results cast text, an unlabeled stream of positive examples, as a fundamentally impoverished signal from which to acquire a language.

Probability appeared to offer a way out. Gold's setting demanded the exact identification of the entirety of a boolean grammar: a string is either in the language or not. However, learning from positive data *is* conceivable if the problem of grammar induction is relaxed to a probabilistic setting where the goal is no longer to learn a strict binary distinction between grammatical and agrammatical strings, but rather the probability distribution over strings. This reorientation has deep roots in Harris's (1954) distributional hypothesis, that a linguistic unit is characterized by the contexts in which it occurs (Wittgenstein, 1953; Firth, 1957). In modern natural language processing, the distributional hypothesis is operationalized through distributed representations (Hinton, 1986; Bengio et al., 2003; Mikolov et al., 2013; Levy & Goldberg, 2014), in which words and their contexts are encoded as real-valued vectors whose geometry reflects distributional similarity. Language models built on such representations estimate distributions over strings directly, and large language models (LLMs) trained on trillions of tokens of text continue in this tradition (Grattafiori et al., 2024; Groeneveld et al., 2024; Yang et al., 2025, *inter alia*). Distributional methods of this kind show that a remarkable amount about linguistic form, and even about meaning (Boleda, 2020), can be recovered from the statistical structure of text alone.

In the probabilistic setting, not everyone is convinced that form suffices. Bender & Koller (2020) pushed against this optimism with an influential thought experiment: an octopus that eavesdrops only on the linguistic forms exchanged between two speakers across an undersea cable, never observing the world they share, cannot, Bender and Koller argue, learn meaning, which is grounded in communicative intent and the extralinguistic world (Harnad, 1990). In some respects this skepticism has proven prescient, as the most capable LLMs are no longer trained on text alone. A post-

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

training stage now routinely augments the maximum-likelihood estimation of a text distribution with extralinguistic supervision; most prominently, reinforcement learning from human feedback (RLHF) optimizes models against human preference judgments (Christiano et al., 2017; Ouyang et al., 2022), injecting a signal about communicative success that raw corpora do not contain. As the frontier continues to move beyond the purely textual, the original question grows sharper rather than moot. We pose it in its probabilistic form: given the distribution of language use, what are the in-principle limits on what *any* system—however much text it observes—can infer about language, and in particular about the *meanings* that speakers intend, from form alone?

Before answering this question, we first remark that *meaning* admits many conceptualizations. For example, meaning can be understood in terms of *reference* (the entities that the utterance picks out in the world); or *sense* (the "thought" an utterance evokes, independent of reference) (Frege, 1892). In addition, meaning can be analyzed *conventionally*—on utterances in-and-of-themselves—or *pragmatically*, taking extralinguistic context into account (Grice, 1957; Sperber & Wilson, 1986/1995). Wittgenstein (1953) provides an example of this distinction, in which a builder shouts "Slab!" to his coworkers on a construction site: the *conventional* meaning of the utterance is (simply) a slab. However, in context, the speaker's intended (or *pragmatic*) meaning is instead "Bring me the stone slab." Human language use is fundamentally pragmatic (Yule, 1996). Indeed, every (human-produced) natural language utterance—including those in an LLM's training data—was said in some social, physical, and cultural context in which a speaker *intended a meaning* and, informed by context, produced an utterance from which a listener could recover that meaning (Grice, 1975). Because speaker intent is so key to communication between a speaker and a listener, then critically, speaker intents also underlie the *distribution* of linguistic forms at scale, which aggregates over these single instances of communication. Despite how foundational speaker intent is to language use, whether systems trained on the distribution of forms alone can represent intended meanings has been understudied in comparison to (conventional) referential or sentential notions of meaning (Merrill et al., 2021; Mandelkern & Linzen, 2024; Baggio & Murphy, 2026). For this reason, we privilege *intended meaning* in our analysis of meaning representation. Our analysis reveals that systems trained on forms (divorced from context) learn provably impoverished representations of intended meaning.

2 A Model of Communication

Our theoretical results in §3 rest on a model of (human) communication in context. In this section, we present the model, which is based loosely on Shannon's communication channel (Shannon, 1948) and formalizes communication between a speaker and a listener.

Notation. We denote random variables in uppercase (X), their realizations in lowercase (x), and their supports in calligraphic type ($\mathcal{X}$); thus $x \in \mathcal{X}$ is one particular value among the set $\mathcal{X}$ of possible values. Our model of communication involves the following random variables:

- M: the **meaning** the speaker intends. Its support $\mathcal{M}$ may be finite, e.g., a closed set of pronouns, or continuous, e.g., a space of colors. To measure distances, we equip $\mathcal{M}$ with a metric d (§3.2).
- C: the **context** in which communication takes place, encompassing everything other than the current utterance, both extralinguistic, e.g., speaker identities, and linguistic, e.g., the prior conversation history. We assume the context is shared by speaker and listener.
- U: the **utterance**: the linguistic form the speaker produces to convey M in context C.
- $\widehat{M}$: the listener's **inferred meaning**. $\widehat{M}$ shares the support $\mathcal{M}$ with M, and communication succeeds to the extent that $\widehat{M}$ matches M.
- Z: a **representation** of the utterance, e.g., the activations of an LLM run on U. Representations are produced by a featurizer $g\colon \mathcal{U} \to \mathcal{Z}$ and mapped back to meanings by a decoder $f\colon \mathcal{Z} \to \mathcal{M}$ (§3).

We write $\mathbb{P}$ for a probability measure over the product space of all random variables. We write $\mathrm{H}(M)$ for the Shannon entropy of M, $\mathrm{H}(M \mid U)$ for the conditional entropy of M given U, and $\mathrm{I}(M;U) = \mathrm{H}(M) - \mathrm{H}(M \mid U)$ for their mutual information; conditional variants such as $\mathrm{I}(M;C \mid U)$ are defined analogously.

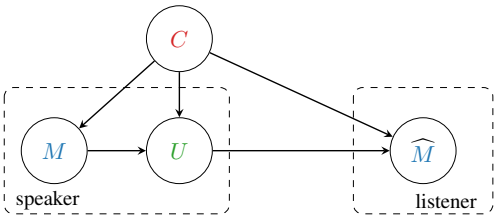


Figure 1: **Communication model.** During communication, a context C influences a speaker's intended meaning, given by random variable M, which the speaker encodes into an utterance U given that same context. Under the same context C, the listener decodes the utterance U into their inferred meaning $\widehat{M}$. Communication is considered successful if $M \approx \widehat{M}$.

The communication model. The graphical model in Figure 1 wires together the variables we just defined. In particular, Figure 1 realizes a dyadic model of communication based on Shannon's communication channel (Shannon, 1948). During a single dialogue turn, a speaker (left dashed box) intends a meaning M in context C and produces an utterance U modulated by the context, where the context can contain both *extralinguistic* factors, such as speaker identities, and *linguistic* ones, such as the prior conversation history (Grice, 1975; Yule, 1996). Given the same context as the speaker, the listener (right dashed box) reconstructs the meaning of the utterance.

We model language use in a speech community as made up of many such single instances of dyadic communication, each between a single speaker and listener. Language use can therefore be described by a *distribution* over utterances, contexts, and meanings produced by various speaker–listener pairs (Frank & Goodman, 2012). We are interested in information-theoretic properties of the entire system, so we abstract away from single communicative instances, modeling their components as random variables and communication as a generative process over them. The graphical model in Figure 1 encodes the following relationships in its edges:

- $C \to M$: meanings arise *in* context; what a speaker intends to convey may depend on the situation, including the topic under discussion, the interlocutors, and their goals.
- $M \to U$ and $C \to U$: the speaker chooses an utterance form given both their intended meaning and the context, e.g., leaving unsaid what the context already makes clear.
- $U \to \widehat{M}$ and $C \to \widehat{M}$: the listener forms their inference from the utterance they perceive and the context they share with the speaker.

The model encodes two key assumptions. First, we assume that the same context is available to the speaker and listener—in practice, this is often not the case: when the interlocutors' contexts differ, they may undergo a process of pragmatic repair to establish a common ground (Clark & Brennan, 1991). Indeed, we do not analyze the multi-turn case at all, focusing instead on single-turn one-way communication from speaker to listener. Second, the linguistic channel may contain noise. For simplicity, our exposition assumes the channel is noiseless; however, our main results directly apply to the noisy channel setting, which we will revisit briefly in §3.

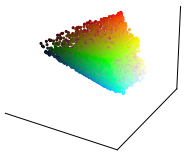
3 Bounding Meaning Inference from Text-Based Representations

Now given our model of communication (Figure 1), we set up the problem of intended meaning representation in a listener who only accesses utterance forms. While our exposition focuses on language models—importantly, text-only LLMs that do not observe context—the following applies to all systems that only observe ungrounded forms. To make precise our notion of meaning representation, here, we first address how we operationalize meanings, followed by representation.

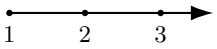
We model meanings as a metric space $(\mathcal{M}, d)$ (Gärdenfors, 2000). Our theory handles both finite meaning sets $\mathcal{M}$, e.g., the closed set of pronouns, or infinite meaning sets, such as continuous color spaces or the natural numbers. Meanings $\mathcal{M}$ can, moreover, be bounded, as in the case of colors (below, **(a)**), which are constrained by human visual perception; or unbounded, as in the case of the natural numbers (below, **(b)**).

In **(b)**, even though the natural numbers are unbounded, humans rarely communicate about extremely large numbers. Instead, cross-linguistically, the likelihood of communicating about a natural number, described by $\mathbb{P}(M)$ in our notation, empirically falls off according to an inverse-square law (Dehaene & Mehler, 1992; Piantadosi, 2014). Then, critically, $\mathbb{P}(M)$ is *tight*, that is, for every small number

$\delta > 0$, there exists a compact (closed and bounded) subset of meanings $\mathcal{M}' \subseteq \mathcal{M}$ that make up most of the probability mass, i.e., such that $\mathbb{P}(M \in \mathcal{M}') > 1 - \delta$. We assume that in language use, meaning distributions are tight, behaving generally like the natural numbers. To keep the presentation simple, in the main text we will present only the *bounded* meanings case, which already meets the tightness assumption (§3.1, 3.2), leaving the more complicated unbounded case to the Appendix.



(a) Bounded $\mathcal{M}$ (colors)



(b) Unbounded $\mathcal{M} = \mathbb{N}$

Learning to represent meaning. What does it mean for a system like an LLM to (learn to) represent speaker meaning? Language models take linguistic forms as input and are trained to output a distribution over the next token. As a byproduct of this training, these models learn a set of feature maps on linguistic forms, given by the models' internal activations under a textual input. In our paper's notation, LLM activations implement a function $g\colon \mathcal{U} \to \mathcal{Z}$, where g maps the space of *forms* ($\mathcal{U}$) to a *representation space* ($\mathcal{Z}$, typically $\mathbb{R}^d$). We propose that a necessary condition for LLM activations to *represent* speaker meaning is that the representation space $\mathcal{Z}$ permits successful inference of the original meaning. That is, there should exist some $f\colon \mathcal{Z} \to \mathcal{M}$ that approximately maps representations in $\mathcal{Z}$ back to true meanings in $\mathcal{M}$.

The problem of learning a good meaning representation can be distilled into learning a $g : \mathcal{U} \to \mathcal{Z}$ such that approximate meaning recovery from $\mathcal{Z}$ is possible: that is, such that there exists $f : \mathcal{Z} \to \mathcal{M}$ where $d(f(g(u)), m) \leq \epsilon$. We consider two cases in particular: $\epsilon = 0$, corresponding to exact recovery, or *classification*, and $\epsilon > 0$, corresponding to ϵ-accurate recovery in a *regression* setting; the former is only applicable for finite $\mathcal{M}$. In the following sections, for each case, we will upper-bound the probability of successful meaning recovery from a representation of the utterance (Theorems 1 and 2). We emphasize that our definitions (and, by extension, our theoretical results) apply to *any* text featurizer g, including the internal activations of a perfect learner trained on infinite data. In short, our results put forth a theoretical limit on meaning representation from text. Relevant to a machine learning audience, this limit also constrains the extent to which meanings can be *learned* from text: what one cannot represent, one cannot learn.

Meaning in utterance vs. context To build intuition about meaning representation in a purely text-based system, first note that both context and utterance contribute to meaning inference in communication. The listener's overall ability to infer meaning can be described by the conditional entropy $\mathrm{H}(M \mid U, C)$ of meanings M given utterances U and context C: if $\mathrm{H}(M \mid U, C) \approx 0$, then it is generally possible to recover the meaning in context. But what of intended meaning can be recovered from utterances alone? Here, our quantity of interest is the conditional entropy $\mathrm{H}(M \mid U)$ of meanings given utterances (without context), which can be decomposed as follows:

$$\mathrm{H}(M \mid U) = \mathrm{H}(M \mid U, C) + \mathrm{I}(M; C \mid U). \tag{1}$$

The RHS consists of $\mathrm{H}(M \mid U, C)$: irreducible meaning ambiguity in communication; and $\mathrm{I}(M; C \mid U)$: what context contributes to meaning independently of the utterance. We assume that in natural language, $\mathrm{H}(M \mid U, C)$ is generally small—communication is typically understood. Then, remaining meaning ambiguity given the utterance is explained by context, seen by the term $\mathrm{I}(M; C \mid U)$. Next, in Theorems 1 and 2, we will see how the terms in Equation (1) govern how well *any* representation of a text utterance, e.g., its activations under an LLM, represent speaker meaning.

3.1 Exact recovery for finite meaning spaces

We consider a classification task that aims to recover the true meaning $M = \widehat{M}$ from a representation of the utterance. We assume the random variable M over meanings has finite categorical support $\mathcal{M}$ (we analyze the countably infinite case in Appendix G and the uncountable case in the next section). The goal in this task is to find some classifier $f : \mathcal{Z} \to \mathcal{M}$ that maps representations Z to their correct meanings. Then, using Equation (1), we can upper-bound the probability that *any* f is correct:

Theorem 1. *Let $g\colon \mathcal{U} \to \mathcal{Z}$ be a map from utterances ($\mathcal{U}$) to representations ($\mathcal{Z}$). Let $f\colon \mathcal{Z} \to \mathcal{M}$ map representations to the original meanings ($\mathcal{M}$). Let the probability that f is correct be $p_e :=$*

$\mathbb{P}[f(g(U)) = M]$. *Then,*

$$\sup_{f \in \mathcal{M}^{\mathcal{Z}}} p_e \leq 1 - \frac{\mathrm{H}(M \mid U) - \mathrm{H}_2(p_e)}{\log(|\mathcal{M}| - 1)} = 1 - \frac{\overbrace{\mathrm{H}(M \mid U, C)}^{\textit{irreducible ambiguity}} + \overbrace{\mathrm{I}(M; C \mid U)}^{\textit{meaning from context only}} - \mathrm{H}_2(p_e)}{\log(|\mathcal{M}| - 1)}.$$

Proof. See Appendix F. ■

Theorem 1 bounds the probability of correct classification, $p_e \stackrel{\text{def}}{=} \mathbb{P}[f(g(U)) = M]$, when the meaning space $\mathcal{M}$ is finite and categorical.[1] While categorical meaning spaces are common (e.g., kinship terms (Kemp & Regier, 2012)), many meaning spaces of interest may however be *uncountably infinite*—for instance, the continuous space of colors. In the next section, we derive an analogous result for a *regression* setting on potentially uncountable spaces.

3.2 Approximate recovery for infinite meaning spaces

Theorem 1 naturally extends to a regression setting over a potentially uncountable meaning space. Continuous meaning spaces fall into this category. Here, instead of considering the probability that f exactly matches the target meaning, under a regression objective we let p_e be the likelihood that $f(Z)$'s meaning estimate is strictly within ϵ of the ground truth, or $d(\widehat{M}, M) < \epsilon$.

Reformulating the problem to the setting of ϵ-accurate recovery requires some extra considerations. Notably, the proof of Theorem 1 relied on Fano's inequality, which applies to random variables $(M, \widehat{M})$ over *finite* alphabets, so it cannot be used here for uncountably infinite meaning spaces. The workaround is to reduce the regression problem to a *classification problem* where Fano's inequality can be directly applied. Formally, let the binary event A be the event that ϵ-accurate recovery holds, i.e., $A = \mathbb{1}[d(\widehat{M}, M) < \epsilon]$. We want to construct the event B that some classification problem, derived from A, is solved. Further, we want to design event B such that $A \Rightarrow B$; then, $\mathbb{P}(A) \leq \mathbb{P}(B)$. That way, we can upper-bound $\mathbb{P}(B)$ using Fano's inequality in the setting of Theorem 1, which will bound the likelihood of ϵ-accurate recovery, $\mathbb{P}(A)$, by extension.

We now construct the event B. Let $\{m_1, \ldots, m_{N(\epsilon)}\}$ be a maximal 2ϵ-packing of $\mathcal{M}$. That is, $d(m_i, m_j) \geq 2\epsilon$ for all $i \neq j \in 1 \ldots N(\epsilon)$, where $N(\epsilon)$ is the packing number. We set up a discrete random variable J_ϵ over these packing points, where J_ϵ is being classified in the setup of Theorem 1. To set up the distribution of J_ϵ, we make use of a "quantizer" function $q : \mathcal{M} \to \{m_n\}_{n=1}^{N(\epsilon)}$ that maps points in $\mathcal{M}$ to the nearest packing point in $\{m_n\}_{n=1}^{N(\epsilon)}$, where $J_\epsilon \stackrel{\text{def}}{=} q(M)$.

Now, let B be the event that packing points are correctly classified given the corresponding utterance, i.e., $B = \mathbb{1}[\widehat{J_\epsilon} = J_\epsilon]$, where $\widehat{J_\epsilon}$ is the estimated packing point. By definition of 2ϵ-packing, event $A \Rightarrow B$. That is, if for every $m \in \mathcal{M}$ the decoder's estimate is less than ϵ, or $d(f(g(u)), m) < \epsilon$, then we are guaranteed to correctly classify each packing point by taking the closest one to the decoder's output. Since $A \Rightarrow B$, then as desired, $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Now, we can upper-bound $\mathbb{P}(A)$—the likelihood of ϵ-accurate recovery—via an upper bound on $\mathbb{P}(B)$. We assume that $\mathcal{M}$ is bounded (for unbounded $\mathcal{M}$, see Appendix I). This yields the theorem:

Theorem 2 (ϵ-accurate regression)**.** *Let $g : \mathcal{U} \to \mathcal{Z}$ be a map from utterances ($\mathcal{U}$) to representations ($\mathcal{Z}$). Let $(\mathcal{M}, d)$ be a bounded metric space and $N(\epsilon)$ be its maximal 2ϵ-packing number. Abbreviate the probability that f is ϵ-accurate as $p_e^\epsilon = \mathbb{P}[d(f(g(U)), M) < \epsilon]$. Then,*

$$\sup_{f \in \mathcal{M}^{\mathcal{Z}}} p_e^\epsilon \leq 1 - \frac{\mathrm{H}(J_\epsilon \mid U) - \mathrm{H}_2(B)}{\log(N(\epsilon) - 1)} = 1 - \frac{\overbrace{\mathrm{H}(J_\epsilon \mid U, C)}^{\textit{irreducible ambiguity}} + \overbrace{\mathrm{I}(J_\epsilon; C \mid U)}^{\textit{meaning from context only}} - \mathrm{H}_2(B)}{\log(N(\epsilon) - 1)}.$$

Proof. See Appendix H. ■

[1] If the linguistic channel is noisy, then the inequality is strict.

Theorems 1 and 2 tell us that speaker meaning inference from *any* representation of the utterance—even that of a perfect learner trained on infinite data—cannot exceed a function of the purely linguistic ambiguity $\mathrm{H}(M \mid U)$. $\mathrm{H}(M \mid U)$ decomposes into a sum of two terms: $\mathrm{H}(M \mid U, C)$ or irreducible meaning ambiguity in communication, and $\mathrm{I}(M; C \mid U)$, i.e., what context resolves about meaning beyond the utterance. That is, whether communication is *low*- or *high*-context, governed by $\mathrm{I}(M; C \mid U)$, adjusts the ceiling on p_e.

4 Experiments

In the previous section, we put forth theoretical bounds for speaker meaning decodability from any representation of the utterance. Now, we show that these bounds hold in practice. As Theorem 1 requires access to ground truth speaker meanings, we will rely on simulated and real datasets for which intended meanings are directly available. We test both settings: (Theorem 1) meanings that are *categorical* and (Theorem 2) *continuous*. We first verify Theorems 1 and 2 in simulation using artificial languages sampled from pre-defined distributions. Then, we turn to natural language data: for categorical meanings, we consider *zero-pronoun resolution* in Mandarin Chinese, and for continuous meanings, we consider a *color-naming* task using Monroe et al.'s (2017) color reference dataset. In all cases, experiments are consistent with the theoretical bounds. We briefly sketch the task setups in this section; **for details see Appendices B and C**.

4.1 Artificial languages

We start by testing Theorems 1 and 2 on artificial languages sampled from pre-defined distributions $\mathbb{P}(M, U, C)$. Crucially, each language is constructed such that $\mathrm{H}(M \mid U, C) = 0$, which we call the *valid language assumption*: meaning can be perfectly decoded from the utterance in context, rendering communication successful. We sample a train and test dataset from $\mathbb{P}(M, U, C)$, then retain only the utterance-meaning pairs to train a decoder for meaning inference. Note that, since utterances are not constrained to be natural language, we do not use pre-trained LLMs for the decoder: instead, we opt for MLPs, moving on to LLMs for the natural language tasks in the next section.

Discrete meanings. We construct artificial languages by defining $\mathbb{P}(M, U, C)$, where $\mathcal{M}$ and $\mathcal{C}$ are finite sets and $\mathcal{U}$ is an alphabet of symbols. Figure 2A illustrates this setting. In brief, in a (categorical) context C (● in Figure 2A), the speaker intends a categorical meaning (e.g., ■) and produces a single symbol in $\mathcal{U}$ (e.g., "a"). The listener tries to decode back the original categorical meaning (■) given the utterance ("a"). We ask whether a map can be trained to predict M from U alone.

To answer this question, we construct a family of six artificial languages. The marginal distributions $\mathbb{P}(M)$, $\mathbb{P}(C)$, and $\mathbb{P}(U)$ over meanings, contexts, and utterances are held uniform over $|\mathcal{M}| = |\mathcal{C}| = 10$, $|\mathcal{U}| = 15$ values, and we systematically vary $\mathrm{I}(M; U) \in [0, \log_2 |\mathcal{M}|]$. Lastly, on each language, we validate Theorem 1 by training an MLP to infer meanings M from utterances U, and verify the best test accuracy (across hyperparameters and five random seeds) is lower than the theoretical bound.

Continuous meanings. Here, we construct artificial languages that communicate over a *continuous* meaning space. For simplicity, contexts and utterances remain categorical, with meanings drawn from a two-dimensional continuous uniform distribution $\mathbb{P}(M) = \mathrm{Unif}[-1, 1]^2$. Figure 2B illustrates this setting: in a shared categorical context (● in the figure), the speaker intends a meaning drawn from $\mathrm{Unif}[-1, 1]^2$ (here, $[-0.4, 0.2]$) and produces an utterance ("b"). Given the utterance, the listener approximately decodes the meaning (e.g., $[-0.43, 0.1]$).

We construct a family of languages $\mathbb{P}(M, U, C)$ that satisfy the valid language constraint ($\mathrm{H}(M \mid U, C) = 0$) and systematically vary across $\mathrm{I}(J_\epsilon; U)$, holding all else equal. Details on how we construct the languages are given in Appendix B. Finally, to validate Theorem 2, we train an MLP to predict M from U, minimizing the MSE loss. We then compare the best-case empirical test ϵ-accuracy across hyperparameter sweeps to the theoretical upper bound.

4.2 Natural language

Moving on from artificial languages to natural language, we use zero-pronoun resolution in Mandarin Chinese (Wang et al., 2018) as a testbed for the discrete Theorem 1 and a color naming task (Monroe

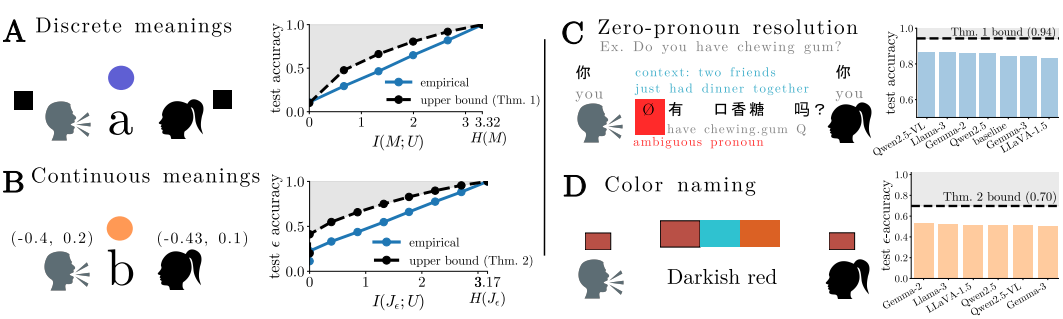


Figure 2: **Experiments. A:** *(Left)* Artificial languages over a categorical meaning space. In a shared categorical context $C = \bullet$, the speaker (left) intends meaning $M = \blacksquare$ and produces utterance $U =$"a". The listener decodes U in context back into the original meaning. *(Right)* For a family of six artificial languages (dots on the curves) that vary $\mathrm{I}(M;U)$ (x-axis), the theoretical upper bound (black) exceeds the best empirical test accuracy across runs (blue), validating Theorem 1. **B:** *(Left)* Artificial languages over continuous meanings in the ϵ-recovery setting. In shared categorical context $C = \bullet$, the speaker intends a 2D coordinate $M \sim \mathrm{Unif}[-1,1]^2$ and produces the utterance $U =$"b". The listener approximates M given utterance in context. *(Right)* Theorem 2 holds across eight artificial languages varying over $\mathrm{I}(J_\epsilon;U)$, seen by the theoretical bound exceeding the empirical test ϵ-accuracy ($\epsilon = 0.4$). **C:** *(Left)* Zero-pronoun resolution in Mandarin Chinese. In a context where the speaker and listener just had dinner together, the speaker asks the listener whether she has chewing gum, dropping the pronoun (you) from the utterance. The listener infers the true pronoun in context. *(Right)* For zero-pronoun resolution from LLM and VLM activations, Theorem 1 holds: the predicted accuracy ceiling (0.93, dashed line) exceeds the best empirical test accuracies across runs (blue bars). **D:** *(Left)* Color naming in context. Given a shared context C of three color chips, the speaker intends the target chip $M = \blacksquare$ and says $U =$"Darkish red", which the listener decodes in context to click on the target. *(Right)* Theorem 2's ceiling (0.70, dashed line) for ϵ-accurate recovery ($\epsilon = 0.2$) exceeds all models' best-case empirical performances over runs (orange bars).

et al., 2017) for the continuous Theorem 2. Here, since we are dealing with natural language, we use pre-trained LLM activations for the utterance featurizer g, and an MLP for the decoder f mapping LLM activations to the meaning. We first list the models, then sketch the tasks.

Models. For the utterance featurizer $g : \mathcal{U} \to \mathcal{Z}$, we use the activations of contemporary LLMs. As a point of comparison, we also consider vision-language models (VLMs) that may also be instruction-tuned, as they are trained with extralinguistic supervision (visual and human preferences). We test six models ranging from 2B-14B parameters: three LLMs, Qwen-2.5 14B, Llama-3 8B, Gemma 2 2B, and three VLMs, Qwen-2.5 VL (7B, instruction-tuned), LLaVa 1.5 (7B), and Gemma 3 (12B, instruction-tuned). For each model, we aim for the best meaning inference performance possible by searching over the last-token representations[2] of all layers.

Zero-pronoun resolution in Mandarin Chinese. Mandarin Chinese is a "pronoun-drop" (pro-drop) language: in pro-drop languages, a pronoun can be excluded from a grammatical sentence if the pronoun's identity can be inferred from surrounding context. For instance, Chinese grammar licenses both examples **(a)** and **(b)** below to mean "do you have chewing gum?", where, notably, **(b)** shows a "zero" pronoun (∅) in place of the second-person singular 你 ('you').

(1) **(a)**

你 有 口香糖 吗?
nǐ yǒu kǒuxiāngtáng ma

2SG have chewing.gum Q

(b)

Ø 有 口香糖 吗?
yǒu kǒuxiāngtáng ma

have chewing.gum Q

The task to infer the dropped pronoun from the utterance is known as *zero-pronoun resolution*, shown in Figure 2C: in a context where the speaker and listener just had dinner together, the speaker asks

[2]We use the last token representation as it is the only one to attend to the entire sequence.

the listener if she has chewing gum. Though the pronoun is dropped from the utterance, the listener, given context, infers the speaker asks *her* for chewing gum, recovering the true pronoun "你" (you).

We use zero-pronoun resolution to test Theorem 1, repurposing Wang et al.'s (2018) Chinese TV subtitles dataset. After preprocessing (Appendix B), the dataset contains $N = 904{,}996$ short utterances. Of these utterances, 20.3% ($N = 184{,}092$) drop the pronoun (as in example **(b)**). The context C, including all audiovisual TV content apart from the utterance, is not provided in the dataset, but we assume that Chinese speakers recover M given U and C with little difficulty (valid language assumption). The final number of pronoun classes in the data is $|\mathcal{M}| = 12$ (see Appendix B for details); the number of possible forms is $|\mathcal{U}| = 13$ (possible pronouns + dropped pronoun).

We extracted frozen LLM embeddings $g : \mathcal{U} \to \mathbb{R}^d$ of the utterances. Then, we trained an MLP $f : \mathbb{R}^d \to \Delta^{11}$ from the LLM embedding to predict the true pronoun, minimizing the cross-entropy loss. We compare against a strong **baseline** that directly selects the pronoun in the sentence if expressed and guesses the most frequent one 我 ('I') if dropped.[3] Lastly, for each model, we verify that the best-case empirical test accuracy across hyperparameters is lower than the theoretical bound.

Monroe et al.'s (2017) color naming task. We use Monroe et al.'s (2017) colors dataset as our testbed for ϵ-accurate regression (Theorem 2). The dataset is collected from humans playing a referential game (Lewis, 1969). During a single round of the task, a speaker and listener are presented the same three color chips (see Figure 2D), where the speaker must communicate the color of a *target* chip to the listener; only the speaker knows the identity of the target chip in advance. The speaker sends a message denoting the target color, e.g., "Darkish red", to the listener, who then clicks on their guess of the target among the three chips. Communication is considered successful if the listener clicks on the correct target, given the speaker utterance and context of three chips.

We filter the dataset for communicative successes, satisfying the valid language constraint $\mathrm{H}(M \mid U, C) = 0$. This yields $N = 38975$ rounds of the task from various speaker-listener pairs. Each round contains representations for three color chips (*context*) in CIELab (a 3D color space like RGB where Euclidean distances mirror human perceptual distances) (Gualdoni & Boleda, 2024), the target chip (*intended meaning*), and speaker message in natural language (*utterance*). Finally, for each model and ϵ, we collect the best-case test ϵ-accuracies over all runs, and verify that these lie below the theoretical upper bound in Theorem 2. For details on estimating the upper bound, see Appendix B.

4.3 Results

Experiments confirm the bounds on meaning inference performance predicted by Theorems 1 and 2.

Artificial languages. Figure 2A (right) shows, for artificial languages whose $\mathrm{I}(M; U)$ spans 0 to $\log_2(|\mathcal{M}|)$, the empirical test accuracy on meaning inference from utterance alone (blue curve) with the theoretical bound (dashed black curve). Two observations: first, the bound holds in practice for all tested languages (the blue curve is lower than black); second, as expected, as the mutual information between meaning and utterance increases, the theoretical bound on meaning inference performance, as well as the empirical performance, both increase. Figure 2B (right) shows a similar trend for the continuous case (for $\epsilon = 0.4$; other values of ϵ can be found in Appendix D). Like the discrete case, the bounds hold for all languages spanning $\mathrm{I}(J_\epsilon; U)$.

Zero-pronoun Resolution. Figure 2C shows that in all models (x-axis), the best meaning inference performance (y-axis) over all hyperparameters and random seeds does not exceed the bound predicted by Theorem 1 (0.94, horizontal line). Interestingly, using LLMs or VLMs did not considerably improve performance beyond the baseline (accuracy 0.85). This observation suggests that, for this dataset, the utterance alone carries little information about the dropped pronoun; LLMs and VLMs trained on web-scale data, moreover, do not overcome this bottleneck. Instead, we hypothesize that visual and social context (C) from the TV show (Wang et al., 2018) enable pronoun inference.

Color naming. Figure 2D shows, for $\epsilon = 0.2$, the predicted ceiling on meaning inference performance (dashed line); the bound holds for all models. All ϵ spanning $[0.01, 0.75]$ are shown in Figure 6; the bound held in all cases, though is tighter for larger ϵ, see Appendix E for discussion.

[3] Because only 20.3% of the dataset showed a dropped pronoun (i.e., 79.7% of the pronouns are explicit), we expect baseline performance to achieve $\approx 80\%$ accuracy.

5 Related Work

A sizeable and growing body of work asks whether LLMs—their behavior or internal activations—can instantiate models or theories of language (Bender & Koller, 2020; Baroni, 2022; Piantadosi, 2024; Boleda, 2025; Futrell & Mahowald, 2026). The present work was inspired by these broader debates. Theorems 1 and 2 suggest that text-only LLMs should not be used to instantiate a *theory* of meaning inference during naturalistic communication, given context is a crucial missing piece. However, our findings would still be consistent with using text-only LLMs' activations to *model* aspects of linguistic representation that can be derived from utterance alone.

Most similar to our work, Merrill et al. (2021) also ask whether ungrounded LLMs are able to acquire semantics. In brief, Merrill et al. (2021) formally show that there exist languages for which semantic equivalence (whether two utterances have the same *denotation*) cannot be recovered from ungrounded linguistic input, even by a perfect learner with access to unlimited data and an "oracle" module providing semantic information. Our work addresses a similar high-level question and arrives at qualitatively similar conclusions, but we take a different route. In particular, (1) we consider pragmatic meaning while Merrill et al. (2021) consider denotational meaning over sets; (2) our definition of successful meaning inference is metric-based (ϵ-accurate recovery), thus gradable, while Merrill and colleagues require exact set-matching; (3) our analytic framework is *information-theoretic* while theirs is based on computability; (4) we establish bounds directly on meaning decodability from a representation of text that can be empirically validated, while their results (more abstractly) establish the non-computability of semantic representation for a class of context-dependent languages. Intriguingly, the two works which diverge radically in method and conceptualization of meaning both recover (extralinguistic) grounding as the missing ingredient for meaning representation.

Our main result (Theorems 1 and 2), i.e., that text-based systems learn provably impoverished meaning representations, aligns with empirical findings that base LLMs perform worse than humans in implicature (Ruis et al., 2023) and presupposition (Sravanthi et al., 2024; Chang & Bergen, 2024) (see Ma et al. (2025) for a survey). A number of positive results in pragmatic meaning inference have also been reported notwithstanding (Andreas, 2022; Hu et al., 2023; Chang & Bergen, 2024); Theorems 1 and 2 suggest that high mutual information between surface forms and meanings may explain positive empirical findings in text-only LLMs' representation of pragmatic meaning.

6 Conclusion

We have put forth theoretical limits on meaning inference from representations of utterance alone (Theorems 1 and 2). While Theorems 1 and 2 apply to *any* representation of the utterance, we verified them for the internal activations of contemporary LLMs and VLMs, popular choices to model meaning in practice. With experiment corroborating theory, our take-home message is that there is no free lunch in meaning inference when meanings are gleaned through a linguistic channel. In natural settings, humans also make use of context as a conduit for meaning inference (Beuls & Van Eecke, 2024); however, in text-only systems, utterance forms bottleneck meaning representation, no matter whether models were trained on trillions of tokens of text or with extralinguistic supervision.

Limitations Our computational resources did not permit testing models larger than 14B. It is plausible that state-of-the-art LLMs would climb closer to the theoretical ceiling predicted by Theorems 1 and 2. In addition, context provides a backchannel to infer meaning. An extension of this work can consider the extent to which adding context aids meaning inference, deriving similar representational limits under enriched contexts. These limits could pertain to, e.g., instruction-tuned models modified with a system prompt. Finally, we only consider meaning representation in a weak sense, i.e., statistical decodability. Future work should analyze meaning representation from a *causal* lens, where decodability is a necessary first step, but not sufficient, to show whether systems trained on forms alone *use* representations of meaning.

References

Jacob Andreas. Language models as agent models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp.

5769–5779, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

Dana Angluin. Inductive inference of formal languages from positive data. *Information and Control*, 45(2):117–135, 1980.

Giosuè Baggio and Elliot Murphy. On the referential capacity of language models: An internalist rejoinder to mandelkern & linzen. *Computational Linguistics*, pp. 1–10, June 2026. ISSN 0891-2017.

Marco Baroni. *On the Proper Role of Linguistically Oriented Deep Net Analysis in Linguistic Theorising*. CRC Press, 2022.

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5185–5198, Online, 2020. Association for Computational Linguistics.

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

Katrien Beuls and Paul Van Eecke. Humans learn language from situated communicative interactions. what about machines? *Computational Linguistics*, 50(4):1277–1311, December 2024.

Gemma Boleda. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6 (Volume 6, 2020):213–234, January 2020. ISSN 2333-9683, 2333-9691.

Gemma Boleda. LLMs as a synthesis between symbolic and distributed approaches to language. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 9365–9379, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7.

Tyler A. Chang and Benjamin K. Bergen. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350, March 2024. ISSN 0891-2017.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Herbert H. Clark and Susan E. Brennan. *Grounding in communication*, pp. 127–149. American Psychological Association, Washington, DC, US, 1991. ISBN 978-1-55798-121-9.

Stanislas Dehaene and Jacques Mehler. Cross-linguistic regularities in the frequency of number words. *Cognition*, 43(1):1–29, January 1992. ISSN 0010-0277.

John Rupert Firth. A synopsis of linguistic theory, 1930–1955. In *Studies in Linguistic Analysis*, pp. 1–32. Blackwell, 1957.

Michael C. Frank and Noah D. Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012.

Gottlob Frege. Uber sinn und bedeutung. *Zeitschrift für Philosophie Und Philosophische Kritik*, 100 (1):25–50, 1892.

Richard Futrell and Kyle Mahowald. How linguistics learned to stop worrying and love the language models. *Behavioral and Brain Sciences*, 49:e198, January 2026. ISSN 0140-525X, 1469-1825.

E. Mark Gold. Language identification in the limit. *Information and Control*, 10(5):447–474, 1967.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya

Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran

Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024.

H. Paul Grice. Logic and conversation. In Donald Davidson (ed.), *The logic of grammar*, pp. 64–75. Dickenson Pub. Co., 1975.

Herbert Paul Grice. Meaning. *Philosophical Review*, 66(3):377–388, 1957.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. Olmo: Accelerating the science of language models, 2024.

Eleonora Gualdoni and Gemma Boleda. Why do objects have many names? a study on word informativeness in language use and lexical systems. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18150–18163, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

Peter Gärdenfors. *Conceptual Spaces: The Geometry of Thought*. The MIT Press, March 2000. ISBN 978-0-262-27355-8.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1–3): 335–346, 1990.

Zellig S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

Geoffrey E. Hinton. Learning distributed representations of concepts. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, pp. 1–12. 1986.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4194–4213, Toronto, Canada, July 2023. Association for Computational Linguistics.

Charles Kemp and Terry Regier. Kinship categories across languages reflect general communicative principles. *Science*, 336(6084):1049–1054, 2012. ISSN 1095-9203.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

David Lewis. *Convention: A Philosophical Study*. Harvard University Press, Cambridge, MA, 1969.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8679–8696, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Matthew Mandelkern and Tal Linzen. Do language models' words refer? *Computational Linguistics*, 50(3):1191–1200, September 2024.

William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. Provable limitations of acquiring meaning from ungrounded form: What will future language models understand? *Transactions of the Association for Computational Linguistics*, 9:1047–1060, September 2021. ISSN 2307-387X.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Will Monroe, Robert X.D. Hawkins, Noah D. Goodman, and Christopher Potts. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338, 2017.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Steven T. Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130, October 2014. ISSN 1069-9384.

Steven T. Piantadosi. Modern language models refute Chomsky's approach to language. *From Fieldwork to Linguistic Theory*, pp. 353–414, July 2024.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 20827–20905. Curran Associates, Inc., 2023.

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

Dan Sperber and Deirdre Wilson. *Relevance: Communication and Cognition*. Blackwell, Oxford, 1986/1995.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12075–12097, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, Shuming Shi, T. Zhang, Yvette Graham, and Qun Liu. Translating pro-drop languages with reconstruction models. In *AAAI Conference on Artificial Intelligence*, 2018.

Ludwig Wittgenstein. *Philosophical Investigations*. Wiley-Blackwell, New York, NY, USA, 1953.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025.

G. Yule. *Pragmatics*. Oxford Introduction to Language Study ELT. OUP Oxford, 1996. ISBN 978-0-19-437207-7.

CONTENTS

A Generative AI Disclosure **16**

B Task setup **16**

B.1 Artificial languages: discrete meaning 16

B.2 Artificial languages: continuous meaning 17

B.3 Mandarin pro-drop dataset (Wang et al., 2018) 18

B.4 Color Naming (Monroe et al., 2017) 19

C Training details **20**

D Additional artificial language results (continuous) **21**

E Additional color-naming results **22**

E.1 Other values of ϵ 22

E.2 Tightness of Theorem 2 22

F Proof of Theorem 1 **24**

G Extension of Theorem 1 to countably infinite meanings **24**

H Proof of Theorem 2 **25**

I Extension of Theorem 2 to unbounded meanings **26**

A GENERATIVE AI DISCLOSURE

Claude Code was used to clean the Monroe et al. (2017) colors dataset and to automate parts of the coding pipeline, especially the data pre-processing. Claude Code was used to polish the writing in §2 as well as perform mechanical tasks in the LaTeX, e.g., ensuring uniformity in the bibliography formatting and strict usage of the LaTeXmacros. Everything else is human-written. Claude Code also checked the proofs of both theorems. The authors take responsibility for all the content in this article.

B TASK SETUP

B.1 ARTIFICIAL LANGUAGES: DISCRETE MEANING

Our goal is to construct a family of distributions $\mathbb{P}(M, U, C)$ such that they (1) are a valid language ($\mathrm{H}(M \mid U, C) = 0$) and (2) we can systematically control how informative the utterance is about the meaning $\mathrm{I}(M; U)$. For simplicity, we set context C to follow a uniform categorical distribution over $|\mathcal{C}| = 10$ categories. Similarly, meanings M are drawn from a (separate) uniform categorical distribution over $|\mathcal{M}| = 10$ categories. We set the utterances U to uniformly vary over $|\mathcal{U}| = 15$ categories; the goal is then to design the joint distribution over M, U, and C such that our constraints are satisfied.

We now design the mappings between M, U, and C. Let $p := \mathrm{I}(M; U)/\log_2 |\mathcal{M}|$ be the target *fraction* of maximum information. We split the utterance alphabet $\mathcal{U}$ into two disjoint families, corresponding to "informative" and "context-dependent" regimes. We let the probability that we are in the informative regime be realized by p, described below:

Informative utterances ($u \in \{0, \ldots, |\mathcal{M}| - 1\}$). We set aside $|\mathcal{M}|$ utterances that are context-*independent*. These utterances correspond to the informative regime, where given u, the meaning is $M = u$ deterministically and independently of c. Then, for every context c,

$$\mathbb{P}(C = c,\ M = u,\ U = u) \ = \ \tfrac{1}{|\mathcal{C}|} \ \cdot \ \tfrac{p}{|\mathcal{M}|}.$$

The expression on the RHS comes from multiplying $\mathbb{P}(C = c)$ with the probability p that we are in the informative regime and the probability $1/|\mathcal{M}|$ that $U = u$ in the informative regime.

Context-dependent utterances ($u \in \{|\mathcal{M}|, \ldots, |\mathcal{U}| - 1\}$). The remaining utterances are context-*dependent*. Each context-dependent utterance carries 0 bits about M unconditionally but recovers M perfectly when combined with c. That is, given u alone, the meaning is *uniform*; given u and c, the meaning is recoverable. We construct the mapping from u, c to m by setting up a deterministic cyclic shift with c acting as a "key":

$$\mathbb{P}(C = c,\ M = (u - |\mathcal{M}| + c) \bmod |\mathcal{M}|,\ U = u) \ = \ \tfrac{1}{|\mathcal{C}|} \cdot \tfrac{1-p}{|\mathcal{U}| - |\mathcal{M}|}.$$

For example, consider that $\mathcal{U} = \{1, 2, \cdots, 15\}$ and $\mathcal{M} = \{1, 2, \cdots, 10\}$. If $u = 11$, and $c = 2$, then $m = 3$ deterministically. The expression on the RHS is given by $\mathbb{P}(C = c)$ multiplied by the probability $1 - p$ that we're in the context-dependent regime and the probability that $U = u$ in that regime.

By construction, each utterance-in-context pair (u, c) maps to a single meaning ($\mathrm{H}(M \mid U, C) = 0$). Also by construction, the marginal $\mathrm{I}(M; U)$ evaluates to exactly $p \cdot \log_2 |\mathcal{M}|$. Sweeping p over six evenly-spaced values in $[0, 1]$ yields the six target information levels shown on the x-axis of Figure 2.

Finally, to construct train and test sets, for each target mutual information, we draw $N = 10{,}000$ training examples and $N = 5000$ disjoint sets from the joint distribution $\mathbb{P}(M, U, C)$.

Validating Theorem 1. Now, we describe how to verify the bound in Theorem 1. We realize the map $f \circ g : U \mapsto \widehat{M}$ by a small two layer MLP (see Appendix C for hyperparameter and training details).[4] First, we split the data into 80/20 train-test splits. After training this decoder to

[4]NB: to avoid confusion, since we are not using LLMs in this setting, we directly use an MLP for $f \circ g$ (in the main text, g was defined as a featurizer and f as a decoder). In the experiments on natural language, we use LLM activations for the featurizer g and an MLP for the decoder f.

predict ground-truth meanings given utterance, we verify that the best-case empirical test accuracy across hyperparameter settings and 5 random seeds (LHS of Theorem 1) is lower than the theoretical upper-bound (RHS).

To estimate the RHS, we solve for the theoretical best performance p_e satisfying the bound, then verify that the empirical performance is lower. Because the joint distribution $\mathbb{P}(M, U, C)$ was constructed analytically, we directly compute RHS terms $\mathrm{H}(M \mid U)$ and $\log_2(|\mathcal{M}| - 1)$. We then find the maximum p_e satisfying the upper bound by numerical approximation, using an off-the-shelf root-finding algorithm (`scipy.optimize.brentq`). Finally, we compare this upper bound against the empirical test accuracy for each artificial language.

B.2 Artificial languages: continuous meaning

Our goal is to verify the ϵ-accurate regression bound (Theorem 2) in a continuous meaning space where we can *tune* the mutual information $\mathrm{I}(J_\epsilon; U)$ between the packing-bin label J_ϵ and the observed utterance U across the interval $[0,\ \mathrm{H}(J_\epsilon)]$.

We first construct J_ϵ given ϵ. Fix $\epsilon > 0$. We first construct the 2ϵ-packing by randomly sampling $N = 20{,}000$ i.i.d. points, and greedily selecting packing points in the order of that sample. We repeat this process 100 times and retain the maximal 2ϵ-packing. We then map meanings m to their nearest packing point via the quantization function q (described in §3). For $\epsilon \in \{0.2, 0.4, 0.6\}$ we obtain $N(\epsilon) := |J_\epsilon| \in \{25, 9, 5\}$, and $\mathrm{H}(J_\epsilon) = \log_2 |J_\epsilon| \in \{4.64, 3.17, 2.32\}$ bits (see below for why $\mathrm{H}(J_\epsilon)$ equals $\log_2 |J_\epsilon|$ exactly in this setup).

Constructing the meaning, utterance, and context distributions. We start by defining the meanings' marginal distribution, and work with a $N = 20{,}000$ sample. In short, we set this distribution to be a noisy version of *the packing*: we sample $J_i \in \{0, \ldots, N(\epsilon) - 1\}$ uniformly and set

$$M_i = m_{J_i} + \eta_i, \qquad \eta_i \sim \mathrm{Unif}\big(B_2(0, \delta)\big), \tag{2}$$

where $B_2(0, \delta)$ is the 2-D ball of radius $\delta = 0.1$. The per-bin count is rounded so that the total N is divisible by both $|J_\epsilon|$ and the context alphabet size $|\mathcal{C}|$; this makes bin frequencies exactly uniform and hence $\mathrm{H}(J_\epsilon) = \log_2 |J_\epsilon|$.

Now, we move on to design the joint $\mathbb{P}(M, U, C)$. We keep contexts categorical for simplicity, fixing $|\mathcal{C}| = 10$. We have $|U| = N/|\mathcal{C}| = 2000$. Every meaning m receives a unique slot $(u, c) \in \{0, \ldots, |\mathcal{U}| - 1\} \times \{0, \ldots, |\mathcal{C}| - 1\}$ from Algorithm 1, i.e., meanings are identified given utterance and context (valid language assumption, $\mathrm{H}(M \mid U, C) = 0$). The overall intuition is very similar to the discrete languages case: context C acts as a "key" that disambiguates the meaning given the utterance. The problem of tuning $\mathrm{I}(J_\epsilon; U)$ now turns into how we distribute utterance, context pairs to meanings—like the last section, we portion out "informative" and "context-dependent" utterances for every meaning.

We tune $\mathrm{I}(J_\epsilon; U)$ via a scalar $p \in [0, 1]$ as follows:

Informative utterances We consider the "bins" of meanings that the quantizer q assigns to each packing point $J_\epsilon = j$. For each J_ϵ-bin j containing n_j meanings, we take $n_{\text{inf}} = \lfloor p\, n_j / |\mathcal{C}| \rfloor \cdot |\mathcal{C}|$ meanings and pack them into *dedicated* "U-groups". Each U-group contains $|\mathcal{C}|$ meanings drawn only from bin j. In these groups, the utterance perfectly identifies J_ϵ alone, i.e., the mutual information with J_ϵ is maximal.

Context-dependent utterances Then, we pool the remaining n_{amb} meanings across all bins and assign them to *shared* U-groups in a round-robin way: after sorting the pool by J_ϵ-bin, meaning at position i is assigned $u = u_{\text{offset}} + (i \bmod n_{\text{amb}}/|\mathcal{C}|)$ and $c = \lfloor i/(n_{\text{amb}}/|\mathcal{C}|) \rfloor$. Then, round-robin spreads each bin uniformly across all shared U-groups, so $\mathbb{P}(J_\epsilon \mid U)$ in a shared U-group equals the marginal $\mathbb{P}(J_\epsilon)$ (up to ± 1-count boundary effects when $|\mathcal{C}|$ does not divide n_j).

By construction the informative contribution to $\mathrm{I}(J_\epsilon; U)$ is exactly $p \times \mathrm{H}(J_\epsilon)$ and the shared contribution is zero. Sweeping $p \in \{0, 1/8, 2/8, \ldots, 1\}$ therefore sweeps $\mathrm{I}(J_\epsilon; U)$ across $[0, \mathrm{H}(J_\epsilon)]$ while keeping the valid language constraint.

Algorithm 1 Build (U, C) from J_ϵ with tunable $\mathrm{I}(J_\epsilon; U) \approx p\,\mathrm{H}(J_\epsilon)$.

Require: Quantizer $q : \mathcal{M} \to \mathbb{N}_{1\ldots|J_\epsilon|}$, J_ϵ; context size $|C|$ where $|\mathcal{C}| \mid N$; parameter $p \in [0, 1]$.

Initialize $U[m] = C[m] = -1$ for all m; $u \leftarrow 0$; $\mathcal{A} \leftarrow \emptyset$.
for all j in $\{0, \ldots, |J_\epsilon| - 1\}$ **do** ▷ Iterate over packing bins
  $\mathrm{bin}_j \leftarrow$ random permutation of $\{m : q(m) = j\}$
  $n_j^{\inf} \leftarrow \lfloor p\,|\mathrm{bin}_j|/|C| \rfloor \cdot |C|$
  for $s = 0, |\mathcal{C}|, 2|\mathcal{C}|, \ldots, n_j^{\inf} - |\mathcal{C}|$ **do**
    for $c = 0, \ldots, |\mathcal{C}| - 1$ **do**
      $U[\mathrm{bin}_j[s + c]] \leftarrow u$; $C[\mathrm{bin}_j[s + c]] \leftarrow c$
    end for
    $u \leftarrow u + 1$
  end for
  Append $\mathrm{bin}_j[n_j^{\inf} :]$ to $\mathcal{A}$
end for
$|\mathcal{A}| = n_{\text{amb}}$; $n_{\text{amb}}^U \leftarrow n_{\text{amb}}/|\mathcal{C}|$
for $i = 0, \ldots, n_{\text{amb}} - 1$ **do** ▷ Round-robin: spreads each bin across all shared U-groups
  $U[\mathcal{A}[i]] \leftarrow u + (i \bmod n_{\text{amb}}^U)$; $C[\mathcal{A}[i]] \leftarrow \lfloor i/n_{\text{amb}}^U \rfloor$
end for
return U, C

Validating Theorem 2. For each artificial language, we validate the bound. To do so, we portion the data into 80/20 train-test splits, and train a two-layer MLP to predict the continuous ground-truth meaning from the utterance, minimizing the MSE loss. The LHS of Theorem 2 is given by the test ϵ-accuracy, i.e., the fraction of test datapoints predicted within-ϵ of the ground truth. For training and hyperparameter details, see Appendix C.

To estimate the upper bound on ϵ-accurate recovery, we first estimate the maximum $\mathbb{P}(B)$ that satisfies Theorem 2, where, recall, $\mathbb{P}(B)$ is the likelihood of correct classification over packing points, and $\mathbb{P}(B) \geq$ Theorem 2 LHS by construction. This entails estimating $\mathrm{H}(J_\epsilon)$ via the plug-in entropy of the empirical bin frequencies of the packing. We estimate $\mathrm{I}(J_\epsilon; U)$ using the plug-in mutual information estimator. Finally, we numerically solve for the upper bound and verify that the empirical test ϵ-accuracy falls below this bound.

B.3 Mandarin pro-drop dataset (Wang et al., 2018)

Meaning, context, and utterance distribution. We repurpose a dataset of $N = 2{,}150{,}945$ short Mandarin TV subtitles from Wang et al. (2018). We start by filtering the dataset to utterances containing exactly one (explicit or dropped) pronoun, after which there are $N = 904{,}996$ utterances. We then construct a dataset of (utterance, true pronoun) pairs (the true pronouns are already provided by Wang et al. (2018)). The meaning inference pipeline (consisting of frozen LLM representation + small MLP decoder) is trained to predict the true pronoun from the utterance.

We considered 12 personal pronouns in Mandarin Chinese. Singular pronouns include "我" ('I'), "你" ('you'), "您" ('you', honorific), 他 ('he'), 她 ('she'), 它 ('it'), along with their plural forms (singular form +们). We normalized each possessive pronoun (base pronoun + 的) to their base singular or plural form. The more colloquial inclusive first-person plural 咱们 ('we', which explicitly includes the interlocutor), was also present in the dataset. The distribution of true pronouns is skewed towards the first-person singular (37.6%), see Figure 3; in contrast, there were no formal second-person plurals (您们) in the dataset, which tend to be rare. This yielded 12 final pronouns.

Validating Theorem 1. For each speaker utterance, we extract its frozen LLM embedding $g : \mathcal{U} \to \mathbb{R}^d; u \mapsto z$. Then, we learn a map $f : \mathbb{R}^d \to \Delta^{11}$ from the LLM embedding to a probability distribution over the 12 possible pronouns. The decoder f is an MLP of up to two layers trained to minimize the cross-entropy loss between the ground truth pronoun and the inferred one. At test time, the inferred pronoun is taken to be the most likely class under f's probability distribution. We include a strong **baseline** that directly selects the pronoun in the sentence if expressed and guesses the most frequent one 我 ('I', first-person) if dropped. Because only 20.3% of the dataset showed a dropped

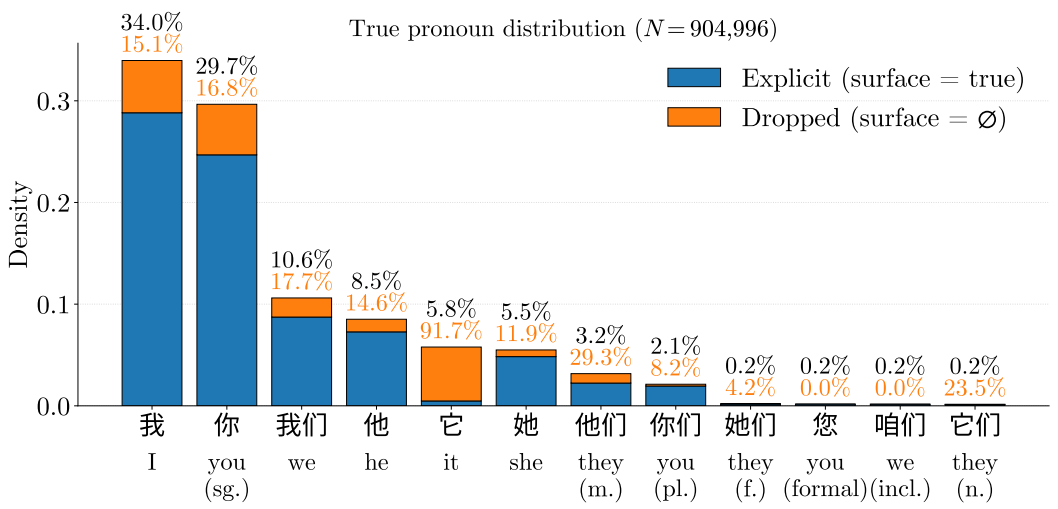


Figure 3: **Mandarin pronouns distribution preprocessed from Wang et al. (2018).** Percentage annotations (black) are each pronoun's percentage of the total. The orange percentage annotations are the percent of each pronoun that is dropped (corresponding to the orange segment in the bars). For example, 我 ('I') is 34% of all true pronouns; it is dropped 15.1% of the time.

pronoun (i.e., 79.7% of the pronouns are explicit), we expect baseline performance to achieve $\approx 80\%$ accuracy. For each model, we estimate Theorem 1's LHS by taking the empirical test accuracy, and verifying that it is lower than the theoretical upper bound.

We estimate the theoretical upper bound (RHS of Theorem 1) using the raw meanings and utterances data. Finally, like in previous tasks, we numerically solve for the largest p_e that satisfies the bound and ensure that this p_e is greater than the best empirical test accuracies across hyperparameters and random seeds.

B.4 Color Naming (Monroe et al., 2017)

Meaning, context, and utterance distribution. We filter the dataset for communicative successes (see Appendix B), such that our valid language requirement $\mathrm{H}(M \mid U, C) = 0$ is met. This cleaned version contains $N = 38975$ rounds of the task from different speaker-listener pairs. Each round contains the RGB codes for the three color chips (*context*), the target chip (*intended meaning*), speaker message in natural language (*utterance*), and listener guess (*communicative success*).

We preprocess the color chips by converting RGB to the three-dimensional CIELAB space, commonly used to model human color perception (Gualdoni & Boleda, 2024) where Euclidean distances correlate to human perceptual distances; we use this CIELAB space to proxy the *meaning space* $\mathcal{M}$, with the metric d given by the Euclidean distance. Then, each color chip, i.e., each meaning m, is represented as a vector in $\mathbb{R}^3$, which we further normalized to $[0, 1]^3$. Note that meanings here live in a bounded metric space as required by Theorem 2.

Validating Theorem 2. For regression, the LHS of Theorem 2 is the empirical fraction of meanings that were predicted within-ϵ. ϵ is a tolerance parameter set by the practitioner; here, we test values of ϵ ranging in $[0.01, 1]$. Similar to the zero-pronoun resolution task, we compute the empirical ϵ-accuracy by using LLM representations of the speaker utterance to infer the target chip color. After analogously extracting the LLM representations of the utterances, we learn an MLP $f : \mathbb{R}^d \to \mathbb{R}^3; z \mapsto \hat{m}$ from the LLM activations to meaning space. The map f is then trained to minimize MSE loss between the ground truth meaning m and the inferred meaning $\hat{m} = f(g(u_n))$. Finally, given error tolerance ϵ, we report the fraction of test samples predicted within-ϵ of the target.

In contrast to the zero-pronoun resolution task, the set of possible referring expressions is unconstrained in the color naming task. Because the number of possible referring expressions can grow

combinatorially with respect to utterance length, the dataset necessarily *undersamples* the utterance space; this issue means that we should not directly estimate $\mathrm{H}(J_\epsilon \mid U)$ from raw utterances, as we did in the previous section. Instead, we set up an auxiliary classification task that predicts J_ϵ from a map $h : \mathcal{U} \to \Delta^{N(\epsilon)-1}$ that maps U to a probability distribution over J_ϵ. Then, we can estimate $\mathrm{H}(J_\epsilon \mid U)$ using $\mathrm{H}(J_\epsilon \mid h(U))$, plugging the latter into the RHS of Theorem 2.

To form h, we further add a classification head on top of the LLM feature vector; see Appendix C for details. On the concatenated train and test sets, we then estimate $\mathrm{H}(J_\epsilon \mid h(U))$ by taking the average cross-entropy loss of the classifier h. In using this strategy to estimate $\mathrm{H}(J_\epsilon \mid h(U))$, we exploit the fact that LLM features map strings to a more "amenable" space that encodes semantic similarities between strings.

C Training details

MLP hyperparameter search. For both **discrete** and **continuous** cases, we implement the decoder as an embedding matrix followed by a two-layer MLP, with fixed hidden dimension throughout the layers. The MLP was trained on the following hyperparameter grid:

- Learning rate $\in \{10^{-4}, 10^{-3}, 10^{-2}\}$
- Hidden width $\in \{64, 128, 256\}$
- Weight decay $\in \{0, 10^{-4}, 10^{-3}\}$.

In the **discrete** case, we trained the MLP for a maximum of 20 epochs with cross-entropy loss, batch size 256, and Adam. In the **continuous** case, for each (ϵ, p) we train the MLP to minimize the MSE loss with Adam, for max 30 epochs. The train/test split is a fixed 80/20 partition of the sampled meanings. In all cases, we report the best test loss over *all epochs*, hyperparameters, and five random seeds.

LLM hyperparameter search. For every (model, layer) pair, we trained nine MLP variants that vary one hyperparameter at a time from a shared reference point (one hidden layer of 256 units, learning rate 10^{-3}, weight decay 10^{-4}):

- Learning rate $\in \{10^{-4}, 10^{-3}, 10^{-2}\}$
- Hidden width $\in \{128, 256, 512\}$
- Weight decay $\in \{0, 10^{-4}, 10^{-3}\}$
- MLP depth $\in \{$linear, 1-hidden, 2-hidden$\}$

For every configuration the probe is trained independently per layer. We used the Adam optimizer with early stopping (patience 8, up to 100 epochs), and mini-batches of 256.

Meaning inference pipeline for natural language experiments. We learn the MLP decoder $f : \mathcal{Z} \to \mathcal{M}$ that maps model activations back to the original meaning space, i.e., inferring the meaning $\widehat{M}$ of the utterance U. In each experiment, we split the data, consisting of (utterance, meaning) pairs, into 80/20 train and test splits. As Theorems 1 and 2 are statements about the *best-case* performance, for each LLM, in training the decoder f, we only report the best test meaning inference performance over all LLM layers, training steps, and hyperparameters.

Zero-pronoun resolution. We drew a 50,000-sentence subsample at random from the total dataset and split it 80/20 into train and test sets (the partition was identical for every model). The original utterances contained spaces between characters, which were removed before tokenization in a preprocessing step.

For each frozen model, we extracted the last-token hidden state at every layer. We then trained MLP probes with cross-entropy loss on the true pronoun class, sweeping the above hyperparameters, and evaluated test accuracy on the held-out set.

D ADDITIONAL ARTIFICIAL LANGUAGE RESULTS (CONTINUOUS)

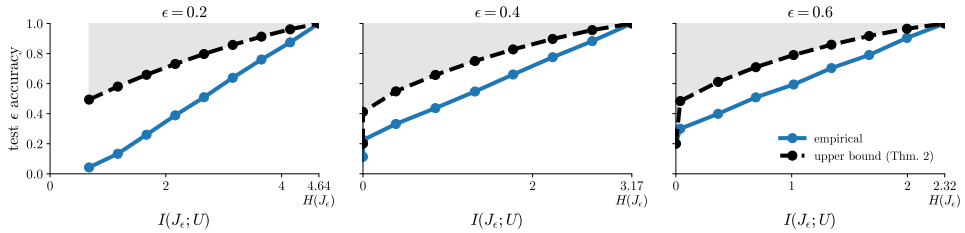


Figure 4: **Artificial languages over continuous meanings; other values of ϵ.** We show that the Theorem 2 bound (black line) holds for various settings of $\epsilon \in \{0.2, 0.4, 0.6\}$.

E ADDITIONAL COLOR-NAMING RESULTS

E.1 OTHER VALUES OF ϵ

Figure 6 shows the Theorem 2 in practice for other values of $\epsilon \in \{0.75, 0.5, 0.3, 0.25, 0.15, 0.125, 0.1, 0.05, 0.01\}$ on the color-naming experiment.

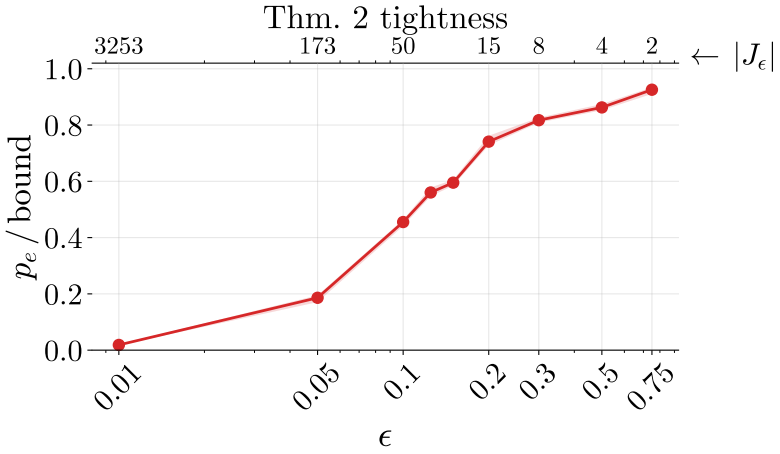


Figure 5: **Theorem 2 tightness, color-naming.** As the parameter ϵ (x-axis), which controls the granularity of the quantization, increases, the tightness of the bound in Theorem 2 (y-axis) improves. The number of bins $|J_\epsilon|$ in the quantization implied by ϵ is shown at the top x-axis; at the right extreme case $\epsilon = 0.75$, the ϵ-accurate regression task reduces to binary classification ($|J_\epsilon| = 2$). At the left extreme, the task is $|J_\epsilon| = 3253$-way classification. We observe that Theorem 2 becomes "useful" in an intermediate regime, $\epsilon \approx 0.1$.

E.2 TIGHTNESS OF THEOREM 2

Theorem 2 holds empirically for each ϵ we tested. However, in Figure 5 one can notice the bound is looser with smaller ϵ, and especially for values of ϵ less than 0.1 ($\epsilon = 0.05, 0.01$). We identify the source of this phenomenon as an increase in $\mathrm{I}(J_\epsilon; U)$: qualitatively, as ϵ decreases, the number of bins in the quantization increases, and a single utterance form corresponds to more bins—this results in inflated $\mathrm{I}(J_\epsilon; U)$.

Qualitatively, *coarser* quantization (larger ϵ) yields a better bound, while *finer* quantization (smaller ϵ) yields a looser bound. The loosening of the bound with ϵ can be explained by a straightforward reason: utterance forms tend to *coarsely* tile the color space, and so the bounds reflect the level of inherent ambiguity in (utterance, meaning "bin") pairs in the quantized meaning space. To provide an intuition: utterances like "purple" may refer to an area of colors. If we decrease ϵ to smaller than the radius of this area, then "purple" is split into many bins. At test time, the classifier $h : \mathcal{U} \to J_\epsilon$ deterministically maps "purple" to a single one of those bins. Because h is deterministically mapping "purple" (U) to the same bin (J_ϵ), we have an increase in the mutual information $\mathrm{I}(J_\epsilon; U)$ between utterances and bin labels relative to $\mathrm{H}(J_\epsilon)$. This results from how language coarsely parcels the meaning space.

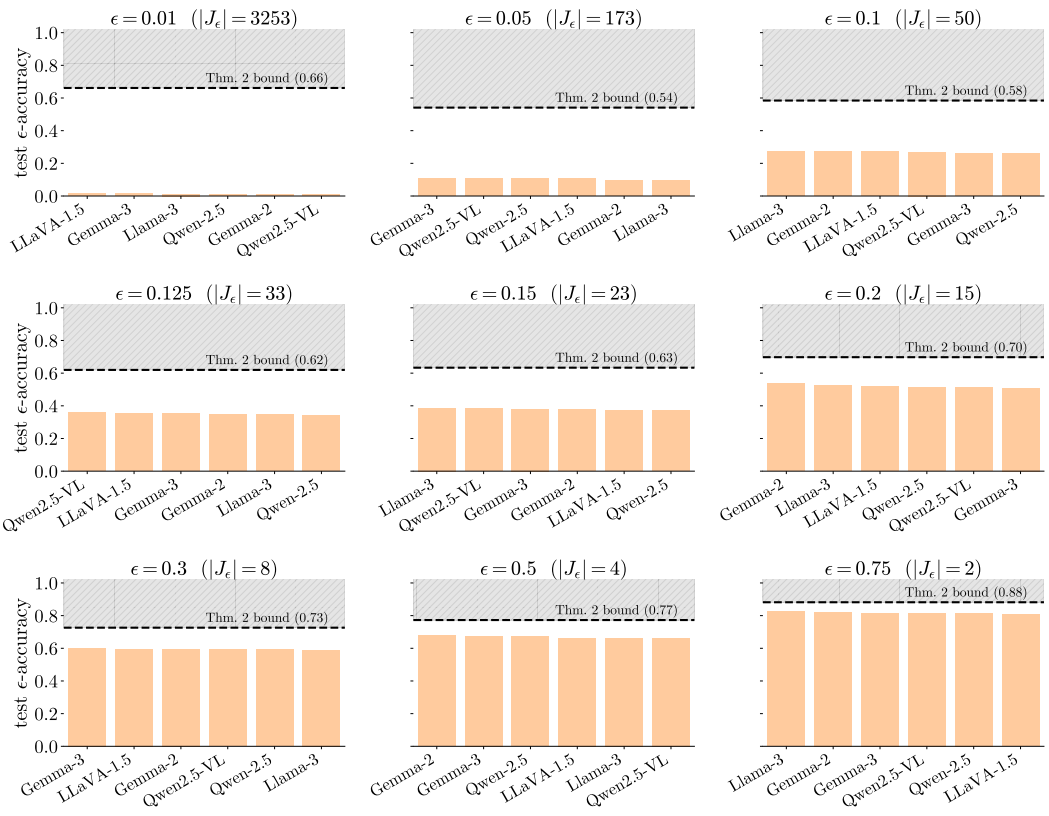


Figure 6: **Color naming, Theorem 2 for all values of ϵ.** Analogue to Figure 2 (left) for each value of ϵ. Bounds hold for all values of ϵ and typically tighten with larger ϵ; for reference, the number of bins $|J_\epsilon|$ in the quantized meaning variable J_ϵ is also shown.

F Proof of Theorem 1

Theorem 1 (Classification error). *Let $g\colon \mathcal{U} \to \mathcal{Z}$ be a map from utterances ($\mathcal{U}$) to representations ($\mathcal{Z}$). Abbreviate the probability that f is correct as $p_e := \mathbb{P}[f(g(U)) = M]$. Then,*

$$\sup_{f \in \mathcal{M}^{\mathcal{Z}}} p_e \leq 1 - \frac{\mathrm{H}(M \mid U) - \mathrm{H}_2(p_e)}{\log(|\mathcal{M}| - 1)} = 1 - \frac{\mathrm{H}(M \mid U, C) + \overbrace{\mathrm{I}(M; C \mid U)}^{\text{meaning from context only}} - \mathrm{H}_2(p_e)}{\log(|\mathcal{M}| - 1)}.$$

Proof. Consider the Markov chain $M \to Z \to f(Z)$, where $f(Z)$ is the decoder's prediction of the original speaker meaning from a representation of text, and p_e is the probability of prediction success. Let $p_i = 1 - p_e$ be the probability the prediction is *incorrect*. By Fano's inequality,

$$\mathrm{H}(M \mid Z) \leq \mathrm{H}_2(p_i) + p_i \log(|\mathcal{M}| - 1)$$

Rearranging, we have

$$\begin{aligned} p_i &\geq \frac{\mathrm{H}(M \mid Z) - \mathrm{H}_2(p_i)}{\log(|\mathcal{M}| - 1)} \\ &\geq \frac{\mathrm{H}(M \mid U) - \mathrm{H}_2(p_i)}{\log(|\mathcal{M}| - 1)} && (\mathrm{H}(M \mid Z) \geq \mathrm{H}(M \mid U)) \\ &= \frac{\mathrm{H}(M \mid U, C) + \mathrm{I}(M; C \mid U) - \mathrm{H}_2(p_i)}{\log(|\mathcal{M}| - 1)} && \text{(Equation (1))} \end{aligned}$$

Theorem 1 follows from substituting p_i with $1 - p_e$ in the above derivation and noting that $\mathrm{H}(M \mid U) = \mathrm{H}(M \mid U, C) + \mathrm{I}(M; U, C)$. Due to the bounded meanings assumption $|\mathcal{M}| < \infty$, the final bound is non-vacuous. ■

G Extension of Theorem 1 to countably infinite meanings

We extend Theorem 1 to countably infinite meaning spaces. First, we show that, for any valid probability distribution $\mathbb{P}(M)$ on a countable set $\mathcal{M}$, we can reasonably approximate its probability mass using a *finite* subset $\mathcal{M}'$ of the support $\mathcal{M}$:

Lemma 1 (Finite approximation.). *Let $\mathcal{M}$ be countable. For all $\delta > 0$, there exists a* finite *subset $\mathcal{M}' \subseteq \mathcal{M}$ such that $\mathbb{P}(M \in \mathcal{M}') \geq 1 - \delta$.*

Proof. Let p_i be $\mathbb{P}(M = m_i)$, for $m_i \in \mathcal{M}$. Since $\mathbb{P}_M$ is a valid probability distribution, we have $\sum_{i=1}^{\infty} p_i = 1$. By convergence of the series, $\lim_{K \to \infty} \sum_{i=1}^{K} p_i = 1$. Thus, for all $\delta > 0$, there exists finite K such that $\sum_{i=1}^{K} p_i \geq 1 - \delta$. Taking $\mathcal{M}' = \{m_1, \cdots, m_K\}$ yields the lemma. ■

We have shown that $\mathbb{P}(M)$, where M is possibly countably infinite, can be approximated using a finite subset of the support. We now use this fact to bound the classification error on countably infinite meaning spaces, by expressing it in terms of classification error on this finite subset.

Corollary 1.1 (Classification error, countably infinite case). *Let $g\colon \mathcal{U} \to \mathcal{Z}$ be a map from utterances ($\mathcal{U}$) to representations ($\mathcal{Z}$). Abbreviate the probability that f is correct as $p_e := \mathbb{P}[f(g(U)) = M]$. Then,*

$$\sup_{f \in \mathcal{M}^{\mathcal{Z}}} p_e \leq 1 - (1 - \delta)\frac{\mathrm{H}(M \mid U, M \in \mathcal{M}') - 1}{\log(|\mathcal{M}'| - 1)}.$$

Proof. We split $\mathcal{M}$ into $\mathcal{M}'$, with probability mass $q := \mathbb{P}(M \in \mathcal{M}') \geq 1 - \delta$, and its complement $\mathcal{M}'^{C} = \mathcal{M} \backslash \mathcal{M}'$ with mass $1 - q = \mathbb{P}(M \in \mathcal{M}'^{C}) \leq \delta$. Let p'_e be the probability of classification accuracy on $\mathcal{M}'$ by some decoder $f \in \mathcal{M}^{\mathcal{Z}}$, and let p'^{C}_e, similarly, be the probability of classification

accuracy on $\mathcal{M}'^C$. We have

$$\begin{aligned} p_e &= qp'_e + (1-q)p'^C_e \\ &\leq qp'_e + 1 - q && (p'^C_e \leq 1) \\ &= 1 - q(1-p'_e) \\ &\leq 1 - (1-\delta)(1-p'_e) && (q \geq 1-\delta) \\ &= (1-\delta)p'_e + \delta. \end{aligned}$$

Then,

$$\sup_{f \in \mathcal{M}^{\mathcal{Z}}} p_e \leq (1-\delta) \sup_{f \in \mathcal{M}^{\mathcal{Z}}} p'_e + \delta. \tag{3}$$

We can bound $\sup_{f \in \mathcal{M}^{\mathcal{Z}}} p'_e$ in the RHS of equation 3 using Theorem 1:

$$\sup_{f \in \mathcal{M}^{\mathcal{Z}}} p'_e \leq 1 - \frac{\mathrm{H}(M \mid U, M \in \mathcal{M}') - \mathrm{H}_2(p'_e)}{\log(|\mathcal{M}'| - 1)} \leq 1 - \frac{\mathrm{H}(M \mid U, M \in \mathcal{M}') - 1}{\log(|\mathcal{M}'| - 1)}. \tag{4}$$

Finally, bounding equation 3 by the RHS of equation 4 yields

$$\sup_{f \in \mathcal{M}^{\mathcal{Z}}} p_e \leq 1 - (1-\delta)\frac{\mathrm{H}(M \mid U, M \in \mathcal{M}') - 1}{\log(|\mathcal{M}'| - 1)}.$$

■

H Proof of Theorem 2

Theorem 2 (ϵ-accurate regression)**.** *Let $g : \mathcal{U} \to \mathcal{Z}$ be a map from utterances ($\mathcal{U}$) to representations ($\mathcal{Z}$). Let $(\mathcal{M}, d)$ be a metric space and $N(\epsilon)$ be its maximal 2ϵ-packing number. Assume $N(\epsilon) < \infty$ for $\epsilon > 0$. Abbreviate the probability that f is ϵ-accurate as $p_e^\epsilon = \mathbb{P}[d(f(g(U)), M) < \epsilon]$. Then,*

$$\sup_{f \in \mathcal{M}^{\mathcal{Z}}} p_e^\epsilon \leq 1 - \frac{\mathrm{H}(J_\epsilon \mid U) - \mathrm{H}_2(B)}{\log(N(\epsilon) - 1)} = 1 - \frac{\overbrace{\mathrm{H}(J_\epsilon \mid U, C)}^{\textit{irreducible ambiguity}} + \overbrace{\mathrm{I}(J_\epsilon; C \mid U)}^{\textit{meaning from context only}} - \mathrm{H}_2(B)}{\log(N(\epsilon) - 1)}. \tag{5}$$

Proof. Let $\widehat{M} = f(Z)$. Let $\{m_n\}_{n=1}^{N(\epsilon)}$ be the maximal 2ϵ-packing of $\mathcal{M}$, meaning $d(m_i, m_j) \geq 2\epsilon$ for all $i \neq j \in \{1 \dots N(\epsilon)\}$. Recall that $J_\epsilon = q(M)$ is a discrete random variable over the $N(\epsilon)$ packing points $\{m_n\}_{n=1}^{N(\epsilon)}$, formed by mapping points $m \in \mathcal{M}$ to their nearest packing point in $\{m_n\}_{n=1}^{N(\epsilon)}$. We construct an auxiliary classification problem where the goal is to estimate $\widehat{J_\epsilon} = J_\epsilon$ given U, the utterances associated with the packing points.

Let the event A be the event that ϵ-accurate recovery is satisfied; i.e., $A = \mathbb{1}[d(\widehat{M}, M) < \epsilon]$. Let B be the event that the auxiliary classification is successful, i.e., $B = \mathbb{1}[\widehat{J_\epsilon} = J_\epsilon]$. By definition of 2ϵ-packing, if $d(\widehat{M}, M) < \epsilon$, then setting $\widehat{J_\epsilon}$ as the closest packing point to $f(Z)$ under the metric d yields $\widehat{J_\epsilon} = J_\epsilon$. In other words, the event $A \Rightarrow B$, and by extension,

$$\mathbb{P}(A) \leq \mathbb{P}(B). \tag{6}$$

Note that the auxiliary classification problem, i.e., the setting of event B, reduces to the classification setting in Theorem 1. We now upper-bound $\mathbb{P}(B)$ using Fano's inequality in the same way as in Theorem 1.

This yields

$$\mathbb{P}(B) \leq 1 - \frac{\mathrm{H}(J_\epsilon \mid Z) - \mathrm{H}_2(B)}{\log(N(\epsilon) - 1)}. \tag{7}$$

Since we assume $N(\epsilon) < \infty$, the bound is non-vacuous.

Because Z is a deterministic function of U, $\mathrm{H}(J \mid Z) \geq \mathrm{H}(J \mid U)$. This yields the bound

$$\mathbb{P}(B) \leq 1 - \frac{\mathrm{H}(J_\epsilon \mid U) - \mathrm{H}_2(B)}{\log(N(\epsilon) - 1)}, \tag{8}$$

which, alongside Equation (6), yields

$$\mathbb{P}(A) \leq 1 - \frac{\mathrm{H}(J_\epsilon \mid U) - \mathrm{H}_2(B)}{\log(N(\epsilon) - 1)}. \tag{9}$$

Finally, using

$$\mathrm{H}(J_\epsilon \mid U) = \mathrm{H}(J_\epsilon \mid U, C) + \mathrm{I}(J_\epsilon; C \mid U)$$

yields the final bound. ■

I Extension of Theorem 2 to unbounded meanings

We assume that the distribution $\mathbb{P}(M)$ over meanings M is *tight* (§3), that is, for any $\delta > 0$, there exists a compact (closed and bounded) subset $\mathcal{M}' \subset \mathcal{M}$ such that $\mathbb{P}(M \in \mathcal{M}') \geq 1 - \delta$. Let the 2ϵ packing number of $\mathcal{M}'$ be $N'(\epsilon)$, and construct J_ϵ on $\mathcal{M}'$ using a quantizer function as in Theorem 2. Then,

Corollary 2.1 (ϵ-accurate recovery, unbounded case)**.** *Let $g\colon \mathcal{U} \to \mathcal{Z}$ be a map from utterances ($\mathcal{U}$) to representations ($\mathcal{Z}$). Let $(\mathcal{M}, d)$ be a metric space and $N(\epsilon)$ be its maximal 2ϵ-packing number. Abbreviate the probability that f is ϵ-accurate as $p_e^\epsilon = \mathbb{P}[d(f(g(U)), M) < \epsilon]$. Then,*

$$\sup_{f \in \mathcal{M}^{\mathcal{Z}}} p_e^\epsilon \leq 1 - (1 - \delta)\frac{\mathrm{H}(J_\epsilon \mid U, M \in \mathcal{M}') - 1}{\log(N'(\epsilon) - 1)}.$$

Proof. Follows from (1) bounded $\mathcal{M}'$ has a finite packing number $N'(\epsilon)$, (2) the proof of Corollary 2.1, and (3) the proof of Theorem 2. ■