Language models (LMs) prefer ambiguous utterances following informative contexts

Yanting Li, Jiaxuan Li, Shiva Upadhye, Noa Attali and Gregory Scontras
Department of Language Science, University of California, Irvine
{yantil5, jiaxul19, upadhyes, noa.attali, g.scontras} @uci.edu

Introduction The sentence *every show wasn't captivating* is ambiguous between the surface interpretation of *no show was captivating* and the inverse interpretation of *not every show was captivating*. How do speakers choose between the ambiguous *every-not* versus the unambiguous *none* or *not-every* utterances? Previous research shows that ambiguity is allowed given more informative context [1-2]. In particular, comprehenders track contextual cues to reason about the scope of the *every-not* ambiguity [3]: if the context sets a high expectation for a random show to be captivating, the inverse interpretation is endorsed more. Meanwhile, pre-trained language models have also demonstrated competence in formal, pragmatic and commonsense inference [5-6], depending on their ability to infer aspects of world knowledge, including event properties [7-8]. Bridging these two lines of work, the present study hypothesizes that LMs prefer the ambiguous *every-not* utterance over the unambiguous alternatives when the context provides informative cues for disambiguation.

Methods: We use naturalistic occurrences of *every-not* utterances (N=375) together with their preceding context from [3,4]. We adapt the materials as illustrated by Table 1. For every occurrence, we consider the original ambiguous *every-not* utterance (1), and construct two unambiguous *none* (2a) and *not-every* (2b) alternatives. We estimate **(i) LM preference** by comparing normalized sentence log-probability of the three utterances given preceding context: LM's preference for an ambiguous utterance is defined as the difference between the log-probability of the ambiguous utterance and the maximum log-probability among the two unambiguous ones. We then estimate **(ii) context informativity** by calculating the LM-estimated entropy over the probability p of the affirmative event, and the probability $1-p$ of the negative event using Eq. 1. The two probabilities are estimated with the normalized sentence probability of the constructed (3a) and (3b) in Table 1. All relevant probability quantities are estimated from GPT-2 [9,10]. Finally, we test the relationship between (i) and (ii) using a linear model. We hypothesize that lower entropy indicates a higher context informativity, leading to a preference for the ambiguous (*every-not*) construction. We also analyzed the **alignment between LM and human** on whether LM-estimated p linearly correlates with human-estimated p' collected in [3], and whether LM preference is predicted by entropy over human estimated event probabilities.

Results Our results show that the LM prefers the ambiguous utterance when contexts are more informative in disambiguating between two possible interpretations (Fig. 1-top-left). This is confirmed by a linear model, where LM estimated entropy over two events significantly predicts the log-probability difference between ambiguous and unambiguous utterances ($t = -1.21, p < .001$). However, there seem to be some discrepancies between humans and LMs. LM's preference for ambiguous utterance is not predicted by humam-estimated event entropy ($t = .09, p = .51$; Fig. 1-top-right). This is likely due to the fact that humans and LMs form different probability estimations of events, where LM estimated probabilities (p) do not correlate with human estimated probabilities (p') ($t = .01, p = .65$; Fig.1-bottom).

Conclusion We find that for *every-not* ambiguity, LMs favor ambiguous utterances when contexts provide strong cues for disambiguation, suggesting that LMs' text generation might be guided by rational inference over context. However, LMs' estimation of event entropy given context does not align with humans'. For future work, we plan to collect humans' preference for producing ambiguous utterances given contexts as speakers, and see if it aligns with the LM preference.

original preceding context: So I guess if we survive the year 2000, we've got the year 2012 right around the corner. @!ZWERDLING: Wow. This is so exhausting. I mean, too bad

type of continuation	text
(1) original *every-not* utterance	every show wasn't captivating
(2) constructed unambiguous alternatives	(2a) no show was captivating / (2b) not every show was captivating
(3) constructed individual events	a random show (3a) was captivating / (3b) wasn't captivating

Table 1: Original and constructed materials for each occurrence of *every-not* utterance from naturalistic corpus.

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \tag{1}$$

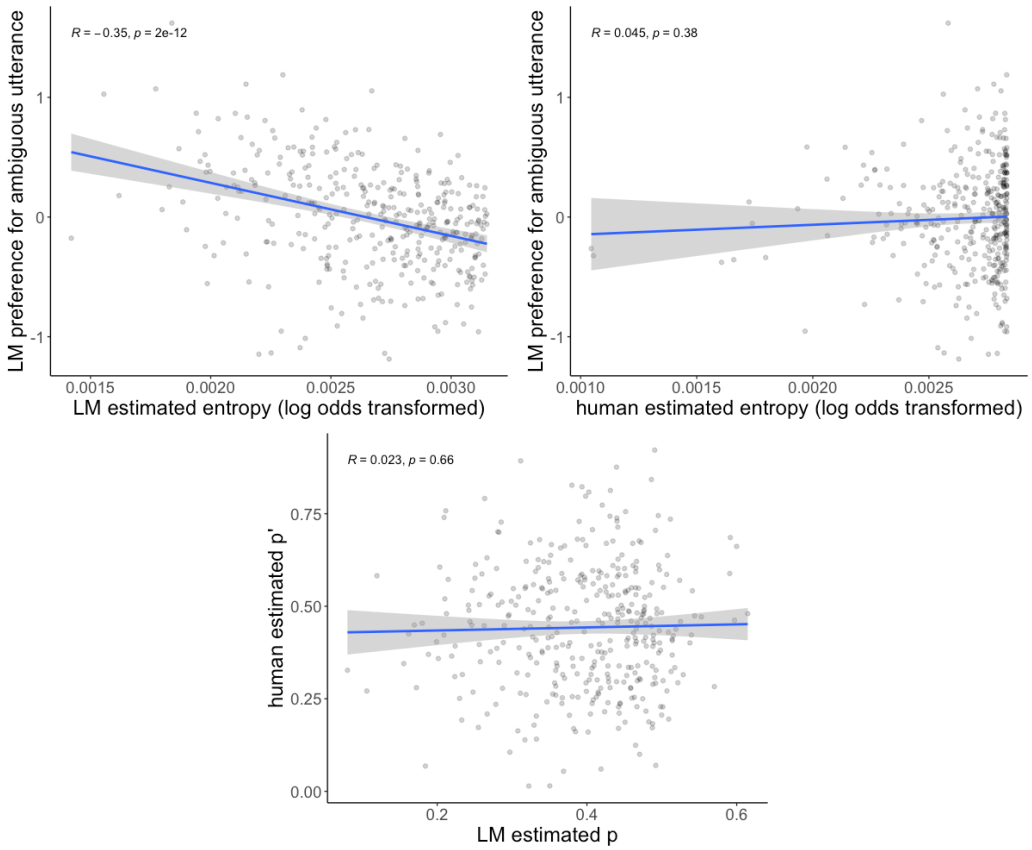

Figure 1: Top-left: The relationships between LM estimated entropy and LM preferences for ambiguous utterances. The entropy is log-odds transformed with a scalar = 0.5. Top-right: The relationships between human estimated entropy and LM preferences for ambiguous utterances. The entropy is log-odds transformed. The entropy is log-odds transformed with a scalar = 0.5. Bottom: The relationship between LM estimated event probability and human estimated event probability.

References. [1] Mahowald et al. (2013) *Cognition*; [2] Piantadosi & Gibson (2012) *Cognition*; [3] Attali (2024) *PhD Dissertation*; [4] Attali et al. (2023) *Experiments in Linguistic Meaning*; [5] Mahowald et al. (2024) *Trends in Cognitive Sciences*; [6] Wei et al. (2022) *arXiv preprint*; [7] Webb et al. (2023) *Nature Human Behaviour*; [8] Kauf et al.(2023). *arxiv*; [9] Radford et al. (2019) *OpenAI blog*; [10] Misra (2022) *arXiv preprint*.