Supplementary Materials: Do language models learn typicality judgments from text?

Kanishka Misra[1], Allyson Ettinger[2], and Julia Taylor Rayz[1]

[1]Purdue University, West Lafayette, IN, USA
[2]University of Chicago, Chicago, IL, USA
[1]{kmisra, jtaylor1}@purdue.edu
[2]aettinger@uchicago.edu

1 Code

The code and data materials used to reproduce the analyses reported in the paper can be found in the following link: https://github.com/kanishkamisra/typicalityprobing. The code is predominantly written using the `minicons`[1] library.

2 Models Studied

3 Stimuli Generation

In this section we describe in detail the language stimuli used in our two experiments. These stimuli have been summarized in Table 2 in the original paper.

3.1 Categories and Exemplars from Rosch (1975)

We collect our categories and their items/exemplars from the typicality ratings collected by Rosch (1975). In total, we have 565 items spanning across 10 different categories. Table A shows the most and least typical items for each of the 10 categories. We use this dataset as the ground truth resource to construct our stimuli for both our experiments.

3.2 Taxonomic Sentence Verification

Recap Our taxonomic sentence verification experiments test whether the model's attribution of hypernyms to noun concepts—by means of word prediction in context—reflects the graded structure of concepts as observed in humans. To this end, we compare the models' log probability of assigning an item (*robin*) its superordinate category (*bird*) to its typicality rating for the category as produced by humans in the original study.

Construction To construct the stimuli for this experiment, we create simple English templates:

$$\{\texttt{A/An/}\phi\}\ \texttt{[ITEM]}\ \{\texttt{is/are}\}\ \{\texttt{a/an/}\phi\}\ \texttt{[CATEGORY]}.$$

where $\{\,.\,\}$ denotes an optional word—with options separated by a '/'—and ϕ is a null character. Check out https://github.com/kanishkamisra/typicalityprobing/python/ for detailed stimuli samples.

[1]https://github.com/kanishkamisra/minicons

Table A: Categories and Exemplars extracted from Rosch (1975). The 'Exemplar' column shows 10 most and 10 least typical items for their corresponding category, as rated by native English speakers.

Category	Exemplar
toy	*doll, top, jack-in-the-box, toy soldier, yo-yo, block, marbles, rattle, stuffed animal, water pistol, ..., bow and arrow, rope, dishes, cards, mitt, horse, gun, animals, tennis racket, books*
bird	*robin, sparrow, bluejay, bluebird, canary, blackbird, dove, lark, swallow, parakeet, ..., duck, peacock, egret, chicken, turkey, ostrich, titmouse, emu, penguin, bat*
sport	*football, baseball, basketball, tennis, softball, canoeing, handball, rugby, hockey, ice hockey, ..., pool, billiards, hunting, jump rope, camping, chess, dancing, checkers, cards, sunbathing*
vegetable	*pea, carrot, green bean, string bean, spinach, broccoli, asparagus, corn, cauliflower, brussels sprouts, ..., escarole, sauerkraut, pickle, baked bean, pumpkin, seaweed, garlic, dandelion, peanut, rice*
tool	*saw, hammer, ruler, screwdriver, drill, nail, tape measure, sawhorse, sandpaper, sander, ..., plaster, wheelbarrow, axe, slide rule, cement, anvil, hatchet, rag, scissor, crane*
fruit	*orange, apple, banana, peach, pear, apricot, tangerine, plum, grape, nectarine, ..., pawpaw, coconut, avocado, pumpkin, tomato, nut, gourd, olive, pickle, squash*
clothing	*pant, shirt, dress, skirt, blouse, suit, slack, jacket, coat, sweater, ..., handkerchief, purse, hairband, ring, earring, watch, cuff link, necklace, bracelet, cane*
vehicle	*automobile, station wagon, truck, car, bus, taxi, jeep, ambulance, motorcycle, street-car, ..., rocket, blimp, skate, camel, feet, ski, skateboard, wheelbarrow, surfboard, elevator*
furniture	*chair, sofa, couch, table, easy chair, dresser, rocking chair, coffee table, rocker, love seat, ..., counter, clock, drape, refrigerator, picture, closet, vase, ashtray, fan, tele-phone*
weapon	*gun, pistol, revolver, machine gun, rifle, switchblade, knife, dagger, shotgun, sword, ..., word, hand, pipe, rope, airplane, foot, car, screwdriver, glass, shoe*

3.3 Category-based Induction

Recap Our category-based induction experiments test whether models reflect typicality effects when making inductive generalizations. That is, are models more likely to generate *"All **birds** can dax"* when given *"**Robins** can dax,"* as opposed to *"**Penguins** can dax"*?

Construction Following Osherson et al. (1990), we use *blank* properties in this experiment – i.e., properties that we estimate to be vaguely unfamiliar to language models (analogous to what Osherson et al. used for humans). To this end, we relied on nonce words such as *dax, slithy, fep, etc.* such that these words do not occur in any model's vocabulary. Since all models in our experiment had the ability to tokenize any word, and not rely on the <unk> token—which had been the dominant method prior to the introduction of wordpieces and byte-pair-encodings—we speculate the usage of nonce words to be a reasonable proxy for *blank* properties.

4 Confound analyses for Category-based Induction

Eliciting judgments from LMs as we do here has two potential confounds that may reduce the robustness of our conclusions.

Premise Order Sensitivity An LM might estimate high probabilities for words in the conclusion simply because it is relying on lexical cues in its premise (Misra et al., 2020), instead of processing the premise *compositionally* and making inferences about items possessing a property. To study this confound, we compute the LMs' average probability for the conclusion sentence when prefixed by a shuffled version of the premise (10 times, with random seeds). We then calculate POS as the difference between this measure (which we call $AS_{shuffled}$) and the original AS score, for each item:

$$\text{POS}_i = AS_i - AS_{shuffled,i} \tag{1}$$

It would be desirable to have POS values that are greater than 0, signifying that the LM is indeed sensitive to the correct word-order structure of the premise. Figure 1 displays the proportion of cases (out of 12,180) where POS was greater than 0, against the models' total number of trainable parameters. We observe that only Incremental LMs show nearly complete sensitivity to the word-order of the premise. Masked LMs on the other hand do show sensitivity in a majority of cases, but are still far below incremental LMs, suggesting that this confound greatly affects their results, and is likely not to affect Incremental LMs. No particular effect of number of parameters was observed.

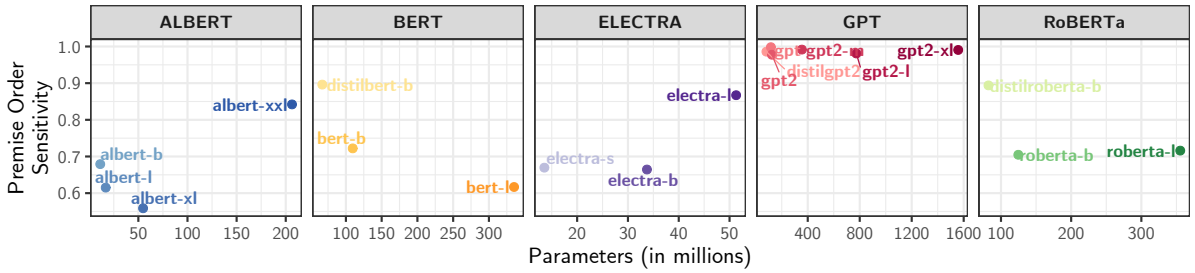


Figure 1: Model's sensitivity to changes in Premise word order.

Taxonomic Sensitivity LMs might tend to repeat the property phrase mentioned in the predicted material with high probability when prefixed by a sentence containing it, i.e., repeating *"can dax"* in the conclusion when already conditioned on the same phrase in the premise (Holtzman et al., 2019), confounding the degree to which the conclusion is generated using the taxonomic relationship between the premise and the conclusion categories. In order to study this tendency, we compute the LMs' probabilities for conclusions consisting of a different category with the exact same property as the original (for instance, *"All fruits are slithy"* given *"Sofas are slithy"*). We call this measure $AS_{flipped}$. Just like in our POS calculation, our TS measure for each item is calculated as the difference between the original measure and the flipped measure:

$$\text{TS}_i = AS_i - AS_{flipped,i} \tag{2}$$

As is the case with POS, it is desirable to have a strongly positive value for sensitivity towards this confound (i.e., TS $>$ 0). Figure 2 shows the the proportion of cases (out of 12,180 for each model) where TS was greater than 0, against the models' total number of trainable parameters. From Figure 2, there is a non-trivial proportion of cases where models produce greater AS scores for the conclusion when the premise concept had no taxonomic relation to the conclusion concept as compared to when it did, suggesting that in many cases models might not be processing based on the taxonomic relation between premise and conclusion and may simply be assigning high probabilities to the property phrase because it was listed in the preceding context – it is common-knowledge that LMs are highly conducive to repeating text already seen in the input context (Holtzman et al., 2019).

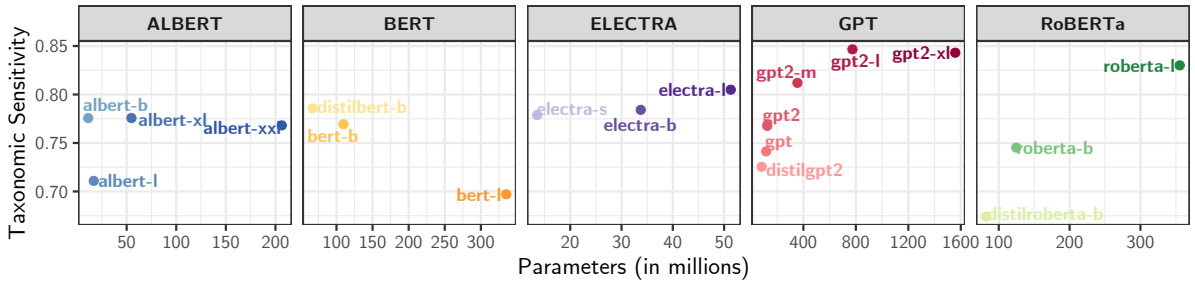


Figure 2: Models' sensitivities to violation of the taxonomic relation between the premise and the conclusion categories.

5 Additional Results

5.1 Category-wise results on Taxonomic Sentence Verification Experiments

See Figure 3.

5.2 Category-wise results on Category-based Induction Experiments

See Figure 4.

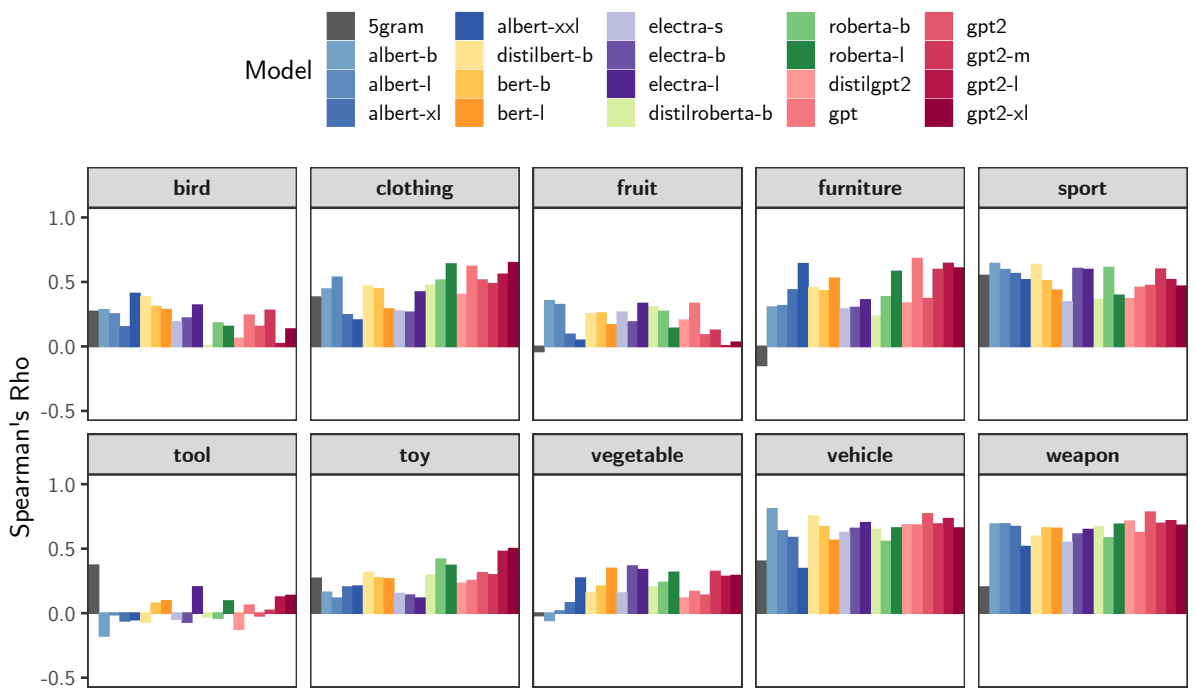
References

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

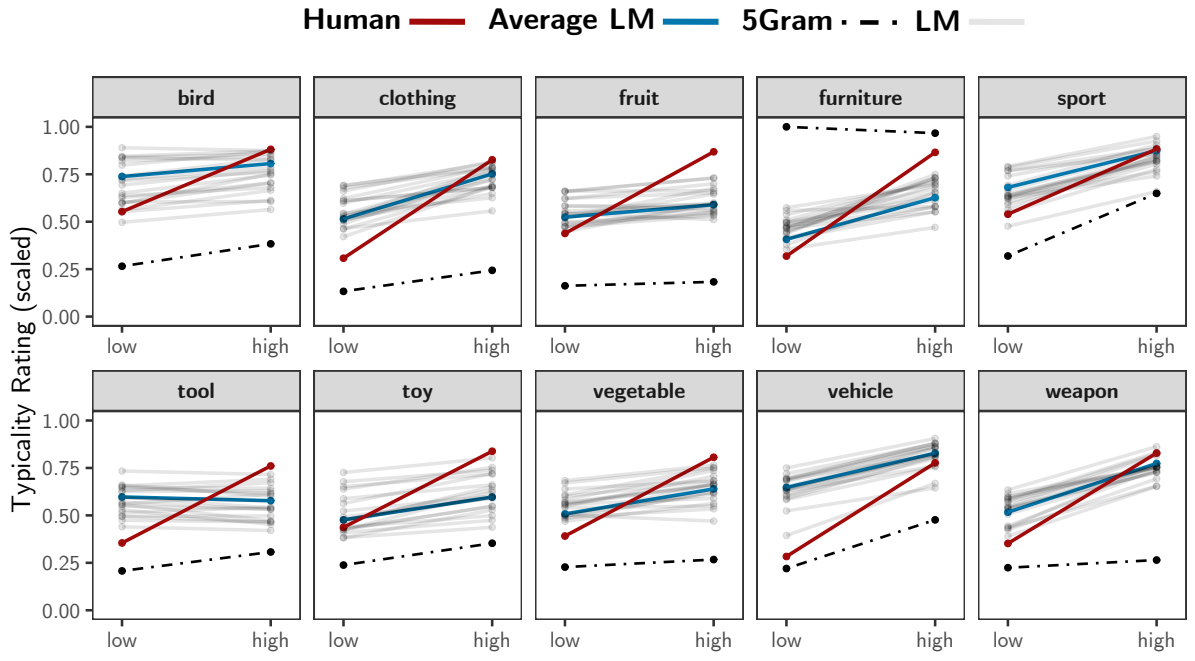
Kanishka Misra, Allyson Ettinger, and Julia Rayz. 2020. Exploring BERT's sensitivity to lexical cues using tests from semantic priming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4625–4635, Online. Association for Computational Linguistics.

Daniel N Osherson, Edward E Smith, Ormond Wilkie, Alejandro Lopez, and Eldar Shafir. 1990. Category-based induction. *Psychological review*, 97(2):185.

Eleanor Rosch. 1975. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192.

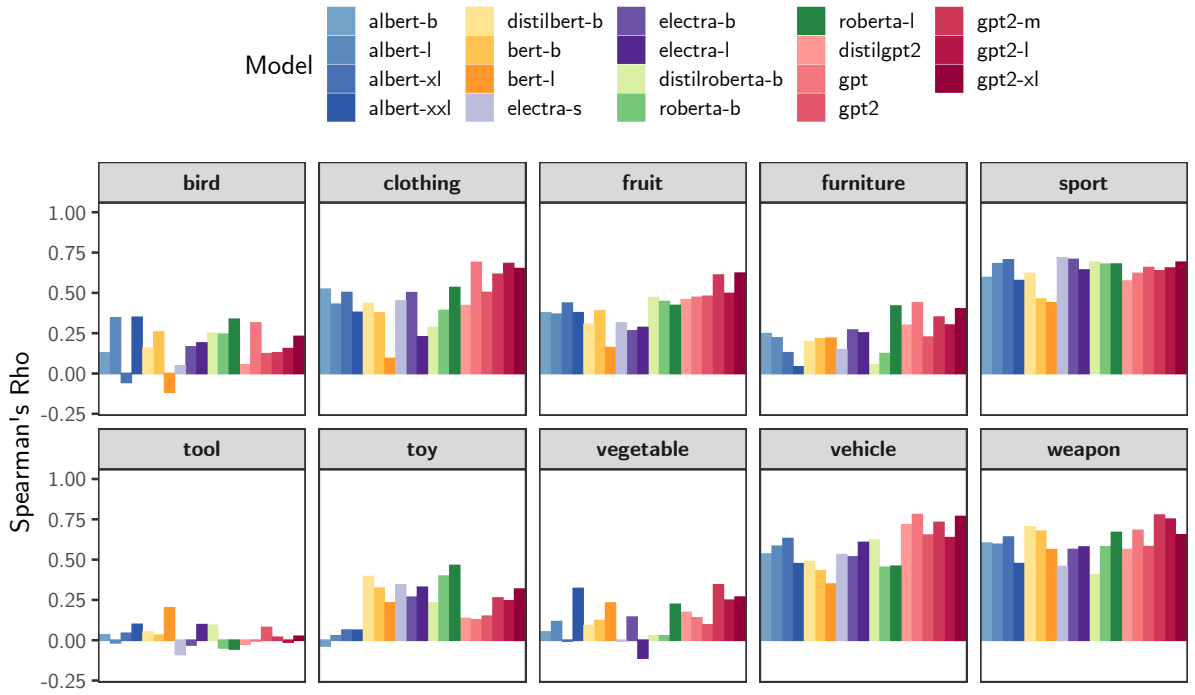


(a)

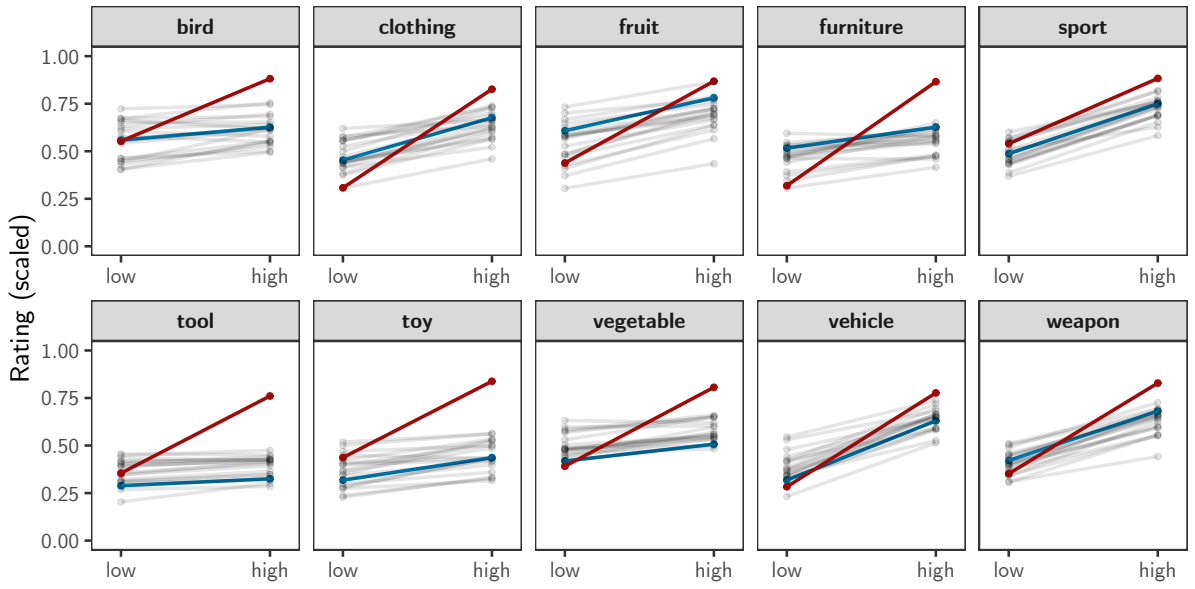


(b)

Figure 3: Category-wise results from the **Taxonomic Sentence Verification** experiments – (a) Spearman's correlation between language model log-probabilities and human typicality ratings; (b) Average scores (log-probability for language models, and raw typicality ratings for humans) assigned to low and high typicality items.



(a)



(b)

Figure 4: Category-wise results from the **Category-based Induction** experiments – (a) Spearman's correlation between adjusted AS-scores and human typicality ratings; (b) Average scores (adjusted AS-scores for language models, and raw typicality ratings for humans) assigned to low and high typicality items.