Can You Learn Semantics Through Next-Word Prediction? The Case of Entailment

Anonymous ACL submission

Abstract

Do LMs infer the semantics of text from co-occurrence patterns in their training data? Merrill et al. (2022) argue that, in theory, probabilities predicted by an optimal LM encode semantic information about entailment relations, but it is unclear whether neural LMs trained on corpora learn entailment in this way because of strong idealizing assumptions made by Merrill et al.. In this work, we investigate whether their theory can be used to decode entailment judgments from neural LMs. We find that a test similar to theirs can decode entailment relations between natural sentences, well above random chance, though not perfectly, across many datasets and LMs. This suggests LMs implicitly model aspects of semantics to predict semantic effects on sentence co-occurrence patterns. However, we find the test that predicts entailment in practice works in the opposite direction to the theoretical test. We thus revisit the assumptions underlying the original test, finding its derivation did not adequately account for redundancy in human-written text. We argue that correctly accounting for redundancy related to *explanations* might derive the observed flipped test and, more generally, improve linguistic theories of human speakers.

1 Introduction

Inspired by the empirical capabilities of language models (LMs) trained on next-word prediction, recent work has examined if and how linguistic meaning might be inferred from raw text (Bender and Koller, 2020; Merrill et al., 2021; Pavlick, 2022; Wu et al., 2023, inter alia). A text corpus is the result of humans using text to communicate information, and doing this efficiently requires following pragmatic principles like avoiding contradictory or redundant sentences. Therefore, training to predict whether sentences can co-occur might lead LMs to represent semantic relationships between sentences (Harris, 1954; Potts, 2020; Michael, 2020).

But does sentence co-occurrence provide enough signal for LMs to learn to represent complex semantic phenomena like entailment? Merrill et al. (2022) derive a simple equation by which the entailment relation between two sentences can be detected using their co-occurrence probability in a corpus generated by speakers who avoid redundancy. Intuitively, non-redundant speakers will rarely utter entailed sentences, so low co-occurrence probability of two sentences is predictive of their entailment relationship. This means that, in principle, learning to model sentence co-occurrence perfectly requires an LM to implicitly model entailment, and entailment classifications can be extracted from the co-occurrence probabilities of such an LM.

However, Merrill et al.'s theoretical result has two caveats. First, it assumes an "ideal" LM that perfectly models the likelihood of texts in a language. Second, it makes the strong (but theoretically motivated; Grice, 1975) assumption that speakers always avoid redundancy. It is thus unclear whether real LMs infer a model of entailment from sentence co-occurrence probabilities in their training data, both because LMs may misestimate probabilities and because the required assumptions about human speakers may be too simplified.

In this work, we empirically evaluate the distributional entailment test from Merrill et al. (2022): can we use it to classify entailment from LM probability estimates? Overall, we find across a wide range of entailment benchmarks and LMs that a variant of the entailment test consistently detects entailment well above random chance. This suggests that LM probability judgments are sensitive to the relationships between sentence meanings that are reflected in sentence co-occurrence patterns, at least to some extent. This further suggests that next-word prediction is a strong enough objective for LMs to acquire at least a partial model of entailment relationships between sentences.

However, this result comes with a surprise.

Across many entailment benchmarks, we find that the direction of the test is *flipped* compared to Merrill et al.'s theoretical test: higher co-occurrence probabilities correlate with entailment when the *opposite* is expected! We take this as evidence against a theory of human speakers based purely on minimizing redundancy. Analyzing natural corpora, we find humans are often more redundant than Merrill et al.'s non-redundant speakers, which could explain the flipped test. We present a preliminary account of how better accounting for explanations (one observed type of redundancy) might predict the flipped test. Overall, our results motivate future work in computational pragmatics accounting for redundancy and are a case study for how the data aggregated about many speakers in LMs can be used to test and develop pragmatic theories.

2 Distributional Semantics and the Entailment Test

There is an old debate in linguistics and NLP about whether distributional semantics—the idea that text co-occurrence patterns can contain semantic information—captures semantics in any true sense (Brunila and LaViolette, 2022). This debate goes back at least to Harris (1954), who argues that sentence co-occurrences patterns in a corpus could be used as data to build a linguistic theory of semantics, but it has been revisited in recent years in terms of LMs. In particular, Bender and Koller (2020)—in disagreement with Harris (1954)—took a strong stance against the claim that LMs "understand" language because understanding requires modeling communicative intent or at least conventionalized semantic denotations, both of which do not appear explicitly in the training data for LMs.

While it is certainly true that LMs are only trained on surface forms, counterarguments to Bender and Koller (2020) have been given for how LMs might be able to reconstruct semantic information from their training data. One line of counterarguments (Potts, 2020; Michael, 2020; Merrill et al., 2022) echoes Harris (1954), positing that sentence co-occurrence probabilities contain information about semantics because speakers aim to be truthful and informative and are thus unlikely to produce contradictory or redundant pairs of sentences. Properly learning which sentences can co-occur (part of LM training) thus amounts to acquiring a *semantic* representation of which sentences are contradictory or redundant with one another. Merrill et al. (2022, CoNLL slides) motivate this claim with the following example:

(1) I have two cats.
  a. **I don't have a cat.*
  b. **I have a cat.*
  c. One is orange.

Example 1a is unlikely to be uttered because it contains a contradiction. More subtly, Example 1b is unlikely because its second sentence is uninformative given the first, even though they are consistent. In contrast, Example 1c is acceptable because it is consistent *and* adds new information. Thus, Example 1 suggests sentence co-occurrence is governed by semantic constraints against inconsistency and redundancy. If strong LMs correctly model such co-occurrences, they might need an implicit model of sentence semantics to determine these properties.

2.1 The Entailment Test

One way to define semantic competency is the ability to resolve entailment relations between pairs of sentences. This simple idea has a long history both in both the philosophy of language (Van Benthem, 1986; Brandom, 2000) and NLP evaluation (Dagan et al., 2010). Drawing on the semantic nature of sentence co-occurrence and its connection to redundancy, Merrill et al. (2022) derive a test to check whether sentence x entails sentence y using their co-occurrence probability in a corpus produced by so-called *Gricean speakers*. If we accept the idea that the ability to evaluate entailment captures semantics in full, this test establishes semantics, can, in principle, be inferred from next-word prediction.

Gricean Speakers. Gricean speakers are a computational model of human speakers implementing fundamental principles for effective communication (the Gricean maxims; Grice, 1975). The maxims say that a speaker should convey as much relevant information as possible without saying too much, among other desiderata. Following standard computational choices in rational theories of pragmatics (Goodman and Frank, 2016), Merrill et al. (2022) operationalized these principles by modeling the probability of a text z produced by a Gricean speaker as a function of the text's information content and cost:

- **Information content:** Sentences with a lot of new information for the listener about the speaker's beliefs are more likely. This penalizes untruthful, uninformative, and redundant

sentences. Let $i_\ell(y \mid x, w)$ be the information y conveys to the listener given beliefs w and context x, which speakers aim to *maximize*.

- **Cost:** Long or complex sentences should be less likely so that speakers do not produce informative, but verbose, text. The model assumes a function $c(y)$ that gives the cost of sentence y, which speakers aim to *minimize*.

Under Merrill et al. (2022)'s model, a Gricean speaker utters y (having said x) with probability

$$p(y \mid x, w) \propto \exp(i_\ell(y \mid x, w) - c(y)).$$

A sequence of sentences $z_1 \cdots z_n$ occurs in a corpus generated by Gricean speakers with probability

$$p(z) = \mathop{\mathbb{E}}_{w}\left[\prod_{i=1}^{n} p(z_i \mid z_{<i}, w)\right].$$

Let \$ denote a special "end-of-text" sentence.

Entailment Test. Assuming a corpus is sampled from a collection of Gricean speakers with different beliefs, Merrill et al. (2022) derive the following measure $\hat{E}_p(x, y)$ for detecting entailment purely using log probabilities of sentence co-occurrences:

$$\begin{aligned}\hat{E}_p(x, y) = &\log p(xy) - \log p(x\$) \\ &- \log p(yy) + \log p(y\$).\end{aligned} \tag{1}$$

A $\sim$0 score means entailment. The first two terms $\approx \log p(y \mid x)$ and the last two $\approx -\log p(y \mid y)$. This gives some intuition for the test: 0 means xy is as redundant as yy, i.e., x entails y (see §A).

3 Evaluating the Entailment Test

Merrill et al. (2022) showed their test could detect entailment from n-gram LMs trained on synthetic data generated by Gricean speakers. Although Gricean speakers capture some principles of how humans speak, they are likely simplistic compared to real language use. Additionally, real LMs may misestimate the co-occurrence probabilities used by the test. For both of these reasons, it is unclear whether the entailment test should correctly detect entailment on natural sentences given LM-estimated probabilities. We thus evaluate the entailment test with probabilities computed by real LMs on natural-language entailment benchmarks.

3.1 Entailment Datasets

We first evaluate the entailment test on existing *broad-coverage* entailment datasets built by crowd workers: RTE (Dagan et al., 2010), MNLI (Williams et al., 2018), WaNLI (Liu et al., 2022), and ANLI (Nie et al., 2020).[1] Unless otherwise mentioned, we always use the training set. We collapse three-way label distinctions (entailment, neutral, contradiction) to entailment or non-entailment. We also evaluate on *targeted synthetic* entailment datasets designed to test specific kinds of entailment à la GLUE diagnostics (Wang et al., 2018): specifically, entailment related to the logical connectives *and*/*or*, the quantifiers *all*/*some*, numbers, passivization, and datives (details in §G). We reported dataset statistics in §I.

3.2 Models

We evaluate the entailment test with probabilities computed by a diverse suite of LMs: GPT-2 (Radford et al., 2019), OPT (Zhang et al., 2022), Llama-1 (Touvron et al., 2023a), Vicuna (Chiang et al., 2023), Llama-2, and Llama-2-Chat (Touvron et al., 2023b). The LMs vary in size, pretraining data, and whether and how they undergo an "alignment" process (i.e., instruction-tuning or RLHF). For each LM family, we use both the smallest and the largest publicly available LM (see §H for a list).

3.3 Evaluation Metric: Flipped ROC-AUC

The entailment test does not directly classify entailment but gives a score where $\sim$0 suggests entailment and higher values suggest non-entailment. This can be converted to a classifier by choosing a decision boundary for entailment, but the choice of a threshold is arbitrary. To evaluate the test, we thus use the standard ROC-AUC metric, which can be understood to evaluate the score holistically across different choices of the threshold. There is an inherent tradeoff between precision and recall with the choice of the threshold, and ROC-AUC provides a consistent way to evaluate without arbitrarily fixing the threshold. Independent of the class imbalance, ROC-AUC ranges from 0 to 100 where 50 is random chance. In many cases, we found that the flipped entailment score (meaning Equation (1) with the sign of each term flipped) detected entailment better than the original score (§4.1). We thus report the ROC-AUC score of the flipped score, which we call *flipped ROC-AUC*.

[1] For ANLI, we use the data collected in the third round.

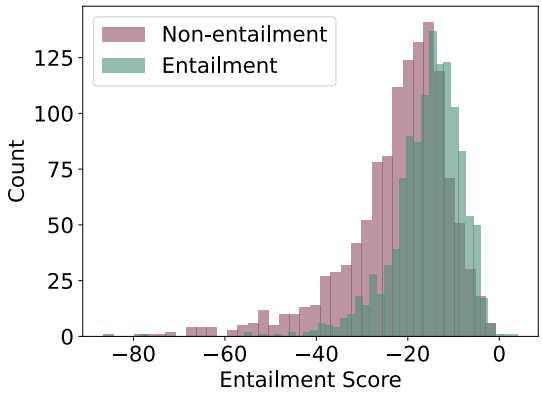

Figure 1: Entailment score $\hat{E}_p(x, y)$ distribution computed with Llama2-70b probabilities on RTE. **The score discriminates the two classes, though imperfectly.**

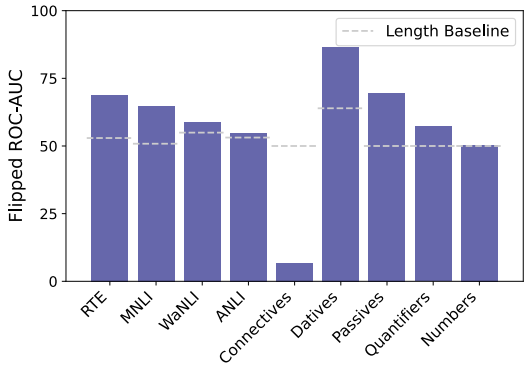

Figure 2: Flipped AUC-ROC scores of the flipped entailment test across datasets using Llama2-70b probabilities. **The flipped test generally performs above random** (=50) and the length baseline, while the original test works better for connectives (represented by <50 Flipped ROC-AUC).

4 Entailment Test Results

Overall, we find the test predicts entailment on the broad-coverage datasets, but only when the test score is *flipped* compared to the theoretical test (i.e., a larger score means entailment). However, the pattern is more complicated for the targeted data, where some constructions follow the flipped trend but others follow the original, unflipped test.

4.1 Flipped Test on Broad-Coverage Data

Figure 1 shows the entailment score $\hat{E}_p(x, y)$ for the RTE training data using Llama2-70b probabilities. The score distinguishes the two classes, but not perfectly. However, the theory predicts smaller $\hat{E}_p(x, y)$ for entailment vs. non-entailment, which is *flipped* in Figure 1 (which we try to account for in §6). We find this holds consistently across the broad-coverage datasets: the flipped entailment test detects entailment above random chance and a length baseline[2] (Figure 2). We also hypothesize the entailment test should be more predictive for better LMs. Using perplexity on the C4 validation set (Raffel et al., 2020) as the proxy for model quality, we plot their correlation in Figure 3. Across broad-coverage datasets, better perplexity is associated with higher flipped ROC-AUC. This suggests LMs that more accurately predict the next token also better model sentence co-occurrence patterns reflecting entailment.

We also evaluate how test performance emerges during training using Pythia-12b checkpoints. Figure 4 shows that ROC-AUC consistently increases as training progresses. Around 1b tokens, flipped ROC-AUC scores on RTE, MNLI, and WaNLI sharply increase together, suggesting the model undergoes a phase transition where general features useful for predicting entailment may be emerging.

[2]Computed by using the premise length, the hypothesis length, or the inverse of each, as the score, whichever of the four yields the best flipped AUC-ROC.

4.2 Varied Pattern for Targeted Phenomena

Figure 2 shows the flipped test works better for datives, passives, and quantifiers. For connectives, the unflipped test better predicts entailment. This suggests that, while the flipped test outperforms the original test in aggregate, the original theory might apply only for *some* constructions. Figure 3 shows the association between LM perplexity and flipped ROC-AUC for the targeted cases. Datives, passives, and quantifiers show a similar trend to the broad-coverage data where lower perplexity associates with higher flipped ROC-AUC, but connectives and numbers mostly follow the original test.

4.3 Learning a Distributional Entailment Test

We have seen that the distributional entailment test of Merrill et al. (2022) can detect entailment, but only when the sign of each term is flipped. We now evaluate this flipped test by comparing it to an oracle test that optimally predicts entailment. Their discrepancies would inform us about realistic LMs and data distributions. We train a small regression model that weights co-occurrence probabilities to predict entailment and inspect the learned weights.

Setup. The original entailment test can be viewed as a linear model with features ϕ and parameters θ:

$$\phi = \langle \underbrace{\log p(xy), \log p(x\$)}_{\textit{Left-hand side (LHS)}}, \underbrace{\log p(yy), \log p(y\$)}_{\textit{Right-hand side (RHS)}} \rangle$$

$$\theta = \langle 1, -1, -1, 1 \rangle.$$

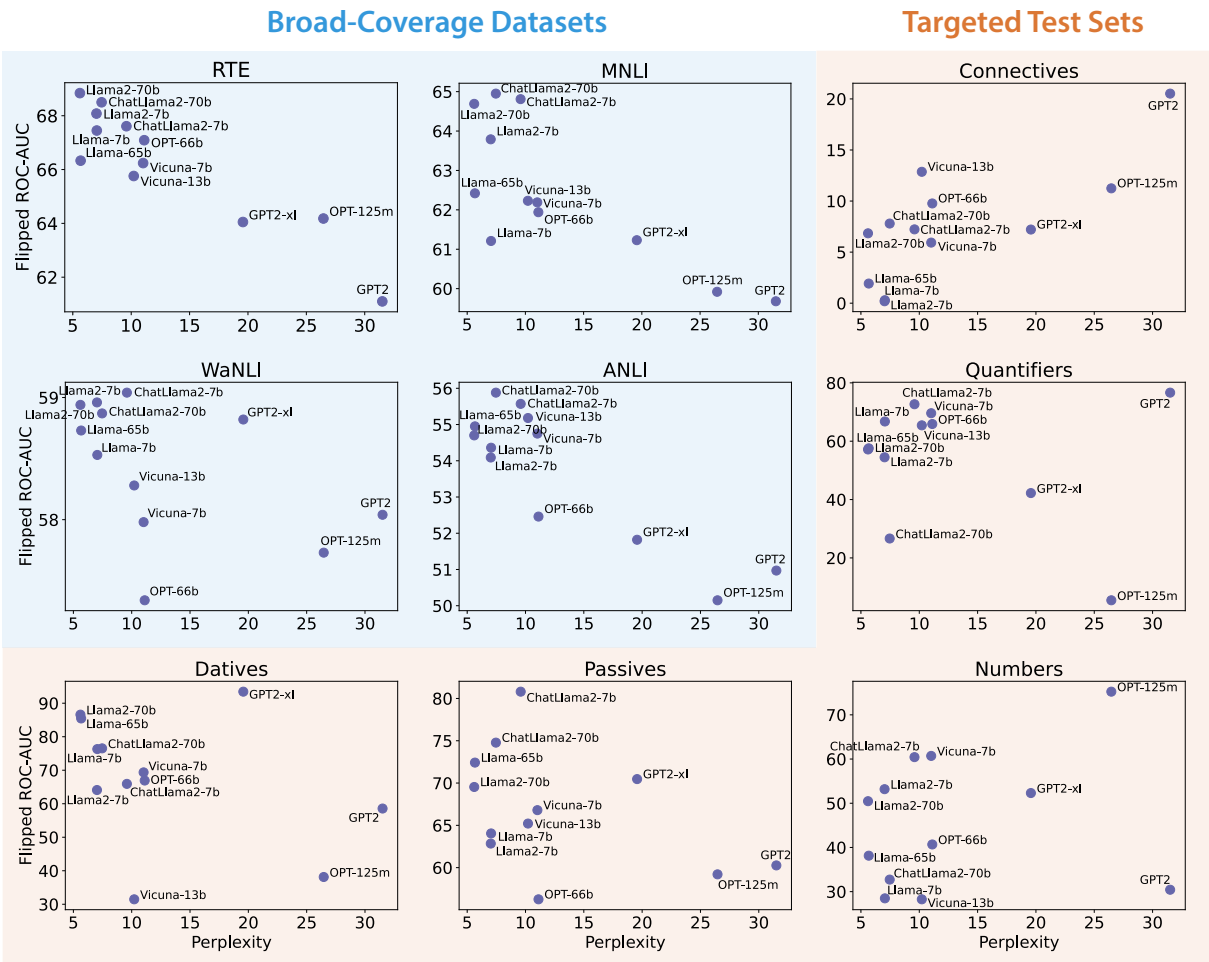

Figure 3: C4 validation perplexity vs. flipped AUC-ROC score for all models on broad-coverage and targeted datasets. Note that the scale of the y-axis differs for each subplot. See Figure 2 for a scale-controlled version of Llama2-70b results. **For broad-coverage datasets, model quality (represented by perplexity) clearly correlates with flipped test performance,** though this is more complicated for the targeted test sets.

Instead of applying the test with parameters θ (original test) or $-\theta$ (flipped test), we now *learn* parameters $\hat{\theta}$ via logistic regression on labeled entailment pairs. This learned test is *not* a standard supervised text classifier: it only gets sentence co-occurrence log-probabilities as input, not text itself.

Results. Figure 5 shows the results for the broad-coverage datasets (other datasets in §F). For the LHS, the negative xy weight matches the positive $x\$$ weight in magnitude, as for the flipped test. For the RHS, the trend is less consistent, but yy and $y\$$ generally get smaller weights than the LHS terms. Nevertheless, in aggregate, yy gets a positive weight of the same magnitude as the negative $y\$$ weight (Figure 6), as for the flipped test.

We interpret the similarity between the flipped and learned tests as evidence for the directional correctness of the flipped test. The main difference between the learned and flipped tests is that the RHS has smaller weights than the LHS for the learned test. This may be due to the transformer's learning biases and not the underlying data: Transformer LMs are prone to in-context copying (Olsson et al.,

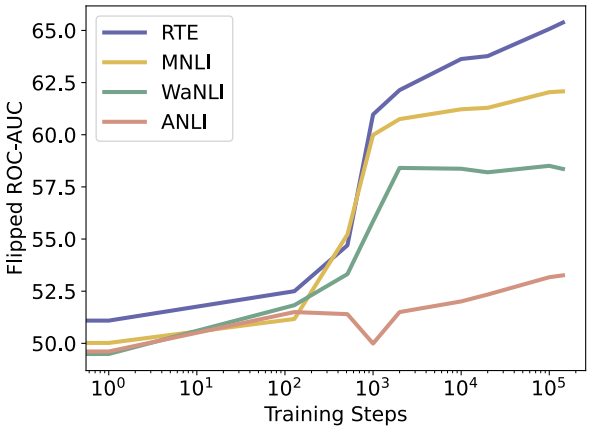

Figure 4: Flipped ROC-AUC of entailment score across Pythia-12b checkpoints. Each step is around 2M tokens.

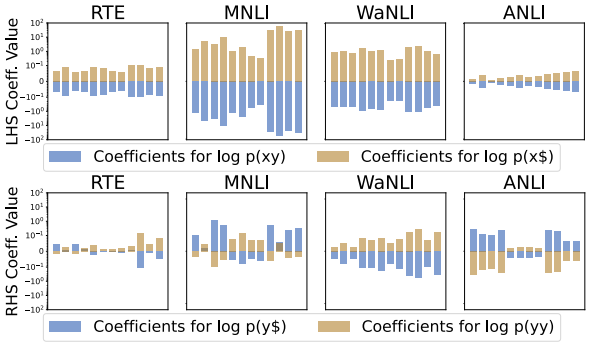

Figure 5: Learned logistic regression coefficients for the log-prob features for the broad-coverage datasets. Each bar represents one LM. For ease of visualization, y-axis is in log scale, except in $[-0.1, 0.1]$ where it is linear.

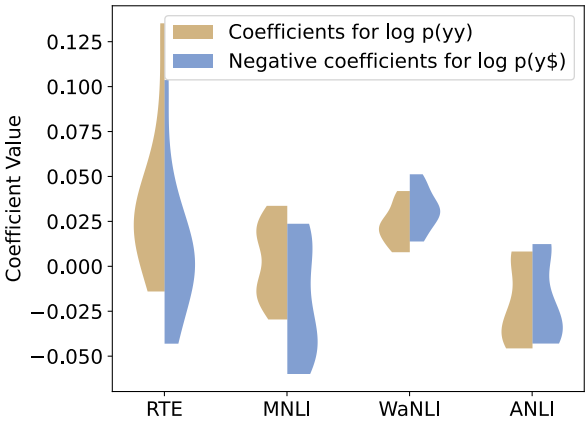

Figure 6: The RHS coefficients, for $\log p(y\$)$ and $\log p(yy)$, marginalized across all LMs.

2022) and thus might overestimate $\log p(yy)$. Reduced RHS weights may correct for this.

5 Corpus Study: Characterizing Naturalistic Linguistic Redundancy

A surprising finding from the previous section is that the entailment test is robustly flipped: entailed continuations tend to be *more likely* than non-entailed ones. This suggests the Gricean speaker assumed to derive the test may be too simplistic to account for humans. In particular, we hypothesize the disconnect may be because human speakers are *explicitly redundant* in certain contexts unlike Gricean speakers, who always avoid redundancy. We thus search for natural instances of *contextually entailed text* in corpora to better understand why real human speakers produce redundant sentences.

Data. To find contextually entailed sentences in different types of discourse, we consider a variety of web domains: Book3 (Gao et al., 2020), Wikipedia (en) (Gao et al., 2020), Multi-News (Fabbri et al., 2019) and Reuters-21578 (Hayes and Weinstein, 1991), Yahoo! Answers Topics (Zhang et al., 2016), and Yelp Reviews (Zhang et al., 2016).

Finding Contextually Entailed Text. For each document in each corpus, we construct premise and hypothesis pairs by choosing six contiguous sentences, with the first five as the premise and sixth as the hypothesis. We use entailment classifiers finetuned from T5 (Honovich et al., 2022) and RoBERTa (Liu et al., 2019) to detect entailment pairs and take the intersection of examples considered entailment by both. We then manually filter to remove incorrect entailment pairs (details in §D).

Results. As Table 1 shows, the frequency of entailed sentences is on the order of at least 10^{-3}. Even this lower bound is several orders of magnitude higher than expected for a Gricean speaker. Quite conservatively, imagine that for each entailed continuation there is at least one alternative of the same length that conveys 10 nats of information, which is quite reasonable given Shannon's lower bound estimate of 0.4 nats/character[3] (Shannon, 1951) and that typical sentences are at least 30 characters. Then the likelihood of producing an entailed sentence should be at most $1/\exp(10) \approx 10^{-5}$. This suggests the data cannot be accounted for by assuming speakers always avoid redundancy.

To better understand what is lost when assuming speakers always avoid redundancy, we inspect examples of contextually entailed text from these corpora. We find there are many reasons speakers produce entailed text. This includes both *repetition* of previous statements (44.44%[4]) and high-level *summaries* or *conclusions* (35.56%). One observed use of repetition is to emphasize an important point:

(2) **Yelp Review:** When he returned with it, he just placed it in front of me on the wet bar- no napkin/coaster, the beer was flat, and contained a FREAKING lemon. ⇒Not an orange- a lemon.

Beyond repetition, we also found examples where a weaker claim follows more specific premises:

(3) **Yelp Review:** Frankly, I'm no oyster aficionado, but after comparing with other

[3]Technically, the Gricean speaker uses semantic information, whereas Shannon's estimate captures *all* information. However, we imagine most information in text is semantic, so these are on the same order of magnitude.

[4]Percentages determined manually; see §E for details.

Data Sources	T5	RoB	$\cap$	$\cap^+$
Book3	0.40	1.31	0.33	0.27
Wikipedia (en)	0.47	1.69	0.30	0.24
Yelp Review	1.53	1.78	0.56	0.50
Multi-News	2.11	2.82	2.11	1.88
Reuters-21578	0.64	1.53	0.51	0.38
Yahoo! Answers	1.63	8.16	0.82	0.82

Table 1: Percentage of sentences entailed by their immediate context. $\cap$ is the intersection of sentences classified as entailment by both T5 and RoBERTa (RoB). $\cap^+$ is the percentage after manual filtering.

> restaurant, it was pretty weak. In comparison to other oyster bars in the area, they were much to liquid-y. That is, they just didn't have enough substance on the whole and also, the taste wasn't really like seawater, it was more salt water than anything. ⇒Fairly disappointed in the oysters.

In Example 3, the final sentences does not restate all the information from any previous sentence but rather makes a weaker claim that summarizes the review. In other cases, we find that the conclusion of logical arguments can behave similarly:

> (4) **Wikipedia:** All of the known sphenacodonts are carnivores except for certain therapsids. Glaucosaurus is plainly not a therapsid . . . And it is just as plainly not a carnivore . . . ⇒So, it is very likely to be an edaphosaur.

With the world knowledge that a glaucosaurus must either be an edaphosaur or a sphenacodont, the final sentence follows logically from the context. Thus, it seems the role of this entailed sentence is to make explicit the conclusion of a logical argument.

In summary, our corpus study reveals that more entailed text is uttered by humans than expected if humans were always avoiding redundancy, as Gricean speakers do. There are many types of entailed text, including both repetition and instances where the entailed text is a summary or conclusion. Next, we will consider how a Gricean speaker might be extended to account for this behavior.

6 Towards Accounting for Redundancy

We have found that, in practice, the flipped entailment test better detects entailment than the original one and that this trend is also supported by an oracle logistic regression analysis (§4). Our corpus study (§5) pointed to a possible explanation: the original test relied on the fact that Gricean speakers always avoid redundancy, but real humans produce redundant text in certain contexts. Quantitatively, the rate of contextually entailed sentences in natural corpora was higher than we would expect if the corpus authors were Gricean speakers. Qualitatively, specific examples suggested humans are redundant both to repeat important information and for the sake of explanation, i.e., they state entailed summaries or conclusions after a more detailed premise. *Prima facie*, such redundancy could lead to a flipped entailment test if entailed continuations, which are fully redundant, become more likely than other continuations. However, it is crucial to have a more concrete theory of *why* speakers are redundant to evaluate this and ideally explain why the test direction varies across constructions. We thus consider some possible angles to extend Gricean speakers to account for redundant speech acts and whether these extensions predict the flipped test.

6.1 Redundancy via Noise Tolerance

Our corpus study showed that one type of redundancy in natural text unaccounted for by Gricean speakers is simple repetition. For example, the speaker in Example 2 repeats the claim that the orange in their beer was not a lemon. Gricean speakers are unlikely to generate such repetition, but they can be extended to do so by assuming there is noise in the communication channel, i.e., listeners may fail to interpret each sentence with some probability (Degen et al., 2019). In this setting, a rational speaker is incentivized to hedge the risk their listener might not understand important information by repeating it twice. We call such a speaker a *noise-tolerant* speaker, which we formalize in §B.

Noise-tolerant speakers can better account for repetition than Gricean speakers, but, if we assume corpora are generated by noise-tolerant speakers, would it explain the flipped direction of the entailment test? The short answer seems to be no. In §B, we derive an extension of the entailment test that "cancels out" noise tolerance by simply repeating the initial sentence in each term n times:

$$\hat{E}_p^n(x, y) \triangleq \log p(x^n y) - \log p(x^n \$) - \log p(y^{n+1}) + \log p(y^n \$).$$

As n increases, this test approximates the original test for a Gricean speaker. Thus, if the source of the flipped test was redundancy introduced by

a speaker's goal of being noise-tolerant, this test should work unflipped. Instead, we find that the *flipped* noise-tolerant test still detects entailment—in fact, better than the original flipped test. Post hoc analysis suggests the better performance may be due to the computational benefit of the additional tokens in the noise-tolerant test prompts. In summary, accounting for noise tolerance does not seem to explain why the test was flipped.

6.2 Redundancy via Explanations

A theory of speakers based on noise tolerance does not seem to explain the flipped entailment test. The noise-tolerant speaker accounts for repetition, but we also saw other kinds of redundancy in the data. In particular, Examples 3 and 4 show redundant sentences can occur at the end of an explanation or logical argument. One account could be that an initial explanation can dramatically lower the processing cost of a later conclusion, and that speakers consider this when selecting utterances. This is not modeled by the Gricean speaker whose processing cost $c(y)$ is independent of the context x. We thus reformulate the cost $c(y \mid x)$ as context-dependent. The impact of x on cost is measured by $\Delta(x, y) \triangleq c(y) - c(y \mid x)$: a large $\Delta(x, y)$ indicates a concise but helpful explanation x before conclusion y. If $\Delta(x, y)$ is large enough, the speaker will prefer to say xy as opposed to just y.

Flipped Test. Let $E(x, y)$ be the desired semantic value of the entailment test. With an explanatory speaker, the test score becomes (see §C):

$$\hat{E}_p(x, y) = E(x, y) + \Delta(x, y) - \Delta(y, y).$$

If we assume $\Delta(x, y)$ dominates $E(x, y)$, the test score can *increase* when x entails y because x will often explain y. This might explain the flipped test pattern. However, to be more complete, this account should be more precise about what factors influence $c(y \mid x)$ and predict why the original test outperformed the flipped test in some cases.

6.3 Discussion

Since we found that the entailment test was flipped in practice and that there are cases where humans are more redundant than Gricean speakers, we explored extensions to the Gricean speaker that could more accurately account for human redundancy and thus better explain the flipped test. We first considered a test that accounts for redundancy due to noise tolerance, finding that this likely could not explain the flipped test. Motivated by §5, we then turned to explanations as another source of human redundancy and showed how accounting for explanations might predict the flipped test.[5] We take this as encouraging evidence for pursuing pragmatic theories that explicitly account for explanations.

Stepping back, we have been able to use LMs as a source of data about sentence co-occurrences to test pragmatics theories and motivate alternatives, in the spirit of Harris (1954)'s idea that corpus data should be the empirical foundation of linguistic theory. A fundamental problem with using corpus data has been data sparsity, but LMs can alleviate this by letting us interpolate the likelihood of arbitrary sentences. We believe this could be a promising paradigm for future research in computational pragmatics to complement human subject experiments.

7 Conclusion

Our results show that sentence co-occurrence probabilities computed by LMs can predict entailment relationships, with a stronger effect for better LMs. This suggests these LMs are implicitly modeling semantic properties of text to some extent in order to predict the next token, in line with Harris (1954)'s proposal that sentence co-occurrences can serve as data for building a theory of semantics. However, the best empirical test for entailment we found was flipped compared to Merrill et al. (2022)'s theoretical test. This suggests a more nuanced theory of pragmatics beyond Gricean speakers is needed to explain how entailment relationships are reflected in sentence co-occurrences. Our corpus study revealed that humans in corpora produce more contextually entailed sentences than idealized Gricean speakers, suggesting pragmatic theories that better handle redundancy might explain our findings.

We took a first step by considering how to model redundancy due to noise tolerance and explanation, but the job is far from done. Rather, our findings call for future work that more completely accounts for the pragmatics of redundancy, especially concerning explanations. This can both advance linguistic theory and serve as a foundation for understanding how meaning can be inferred from a corpus, as well as as the potential limits of distributional semantics and LMs.

[5] Another reason speakers may be redundant, which we have not considered, is to trigger the listener to reanalyze the question under discussion. E.g., Example 2 may prompt the listener to infer the speaker's goal is to express frustration rather than convey the facts of their order.

Limitations

Regarding the theoretical foundations for the entailment test, Merrill et al. (2022) indicate in an erratum that the entailment test may have false positives for rare sentences pairs that are nearly contradictory. Further, the theory may be less applicable to LMs that have undergone an alignment process like RLHF. Overall, these qualifications to the test theory increase the value of our empirical study of whether the test works in practice.

Regarding our analysis of our results, we have assumed that the flipped entailment test pattern reflects differences between Gricean speakers and human speakers in corpora, but it, in principle, systematic estimation errors by LMs could explain the flipped entailment test pattern independent of the distribution of strings in the training corpus.

References

Emily M. Bender and Alexander Koller. 2020. 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*, pages 5185–5198, Online. Association for Computational Linguistics.

Robert Brandom. 2000. *Articulating Reasons: An Introduction to Inferentialism*. Harvard University Press, Cambridge, Mass.

Mikael Brunila and Jack LaViolette. 2022. What company do words keep? revisiting the distributional semantics of J.R. firth & zellig Harris. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4403–4417, Seattle, United States. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. Recognizing textual entailment: Rational, evaluation and approaches–erratum. *Natural Language Engineering*, 16(1):105–105.

Judith Degen, Robert D. Hawkins, Caroline Graf, Elisa Kreiss, and Noah D. Goodman. 2019. When redundancy is useful: A bayesian approach to 'overinformative' referring expressions.

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. 2019. Multi-news: a large-scale multi-document summarization dataset and abstractive hierarchical model.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling.

Noah D. Goodman and Michael C. Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. 2023. Think before you speak: Training language models with pause tokens.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

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

Philip J. Hayes and Steven P. Weinstein. 1991. CONSTRUE/TIS: A system for content-based indexing of a database of news stories. In *Proceedings of the 2nd Conference on Innovative Applications of Artificial Intelligence (IAAI-90), May 1-3, 1990, Washington, DC, USA*, pages 49–64. AAAI Press, Chicago, IL, USA.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. 2021. Provable Limitations of Acquiring Meaning from Ungrounded Form: What Will Future Language Models Understand? *Transactions of the Association for Computational Linguistics*, 9:1047–1060.

William Merrill, Alex Warstadt, and Tal Linzen. 2022. Entailment semantics can be extracted from an ideal language model. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 176–193, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Julian Michael. 2020. To dissect an octopus: Making sense of the form/meaning debate.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Ellie Pavlick. 2022. Semantic structure in deep learning. *Annual Review of Linguistics*, 8(1):447–471.

Christopher Potts. 2020. Is it possible for language models to achieve understanding?

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Claude E Shannon. 1951. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Johan Van Benthem. 1986. *Natural Logic*, pages 109–119. Springer Netherlands, Dordrecht.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Zhaofeng Wu, William Merrill, Hao Peng, Iz Beltagy, and Noah A. Smith. 2023. Transparency helps reveal when language models learn meaning. *Transactions of the Association for Computational Linguistics*, 11:617–634.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2016. Character-level convolutional networks for text classification.

A Test Derivation for Gricean Speakers

As shown by Merrill et al. (2022), the entailment test score $\hat{E}_p(x, y)$ score defined in terms of co-occurrence log-probabilities is equivalent to the following semantic quantity:

$$E(x, y) \triangleq \log \frac{\mathbb{E}_w[\exp(i_\ell(xy \mid w))g(x, w)]}{\mathbb{E}_w[\exp(i_\ell(x \mid w))g(x, w)]},$$

where $g(x, w)$ captures the normalizing factor from the speaker (cf. Merrill et al., 2022).

Proposition 1 (Merrill et al., 2022)**.** *Let p be a Gricean speaker. Then, for any x, y, $\hat{E}_p(x, y) = E(x, y)$.*

Proof. We recount an abbreviated version of the proof from Merrill et al. (2022, Appendices C and H). We use the fact that, for any x, y,

$$\log p(xy) - \log p(x\$) = E(x, y) - c(xy) + c(x\$).$$

Applying this property to both sides of $\hat{E}_p(x, y)$ yields

$$\begin{aligned}
\hat{E}_p(x, y) &= \log p(xy) - \log p(x\$) - \log p(yy) + \log p(y\$) \\
&= E(x, y) - c(xy) + c(x\$) - \cancel{E(y, y)} + c(yy) - c(y\$) \\
&= E(x, y) + \cancel{c(xy^2\$)} - \cancel{c(xy^2\$)}.
\end{aligned}$$

We conclude that $\hat{E}_p(x, y) = E(x, y)$. □

Crucially, $E(x, y)$ is closely related to entailment. If x entails y, then y conveys no information after x, so $E(x, y) = 0$. On the other hand, if $E(x, y) = 0$, then it must either be that a) x entails y or b) y nearly contradicts x, meaning the probability that x, y are consistent is small (Merrill et al., 2022, Erratum). Assuming near contradiction is unlikely, the entailment test (since it computes E) is then effectively a test for entailment defined purely in terms of sentence co-occurrence probabilities.

B Noise-Tolerant Speakers

We now formalize a model of noise-tolerant speakers that can account for repetition (Example 2). Our speaker is inspired by Degen et al. (2019)'s speaker designed to account for overredundant referring expressions but extends better to multiple sentences. We assume each sentence x has some probability ϵ_x of not being interpreted. When anticipating the information a listener gains from a text, a speaker marginalizes over the potential interpretations the listener might form by failing to interpret different sentences:

$$p(z \mid w) \propto \mathop{\mathbb{E}}_{e}[\exp(i_\ell(e \mid w))] \exp(-c(z)),$$

where e is a set of indices for sentences in z that are full comprehended. Formally, e is a subset of z's indices representing a subsequence. Note that $i_\ell(e \mid w)$ is defined in the natural way: it is the information a listener would get from just the sentences of z activated in e and not the other ones. This implicitly depends on z. The distribution of e is determined by ϵ's for each sentence in z:

$$p(e \mid w, z) = \prod_{t=1}^{n} \begin{cases} 1 - \epsilon_{z_t} & \text{if } t \in e \\ \epsilon_{z_t} & \text{otherwise.} \end{cases}$$

B.1 Theoretical Result

The original entailment test does not hold for noise-tolerant speakers, but a straightforward extension does. For any $n \geq 1$, we define the extended test as

$$\begin{aligned}
\hat{E}_p^n(x, y) \triangleq &\log p(x^n y) - \log p(x^n\$) \\
&- \log p(y^{n+1}) + \log p(y^n\$).
\end{aligned} \tag{2}$$

This extended test (with p as a noise-tolerant speaker) approximates the original test for a Gricean speaker, with error vanishing exponentially in n:

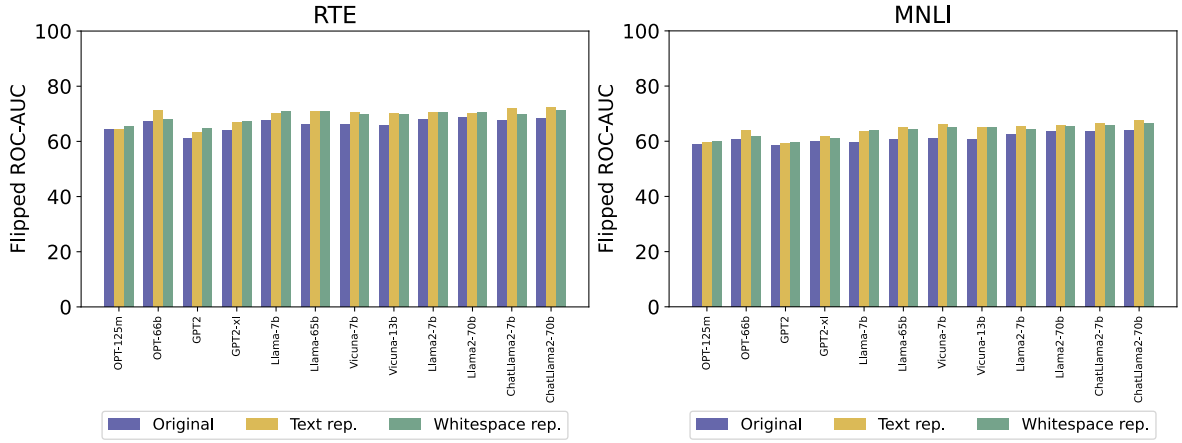

Figure 7: Performance of noise-tolerant (§B) vs. original test on RTE training set and MNLI matched validation set.

Proposition 2. *Let p be a noise-tolerant speaker. As n increases, $\hat{E}_p^n(x, y)$ converges to $E(x, y)$ with error vanishing exponentially in n.*

Proof. The idea is that, unlike a Gricean speaker, a noise-tolerant speaker will produce $p(ab)$ to account for the chance that a was not interpreted. If a repeats several times, the chance a was not interpreted goes to 0.

In order to show that the original test fails with this speaker and work out an alternative, we first work out some basic properties of this speaker's utility. Let $\dot{I}(z \mid w) \triangleq \mathbb{E}_e[i_\ell(e \mid w)]$ be the expected utility of z. We can first characterize the utility of a 2-gram xy under the noisy-channel speaker:

$$\begin{aligned}\dot{I}(xy \mid w) &= \epsilon_x\epsilon_y \cdot 0 + (1-\epsilon_x)\epsilon_y i_\ell(x \mid w) + \epsilon_x(1-\epsilon_y)i_\ell(y \mid w) + (1-\epsilon_x)(1-\epsilon_y)i_\ell(xy \mid w)\\ &= (1-\epsilon_x)\epsilon_y i_\ell(x \mid w) + \epsilon_x(1-\epsilon_y)i_\ell(y \mid w) + (1-\epsilon_x)(1-\epsilon_y)i_\ell(xy \mid w).\end{aligned}$$

We can apply this to get the expected utility of the utterances xx and $x\$$ under the noisy-channel speaker:

$$\begin{aligned}\dot{I}(xx \mid w) &= \epsilon_x^2 \cdot 0 + 2(1-\epsilon_x)\epsilon_x i_\ell(x \mid w) + (1-\epsilon_x^2)i_\ell(x \mid w)\\ &= (1-\epsilon_x^2)i_\ell(x \mid w)\\ \dot{I}(x\$ \mid w) &= (1-\epsilon_x)\epsilon_\$ i_\ell(x \mid w) + (1-\epsilon_x)(1-\epsilon_\$)i_\ell(x \mid w)\\ &= (1-\epsilon_x)i_\ell(x \mid w).\end{aligned}$$

We can now see that the original test does not work under a noise-tolerant speaker. The original entailment theorem worked by checking $i_\ell(y \mid x, s) = i_\ell(x \mid x, s)$ to see whether y is informative after x. Naively applying the original entailment test with a noise-tolerant speaker, however, will use $\dot{I}$ in place of i_ℓ. We can see that this does not represent the same quantity if ϵ_x, ϵ_y are non-negligible:

$$\begin{aligned}\dot{I}(x \mid x, w) &= \epsilon_x(1-\epsilon_x)i_\ell(x \mid w)\\ \dot{I}(y \mid x, w) &= \epsilon_x(1-\epsilon_y)i_\ell(y \mid w) + (1-\epsilon_x)(1-\epsilon_y)i_\ell(y \mid x, w).\end{aligned}$$

However, for the new test, we find the following:

$$\begin{aligned}\dot{I}(x \mid x^n, w) &= \epsilon_x^n(1-\epsilon_x)i_\ell(x \mid w) \approx 0\\ \dot{I}(y \mid x^n, w) &= \epsilon_x^n(1-\epsilon_y)i_\ell(y \mid w) + (1-\epsilon_x^n)(1-\epsilon_y)i_\ell(y \mid x, w) \approx (1-\epsilon_y)i_\ell(y \mid x, w).\end{aligned}$$

For large n, this overcomes the ϵ's since it means the test checks whether $i_\ell(y \mid x^n, w)$ is nonzero for all w (assuming $\epsilon_y < 1$, i.e., a human can possibly evaluate y). □

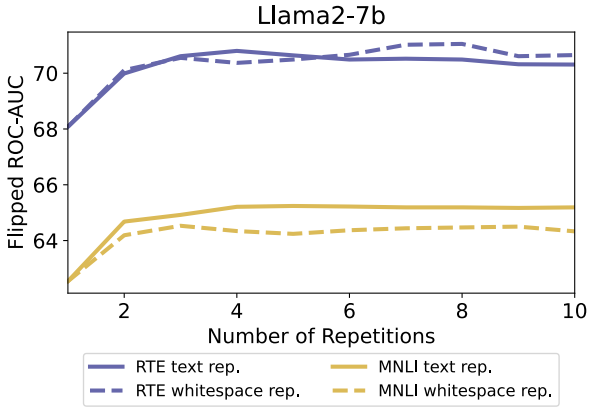

Figure 8: Performance of the noise-tolerance test with different numbers of repetitions (values of n in Equation (2)). The original test is $n = 1$.

B.2 Empirical Results

We compare the noise-tolerant test with $n = 5$ repetitions against the original test, using the RTE training set and the MNLI matched validation set.[6] As shown in Figure 7, the flipped noise-tolerant test consistently detects entailment better than the original flipped test. However, the fact that the test still works better flipped is just as unexpected with the noise-tolerant test as with the original test.

We were thus skeptical whether the boost in performance from the noise-tolerant test was due to more realistic speaker assumptions and aimed to access whether there could be a confounding explanation. In particular, in addition to accounting for ways speakers can be redundant, the noise-tolerant grants the LM additional tokens and thus more steps of computation, which could enable more closely approximating each log-likelihood (Goyal et al., 2023). To control for this, we introduce a "pause token" test where, for each term $\log p(ab)$, spaces are inserted between a and b to add the same number of tokens that would be added by replacing a with a^n.[7] Assuming spaces carry no semantics, the pause token test should measure the same quantity as the original entailment test, but with more compute than the noise-tolerant test.

As shown in Figure 7, the pause token test outperforms the original test, suggesting the computational benefit of additional tokens may explain the test improvement. For many datasets, the pause token test performs slightly worse than the noise-tolerant test, but because the absolute difference is small and not consistent, we do not take this as evidence that the noise-tolerant test provides a benefit beyond more tokens of computation. Further, Figure 8 shows that increasing the number of repetitions yields roughly monotonic but diminishing returns, as might be expected for a computational resource. Overall, we conclude the stronger performance of the noise-tolerant test likely reflects the greater computational power of padding tokens and not better assumptions about human speakers.

C Explanatory Speakers

The only change we make to the speaker to support explanations is generalizing the cost $c(y \mid x)$ to depend on the prior context. We assume that $c(\$ \mid z) = c(\$)$ for all z.

Proposition 3. *Let p be an explanatory speaker. Then, for any x, y,*

$$\hat{E}_p(x, y) = E(x, y) + \Delta(x, y) - \Delta(y, y).$$

Proof. By definition,

$$\begin{aligned}\hat{E}_p(x, y) &= E(x, y) - c(xy) + c(x\$) + c(yy) - c(y\$) \\ &= E(x, y) - c(y \mid x) + c(\$ \mid x) + c(y \mid y) - c(\$ \mid y).\end{aligned}$$

[6] Due to the repetitions multiplicatively increase sequence length, running this test on the MNLI training set, as we do in the other experiments, was not feasible for us.

[7] The tokenizer for Llama models treats 16 consecutive whitespaces as a single token. We hence insert 16 times more whitespaces for Llama-based models to control for the token count.

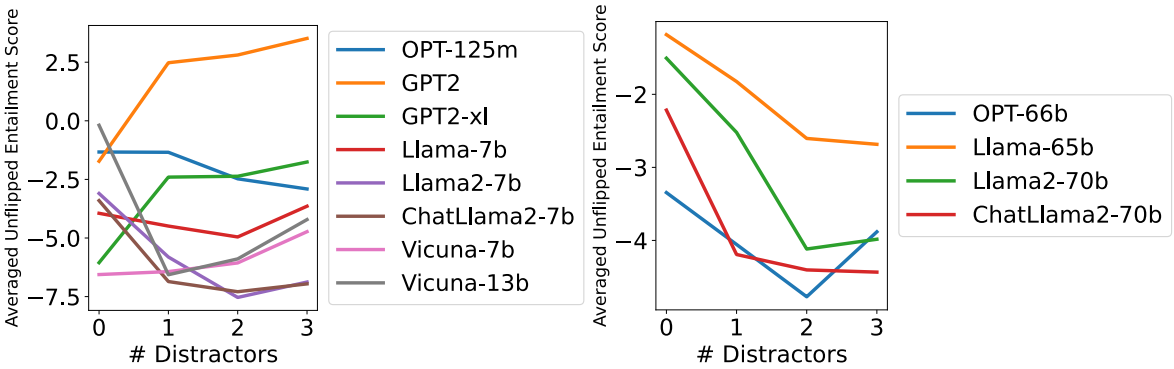

Figure 9: Unflipped entailment test score as a function of the number of distractors in the premise, with <65b models (left) and ≥65b models (right), for the RTE dataset.

These cost terms do not all cancel out (as for Gricean speakers). Instead, we get

$$\begin{aligned}\hat{E}_p(x,y) &= E(x,y) - c(y \mid x) + \cancel{c(\$)} + c(y \mid y) - \cancel{c(\$)} \\ &= E(x,y) - c(y \mid x) + c(y \mid y) + c(y) - c(y) \\ &= E(x,y) + \Delta(x,y) - \Delta(y,y).\end{aligned}$$

□

C.1 Further Details and Experiments

To be convincing, the explanatory speaker account should ideally explain why the original test worked better than the flipped test for some targeted cases like logical connectives and numbers (Figure 2). The connectives could possibly be explained by the fact that the connectives hypotheses introduced new entities that did not occur in the premise (cf. Example 13). Because these entities do not exist in the discourse, it would be infeasible for a listener to reason about whether they are entailed in advance, making semantic priming unlikely. We would thus expect $\Delta(x,y)$ and the test to better match $E(x,y)$ in this case.

The semantic priming account predicts that, for entailed pairs, the test score should reflect how much x semantically primes y. Assuming adding distractors to the premise reduces semantic priming, it thus predicts that the entailment score should decrease as more distractors are added to the premise. We test this by generating entailment pairs with distractors in the premise like the following:

(5) Olivia lives in Paris. James lives in Tokyo. ⇒Olivia lives in France.

As shown in Figure 9, this pattern holds for all ∼70b LMs we considered, although the results for LMs of smaller scales are more inconsistent. We take this as weak evidence that the speakers LMs are modeling (i.e., humans) may be accounting for the reduction in processing time that an explanation can provide.

D Manual Inspection of Entailment Classified by Models

The following are examples of premise-hypothesis pairs which were marked as entailment by both T5 (Honovich et al., 2022) and RoBERTa (Liu et al., 2019). Through manual inspection, however, we find that they were in fact incorrectly classified as such. We include a comprehensive list of those cases as well as reasoning as to why we believe the pair is not entailment.

(6) **Multi-News:** The man survived the fall and the waters.After he was rescued, he noted that a "burning platform" caused a radical change in his behaviour.We too, are standing on a "burning platform," and we must decide how we are going to change our behaviour.Over the past few months, I've shared with you what I've heard from our shareholders, operators, developers, suppliers and from you.Today, I'm going to share what I've learned and what I have come to believe. ⇒I have learned that we are standing on a burning platform.

The premise does not contain information regarding the fact that the narrator had "learned [they] are standing on a burning platform".

(7) **Books3:** That's where you're wrong.I only have negatives.Minus wishes.""E, what are you going on about?"I asked gently, leaning in and wincing as my shirt caught on the dressing. ⇒"I only know what I don't want.

The "I only have negatives" in the premise can be interpreted as the narrator only having things that they don't want. This is in contrast to the knowledge aspect which is brought up in the hypothesis.

(8) **Book3:** I glance over my shoulder.Liam moves fast too, throwing himself at the hill to catch me.It's fine, I can outrun him over distance.All I need is a head start.So I push myself, stumbling on the dry churned-up turf. ⇒Behind me, Liam speeds up.

The premise indicates that "Liam moves fast...to catch me". It does entail that he "speeds up", which is in the hypothesis.

(9) **Wikipedia:** Henry admits he doesn't dance, and encourages Minnie to dance with Sidney.Henry thinks Minnie must find life with him dull, and resolves to learn to dance.He keeps this secret from her in order to surprise her on her birthday.He takes private dancing lessons, instructed by Madame Gavarni and her niece.Minnie seems to grow distant. ⇒Henry thinks she is bored, and looks forward to surprising her with dancing.

While Henry thinking Minnie is bored and planning on surprising her with dancing, him "[looking] forward to [it]" is new information presented in the hypothesis not in the premise.

(10) **Wikipedia:** "Established in 1923, it has a membership of around 230,000 and is open to past and present members of the UK Civil Service and public sector plus organisations that were formerly part of the British Civil Service, for instance Royal Mail and BT.Relatives of existing members may also join.History Boundless by CSMA is a mutual organisation.It was founded as the Civil Service Motoring Association in 1923 by Frank Vernon Edwards, an executive officer in the Ministry of Labour who had an interest in motorcycle trials.CSMA Club was designed to be a small motorsport organisation of around 300 members, but by 1930 the membership was over 5,000." ⇒The membership currently stands over 230,000.

The premise does not indicate that there are more than 230,000 members, which means that the hypothesis is adding additional information not contained in the premise.

(11) **Reuters-21578:** "A spokeswoman for the EC Commission said the detailed 25-page report of alleged malpractices was in response to a similar document issued by U.S. Administration officials in November, and updated a previous EC list.EC External Trade Relations Commissioner Willy De Clercq said its object was to show such actions were not solely taken by trading partners of the U.S. And that "the U.S.Were not innocents in the matter."The report covers the entire field of EC-U.S. Commercial relations and lists more than 30 obstacles ranging from tariff measures, import quotas, customs duties, anti-dumping procedures, fiscal barriers and export subsidies.The Commission said not all the barriers mentioned were necessarily inconsistent with U.S. International obligations, and emphasised many of them could be removed at upcoming international trade talks." ⇒The purpose of the report is to make clear that trade practices which impede exports are not a unique problem only faced by U.S.

The premise describes something different from the hypothesis in that the objective of the report was to show "such actions were not solely taken by trading partners of the U.S." but also participated in by the U.S. itself.

(12) **Yelp Review:** "While it looks decent on the outside and the inside, the food and service were simply terrible.Chicken was very watered down, the salsa was flavorless, and the service make a fast food chain look really good.Just a poor, poor experience at this location overall.If this was

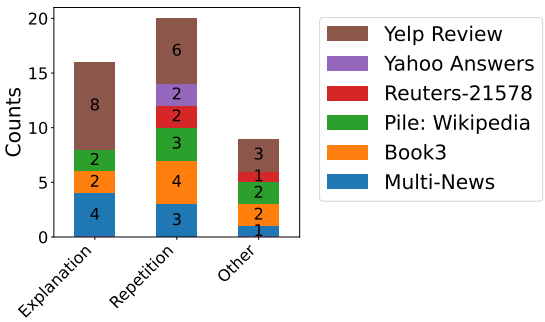

Figure 10: Frequency of occurrences of entailment categories across data sources.

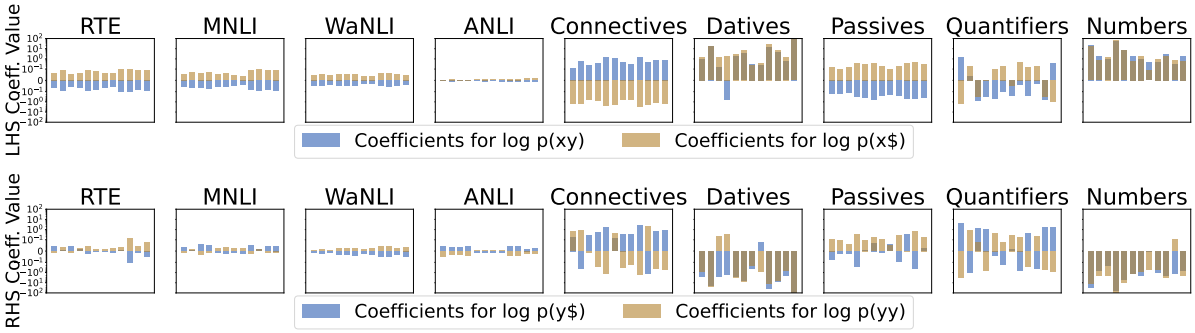

Figure 11: Learned logistic regression coefficients for the log-probability features for the broad-coverage datasets. Each bar represents one LM. For ease of visualization, y-axis is in log scale, except in $[-0.1, 0.1]$ where it is linear; and it is capped at $[-100, 100]$, requiring truncation in a few cases.

the only El Cancun in Charlotte, I would feel the same way many posters do and just never come back. Luckily for me, I live in Rock Hill." ⇒There's an El Cancun here.

The hypothesis introduces new information about another El Cancun location being where the speaker is, which is not present in the premise.

E Manual Classification of Entailment Categories

Based on the filtered manual results described by $\cap^+$ in Table 1, we manually classify results into three categories: explanation, repetition, and other. The classification of explanation and repetition were determined based on the definition in §6.2 and §6.1, respectively. Results of these classifications across data sets are shown in Figure 10.

F Learned Entailment Test for More Datasets

In Figure 11, we show the coefficients for the learned entailment test (§4.3). However, we note a caveat for the targeted evaluation datasets: because they are manually curated, there are simple dataset artifacts that can be used to distinguish between the two classes (for example, some types of hypotheses only exist for entailment instances). When we learned a classifier, such artifacts could be exploited (and we do see that they are exploited in practice). We thus highlight that the interpretation of the relevant coefficients are not straightforward.

G Synthetic Data for Targeted Evaluation

The GLUE diagnostics (Wang et al., 2018) are not a public dataset; hence, we make our synthetic targeted evaluation data based on the GLUE design principles. We create synthetic data following the following templates, where the names and base propositions vary according to a hard-coded list.

Dataset	# Instances
RTE-train	2,490
MNLI-train	392,702
MNLI-validation-matched	9,815
WaNLI-train	102,885
ANLI-train	100,459
Connectives	1,800
Quantifiers	780
Numbers	260
Passives	2,160
Datives	720

Table 2: The number of instances for each dataset we use.

Connectives. The premise $p(a)$ entails $p(a \vee b)$ but not $p(a \wedge b)$:

(13) I saw James.
  a. I saw James or Olivia. ✓
  b. I saw James and Olivia. ✗

Quantifiers. For a non-empty domain, *all* $a\, p(a)$ entails *some* $a\, p(a)$ but not *no* $a\, p(a)$:

(14) All of the crops failed.
  a. Some of the crops failed. ✓
  b. None of the crops failed. ✗

Numbers. Similarly, *at least two* entails *at least one* but not *at least three*:

(15) At least two of the crops failed.
  a. At least one of the crops failed. ✓
  b. At least three of the crops failed. ✗

Passivization. Given a premise with a transitive verb, the reduced passive with the original object as the subject is entailed, but the reduced passive with the original subject as the subject is not:

(16) Olivia saw Mia.
  a. Mia was seen. ✓
  b. Olivia was seen. ✗

Datives. Given a sentence with a direct object and an optional indirect object, the sentence with the indirect object removed is entailed, but the sentence with the direct object is not:

(17) Liam baked Noah a cake.
  a. Liam baked a cake. ✓
  b. Liam baked Noah. ✗

H Language Models We Used

We test a variety of LM families, and for each, we use the smallest and largest public-available variant. Specifically, we use GPT-2 small (117M parameters) and XL (1.5B), OPT 125M and 66B, Llama-1 7B and 65B, Vicuna 7B and 13B, Llama-2 7B and 70B, and ChatLlama-2 7B and 70B.

I Dataset Stastics

We report dataset statistics in Table 2.