What Can String Probability Tell Us About Grammaticality?

Jennifer Hu[*]
Department of Cognitive Science
Johns Hopkins University, USA
jennhu@jhu.edu

Ethan Gotlieb Wilcox
Department of Linguistics
Georgetown University, USA
ethan.wilcox@georgetown.edu

Siyuan Song
Department of Linguistics
The University of Texas at
Austin, USA
siyuansong@utexas.edu

Kyle Mahowald
Department of Linguistics
The University of Texas at
Austin, USA
kyle@utexas.edu

Roger P. Levy
Department of Brain and
Cognitive Sciences
Massachusetts Institute of
Technology, USA
rplevy@mit.edu

Abstract

What have language models (LMs) learned about grammar? This question remains hotly debated, with major ramifications for linguistic theory. However, since probability and grammaticality are distinct notions in linguistics, it is not obvious what string probabilities can reveal about an LM's underlying grammatical knowledge. We present a theoretical analysis of the relationship between grammar, meaning, and string probability, based on simple assumptions about the generative process of corpus data. Our framework makes three predictions, which we validate empirically using 280K sentence pairs in English and Chinese: (1) correlation between the probability of strings within minimal pairs, i.e., string pairs with minimal semantic differences; (2) correlation between models' and humans' deltas within minimal pairs; and (3) poor separation in probability space between unpaired grammatical and ungrammatical strings. Our analyses give theoretical grounding for using probability to learn about LMs' structural knowledge, and suggest directions for future work in LM grammatical evaluation.

1 Introduction

Understanding what probabilistic language models (LMs) can learn about grammar has major ramifications for theories of language learning and structure (Linzen, 2019; Warstadt and Bowman, 2022; Baroni, 2022; Piantadosi, 2024). In the past decade, there have been many efforts to evaluate LMs' grammatical knowledge (e.g., Warstadt et al., 2020; Hu et al., 2020; Linzen et al., 2016; Tjuatja et al., 2025), with some asserting that models have largely achieved grammatical competence (e.g., Mahowald et al., 2024) and others much more skeptical (e.g., Dentella et al., 2023; Lan et al., 2024; Fox and Katzir, 2024).

Some linguistic theories would posit that the ideal competence grammar would assign 0 probability to all ungrammatical strings. But LMs, by their nature, will assign non-zero probability to all strings. And by virtue of what they are designed for (modeling language in real-world contexts), it is not a desirable property of LMs that they assign 0 probability to ungrammatical strings. After all, in any realistic application setting, LMs would need to be able to interpret and handle ungrammatical utterances. If we are willing to accept that LMs will assign non-zero probability to ungrammatical strings, while potentially being able to represent grammatical generalizations in a theoretically meaningful way, then the scientific task of assessing grammatical knowledge in LMs requires working around this property.

Part of the field's uncertainty over LMs' grammatical competence stems from uncertainty over

[*] Work done while at the Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University.

Transactions of the Association for Computational Linguistics, vol. 14, pp. 124–146, 2026. https://doi.org/10.1162/TACL.a.611
Action Editor: Shay Cohen. Submission batch: 12/2024; Revision batch: 6/2025; Published 1/2026.

how to best assess grammatical knowledge in models. Given the success and convenience of prompting methods, a tempting approach is to simply ''ask'' models what sentences are grammatical or not (Dentella et al., 2023; Katzir, 2023), just as is commonly done for humans (Schütze, 2016; Sprouse and Almeida, 2012; Mahowald et al., 2016). But answering ''Is this sentence grammatical?'' requires more than just knowledge of grammar: It requires knowing what *grammaticality* means, as well as other auxiliary abilities such as being able to (truthfully) answer questions. As a result, this method systematically underestimates grammatical competence in LMs (Hu and Levy, 2023; Hu et al., 2024). It's easy to see the problem if we imagine a model trained without ever seeing the word ''grammatical'': It would have the same underlying knowledge of linguistic structure but be unable to answer the question.

An alternate approach is to measure the probabilities that models assign to strings, with the logic that models should assign higher probability to grammatical versus ungrammatical strings. But it is not immediately obvious that this is the best way to assess LMs' grammatical abilities, as grammaticality and probability are fundamentally distinct notions in linguistics (Chomsky, 1957; Berwick, 2018). A well-known illustration of this distinction is the sentence ''Colorless green ideas sleep furiously'' (Chomsky, 1957). The string has low probability (at least when it was originally coined), but, crucially, people still have the intuition that it is grammatical in a way ''*Furiously sleep ideas green colorless'' isn't. This line of thinking might make it seem like assessing models via string probability is fundamentally flawed. Indeed, critics have argued that the distinction between likelihood and grammaticality is ''entirely foreign'' (Katzir, 2023) to LMs, making them unsuitable models of grammatical competence.

However, Fox and Katzir (2024), Lan et al. (2024), and others go on to note that, in some cases, probability may be aligned enough with grammatically that it can be informative. And in practice, probability-based evaluations of grammatical knowledge in LMs use **minimal pairs**—i.e., pairs of sentences that differ only slightly from each other, and which form a grammaticality contrast (Marvin and Linzen, 2018; Futrell et al., 2019; Warstadt et al., 2020; Hu et al., 2020; Wilcox et al., 2023; Hu et al., 2024). Intuitively, researchers construct minimal pairs to isolate a specific grammatical contrast and factor out other properties that might affect string probability, such as length or lexical frequencies. But even this practice has also been criticized by recent work. For example, Leivada et al. (2024a) write: ''If LMs need specific comparisons in order to tell apart grammatical from ungrammatical sentences, this already counts as an inherent discrepancy from humans, who are able to make such judgments without such a comparison''. If string probability offers a window into grammaticality, they argue, then it should be possible to find a threshold on probability that separates grammatical and ungrammatical sentences (Leivada et al., 2024a,b).

Here, we give a formal argument for why the minimal pair approach can be appropriate, and does not necessarily elide the distinction between grammaticality and string probability. Broadly, our framework states that the probability of a string comes from two latent variables: the string's *message* and the string's *grammaticality*. The logic of minimal pair judgments follows naturally from this framework. All else equal, grammatical sentences obtain higher probability than ungrammatical sentences. So if two utterances convey the exact same message, but one is grammatical and one isn't, then the grammatical one should have a higher probability. In practice, this is hard to do, since any utterances that differ in the words they contain will convey at least slightly different messages. But if the messages are *sufficiently close*, then the minimal pair assumption can be used, and comparing the probabilities of the two strings will give insight into grammaticality.

Our framework also makes it clear that the probability of a message can overwhelm the contribution of grammaticality in determining string probability. Uncontroversially, a model that is well-calibrated to the world *should* assign higher probability to the string ''He went to the store'' than the string ''Cordelia went to the store'', despite the fact that both sentences are grammatical, since it is far more probable to express a message about some person with male pronouns than a person with an uncommon name such as Cordelia. But models will face competing pressures if tasked with comparing (a) ''Cordelia went to the store herself'' vs. (b) ''*He went to the store herself''. The former is clearly (more) grammatical, but the latter seems to convey a more probable message.[1] Under our framework, a model that assigns higher probability to (b) than (a) would not necessarily be failing to capture the distinction in grammaticality. Rather, because the pair is not appropriately controlled, the probabilities of the two strings are confounded with the probabilities of their messages.

In the rest of this paper, we first give a formal characterization of string probabilities in corpora and models. Then, we derive three predictions and test them empirically on 280K sentence pairs in English and Chinese. First, we predict a correlation between the log-probability of grammatical and ungrammatical sentences that convey roughly the same message, since their message probability is controlled for. Second, we predict that differences in human acceptability judgments on appropriately controlled minimal pairs will align with differences in log-probability of the strings in the pair. Finally, we predict a lack of separation in string probability space between unpaired grammatical and ungrammatical sentences. While this phenomenon has been taken to indicate a *failure* of models to capture grammaticality (Leivada et al., 2024a,b), we argue that it follows from our framework under reasonable assumptions.

More generally, we see our contribution as providing theoretical grounding for the practice of using minimal-pair probability comparisons to assess LMs' grammatical knowledge. While this practice is widely used in NLP, it has generally been given only brief and informal justification in empirical work (e.g., Warstadt et al., 2020), or rejected altogether (e.g., Leivada et al., 2024a). We use our theoretical analysis and empirical results to motivate recommendations for future work on grammatical evaluation of LMs.

2 Theoretical Framework

2.1 Strings, Messages, and Grammaticality

We consider a word, w, drawn from a vocabulary Σ, as well as strings, $\mathrm{s} \in \Sigma^*$ which are sequences of words. We write w_n for the word at index n in a string $\mathrm{s}_N = [w_1 \dots w_n \dots w_N]$, where $1 \leq n \leq N$. Let S denote a random variable that ranges over strings. Additionally, let M be a random variable that ranges over possible messages $m \in \mathcal{M}$. Finally, let G be a binary random variable. When $G = 1$, the intended message m is realized according to the grammatical rules of the language. When $G = 0$, m is *not* realized according to the grammatical rules of the language—i.e., there is an **error** in the process of realizing the message in string form.

In our framework, the probability of a string s is influenced by possible underlying messages and whether those messages are grammatically realized. We therefore write this probability as:

$$P(\mathrm{s}) = \sum_{\substack{m \in \mathcal{M}, \\ g \in \{0,1\}}} P(\mathrm{s}|m, g) P(g|m) P(m) \qquad (1)$$

Natural language is ambiguous, so strings often have more than one meaning. It is also arguably the case that some meanings can equivalently be realized by more than one choice regarding string realization (e.g., ‘‘Sam gave presents to the children’’ and ‘‘Sam gave the children presents’’ both realizing the same description of a transfer-of-possession event). For simplicity of mathematical treatment, however, our framework treats messages as equivalence classes of meanings plus string realization choices, with the specific underlying meaning probabilistically marginalized out. This is formalized in the following assumption:

Assumption 1. *Deterministic mapping from messages to strings when $G = 1$*

We assume that $P(\mathrm{s}|m, G = 1)$ and $P(m|\mathrm{s}, G = 1)$ is deterministic. That is, given an intended message m, there is only one way to realize m according to the rules of the grammar.

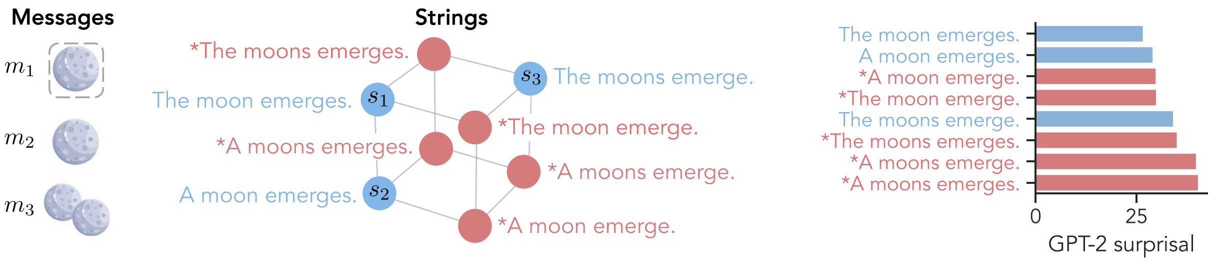
Definition 1. *Grammatical string*

We say that a string s is grammatical if, for some message $m \in \mathcal{M}$, $P(\mathrm{s}|m, G = 1) = 1$. Note that by Assumption 1, for every grammatical string s, $P(m|\mathrm{s}, G = 1) = 1$ for exactly one m.

Definition 2. *Ungrammatical string*

We say that a string s is ungrammatical if there is no m for which $P(\mathrm{s}|m, G = 1) > 0$. By Assumption 1, this is equivalent to saying there is no m for which $P(\mathrm{s}|m, G = 1) = 1$. Note that this does *not* mean that a string s is ungrammatical if for some m, $P(\mathrm{s}|m, G = 0) > 0$, as

[1] In this case, the intended message could be ‘‘She went to the store herself’’ or ‘‘He went to the store himself’’. We discuss intended messages in more detail in Section 4.2.



(a) (b)

Figure 1: (a) Illustration of messages (left) and strings (right) in toy domain. Blue = grammatical strings. Red = ungrammatical strings. (b) Surprisal (negative log probability) assigned to toy strings by GPT-2.

grammatical strings can be generated by an errorful realization of a message.

There is no ‘‘meaning’’ of an ungrammatical string in the sense that there is a unique message associated with a grammatical string (by Assumption 1). But there is a probability distribution over messages associated with an ungrammatical string, and in some cases it will be useful to specify the ‘‘most likely message’’ of an ungrammatical string. The details of how ungrammatical strings relate to messages will depend on an *error model*, which we discuss in Section 2.2.

2.1.1 Toy Example

We now walk through a simple example to illustrate the key intuitions and definitions discussed above. Consider the set of 8 strings formed by crossing two possible values of three syntactic elements: ‘‘{The, A} {moon, moons} {emerge, emerges}’’. These strings are visualized as nodes on a cube in Figure 1a. Each edge between strings represents one edit: In this case, swapping ‘‘The’’/‘‘A’’, or ‘‘moon’’/‘‘moons’’, or ‘‘emerge’’/‘‘emerges’’. Here, there are three strings $\{\mathrm{s}_1, \mathrm{s}_2, \mathrm{s}_3\}$ which can be viewed as the error-free realizations of three messages $\mathcal{M} = \{m_1, m_2, m_3\}$, respectively:

$$\begin{aligned} m_1 &= \exists! x\ [\text{moon}(x) \wedge \text{emerge}(x)] \qquad (2)\\ m_2 &= \exists x\ [\text{moon}(x) \wedge \text{emerge}(x)]\\ m_3 &= \exists x \exists y\ [\text{moon}(x) \wedge \text{emerge}(x)\\ &\qquad \wedge \text{moon}(y) \wedge \text{emerge}(y)\\ &\qquad \wedge (x \neq y)]\\ \mathrm{s}_1 &= \text{``The moon emerges.''}\\ \mathrm{s}_2 &= \text{``A moon emerges.''}\\ \mathrm{s}_3 &= \text{``The moons emerge.''} \end{aligned}$$

Therefore, s_1, s_2, and s_3 are grammatical. However, note that even these grammatical strings could be generated by errorful processes for certain messages: For example, presumably $P(\mathrm{s}_1|m_3, G = 0) > 0$. The other 5 strings in the set can each be viewed as errorful realizations of any of the messages m_1, m_2, or m_3. Furthermore, there is no $m \in \mathcal{M}$ for which any of these strings is the error-free realization. Therefore, the other 5 strings are ungrammatical.

Although for a given m these strings are all errorful realizations, they might not all be equally likely. In our framework, string probability is influenced by the likelihood of (potential) underlying messages. Here, we expect $P(m_1) > P(m_2) > P(m_3)$: i.e., it is most probable (at least on Earth) to express a message about a unique moon, somewhat probable to express a message about a single moon, and improbable to express a message about multiple moons. These differences in message probabilities can conflict with grammaticality in practice. Figure 1b shows surprisals (i.e., negative log probability) assigned by GPT-2 (Radford et al., 2019) to each of the 8 toy strings from Figure 1a. While the 3 grammatical strings have lower surprisals on average than the 5 ungrammatical strings, we also see that the strings which mention a *singular* moon tend to have lower surprisal. For example, GPT-2 assigns higher surprisal to the grammatical ‘‘The moons emerge’’ than the ungrammatical ‘‘*The moon emerge’’.

Another factor that affects the probability of ungrammatical strings is the way they are realized under a reasonable *error model*. We discuss this in more detail in the following section.

2.2 Error Model

The role of errors in speech comprehension and production has been studied extensively (e.g., Levy, 2008; Goldrick, 2011). We adopt a set of minimal working assumptions required for a basic error model. Let $s^{*(m)}$ be the ‘‘grammatical realization’’ of m: i.e., the string s such that $P(\mathrm{s}|m, G = 1) = 1$. By Assumption 1, this grammatical realization is unique. Then, $P(\mathrm{s}|m, G = 0)$ is concentrated in an ‘‘error neighborhood’’ of $s^{*(m)}$ that excludes $s^{*(m)}$. That is, although violating a language’s syntax, an ungrammatical realization of a message m tends to be mostly similar to a grammatical realization of m.

We can then quantify the ‘‘error distance’’ from one string s_1 to another s_2 conditioned on m as a distance, $\mathcal{D}(s_1 \to s_2|m)$. This distance ranges over non-negative integer values, with $\mathcal{D}(s_1 \to s_2|m) = 0$ if and only if $s_1 = s_2$, and $\mathcal{D}(s_1 \to s_2|m) = 1$ if the two strings differ by a single error. We assume that the probability of each error step is some small value centered around ϵ. Thus the number of error steps d is geometrically distributed, $P(d) \approx (1 - \epsilon)\epsilon^d$, and $P(G = 0|m) \approx \epsilon$.

For any specific string $\mathrm{s} \neq s^{*(m)}$, this gives us:[2]

$$P(\mathrm{s}|m, G = 0) \approx \frac{1-\epsilon}{K}\left(\frac{\epsilon}{K}\right)^{\mathcal{D}(s^{*(m)} \to \mathrm{s}|m)-1} \tag{3}$$

for $\mathrm{s} \neq s^{*(m)}$, where K (which we treat as constant) denotes the number of different possible errors that could be made at any given error step. This distance can be thought of as the number of errors required to change s to $s^{*(m)}$, if m is the intended message.

Returning to the notion of ‘‘meaning’’ of ungrammatical strings, while ungrammatical strings do not have a unique message in our framework, they can be thought of as corresponding to messages from nearby grammatical strings. If we assume that errors are relatively rare and the space of messages is relatively sparse, then in many cases the most likely message of an ungrammatical string will be transparent.

2.2.1 Toy Example

Returning to the toy example from Section 2.1.1 and Figure 1a, we can think of each edit between a grammatical string (blue node) and ungrammatical string (red node) as an error. We can see how messages relate to ungrammatical strings. Consider an ungrammatical string such as ‘‘*A moon emerge’’. The most likely message of this string is m_2, corresponding to the message associated with the closest grammatical string, ‘‘A moon emerges’’. While ‘‘*A moon emerge’’ could in principle be an ungrammatical realization of the other two messages (m_1 and m_3), the realization process would involve more errors, making m_1 and m_3 less likely to be the underlying message.

2.3 Minimal Pairs

We now arrive at our definition of minimal pairs.

Definition 3. *Meaning-matched pair*

A meaning-matched pair is a pair of strings $(\mathrm{s}, \mathrm{s}')$ such that (1) s is the grammatical realization of some message m; (2) s′ is ungrammatical; and (3) s′ is a *reasonably likely* ungrammatical realization of m; i.e.:

$$\exists \text{ small } \delta > 0 \text{ s.t. } P(m|\mathrm{s}', G = 0) > 1 - \delta \tag{4}$$

Definition 4. *Minimal pair*

A meaning-matched minimal pair, or **minimal pair** for short, is a meaning-matched pair $(\mathrm{s}, \mathrm{s}')$ such that $\mathcal{D}(\mathrm{s} \to \mathrm{s}'|m) = 1$ when $m = \arg\max_M P(M|\mathrm{s})$.

In other words, a minimal pair is a meaning-matched pair where there is only one thing ‘‘wrong’’ with the ungrammatical string s′.

2.3.1 Toy Example

Recall our simple example from Section 2.1.1 and Figure 1a. Here, the set of meaning-matched pairs is given by every pair of grammatical and ungrammatical strings (i.e., blue and red in Figure 1a). In this case, there are 15 such pairs, formed by pairing each of $\{s_1, s_2, s_3\}$ with each of the five ungrammatical strings. Accordingly, the set of minimal pairs is the subset of meaning-matched pairs which include one of the grammatical strings $\{s_1, s_2, s_3\}$ and another string which is one edge (i.e., error) away from the grammatical string. In this example, there are 7 such pairs.

An example meaning-matched pair which is *not* a minimal pair would be (s_1 = ‘‘The moon emerges’’, s′ = ‘‘*A moons emerge’’), as m_1 is the message associated with s_1, and there are multiple errors needed to generate s′ from m_1.

[2] Equation 3 ignores the possibility that multiple error sequences from m might lead to the same s, which would increase $P(\mathrm{s}|m, G = 0)$ in a way that is ultimately canceled out in the derivations presented in our Appendix.

3 Three Predictions

We now describe three predictions that fall out of our framework, with additional assumptions that we specify along the way. The full derivation for each prediction is given in Section A. After outlining these predictions, the rest of the paper is dedicated to testing them, empirically.

Prediction 1. *Correlation between the log-probability of grammatical and ungrammatical strings within minimal pairs.*

If string probability only depends on grammaticality G, then all ungrammatical strings should receive the same (near-)zero probability. In contrast, our framework states that M also plays a role. We predict that the probability of a grammatical string is primarily determined by the probability of its message, and the probability of an ungrammatical string is primarily determined by the probability of the message of the nearest grammatical neighbor string. We therefore expect to see a correlation between the log-probability of the grammatical string and the log-probability of the ungrammatical string across sets of minimal pairs (**Prediction 1a**).

However, minimal pairs are a theoretical ideal, and in practice not all researcher-constructed minimal pairs will be truly ''minimal''. When we consider pairs where the most probable message of the grammatical and ungrammatical strings are less similar—i.e., when the pairs are ''less minimal''—the contribution of M is less controlled. Therefore, we predict a weaker correlation for pairs that are less minimal (**Prediction 1b**).

Prediction 2. *Correlation between differences in log-probability and human acceptability judgments within minimal pairs.*

Native speaker acceptability judgments vary with both grammatical well-formedness and meaning plausibility (Schütze, 2016). Using our framework, we operationalize (i) with $\log P(m)$, and (ii) with the number of errors. Then if we consider minimal pairs $(\mathrm{s}, \mathrm{s}')$, where the understood message between s and s' is the same, the *difference* in acceptability judgments depends primarily on the error probability of taking s to s', as does the *difference* in string log-probability. We therefore predict that differences in string probability are correlated with differences in human acceptability judgments, within minimal pairs (**Prediction 2a**). As the ''minimalness'' of the pair decreases, the contribution of message M increases, and we expect to find weaker correlation between log-probability differences and acceptability judgment differences (**Prediction 2b**).

Prediction 3. *Potentially poor separation based on probability between grammatical / ungrammatical strings.*

In illustrating the distinction between probability and grammaticality, Chomsky (1957) wrote: ''If we rank the sequences of a given length in order of statistical approximation to English, we will find both grammatical and ungrammatical sequences scattered throughout the list''. By viewing probability as influenced by grammaticality *and* meaning, our framework provides a theoretical basis for Chomsky's prediction: That is, string probability does not separate grammatical and ungrammatical strings. While Leivada et al. (2024a,b) argue that this lack of separation means that it is problematic to use probability to measure grammaticality, we note that this prediction falls directly out of our theoretical framework.

Importantly, the expected degree of separation depends on how grammatical and ungrammatical strings are pooled together. When the strings come from minimal pairs, we would expect to see better separation on the basis of probability, as the message probabilities are controlled for. When there are no constraints on the relationship between the grammatical/ungrammatical strings, however, the distributions of messages associated with the grammatical and ungrammatical strings can be very different from each other, and string probability might achieve poor separation.

We note that so far we have only discussed *pure* string probability, which has been investigated in recent studies (Leivada et al., 2024a,b). There are reasons to expect that grammatical/ungrammatical strings would not be separated by pure probability: When there is no upper bound on the length of grammatical strings and at least some ungrammatical string is assigned non-zero probability (as is the case for LMs), then there must be an infinite set of grammatical strings that are assigned lower probability than that ungrammatical string. Indeed, our framework also suggests that replacing pure string probability with a ''normalizing'' function of string probability and other

Dataset	Language	# items	Reference	Prediction 1	Prediction 2	Prediction 3
BLiMP	English	66993	Warstadt et al. (2020)	✓	✗	✓
SCaMP-P	English	67000	McCoy and Griffiths (2025)	✓	✗	✓
SCaMP-I	English	67000	McCoy and Griffiths (2025)	✓	✗	✓
SyntaxGym	English	1018	Hu et al. (2020)	✓	✗	✓
ZhoBLiMP	Chinese	35400	Liu et al. (2024)	✓	✗	✗
SLING	Chinese	39976	Song et al. (2022)	✓	✗	✗
CoLA	English	8551	Warstadt et al. (2019)	✗	✗	✓
LI	English	1883	Sprouse et al. (2013); Mahowald et al. (2016)	✗	✓	✓
HLL	Chinese	213	Chen et al. (2020)	✗	✓	✗

(a)

Model	HuggingFace ID	Language	# params	Vocab size	Training data
GPT-2	`gpt2`	English	124M	50257	40 GB
Llama-3-70B	`meta-llama/Meta-Llama-3-70B`	English	70B	128256	15T tokens
GPT-2 ZH	`uer/gpt2-chinese-cluecorpussmall`	Chinese	102M	21128	100 GB
Llama-3-8B ZH	`hfl/llama-3-chinese-8b`	Chinese	8B	128256	120 GB

(b)

Table 1: (a) Datasets and (b) models used in our experiments. ‘‘# items’’ = # pairs for each dataset except CoLA, and # sentences for CoLA. ‘‘SCaMP-P/I’’ = plausible/implausible subsets of SCaMP.

grammar-independent features should bring the distributions of messages associated with grammatical and ungrammatical strings closer together, and thereby increase the separability. This prediction provides a novel justification for previous work which has hypothesized that grammaticality and probability are linked through a complex function (Pauls and Klein, 2012; Lau et al., 2017; Tjuatja et al., 2025).

In the rest of this paper, we report the results of three experiments designed to empirically test the three predictions spelled out above.

4 Prediction 1: Correlation Between Grammatical/Ungrammatical Log-Probability Within Minimal Pairs

We now investigate the first set of predictions made by the theory in Section 2: (a) probabilities of grammatical and ungrammatical strings in minimal pairs should be correlated, and (b) this correlation should be weaker for less ‘‘minimal’’ pairs.

Our code and data are all publicly available at `https://github.com/jennhu/probability-grammaticality`.

4.1 Evaluation Materials

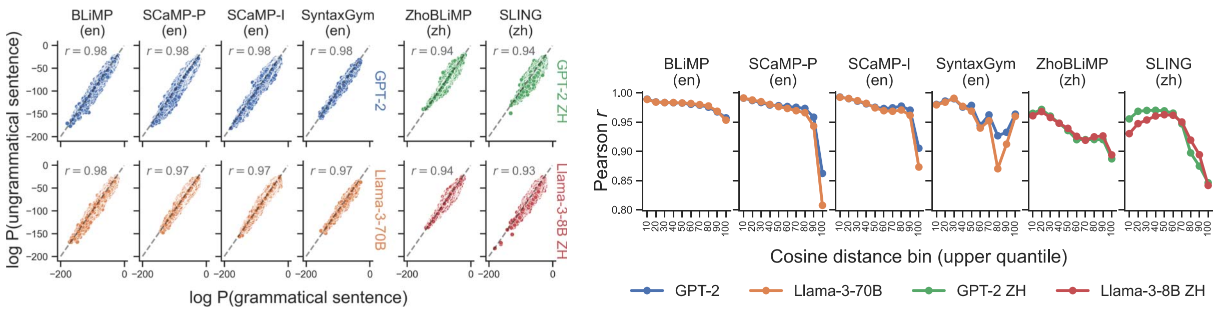
To test Prediction 1, we need a set of paired grammatical and ungrammatical sentences with varying degrees of ‘‘minimalness’’. We test our theory on existing data sets from two typologically different languages, English and Mandarin Chinese, as our theoretical framework is language-agnostic and only makes basic assumptions about the generative process of corpus strings. We evaluate models on five datasets, summarized in Table 1a: BLiMP (Warstadt et al., 2020), SCaMP (McCoy and Griffiths, 2025), and SyntaxGym[3] (Hu et al., 2020) in English, and ZhoBLiMP (Liu et al., 2024) and SLING (Song et al., 2022) in Chinese. Each dataset is proposed to contain ‘‘minimal pairs’’, although the pairs may diverge to varying degrees from the theoretical ideal defined in Definition 4.

One attractive feature of these datasets is that they collectively vary in terms of semantic plausibility. For example, SyntaxGym was manually designed to avoid implausible sentences; BLiMP, because of its templatic generation, includes a mix of plausible and implausible; and SCaMP includes plausible and implausible subsets. This allows us to test our framework on a space of messages $\mathcal{M}$ that is not restricted to probable or commonplace ones: If a pair of sentences shares the same underlying message, then probability can reveal information about grammaticality, regardless of how probable the message itself is *a priori*.

4.2 Measuring Minimalness

To empirically evaluate Prediction 1b, we need a way of quantifying the ‘‘minimalness’’ of a

[3] We use the subset of SyntaxGym compatible with sentence scoring.



(a)

(b)

Figure 2: (a) Prediction 1a: Logprobs of paired grammatical (x-axis) and ungrammatical (y-axis) sentences are correlated. Dashed line: $x = y$. (b) Prediction 1b: Correlation between grammatical and ungrammatical logprobs (y-axis) generally decreases as within-pair cosine distance (x-axis) increases.

putative minimal pair. According to our framework, a natural way to measure the minimalness of a minimal pair $(\mathbf{s}, \mathbf{s}')$ would be to measure the similarity between the message of $\mathbf{s}$ (i.e., $m = \arg\max_M P(M|\mathbf{s})$) and the message of the *closest grammatical string* to $\mathbf{s}'$. That is: If the ungrammatical string $\mathbf{s}'$ were in a ‘‘true’’ minimal pair (according to Definition 4) with another grammatical string $\mathbf{s}^*$, how similar in meaning is $\mathbf{s}$ to $\mathbf{s}^*$?

In practice, this quantity is difficult to systematically estimate, as it involves specifying the closest grammatical string to each ungrammatical string in a dataset of minimal pairs. We approximated this quantity by measuring the similarity between the message of $\mathbf{s}$ and the ‘‘message’’ of $\mathbf{s}'$, taking a usage-based approach to meaning. Namely, we adopted the assumption that sentences that convey similar messages will be closer in a high-dimensional embedding space learned for meaning-based tasks. To quantify the minimalness of a pair, we therefore measured the cosine similarity (in practice, the cosine distance) between the embeddings of the grammatical and ungrammatical sentences in the pair.[4]

In order to compute correlations at different levels of minimalness (as required to test Predictions 1b and 2b), we grouped pairs into 10 equally-sized bins based on the within-pair cosine distance.

4.3 Computing String Probability

We compute the probability of a sentence by aggregating the probability of each token conditioned on all previous tokens using autoregressive language models. In practice, we do this by computing the log probability of each token conditioned on its left context, and summing these values to get the log probability of the full sentence.

The predictions of our framework apply to any probabilistic model that has learned a reasonably accurate distribution of $P(S)$. We felt this to be a reasonable assumption for moderately-sized Transformer models trained on Internet-scale text. Although an interesting avenue for further research, directly testing the influence of specific factors such as model size was not the key motivation of our experiments, so we simply chose two models for each language, covering multiple sizes and model families. We evaluated two open-source base (i.e., not fine-tuned) models of varying sizes on each dataset (see Table 1b for details). We evaluated GPT-2 (Radford et al., 2019) and Llama-3-70B (AI@Meta, 2024) on the English datasets. For the Chinese datasets, we evaluated a GPT-2 model trained on CLUECorpusSmall (Xu et al., 2020), which we refer to as GPT-2 ZH (Zhao et al., 2019), and Llama-3-8B ZH (Cui et al., 2023).

4.4 Results

The results (shown in Figure 2) largely confirm our first set of predictions. As suggested by Prediction 1a, Figure 2a shows a strong positive correlation between the log probability of the ungrammatical and grammatical sentences in each pair, across both datasets and model sets. Furthermore, Figure 2b shows that the Pearson r correlation between grammatical and ungrammatical log probability decreases as the pairs are less

[4] We used `sentence-transformers` models `all-mpnet-base-v2` to embed English sentences and `uer/sbert-base-chinese-nli` for Chinese.

controlled for meaning (i.e., as the cosine distance increases), as suggested by Prediction 1b. These patterns hold for multiple models, datasets, and languages.

5 Prediction 2: Probability Differences Align with Acceptability Differences

We now investigate the second set of predictions: (a) human acceptability judgment differences and string log-probability differences should be correlated within minimal pairs, and (b) this correlation should be weaker for pairs that are less minimal.

5.1 Evaluation Materials

To test Prediction 2, we need human acceptability judgments of each sentence in isolation. Our evaluation datasets are summarized in Table 1a. We only used existing data and did not collect any new human judgments. The English dataset (LI) includes pairs from Sprouse et al. (2013) and Mahowald et al. (2016). Both studies randomly sampled paired sentences (grammatical vs. ungrammatical/questionable) from *Linguistic Inquiry* journal papers and collected acceptability judgments for each sentence from native English speakers. The Chinese dataset (HLL) includes pairs from Chen et al. (2020), where the authors collected acceptability judgments for each sentence within related groups (pairs, triples, or n-tuples) from a Chinese syntax textbook (Huang et al., 2009).[5]

Each of the three data sources we use originally measured acceptability judgments on a 7-point Likert scale for each sentence in isolation. We z-scored (centered and scaled) all Likert score ratings within participants and then calculated the mean z-score for each sentence. For each pair, the difference in mean z-scores between the grammatical and ungrammatical sentences indicates the human acceptability judgment difference.

Filtering for Meaning-matched Pairs. While the grammatical and ungrammatical sentences in both datasets are presented by their original authors as ‘‘pairs’’, they vary widely in terms of how similar they are to each other. For example, the LI dataset contains the pair formed by grammatical sentence ‘‘The apples fell just a short fall to the lower deck, and so were not too badly bruised’’ and ungrammatical sentence ‘‘*The submarine emerged an abrupt emergence’’. This sort of pair fails to meet our definition of ‘‘meaning-matched’’ pair (Definition 3): any reasonable value of δ would exclude this pair, as the ungrammatical sentence is extremely unlikely to be an errorful realization of the message of the grammatical sentence. These pairs are also potentially problematic for our analyses. We assume that messages do not vary dramatically with grammaticality (Assumption 3; Section A). But, when grammatical/ungrammatical sentences are paired in the way that LI and HLL were curated, there could be systematic differences in how messages are distributed across different values of grammaticality. That is, for radically non-meaning-matched pairs, the ungrammatical variant could map onto a systematically higher-probability or lower-probability message than the grammatical one. This is less of a concern when the pairs are tightly matched for meaning.

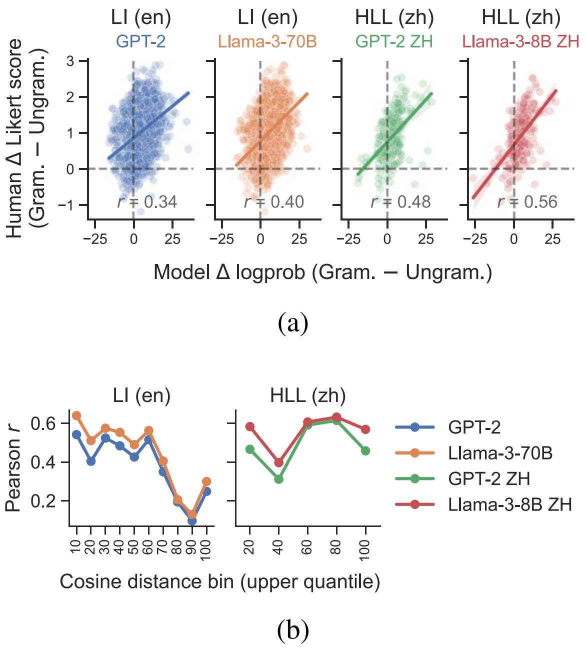
We therefore only kept sentence pairs which are reasonably ‘‘meaning-matched’’. We defined an empirical threshold that guards against pairs like the one above, but still allows for enough variability in minimalness to test our predictions. In practice, we did this by only keeping pairs where the Levenshtein edit distance between the strings was below the 75th quantile across all pairs.

5.2 Results

The results for these analyses are displayed in Figure 3. Figure 3a shows that human Likert score differences are correlated with log probability differences (Prediction 2). We also note that similar results have been reported for other languages (Suijkerbuijk et al., 2025).

Figure 3b shows that the degree of alignment (Pearson r correlation coefficient) decreases as cosine distance increases for the English LI dataset, as predicted by Prediction 2b. However, for the Chinese HLL dataset, we do not find clear evidence for Prediction 2b. There could be several reasons for the differences between the LI and HLL results. First, we note that LI has a substantially wider spread of cosine distances between

[5] For the groups with more than two sentences, we created pairs by juxtaposing all possible grammatical vs. ungrammatical/questionable pairings. That is, for a given group with grammatical sentences $\{g_1, \ldots, g_n\}$ and ungrammatical sentences $\{u_1, \ldots, u_m\}$ we create all the possible pairs (g_i, u_j) for $i \in [1, n]$ and $j \in [1, m]$; usually there are 2–4 sentences in a group.



(a)

(b)

Figure 3: (a) Prediction 2a: Deltas between grammatical and ungrammatical strings are correlated between models and humans. (b) Prediction 2b: Correlation between model and human deltas (y-axis) generally decreases as within-pair cosine distance (x-axis) increases. Note this pattern is only apparent in the English data.

the 70th and 95th quantiles, which is where the clearest drop in correlations is seen. It could also be the case that humans' acceptability judgments might reflect slightly different factors in Chinese versus English. Another potential cause might be practical issues with the models trained on Chinese data: e.g., a lower-quality sentence embedding model might not faithfully represent the similarity in underlying message between sentences.

6 Prediction 3: Poor Separation of Grammatical/Ungrammatical Strings

We now test our final prediction: potentially poor separation between the probability of grammatical and ungrammatical strings. As discussed in Section 3 and Section A.3, our framework predicts that the limiting factor for such separation is the variance of messages associated with grammatical and ungrammatical string sets. Therefore, in addition to examining raw string probability, in this section we introduce several normalizing transformations that reduce variance in messages and should thereby also increase separability. Specifically, we propose a novel scoring function that represents the Bayes factor between different generative processes of observed strings. And we give a novel derivation for the Syntactic Log Odds Ratio (SLOR; Pauls and Klein, 2012; Lau et al., 2017), as equivalent to the average Pointwise Mutual Information between a word and its preceding context. We find that neither raw string probability nor any normalized string probabilities result in good separation of grammatical/ungrammatical strings, in line with our prediction.

6.1 Transformations of Probability

In addition to the notation introduced in Section 2.1, we define a language model $p_\theta(\boldsymbol{s}) : \Sigma^* \to \mathbb{R}$ as a function from strings to probabilities. Here, we use w to refer to *tokens* instead of words. In practice, the LMs we work with are autoregressive, assigning probabilities to tokens given their preceding context. That is, they are functions of the type $\mathbf{s}_N \to \mathbb{R}^N$, mapping strings of length N to N-dimensional vectors of probabilities. We consider linking functions $f : \mathbb{R}^N \to \mathbb{R}$ that map these vectors of token probabilities $p_\theta(w_n \mid \boldsymbol{s}_{<n})$ to scores. Below, we enumerate several candidates for this function. For brevity, we will write $f(\mathbf{s})$ instead of $f(p_\theta(\mathbf{s}))$.

Metric 1. *Probability*

A natural starting point is to test whether raw probability $p_\theta(\mathbf{s})$ can separate grammatical and ungrammatical strings in a language. Equivalently, we consider the log of this joint probability, or the sum of the log probabilities assigned to each word.

$$\begin{aligned} f(\mathrm{s}) &= \log p_\theta(\mathrm{s}) \\ &= \sum_{n=1}^{N} \log p_\theta(w_n \mid \boldsymbol{s}_{<n}) \end{aligned} \tag{5}$$

The ability of this metric to separate grammatical and ungrammatical sentences has been explicitly investigated (Lau et al., 2017; Leivada et al., 2024a), and is also the implicit standard for minimal pair comparisons (e.g., Warstadt et al., 2020).

Metric 2. *Bayes Factor: Uniform Distribution*

When determining whether a sentence is grammatical, comprehenders may consider the data-generating process that was likely to produce the string. One rational approach would be to evaluate the competing evidence for two hypotheses: that the string was produced by the grammar, and that the string was produced by a non-grammatical generative process. We instantiate this intuition by considering the **Bayes Factor**, or the ratio of the likelihoods, of the sentence under two hypotheses.

Let H_G denote the hypothesis that the grammar is the generating process, which we estimate with an LM's distribution p_θ. And let H_{uniform} denote the hypothesis that the generating process is simply a uniform distribution over the vocabulary Σ. Given an observed string $\boldsymbol{s}$, we can define the log Bayes factor between H_G and H_{uniform} as:

$$\begin{aligned} f(\boldsymbol{s}) &= \log p(\boldsymbol{s}|H_G) - \log p(\boldsymbol{s}|H_{\text{uniform}}) \\ &= \log p_\theta(\boldsymbol{s}) + N \log |\Sigma| \end{aligned} \tag{6}$$

Metric 3. *Bayes Factor: Unigram Distribution*

We also consider the log Bayes Factor between hypothesis H_G and the hypothesis H_{unigram} under which the data was generated by sampling from a unigram distribution $p(\cdot)$:

$$\begin{aligned} f(\boldsymbol{s}) &= \log p(\boldsymbol{s}|H_G) - \log p(\boldsymbol{s}|H_{\text{unigram}}) \\ &= \log p_\theta(\boldsymbol{s}) - \sum_{n=1}^{N} \log p(w_n) \end{aligned} \tag{7}$$

This is equivalent to the ''Norm LP (Sub)'' metric proposed by Lau et al. (2017), or, equivalently, SLOR without the length-normalization factor.

Metric 4. *Statistical Association (SLOR)*

Next, we consider the average statistical association between a word and its context. To instantiate this hypothesis, we use the **pointwise mutual information (PMI)** between a word and its preceding context. The PMI between realizations of two random variables is the log ratio of their joint probability assuming dependence and independence. Using average PMI, we derive the following transformation:

$$\begin{aligned} f(\boldsymbol{s}) &= \frac{1}{N} \sum_{n=1}^{N} \text{PMI}(w_n; \boldsymbol{s}_{<n}) \\ &= \frac{1}{N} \sum_{n=1}^{N} (\log p_\theta(w_n \mid \boldsymbol{s}_{<n}) - \log p(w_n)) \end{aligned} \tag{8}$$

where $p(w_n)$ is the unigram (i.e., frequency) estimate of word w_n. This metric is equivalent to the **Syntactic Log-Odds Ratio (SLOR)** proposed by Pauls and Klein (2012), which has also been investigated by Lau et al. (2017), although the connection to PMI has not been previously established.

Metric 5. *Mean Probability*

As a simple variation of Equation (5) that controls for length, we consider mean log probability.

$$f(\boldsymbol{s}) = \frac{1}{N} \sum_{n=1}^{N} \log p_\theta(w_n \mid \boldsymbol{s}_{<n}) \tag{9}$$

6.2 Evaluation Materials

To test Prediction 3, we no longer need paired sentences (as was the case for Predictions 1 and 2), but instead simply need large sets of grammatical and ungrammatical sentences from which to compute the relevant metrics. We evaluate models on five English datasets that contrast ungrammatical and grammatical sentences: the three minimal-pair datasets used to test Prediction 1 (BLiMP, SCaMP, and SyntaxGym), the LI dataset used to test Prediction 2, and CoLA (Warstadt et al., 2019).[6] See Table 1a for a summary.

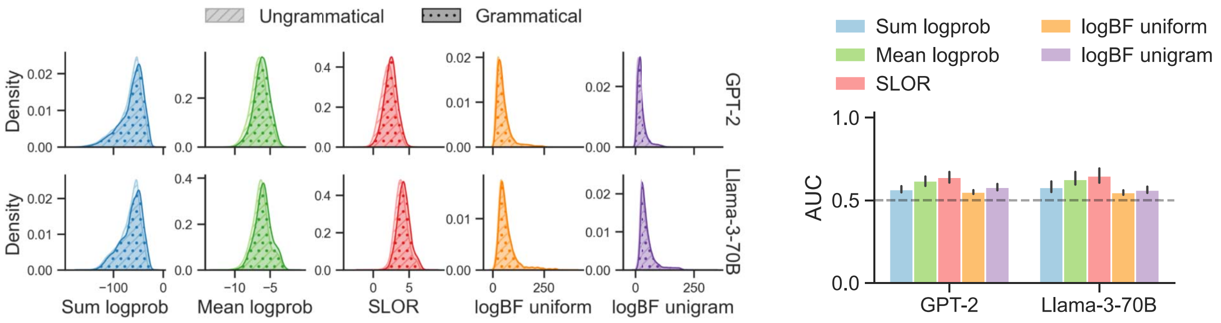
6.3 Computing Separation

To quantify the degree of separation for a given linking function f, we first pool all grammatical and ungrammatical strings from the dataset into one flat set. We then compute all scores $f(\boldsymbol{s})$ for each string $\boldsymbol{s}$ in this set, and compute a receiver operating characteristic (ROC) curve for these scores, treating grammatical strings as class 1 and ungrammatical as class 0. We use the area under the ROC curve (AUC) as our measure of separability, where AUC = 0.5 indicates no separability, and AUC = 1 indicates perfect separability.

We evaluate the same two LMs used to evaluate the English datasets for Predictions 1 and 2: GPT-2 and Llama-3-70B (see Table 1b). To obtain token frequency measurements for SLOR, we sought to estimate the distribution of tokens in each model's training data.[7] Since we do not have access to this data, we used the HuggingFace FineWeb

[6] Here, we focus on English due to the computational demands of token-frequency estimation for SLOR.

[7] Our SLOR metric is computed over *tokens*, not words.



(a)

(b)

Figure 4: Evaluation of Prediction 3. (a) Distributions of scores are highly overlapping across grammatical and ungrammatical sentences (pooled across datasets). (b) Poor separability (area under receiver operating characteristic curve, or AUC) achieved by each model and probability transformation. Horizontal line at 0.5 indicates no separation. For dataset-specific results, see Section B, Figures 5 and 7.

dataset (Penedo et al., 2024) as a representative sample of a high-quality Internet text corpus. We used each model to tokenize 1 million items from the `sample-10BT` sample of FineWeb, and then used this sample to estimate the token frequency distribution for each model.

6.4 Results

Figure 4a shows the scores assigned by both models to grammatical and ungrammatical sentences, collapsed across datasets. There appears to be substantial overlap between both sentence groups, across all metrics. This is confirmed by the ROC curves, shown in Section B / Figure 6, as well as the corresponding AUC scores (which do not exceed 0.75), shown in Figure 4b. Overall, mean logprob and SLOR consistently outperform the other metrics, suggesting that length normalization helps improve separability somewhat.

We also note that the models achieve high accuracy (∼80%) on standard minimal pair comparisons (Section B, Figure 8). So, by a minimal pair standard, models are sensitive to the relevant grammatical manipulations, but this is not reflected in (simple transformations of) string probability, as predicted by our theoretical proposal.

7 Discussion

It is uncontroversial that string probability is not the same as grammaticality. But that does not mean that string probability cannot reveal information about a probabilistic model's underlying grammatical knowledge. Here, we argued that these probabilities are determined by a combination of the probability of the message and the grammaticality of the string. Our theoretical framework shows that some prior critiques of minimal-pair analysis—e.g., that probability does not robustly separate grammatical and ungrammatical strings (Leivada et al., 2024a,b)—fall out of simple assumptions about the generative process underlying linguistic corpus data.

An offshoot of our analysis is that studies of grammaticality in LMs that use minimal pairs that are *not* tightly controlled (e.g., Vázquez Martínez et al., 2023) risk underestimating the grammatical competence of models by failing to control for the influence of M. In other words, a model could seemingly not differentiate between grammatical and ungrammatical strings, if the messages across grammatical and ungrammatical strings are not well-controlled. As we have argued here, this does not necessarily imply that the model has not learned generalizations about grammatical rules. If we are interested in isolating the model's sensitivity to grammaticality, we have to use carefully designed procedures to factor out M.

One such procedure is minimal pair string comparisons, where the members of the minimal pair are closely matched in M but are hypothesized to differ in G. While controlled minimal pair comparisons are hardly new in NLP (e.g., Marvin and Linzen, 2018; Futrell et al., 2019), we have provided new theoretical grounding for these practices. In addition, our work lays the foundation for using computational techniques to isolate the effects of M and G, which has been explored in recent work (Stańczak et al., 2024).

Our analyses also raise new questions regarding LMs' grammatical knowledge. The poor separation achieved by state-of-the-art LMs in Section 6 feels counterintuitive: If LMs virtually always produce grammatical strings (under standard sampling procedures), then why is there so much overlap between the probabilities assigned to grammatical and ungrammatical strings? This tension between *discriminative failures* and *generative abilities* could be seen as a specific realization of the ''generative AI paradox'' (West et al., 2024), and also connects to recent work demonstrating that language identification is impossible except in highly constrained cases (Gold, 1967; Angluin, 1980), whereas language generation is possible for any countable list of languages (Kleinberg and Mullainathan, 2024).

Leivada et al. (2024b) argue that comparing isolated acceptability judgments in humans against minimal-pair probability differences in models is not comparing ''apples with apples'', and thus unfair. We hope that our theoretical model can bring greater clarity to the issue of what makes a *fair* comparison. When comparing the (cognitive) abilities of two groups—humans and models, younger and older children, or even two different animal species—we maintain that the researcher must design assessments with that group's (cognitive) computational architecture in mind. Applying the same evaluation method, which might impose auxiliary challenges on one group but not the other, can artificially increase apparent intergroup differences (Firestone, 2020; Lampinen, 2024; Hu and Frank, 2024). For example: The intelligence of a squirrel should not be judged based on its ability to solve a Rubik's cube.[8] Similarly, using metalinguistic judgments or isolated string probability as a window into grammaticality ignores the reality of what LMs are, and what they are trained to do—namely, to maximize the probability of strings from a corpus. More broadly, linguistic theory continues to suggest new ways to evaluate LMs, just as modern LMs provide new tools for studying the relationship between probability and grammaticality.

Acknowledgments

We would like to thank the editors and reviewers for their helpful feedback. This work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence. Kyle Mahowald acknowledges funding from Open Philanthropy and NSF CAREER grant 2339729 through the Directorate for STEM Education (EDU). Roger Levy acknowledges support from NSF grant BCS-2121074 and a grant from the Simons Foundation to the Simons Center for the Social Brain at MIT. The moon icon in Figure 1a was made by vectorsmarket15 from flaticon.com.

References

AI@Meta. 2024. Llama 3 Model Card.

Dana Angluin. 1980. Inductive inference of formal languages from positive data. *Information and Control*, 45(2):117–135. https://doi.org/10.1016/S0019-9958(80)90285-5

Marco Baroni. 2022. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. In Shalom Lappin and Jean-Philippe Bernardy, editors, *Algebraic Structures in Natural Language*. Taylor & Francis. https://doi.org/10.1201/9781003205388-1

Robert C. Berwick. 2018. Syntactic structures after 60 years. In Norbert Hornstein, Howard Lasnik, Pritty Patel-Grosz, and Charles Yang, editors, *The Impact of the Chomskyan Revolution in Linguistics*, pages 177–194. De Gruyter Mouton. https://doi.org/10.1515/9781501506925-181

Zhong Chen, Yuhang Xu, and Zhiguo Xie. 2020. Assessing introspective linguistic judgments quantitatively: The case of the syntax of Chinese. *Journal of East Asian Linguistics*, 29:311–336. https://doi.org/10.1007/s10831-020-09210-y

Noam Chomsky. 1957. *Syntactic Structures*. Mouton. https://doi.org/10.1515/9783112316009

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca. *arXiv preprint arXiv:2304.08177*.

Vittoria Dentella, Fritz Günther, and Evelina Leivada. 2023. Systematic testing of three Language Models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National*

[8] This example is due to Andrew Lampinen.

Academy of Sciences, 120(51):e2309583120. Publisher: Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.2309583120, PubMed: 38091290

Chaz Firestone. 2020. Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 117(43):26562–26571. Publisher: Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.1905334117, PubMed: 33051296

Danny Fox and Roni Katzir. 2024. Large language models and theoretical linguistics. *Theoretical Linguistics*, 50(1–2):71–76. https://doi.org/10.1515/tl-2024-2005

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1004

E. Mark Gold. 1967. Language identification in the limit. *Information and Control*, 10(5):447–474. https://doi.org/10.1016/S0019-9958(67)91165-5

Matthew Goldrick. 2011. Linking speech errors and generative phonological theory. *Language and Linguistics Compass*, 5(6):397–412. https://doi.org/10.1111/j.1749-818X.2011.00282.x

Jennifer Hu and Michael Frank. 2024. Auxiliary task demands mask the capabilities of smaller language models. In *First Conference on Language Modeling*.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.158

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.306

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121. Publisher: Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.2400917121, PubMed: 39186652

C.-T. James Huang, Y.-H. Audrey Li, and Yafei Li. 2009. *The Syntax of Chinese*. Cambridge Syntax Guides. Cambridge University Press.

Roni Katzir. 2023. Why large language models are poor theories of human linguistic cognition: A reply to Piantadosi. *Biolinguistics*, 17. https://doi.org/10.5964/bioling.13153

Jon Kleinberg and Sendhil Mullainathan. 2024. Language Generation in the Limit. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://doi.org/10.52202/079017-2111

Andrew Lampinen. 2024. Can language models handle recursively nested grammatical structures? A case study on comparing models and humans. *Computational Linguistics*, pages 1–35. https://doi.org/10.1162/coli_a_00525

Nur Lan, Emmanuel Chemla, and Roni Katzir. 2024. Large language models and the argument from the poverty of the stimulus. *Linguistic Inquiry*, pages 1–28. https://doi.org/10.1162/ling_a_00533

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241. https://doi.org/10.1111/cogs.12414, PubMed: 27732744

Evelina Leivada, Vittoria Dentella, and Fritz Günther. 2024a. Evaluating the language abilities of large language models vs. humans:

Three caveats. *Biolinguistics*, 18. https://doi.org/10.5964/bioling.14391

Evelina Leivada, Fritz Günther, and Vittoria Dentella. 2024b. Reply to Hu et al.: Applying different evaluation standards to humans vs. large language models overestimates AI performance. *Proceedings of the National Academy of Sciences*, 121(36):e2406752121. Publisher: Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.2406752121, PubMed: 39186655

Roger Levy. 2008. A noisy-channel model of human sentence comprehension under uncertain input. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 234–243, Honolulu, Hawaii. Association for Computational Linguistics. https://doi.org/10.3115/1613715.1613749

Tal Linzen. 2019. What can linguistics and deep learning contribute to each other? Response to Pater. *Language*, 95(1). https://doi.org/10.1353/lan.2019.0001, https://doi.org/10.1353/lan.2019.0015

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535. Cambridge, MA Publisher: MIT Press. https://doi.org/10.1162/tacl_a_00115

Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, Rui Wang, and Hai Hu. 2024. ZhoBLiMP: A systematic assessment of language models with linguistic minimal pairs in Chinese.

Kyle Mahowald, Jeremy Hartman, Peter Graff, and Edward Gibson. 2016. SNAP judgments: A small N acceptability paradigm (SNAP) for linguistic acceptability judgments. *Language*, pages 619–635. https://doi.org/10.1353/lan.2016.0052, https://doi.org/10.1353/lan.2016.0051

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540. https://doi.org/10.1016/j.tics.2024.01.011, PubMed: 38508911

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1151

R. Thomas McCoy and Thomas L. Griffiths. 2025. Modeling rapid language learning by distilling Bayesian priors into artificial neural networks. *Nature Communications*, 16(1):4676. https://doi.org/10.1038/s41467-025-59957-y, PubMed: 40393968

Adam Pauls and Dan Klein. 2012. Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 959–968, Jeju Island, Korea. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlícek, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The FineWeb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Steven T. Piantadosi. 2024. *Modern Language Models Refute Chomsky's Approach to Language*, chapter 15. Language Science Press, Berlin.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report.

Carson T. Schütze. 2016. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistics Methodology*. Number 2 in Classics in Linguistics. Language Science Press, Berlin. https://doi.org/10.26530/OAPEN_603356

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. SLING: Sino linguistic evaluation of large language models. In *Proceedings of the 2022 Conference on*

Empirical Methods in Natural Language Processing, pages 4606–4634, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.305

Jon Sprouse and Diogo Almeida. 2012. Assessing the reliability of textbook data in syntax: Adger's core syntax. *Journal of Linguistics*, 48(3):609–652. https://doi.org/10.1017/S0022226712000011

Jon Sprouse, Carson T Schütze, and Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua*, 134:219–248. https://doi.org/10.1016/j.lingua.2013.07.002

Karolina Stańczak, Kevin Du, Adina Williams, Isabelle Augenstein, and Ryan Cotterell. 2024. The causal influence of grammatical gender on distributional semantics. *Transactions of the Association for Computational Linguistics*, 12:1672–1685.

Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L. Frank. 2025. BLiMP-NL: A corpus of dutch minimal pairs and acceptability judgments for language model evaluation. *Computational Linguistics*, pages 1–35. https://doi.org/10.1162/coli_a_00559

Lindia Tjuatja, Graham Neubig, Tal Linzen, and Sophie Hao. 2025. What goes into a LM acceptability judgment? Rethinking the impact of frequency and length. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2173–2186, Albuquerque, New Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.109

Héctor Vázquez Martínez, Annika Lea Heuser, Charles Yang, and Jordan Kodner. 2023. Evaluating neural language models as cognitive models of language acquisition. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 48–64, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.genbench-1.4

Alex Warstadt and Samuel R. Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In Shalom Lappin and Jean-Philippe Bernardy, editors, *Algebraic Structures in Natural Language*. Taylor & Francis. https://doi.org/10.1201/9781003205388-2

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8. Publisher: MIT Press. https://doi.org/10.1162/tacl_a_00321

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641. Cambridge, MA. Publisher: MIT Press. https://doi.org/10.1162/tacl_a_00290

Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D. Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, Benjamin Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi. 2024. The generative AI paradox: ''What it can create, it may not understand''. In *The Twelfth International Conference on Learning Representations*.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2023. Using computational models to test syntactic learnability. *Linguistic Inquiry*, pages 1–44.

Liang Xu, Xuanwei Zhang, and Qianqian Dong. 2020. CLUECorpus2020: A large-scale Chinese corpus for pre-training language model. ArXiv preprint.

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. UER: An open-source toolkit for pre-training models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 241–246, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-3041

A Derivation of Predictions

Below we provide full derivations of each prediction from our framework. We first lay out several assumptions we make use of in our derivations. We believe these assumptions are generally plausible; moreover, cases where they do not apply might be informative for understanding LM behavior.

Assumption 2. *Most realizations of intended messages are grammatical.*

Formally, $P(G = 0|m)$ is relatively small.

Assumption 3. *Probability of grammatical realization does not vary dramatically with intended message.*

Formally, we will require that the covariance of $\log P(G = 0|m)$ and $\log P(m)$ is smaller than the variance of $\log P(m)$.

Taking Assumptions 2 and 3 together, we will write $P(G = 0|m) \approx \epsilon$, and thus $P(G = 1 \mid m) \approx (1 - \epsilon)$, for some $\epsilon > 0$. This leaves implicit that ϵ potentially varies with m,[9] but this variability is relatively small, compared to variability in the overall probability of m. We will additionally assume that $\epsilon \ll (1 - \epsilon)$.

Assumption 4. *The strings of interest are not in a dense ''error neighborhood'' of strings that grammatically encode messages of considerably higher probability than the strings' preferred messages.*

For any string of interest $\mathbf{s}$, let $\mathcal{M}^d$ denote the set of messages $\{m_i : \mathcal{D}(\mathbf{s}^{*(m_i)} \to \mathbf{s}|m_i) = d\}$, let $m^* = \arg\max_m P(m|\mathbf{s})$, and let $P(\mathcal{M}^d) = \sum_{m \in \mathcal{M}^d} P(m)$. Formally, we will require that $\frac{P(\mathcal{M}^d)}{K^d} \ll \frac{1}{\epsilon} P(m^*)$ for all d.

Assumption 5. *Regions distant in ''error space'' are not dramatically higher in message probability relative to their error distance.*

Formally, we require that $\frac{P(\mathcal{M}^d)}{K^d}$ does not grow exponentially fast with rate $\frac{1}{\epsilon}$.

Assumption 6. *For minimal pairs, intended message probabilities for grammatical strings of interest do not vary dramatically in how characteristic they are of message probabilities in the immediate error neighborhood of the ungrammatical string in the minimal pair.*

Formally, for a minimal pair $(\mathbf{s}, \mathbf{s}')$ with intended message $m^* = \arg\max_m P(m|\mathbf{s})$, $\frac{P(\mathcal{M}^1)}{P(m^*)}$ can be treated as constant.

A.1 Prediction 1

Prediction 1. *Correlation between the log-probability of grammatical and ungrammatical strings within a minimal pair after controlling for meaning.*

Let $(\mathbf{s}, \mathbf{s}')$ be a minimal pair, where $m^* = \arg\max_m P(m|\mathbf{s})$. The probability of a string is given by Equation (1), reproduced below:

$$P(\mathbf{s}) = \sum_{m \in \mathcal{M}, g \in \{0,1\}} P(\mathbf{s}|m, g) P(g|m) P(m) \tag{10}$$

[9] For example, a message that would be realized by a triply-center-embedded sentence might be more likely to involve errorful realizations.

By Assumption 1, when $G = 1$ we only need to consider probability from m^*, and $P(\mathbf{s}|m^*, G = 1) = 1$. Therefore, we can simplify this as:

$$\begin{aligned} P(\mathbf{s}) = P(G = 1|m^*)P(m^*) \\ + \sum_{m \in \mathcal{M} \setminus \{m^*\}} P(\mathbf{s}|m, G = 0)P(G = 0|m)P(m) \end{aligned} \tag{11}$$

By Assumption 3, we have $P(G = 0|m) \approx \epsilon$ and $P(G = 1 \mid m) \approx (1 - \epsilon)$, so we can further write:

$$\begin{aligned} P(\mathbf{s}) \approx (1 - \epsilon)P(m^*) \\ + \epsilon \sum_{m \in \mathcal{M} \setminus \{m^*\}} P(\mathbf{s}|m, G = 0)P(m) \end{aligned} \tag{12}$$

Now, consider the second term. The intuition is that we will group the possible messages according to how many ''error steps'' d their grammatical realization string is from $\mathbf{s}$. Using this grouping, we rewrite the sum as

$$\sum_{m \in \mathcal{M} \setminus \{m^*\}} P(\mathbf{s}|m, G = 0)P(m) = \sum_{d \in \{1,2,\dots\}} \sum_{m_i^d \in \mathcal{M}^d} P(\mathbf{s}|m_i^d, G = 0)P(m_i^d) \tag{13}$$

Applying Equation (3), we can write:

$$P(\mathbf{s}|m_i^d, G = 0) \approx \frac{1 - \epsilon}{K} \left(\frac{\epsilon}{K}\right)^{d-1} \tag{14}$$

and we can simplify Equation (12) to:

$$P(\mathbf{s}) \approx (1 - \epsilon)P(m^*) + (1 - \epsilon) \sum_{d \in \{1,2,\dots\}} \left(\frac{\epsilon}{K}\right)^d P(\mathcal{M}^d) \tag{15}$$

Invoking Assumption 4, we can bound the second term by a value much smaller than the first:

$$(1 - \epsilon) \sum_{d \in \{1,2,\dots\}} \left(\frac{\epsilon}{K}\right)^d P(\mathcal{M}^d) \ll (1 - \epsilon) \sum_{d \in \{1,2,\dots\}} \epsilon^{d-1} P(m^*) \tag{16}$$

Therefore, the whole second term in Equation (15) can be dropped in the approximation, giving us:

$$P(\mathbf{s}) \approx (1 - \epsilon)P(m^*) \tag{17}$$

and since ϵ is small, we further approximate:

$$\log P(\mathbf{s}) \approx \log P(m^*) \tag{18}$$

Now, let us turn to the ungrammatical member of the pair $P(\mathbf{s}')$. To begin, the probability of the ungrammatical member of the pair $P(\mathbf{s}')$ is the same as in Equation (1). By Definition 2, we can simplify this as:

$$P(\mathbf{s}') = \sum_{m \in \mathcal{M}} P(\mathbf{s}'|m, G = 0)P(G = 0|m)P(m) \tag{19}$$

We will use similar machinery as above, giving us an expression equivalent to the second term of Equation 15, and we further drop the factor $(1 - \epsilon)$:

$$P(\mathbf{s}') \approx \sum_{d \in \{1,2,\dots\}} \left(\frac{\epsilon}{K}\right)^d P(\mathcal{M}^d) \tag{20}$$

Invoking Assumption 5, this sum decreases exponentially quickly with d and so we drop the terms where $d > 1$, giving us:

$$P(\mathrm{s}') \approx \left(\frac{\epsilon}{K}\right) P(\mathcal{M}^1) \tag{21}$$

Now, define $A = \frac{1}{K}\frac{P(\mathcal{M}^1)}{P(m^*)}$, which by Assumption 6 we will treat as constant.

$$P(\mathrm{s}') \approx \epsilon \cdot A \cdot P(m^*) \tag{22}$$
$$\log P(\mathrm{s}') \approx \log \epsilon + \log P(m^*) + \log A \tag{23}$$

The formula for the covariance of two random variables X and $Z = X + Y + c$ for a constant c is given by:

$$\rho_{X,Z} = \frac{Var(X) + Cov(X,Y)}{\sqrt{Var(X)(Var(X) + Var(Y) + 2Cov(X,Y))}} \tag{24}$$

In our case, let $X = \log P(m^*)$ and $Y = \log \epsilon$. Then X corresponds to $P(\mathrm{s})$ by Equation (18) and Z corresponds to $P(\mathrm{s}')$ by Equation (18). By Assumption 3, we assume that $Var(X)$ is larger than the magnitude of $Cov(X, Y)$, so $\rho_{X,Z}$ (and the resulting correlation) will be positive.

A.2 Prediction 2

Prediction 2. *Correlation between differences in log-probability and human acceptability judgments.*

Consider two strings, s and s′. We assume that the human acceptability judgment of a string Acc(s) reflects two factors: (i) the plausibility of the intended message, and (ii) how well a linguistic form conveys the intended message.

We operationalize (i) with $\log P(m^*)$, where $m^* = \arg\max_m P(m|\mathrm{s})$ is the inferred message of the string; and (ii) with the number of errors $d = \mathcal{D}(\mathrm{s}' \to \mathrm{s}|m^*)$. So we have

$$\log P(\mathrm{s}') \approx \log P(m^*) + d \log \frac{\epsilon}{K} \tag{25}$$

where the two terms on the right-hand side reflect factors (i) and (ii). If the weights of these factors are W_i and W_{ii} respectively, then, calling the log-probability ‘‘error cost’’ $E = d \log \frac{\epsilon}{K}$, we can write the acceptability of s′, Acc(s′), as:

$$\mathrm{Acc}(\mathrm{s}') \approx W_i \log P(m^*) + W_{ii} \log E \backslash \tag{26}$$

As a special case, when the message is grammatically realized, the error distance $E = 0$ and $\mathrm{Acc}(\mathrm{s}') \approx W_i \log P(m^*)$.

Based on these assumptions, if one compares the acceptabilities of the grammatical string s and the ungrammatical string s′ in a minimal pair, factor (i) cancels out in the acceptability difference:

$$\begin{aligned} \mathrm{Acc}(\mathrm{s}) &\approx W_i \log P(m^*) \\ \mathrm{Acc}(\mathrm{s}') &\approx W_i \log P(m^*) + W_{ii}E \\ \mathrm{Acc}(\mathrm{s}) - \mathrm{Acc}(\mathrm{s}') &\approx W_{ii}E \end{aligned} \tag{27}$$

and the acceptability difference should correlate with the sentence log-probability difference, E, which is Prediction 2a. For pairs that are not meaning-matched, however, we have two different meanings and so factor (i) will not cancel out, weakening the correlation between log-probability difference and acceptability difference. If we further assume that larger differences in message tends to imply larger differences in message log-probability, then correlations between acceptability difference and log-probability difference will drop with larger differences in message, which is Prediction 2b.

A.3 Prediction 3

Prediction 3. *Potentially poor separation based on probability between grammatical/ungrammatical strings.*

We consider a pooled set of grammatical and ungrammatical strings. We can measure the ''separation'' based on $P(P(g) > P(u))$, where g is a grammatical string drawn from the pool, u is an ungrammatical string drawn from the pool, and the two draws are independent. Note that this is equivalent to the area under the receiver operating characteristic (ROC) curve for the probabilities assigned to each string in the set. If this quantity is 1, this means that $P(g)$ is always greater than $P(u)$.

We first analyze the case for minimal pairs, where there is the greatest reason to expect good separation *a priori*. Using similar logic as above, we have:

$$P(g) \approx (1-\epsilon)P(m_g) \tag{28}$$

$$P(u) \approx \frac{\epsilon}{K}P(m_u) \tag{29}$$

where m_g is the unique message associated with g, and m_u is the highest probability message given u, i.e., $m_u = \arg\max_m P(u|m)$.

Therefore,

$$\log \frac{P(g)}{P(u)} \approx \log P(m_g) - \log P(m_u) - \log \epsilon + \log K \tag{30}$$

and so, approximately, $P(g) > P(u)$ iff:

$$\log P(m_g) - \log P(m_u) > \log \epsilon - \log K \tag{31}$$

When the population of strings consists of minimal pairs, m_g and m_u are drawn from the same distribution $P(m)$, so $\log P(m_g) - \log P(m_u)$ is symmetric around 0. Furthermore, since $0 < \epsilon < 1$, and $K \geq 1$, we know that $\log \epsilon - \log K$ must be negative. For any symmetric distribution, with $\mu = 0$, and PDF f, $\int_{-\infty}^{0} f(x)dx < 0.5$. Therefore, it is guaranteed that the AUC $P(P(g) > P(u)) > 0.5$. However, its actual value depends greatly on the variance of $\log P(m)$, and might well not be much above 0.5.

This analysis also offers a way of understanding why and under what conditions minimal pairs offer good testing grounds for LMs' grammatical capabilities. ''Better'' LMs can be expected to place higher probability on strings that (i) are grammatical, and (ii) convey plausible meanings. Factor (i) means that ''better'' LMs will have lower values of ϵ; factor (ii) will, if anything, tend to narrow the range of log-probabilities of plausible meanings (pushing them to higher values), and hence reduce the variance of $\log P(m)$. Among LMs clearing a reasonable minimum bar of quality, K is in contrast best thought of as a property of the grammatical strings in the set of minimal pairs, not as a property differing among LMs, so it is a constant for purposes of this analysis. Therefore, both factors (i) and (ii) will increase AUC.

This analysis can be broadened to populations of strings beyond just sets of minimal pairs: In those cases, m_g and m_u are not drawn from the same distribution, so there is even less reason to expect high AUC. Indeed, if $P(m_u)$ tends to be higher than $P(m_g)$, we might even see AUC < 0.5.

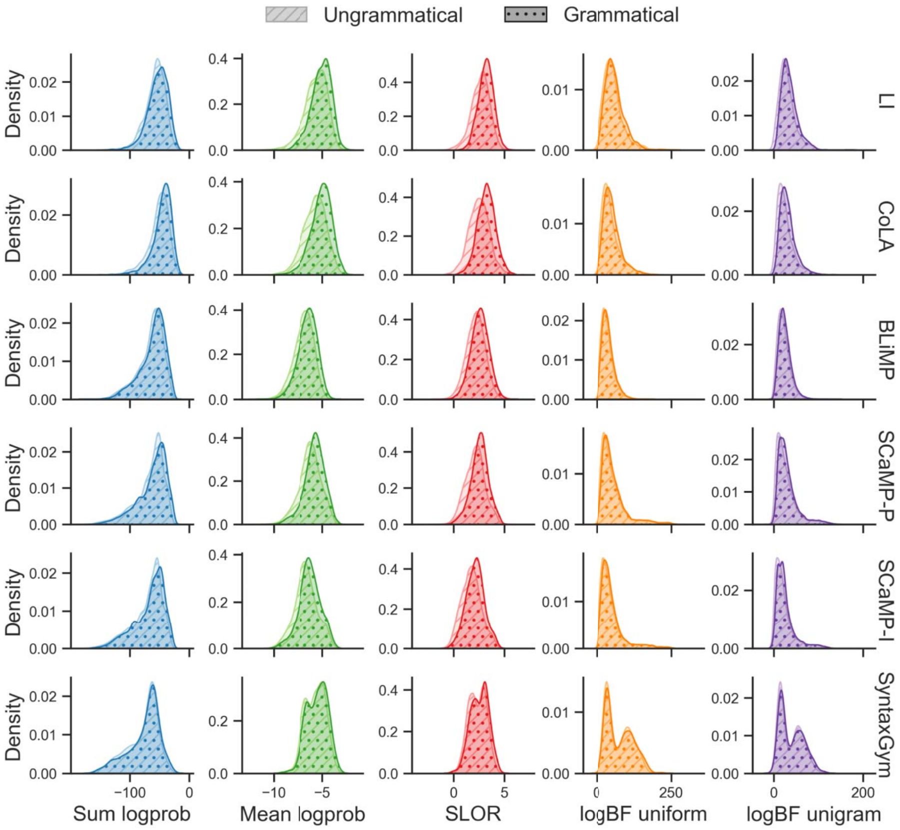
Generalizing to Transformations of String Probability. As shown above, for any LM that achieves $1 - \epsilon > \frac{\epsilon}{K}$, the limiting factor on AUC is the distributions of $\log P(m_g)$ and $\log P(m_u)$. For the case of minimal pairs, replacing the two string log-probability terms on the left-hand side of Equation (31) with random variables that are lower-variance will generally increase AUC: the two terms are drawn from the same distribution, thus their expectation is zero, and thus the lower their variance the greater the probability will tend to be that the inequality is satisfied.

Two transformations of interest described in Section 6.1 can be analyzed fairly straightforwardly: dividing string probability by probability under a uniform per-token distribution, and dividing string probability by its probability under a unigram per-token distribution (our Metrics 2 and 3). These transformations effectively replace LM log-probability with the difference of uniform or unigram log-probability from LM log-probability. In general, for two random variables X and Y with standard deviations σ_X and σ_Y and correlation ρ, we have $\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y - 2\rho\sigma_X\sigma_Y$. Therefore, $\sigma^2_{X-Y} < \sigma^2_X$ if $\sigma_Y < 2\rho\sigma_X$. That is, the random variables must be positively correlated, and Y's variance must not be too large relative to that of X. These criteria are good candidates for being met by both transformations. Among grammatical strings, those with lower LM probabilities tend to be longer and haver rarer words, so the distribution of grammatical string probabilities under an LM is likely to be positively correlated with the distributions of grammatical string probabilities under both uniform and unigram distributions. And, since LMs are better models than uniform and unigram models, LM log-probabilities among a grammatical set of strings will tend to have lower variance—clustered more tightly toward 0—than the log-probabilities of uniform or unigram distributions over the same set of grammatical strings.

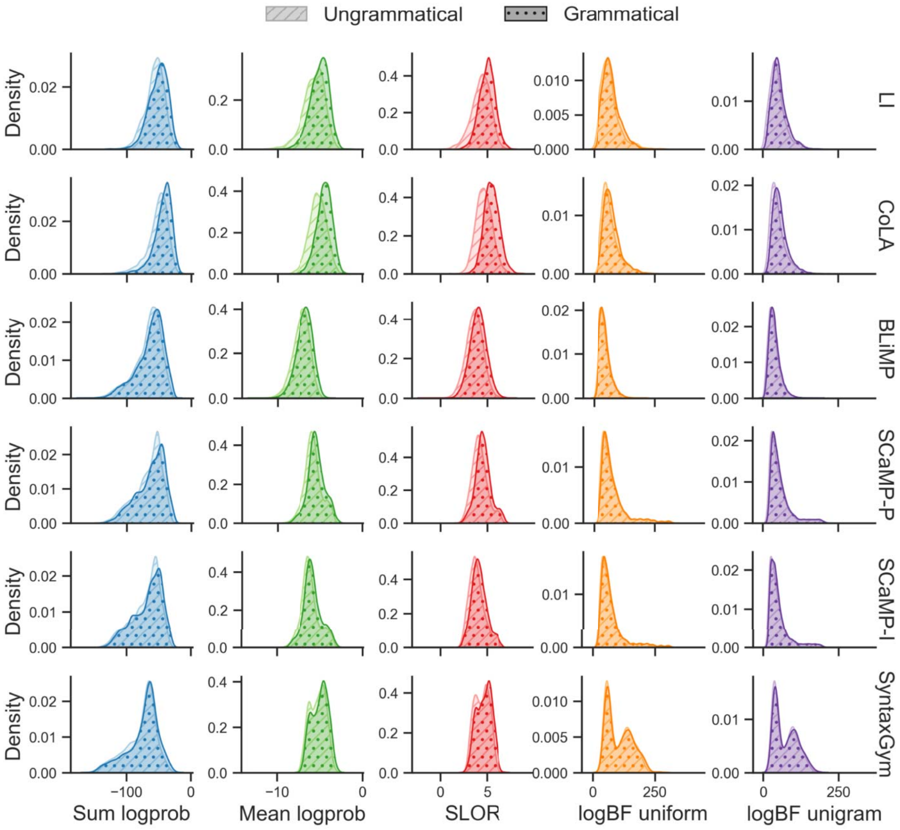
We have found Metrics 4 and 5 less straightforward to analyze, but speculate that a generally similar approach may be feasible.

No corresponding general results can be offered for analysis of AUC among pools of strings not consisting of sets of minimal pairs, because of the differences in distributions of $P(m_g)$ and $P(m_u)$. Indeed, sets of strings adversarial to a metric could be constructed; for example, selectively including only grammatical strings involving frequent words and only ungrammatical strings involving rare words would be adversarial to Metric 3. However, we expect that these approaches will generally tend to improve AUC over that obtained by using raw string probabilities, when sentence sets are not adversarially constructed.

B Additional Figures for Prediction 3

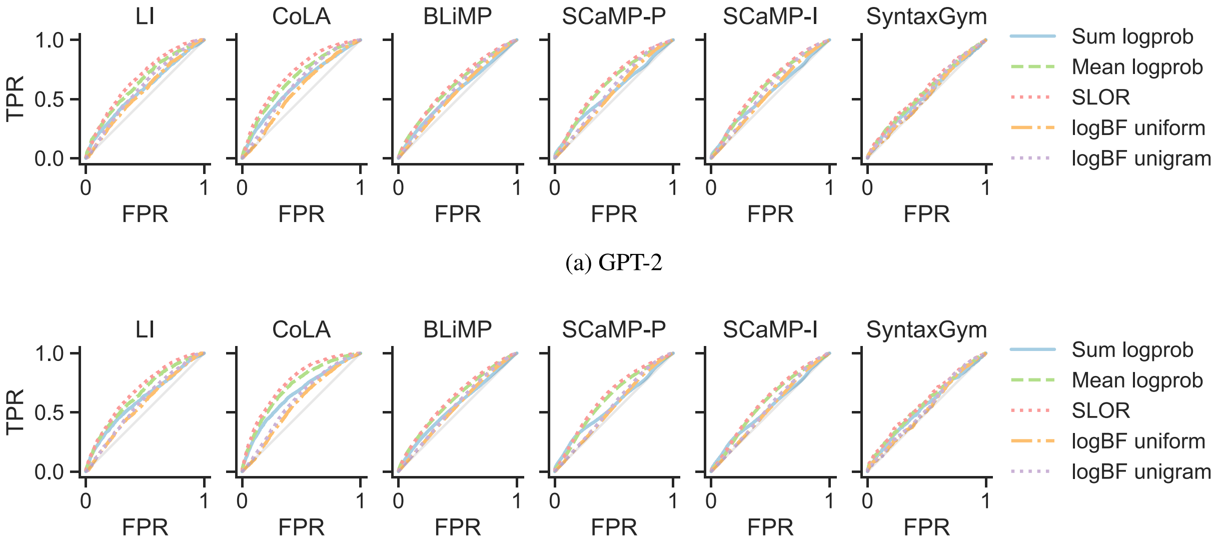


(a) GPT-2



(b) Llama-3-70B

Figure 5: Score distributions for grammatical and ungrammatical sentences from each English dataset.



(a) GPT-2

(b) Llama-3-70B

Figure 6: ROC curves achieved by models on each English dataset.

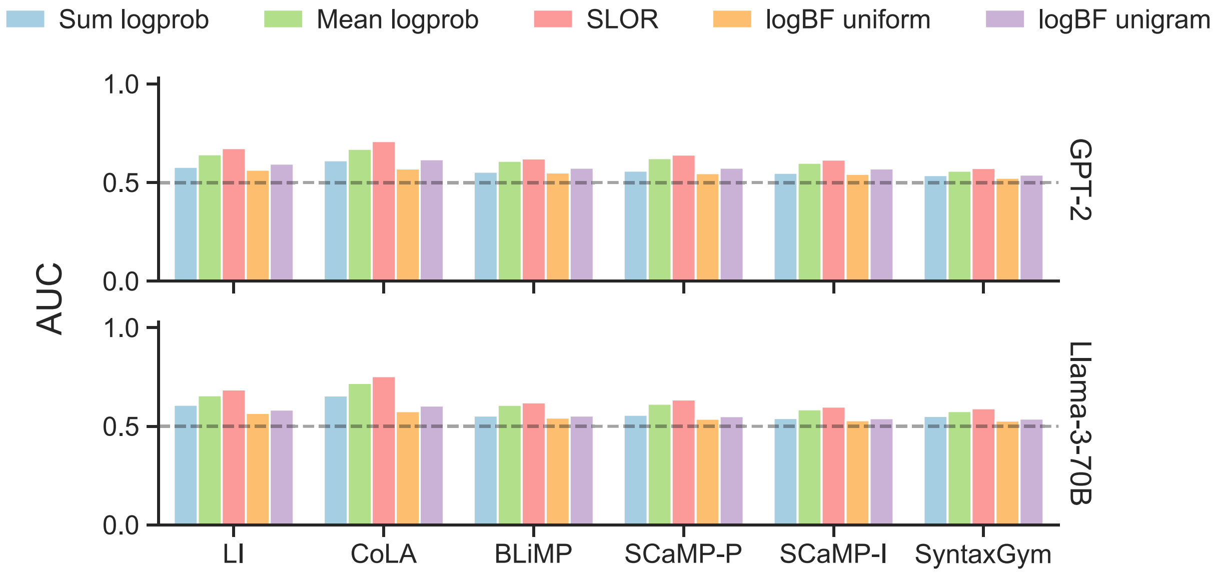


Figure 7: AUC scores for each model (rows), metric (hue), and dataset (x-axis), corresponding to ROC curves in Figure 6.

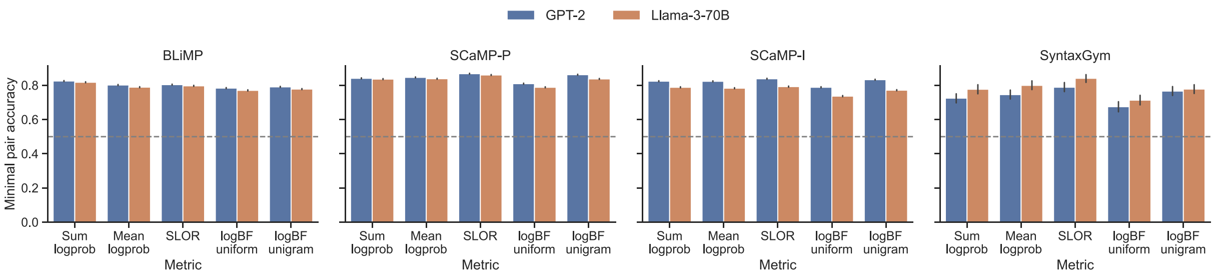


Figure 8: Accuracy (i.e., proportion of pairs where the grammatical sentence receives a higher score) achieved on English minimal pair datasets, using the five scoring functions discussed in Section 6.1.