Beyond the best guess: Distributional predictions from large language models for human forced-choice data

Michael Franke[*]
Department of Linguistics
University of Tübingen
mchfranke@uni-tuebingen.de

Polina Tsvilodub
Department of Linguistics
University of Tübingen
polina.tsvilodub@gmail.com

Fausto Carcassi
ILLC
University of Amsterdam
fausto.carcassi@gmail.com

Abstract

fill me

1 Introduction

[MF: would be good to add references here]

The invention of deep neural transformer architectures (Vaswani et al., 2017) enabled a new generation of powerful large language models (LLMs) (Devlin et al., 2019; Chung et al., 2022; OpenAI, 2023; Touvron et al., 2023) which excel on standard benchmarks and promise to serve as foundation models for a vast and diverse set of applications (Bommasani et al., 2021). Recent works increasingly go beyond using LLMs based on single-run input-output behavior, and instead utilize LLMs as a part of a larger computational process. Examples include sophisticated prompting strategies (e.g., Liu et al., 2022), or structured reasoning models (e.g., Creswell, Shanahan, and Higgins, 2022; Gao et al., 2023; Paranjape et al., 2023). Information from LLMs is also used to rank or numerically score options in open-ended applications, e.g., to mimic human judgements of relevance or interestingness (e.g., Park et al., 2023; Zhang et al., 2023). Other work, uses LLMs as part of bigger programs to build towards something more akin to explanatory cognitive models (e.g., Wong et al., 2023).

For all of these applications, it is crucial to understand what LLMs can or cannot reliable do. The prevalent approach towards characterizing the capabilities of LLMs relies on benchmark testing, which usually consists in assessing the accuracy of LLM predictions in tasks where a designated “target answer” exists, averaged over many instances of this task. Benchmark-driven assessments are very useful to systematically study the effects of scale, e.g., in an engineering-oriented context (e.g., BIG-bench authors, 2023). Other recent work is more psychologically oriented and asks to what extent LLM performance is “human-like.” LLM performance is therefore compared to human choice behavior in psychological experiments to investigate whether LLMs predict patterns of human answer behavior qualitatively (e.g., Binz and Schulz, 2023; Hagendorff, 2023; Shiffrin and Mitchell, 2023). Going further, there is also work investigating whether LLMs can make adequate *quantitative predictions*. For example, work at the interface between NLP and computational psycholinguistics (Marvin and Linzen, 2018; Hu et al., 2020) has evolved into investigations of whether quantitative predictions by LLMs match quantitative aspects in human experimental data, such as reading times (Wilcox, Vani, and Levy, 2021) or the amplitude of common EEG signals like the N400 component (Lindborg and Rabovsky, 2021).

This work seeks to extend the investigation of the human-likeness of quantitative information derived from LLMs even further, by exploring whether LLM-derived measures can feed into probabilistic models predicting the full distribution of human answers in a forced-choice tasks. The question after the human-likeness of quantitative LLM-derived information matters for applications which use numerical scores to rank or weigh options (e.g., Park et al., 2023; Zhang et al., 2023). Moreover, to the extent that LLMs are used as part of explanatory “neuro-symbolic models” of information processing (Garcez and Lamb, 2020), understanding whether and how LLMs might yield full-fledged distributional predictions is important, e.g., to explore their integration into probabilistic (cognitive) models (c.f., Frank, 2023).

[*] Corresponding author.

A major contribution of this work is methodological. By embedding numerical scores from LLMs into probabilistic models for human multiple-choice data, we can use Bayesian data analysis to train, test and compare different LLM-derived predictor scores. [MF: describe high-level results here].

The paper is structured as follows. Section 2 further motivates the desire to assess quantitative information from LLMs, and it introduces the distinction between item-level and condition-level predictions. Section 3 looks at human data from a simple but non-trivial forced-choice experiment, namely a text-based reference game, for which Section 4 introduces a salient probabilistic cognitive model (Frank and Goodman, 2012). Section 5 investigates whether LLM-derived item-level predictions are adequate to capture the human data at the item-level and finds that they are not, even though a condition-level predictor from the RSA model is. Section 6 then discusses different ways of deriving probabilistic predictions from LLMs at the condition-level and compares them against the human data and each other. We find that not every way of constructing condition-level predictions from item-level scores is equally good. [MF: grant conclusion?]

2 Motivation: Why care for distributional predictions?

The prevalent approach to assessing LLM performance is based on benchmark-testing, using large data sets for tasks in which a designated goal answer exists. Tasks often contain multiple-choice options, and accuracy for a task is determined with a "winner-takes-all" (WTA) strategy (e.g., BIG-bench authors, 2023). Let $\{I_1, \dots, I_m\}$ be m be instances of the same task, or items belonging to the same (logical) condition in a behavioral experiment. Each item $I_k = \langle x_k, \langle y_{k1}, \dots, y_{kl} \rangle \rangle$ consists of an input prompt x_k, which is a string of text, and l choice options $\langle y_{k1}, \dots, y_{kl} \rangle$, all of which are strings as well, possibly composed of $|y_{ki}|$ words, $y_{ki} = w_{ki1}, \dots, w_{ki|y_{ki}|}$. Among the choice options is a designated *target option* y_{ki^*} that is assumed to be the "true" or to-be-selected goal answer. The most obvious *item-level score* an (autoregressive) LLM provides for each choice option y_{ki} is its log-probability:[2]

$$S\left(y_{ki}, x_k\right) = \log P_{\text{LLM}}\left(y_{ki} \mid x_k\right) = \sum_{j=1}^{|y_{ki}|} \log P_{\text{LLM}}\left(w_{kij} \mid x_k, w_{ki1}, \dots, w_{ki(j-1)}\right) .$$

The WTA approach considers item I_k to be answered correctly if the target option y_{ki^*} is the option that maximizes the item-level score. The WTA-based accuracy of the LLM on the given task is the proportion of items which are answered correctly.

[MF: the following paragraphs should be shortened and sharpened] While generally a pragmatic and useful approach, there are several reasons why going beyond the best guess can be beneficial. Consider first a high-level conceptual argument. If task performance is categorical (in the most extreme case: binary), somewhere along the path from a numerical item-level score to accuracy a discontinuity has to be introduced. Every such discontinuity bottleneck entails loss of information. As this information might be useful, discontinuity should ideally happen as late as possible; or even better: not at all, unless it is irrelevant. Consequently, the question becomes: when is the additional information inherent in the scores (not) relevant for assessing LLM performance?

The WTA approach implicitly assumes that item-level choices are resolved by a greedy-like choice of the best alternative. It is therefore a good measure of accuracy when the LLM's main use cases are of a greedy-like nature as well. When applications rely on a less extreme sampling based strategy, the WTA approach can be generalized to softmax sampling. The probability of choosing option y_{ki} for item I_k is:

$$P_{\text{choose}}(y_{ki}) \propto \exp\left[\alpha \, S\left(y_{ki}, x_k\right)\right] ,$$

and the accuracy is the average probability of choosing the designated target option y_{ki^*}. For $\alpha \to \infty$, this softmax strategy converges to the WTA strategy. But in general, these two approaches can yield even categorically different predictions.

Imagine that there are two options, and that the target option's score is a small ϵ better in 80% of the task's items, and otherwise worse. The WTA-based accuracy is 0.8. This number is useful as a performance measure for applications in which the LLM is used in exactly the way the WTA strategy describes, e.g., any implementation which is outcome equivalent to greedy decoding with rejection sampling on a domain that contains only the available options. For such a case, it never matters how much worse the goal answer is scored in the 20% of the cases where it is not the maximum. As only the best option will be chosen, that information is irrelevant. But if an application uses anything other than

[2] More elaborate item-level scores include corrections for variable length of answer options (e.g., Brown et al., 2020) or variation in base rate among answer options (e.g., Holtzman et al., 2021). From the point of view of experimental psychology, these corrections are *post hoc* fixes to improperly balanced experimental materials. For the purposes of this paper, where answer options are all equally long and commensurable, these corrections may be temporarily ignored for simplicity.

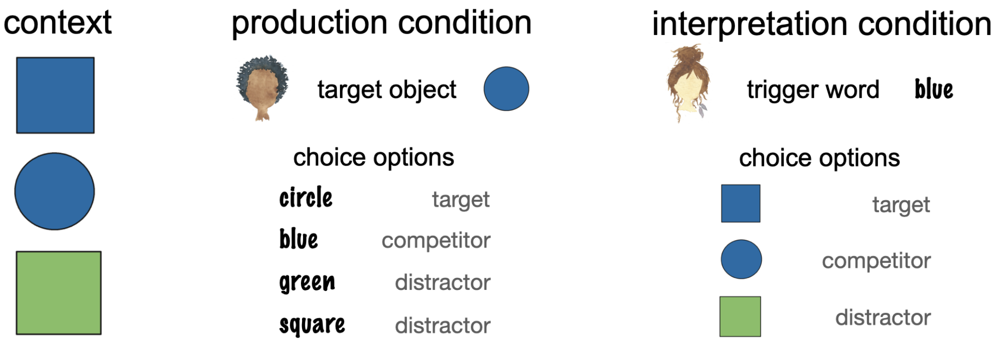


Figure 1: Structure of a reference game with human participants. Each trial consists of a set of objects, the so-called context. In production trials, participants choose a single word to describe a target object from the context. In interpretation trials, an object is selected as the likely object a trigger word referring to.

greedy-like responses, the accuracy score of 0.8 may be misleading. If the remaining 20% of the items are such that the non-goal option is almost infinitely better, it would be chosen under a pure sampling strategy, where $\alpha = 1$, with virtual certainty, so the softmax-based accuracy would around 0.4.[3] This is a categorical shift from predominantly goal answer to predominantly non-goal answer. The argument holds equally if numbers for the two options are reversed, so that there is no way of saying which of the two measures of accuracy would generally be more favorable for selecting the target option.

Differences in accuracy scores between a WTA-based or a more general softmax-based strategy depend on the variation in item-level scores, in particular the relation between score-ordering and score-differences. (The example in the previous paragraph was set up deliberately to have one score-ordering for very small score-differences, and another for huge score-differences.) This raises an important general question, namely how to deal with item-level variation in a system's predictions in the first place. Data from forced-choice tasks with human participants will likely show item-level variation as well, which is why hierarchical modeling accommodating item-level differences is recommended in statistical analyses (e.g., Jaeger, 2008; Barr et al., 2013). But whether item-level variation in LLM scores aligns with differences in human choice rates for each item is an empirical question that has to be addressed by explicit modeling and comparison; an exercise that this paper starts to take on. On *a priori* grounds, the WTA-based strategy may be deemed implausible as a predictor of human-level choice behavior for a simple item: it is unlikely that, for each item, there is a single choice option that (almost) *all* participants will choose.

To the extent that machine predictions show extreme, unsystematic, erratic item-level variation which is quite unlike that of human participants in comparable experiments, an alternative strategy suggests itself, namely to *not* judge a system by its average item-level predictions, but based on its task- or condition-level predictions obtained from different ways of aggregating over item-level variation. Section 6 therefore introduces and compares different ways of deriving condition-level predictions from (variable) item-level scores. Properly constructed condition-level predictions might be more aligned with average human choice rates, despite the fact that each complex system (machine and human) shows quite different underlying item-level variation. In other words, we should ask whether modern LLMs can predict aggregate human data by *some* process of aggregating its item-level scores, because it might turn out that predicting aggregates is the best the LLMs can do (in particular cases).

3 Experiment: Reference games

Reference games are an established, well-understood and austere experimental paradigm to test human decision making in abstract communicative tasks (e.g., Frank and Goodman, 2012; Degen, Franke, and Jäger, 2013; Qing and Franke, 2015; Frank, 2016; Sikos et al., 2021). A reference game consists of two players, a speaker and an interpreter, who jointly observe a set of objects, usually referred to as context (see Figure 1). In the **production condition**, the speaker is assigned a *target object* from the context set which they have to describe to the interpreter. In the **interpretation condition**, the interpreter observes a description, here called *trigger word*, and chooses one of the objects from the context set. The goal of the game is, for the speaker, to choose a description that enables the interpreter to choose the target object; and, for the interpreter, to guess correctly which object the speaker had in mind.

[3] The probability of the target option in the 80% of items where the goal answer is slightly better is 0.5 in the limit of $\epsilon \to 0$, and it is virtually 0 in the remaining 20% of the cases. This gives an expected rate of: $^4/_5\ ^1/_2 + {}^1/_5\ 0 = {}^2/_5$.

The example in Figure 1 is a standard case, which we will use throughout, where choices are informative about the pragmatic reasoning that decision makers engage in. In this example, there are two features that differ across three objects (here shape and color). One object shares both its color and shape with one other object, while the two other objects have one unique feature (e.g., being the only circle, or the only green object). In a critical production trial, the target object to describe is one of the two objects with a unique feature. The speaker has four words to choose from. The **target utterance** is the word which uniquely describes the target object. The **distractor utterance** is the word that is true of the target object, but also true of another object. The other utterances, both of which are false of the target are **competitor utterance**. In a critical interpretation trial, the trigger word is one that is true of two of the three objects. If participants engage in pragmatic thought, they might reason that *if* the speaker had wanted to refer to one of the two objects of which the trigger word is true (blue square and blue circle in Figure 1), the speaker could have used a more informative word for exactly one of those two objects ("circle"), so they are more likely to refer to the **target object** (the blue square in Figure 1). The **competitor object** is the other object of which the trigger word is true. The **distractor object** is the object of which the trigger word is false.

We replicated a simple reference game with human participants in which each trial instantiated the structure of the example shown in Figure 1. While previous reference games with human participants used pictorial representations of objects, and sometimes even pictorial representations of messages, we implemented a text-only version in order to be able to compare the predictions of LLMs for the exact same stimuli. The experiment was realized as an online task using `magpie` (Franke, Ji, et al., 2023).[4]

3.1 Participants

A total of 302 participants were recruited via Prolific for monetary compensation (£0.45, corresponding to roughly £15.40 per hour). All participants self-identified as native speakers of English.

3.2 Materials & design

We created 100 different items as stimulus material via a stochastic process. Each item is a different textual description of a reference game with the same logical structure as the example from Figure 1. For each item, the context consisted of three objects. Objects are defined by a triple of properties, namely a color, a shape and a texture. For each property, there were four possible values, e.g., blue, green, red, and orange for color. The sampled items differed in terms of the properties of the objects in the context set, and in terms of the order in which the objects and expression alternatives were presented in the text. Figures 6 and 7 from Appendix A show example screenshots from the experiment.

3.3 Procedure

For each participant the experiment sampled four different items. Participants first played two of these in the production condition, then the other two in the interpretation condition.

3.4 Results

The overall distribution of choices that correspond to the target, competitor, and distractor states is shown in Figure 2.[5] It is interesting that the distractor options were chosen rather often. We also see that the number of target choices is higher in the production condition than in the interpretation condition. This is in line with previous experimental results on human reference games. For example, in previous forced-choice reference games with human participants with pictorial presentations of objects, Qing and Franke (2015) observed the following proportions: $\langle 0.882, 0.118, 0 \rangle$ in the production and $\langle 0.492, 0.506, 0.003 \rangle$ in the interpretation condition (for 288 observations in each condition).

4 Model predictions from probabilistic pragmatics

Data from reference games with human participants have been variously analyzed with probabilistic models using inspiration from behavioral game theory (e.g., Degen, Franke, and Jäger, 2013), probabilistic Bayesian modeling (e.g., Frank and Goodman, 2012) or other forms of probabilistic modeling (e.g., Gatt et al., 2013). Common to these approaches is that they derive or define, based on some explicit conceptual motivation, a parameterized stochastic

[4] The code for the experiment can be found at https://github.com/magpie-ea/magpie3-text-refgame, and a live version of the experiment can be tested at https://magpie-ea.github.io/magpie3-text-refgame/.

[5] The production condition actually has two distractor choices. Here and in the following, these are lumped together as a single category, also when modeling random errors in later models. This is a simplification but does not change anything of substance.

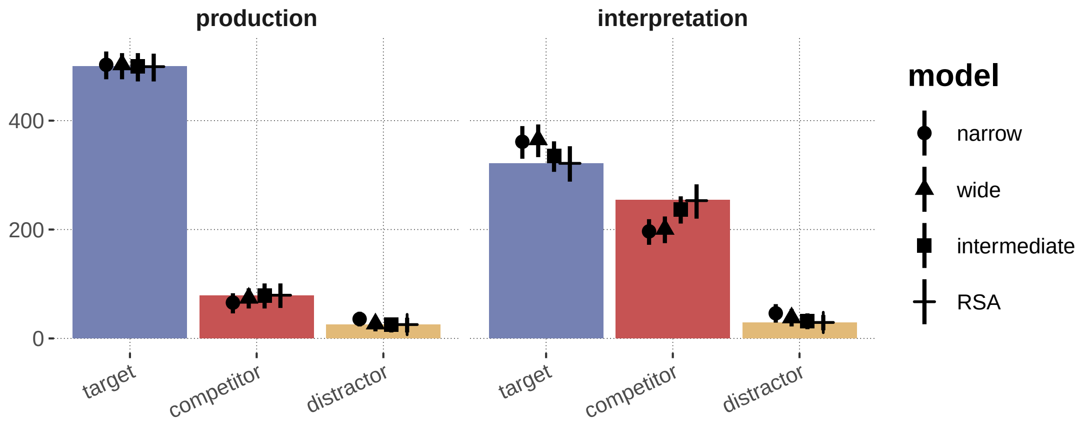


Figure 2: Counts of choices from reference games with human participants (colored bars), with summary statistics from the posterior predictive distribution of four models (shapes and error bars). Shapes show the mean of the posterior predictive distributions of the RSA model and three further models derived from item-level LLM-supplied scores. Error-bars show corresponding 95% credible intervals of the posterior predictive.

speaker policy, $P_S(u \mid s; \theta_S)$, modulated by parameters θ_S, for a speaker's choice of expression or utterance u given a referent or state s, which the speaker wants to communicate; and a stochastic listener policy, $P_L(s \mid u; \theta_L)$, capturing the probability of choosing a referent s for utterance u.

As a concrete example, we introduce the Rational Speech Act (RSA) model first described in this form by Frank and Goodman (2012) (for overview see Franke and Jäger, 2016; Goodman and Frank, 2016; Stevens and Benz, 2018; Scontras, Tessler, and Franke, 2021; Degen, 2023). The RSA model defines pragmatic reasoning as a sequence of iterated (soft-)optmization of policies, grounding out in literal interpretation. If $\mathfrak{L}(s, u) \mapsto \{0, 1\}$ is a semantic meaning function mapping each pair of state s and utterance u to a (binary) truth-value, and if P_s is a prior over states, a literal listener policy is defined as:

$$P_{L_0}(s \mid u) \propto \mathfrak{L}(s, u) P(s) .$$

The pragmatic speaker soft-optimizes the choice of utterance to minimize the literal listener's surprisal for the to-be-communicated state:

$$P_S(u_i \mid s, \alpha) \propto \exp\left[\log P_{L_0}(s \mid u_i)\right] .$$

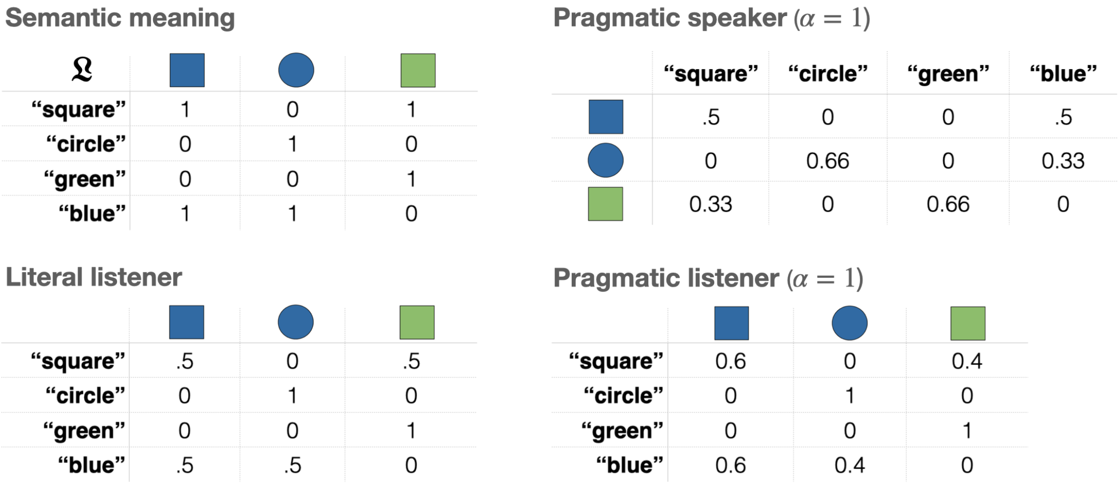
Finally, the pragmatic listener is defined as the policy resulting from applying Bayes rule, solving the inverse-problem for the previously defined speaker policy:

$$P_L(s \mid u, \alpha) \propto P_S(u \mid s, \alpha)\ P(s) .$$

Figure 3 gives example calculations (assuming flat priors and $\alpha = 1$) for the reference game from Figure 1. For $\alpha = 1$, the model predicts that the probabilities of target, competitor and distractor options are $\langle {}^2/_3, {}^1/_3, 0 \rangle$ for the production, and $\langle 0.6, 0.4, 0 \rangle$ for the interpretation condition. Increasing α will increase the odds of target over competitor choices. Yet, the model also predicts probability zero for distractor choices, so that the human data shown in Figure 2, where the distractor option was chosen in both conditions, would immediately rule out the model entirely. It is therefore common to include a random error probability for each choice (e.g., Lee and Wagenmakers, 2015), such as via Laplace smoothing: if $P_r(R_l, C, \alpha_c)$ is the RSA model's probabilistic prediction for response category R_l (target, competitor, or distractor) for condition C (production or interpretation) and condition-specific optimality α_c, the random-error smoothed prediction is:

$$P_r(R_l, , C; \alpha_c, \epsilon_c) \propto P_r(R_l, C; \alpha_c) + \frac{\epsilon_c}{3} ,$$

where ϵ_c is a (condition-specific) parameter giving the probability that a choice was made by randomly guessing. The result is a four-parameter model, one pair of parameters per condition, which provides a likelihood function for the categorical choice data and so can be fitted to the data and compared against other probabilistic models providing a likelihood function for the same data.



Semantic meaning

$\mathfrak{L}$	🟦	🔵	🟩
"square"	1	0	1
"circle"	0	1	0
"green"	0	0	1
"blue"	1	1	0

Pragmatic speaker ($\alpha = 1$)

	"square"	**"circle"**	**"green"**	**"blue"**
🟦	.5	0	0	.5
🔵	0	0.66	0	0.33
🟩	0.33	0	0.66	0

Literal listener

	🟦	🔵	🟩
"square"	.5	0	.5
"circle"	0	1	0
"green"	0	0	1
"blue"	.5	.5	0

Pragmatic listener ($\alpha = 1$)

	🟦	🔵	🟩
"square"	0.6	0	0.4
"circle"	0	1	0
"green"	0	0	1
"blue"	0.6	0.4	0

Figure 3: Example of predictions from the RSA model (with $\alpha = 1$). The semantic meaning is shown as a matrix of binary truth-values. The policies of literal listener, pragmatic speaker and listener are calculated for uniform priors over state (referents), assume $\alpha = 1$ and are shown as row-stochastic matrices.

Parameterized predictions like $P_r(R_l, , C; \alpha_c, \epsilon_c)$ can be assessed in the light of the empirical data with the usual tools of Bayesian data analysis (e.g. Gelman et al., 2014; McElreath, 2016; Lambert, 2018). Let $\alpha_c \sim$ log-Normal(0.5, 1) have a reasonably wide log-Normal prior, and let $\epsilon_c \sim$ Beta(1, 5) have a Beta prior favoring small values. Using Stan (Stan Development Team, 2023) for Bayesian inference, we obtain estimates of posterior credible values of model parameters (summary statistics of which are shown in Table 1). [MF: include information on MCMC, samples, warm-up etc.; follow Kruschke recipe]

To assess goodness-of-fit, we can use the *posterior predictive distribution*, i.e., the model's predictions about data of the same size and structure as the training data. [MF: maybe insert formal definition of post-pred distribution?] As a minimal bar, we would require a model trained on observed data D_{obs} to not be surprised by D_{obs}. Figure 2 shows summary statistics (means and 95% credible intervals) for the posterior predictive distribution of the RSA model (among other models to be introduced later). We see that the RSA model passes this "visual posterior predictive check" (Kruschke, 2015) for both conditions. To corroborate the visual impression, Table 2 shows sample-based estimates of Bayesian posterior predictive p-values, using likelihood of the observed data as a test statistics. [MF: maybe insert formal definition of BPPPV?] These values approximate the probability that a model trained on D_{obs} would predict future data of the same size and format that is at least as unlikely as the data D_{obs} itself. Low values therefore indicate that the trained model would be highly surprised by the data it was trained on; an indication that it failed to capture something essential about the training data. The results from Table 2 show that the RSA model passes this numerical test of model adequacy.

The following sections apply these test of Bayesian model criticism also to models built around predictor values from LLMs. Section 5 first looks at the data from each item individually, before Section 6 investigates different ways for deriving condition-level predictions.

5 Item-level predictions from LLMs

Figure 2 shows the condition-level counts for choices for all of the 100 different items combined. An LLM, on the other hand, first and foremost makes predictions about each item. The items of the reference game experiment differ in the task-relevant features (color, shape, texture), as well as the order of presentation of states and words. We expect both human and LLM predictions to vary between different items: e.g., humans seem to have preferences for some features (Qing and Franke, 2015), maybe even leading to over-production of informationally irrelvant material (Degen, Hawkins, et al., 2020) [MF: more refs on (color) over-specification]; and machines may be susceptible to the presentation of the order of choice options. [MF: references from the NLP literature on effects of prompt variation?] The main question to be addressed in this section is whether a standard item-level score from LLMs provides a good fit, if we aim at predicting the human data separately for each item, not as a condition-level average.

For a set-up in which the task descriptions and choice options are all text-based, predictions of LLMs at the item-level can be derived as a function of the (log) probabilities assigned to the continuations corresponding to each choice

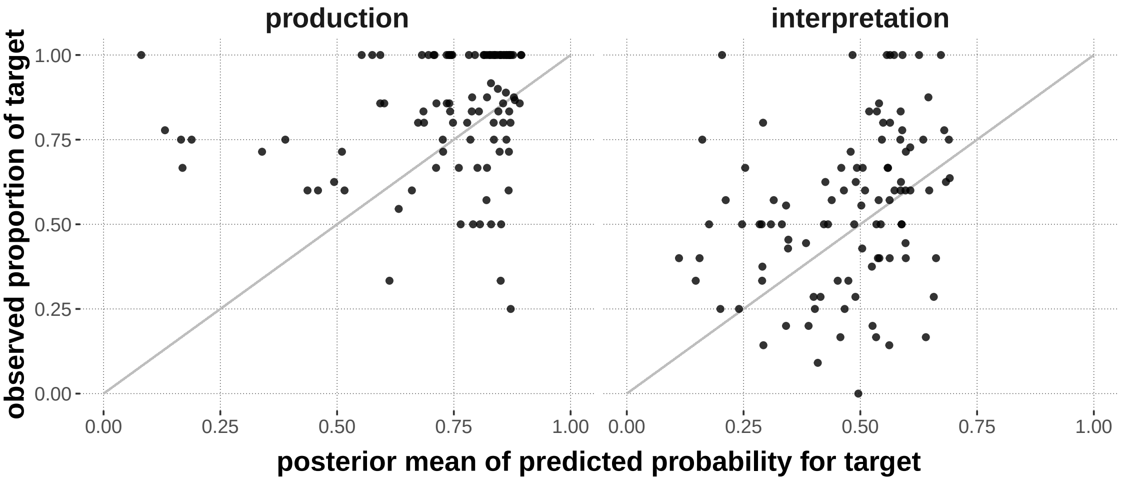


Figure 4: Item-level prediction-observation plot. Each dot represents an item. The x-coordinate represents the mean of the posterior prediction for the target choice probability for the given item. The y-coordinate represents the observed proportion of target choices for that item. The gray line (identity function) is the ideal prediction.

option after processing the task description. Let C be the experimental condition of interest (here: the production or interpretation condition). Let $\{I_1, \dots I_m\}$ be the multi-set of items that occurred in the human experiment for condition C.[6] Each item I_k consists of a task description T_k and l choice options $O_{k1}, \dots, O_{kl}$, which are treated as textual continuations of the task description (see Appendix B for an example). For each item I_k, let $F(R_l, I_k)$ be the choice option O_{ki} that corresponds to *response type* R_l (here: target, competitor, distractor).[7] The *item-level score* $S(O_{ki}, I_k)$ for option $O_{ki} = w_{ki1}, \dots w_{kin}$ of item I_k is defined as the log-probability of continuation O_{ki} after prompt T_k:[8]

$$S(O_{ki}, I_k) = \sum_{j=1}^{n} \log P_{\text{LLM}} \left(w_{kij} \mid T_k, w_{ki1}, \dots, w_{ki(j-1)} \right) .$$

The item-level score of response type R_l is $S(R_l, I_k) = S(F(R_l, I_k))$.

Using item-level scores from GPT-3.5, we fit a model to the human data at the item-level. The data to be explained is the set of all triples of response type counts $\langle t_k, c_k, d_k \rangle$ (target, competitor, distractor), separately for each item I_k. The model's predictions for the data from item I_k from condition C (production or interpretation) are:

$$P_{\text{item}}(R_l, C, I_k, \alpha_c, \epsilon_c) \propto \frac{\exp(\alpha_c \, S(R_l, I_k))}{\sum_{l'} \exp(\alpha_c \, S(R_{l'}, I_k))} + \epsilon_c \, .$$

In words, we use a single pair of parameters α_c and ϵ_c, one for each condition, take the LLM's item-level scores for each item, and derive the response probabilities in the exact same way as before. Priors on parameters are as for the RSA model described in Section 4.

Figure 4 shows that the item-level LLM scores predict variance which is not borne out by the human data. Concretely, the plot shows, for each item, mean posterior estimates of the model's predicted probability of choosing the target option (x-axis), together with the observed proportion of target choices in the human data (y-axis). There is ample variation in the model's predictions, especially visible in the production condition, owing to the fact that the item-level scores of the LLM sometimes clearly favor another option than the target choice. So, the model itself predicts systematic variability at the item level. The human data, too, show variability at the item-level, but there is no (visual) indication that the item-level variability predicted by the LLMs is borne out by the human data.

[6] By using a multi-set, which may contain a single task multiple times, we produce aggregate predictions for exactly the set of item that the participant group saw, which provides the most fitting counterpart to the human data.

[7] In the current set-up the response type "distractor" has two instantiations in the production condition. Since choice options are a single word in the production condition, for simplicity, we lump both of the distractor words together and treat them as a single option by taking the sum of the log-probabilities for both distractor words.

[8] Numerically, we worked with *average* log-probabilities. As in our experiment, all choice options had the same number of words, this is largely irrelevant, except for the effect of the α parameter in the interpretation condition. Using length-corrected log-probabilities is the same as using raw log-probabilities with an α parameter divided by the number of words. In this way, we make the α parameter more meaningfully comparable across conditions.

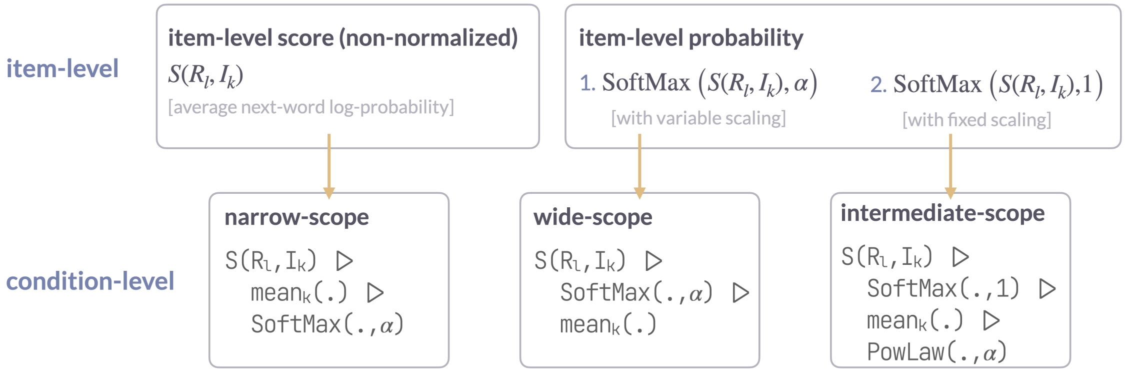


Figure 5: Schematic overview over three different approaches of obtaining condition-level predictions by aggregating over item-level scores. The basic item-level score is average next-word log probability. This can be taken as-is into averaging (narrow-scope aggregation), or first transformed from log- into probability-space, either with optimization before (wide scope) or after averaging (intermediate scope). The notation for the condition-level options uses pseudo-code to more clearly reflect the relevant operator scope differences (where triangles represent the pipe operator).

Sampling-based approximations of Bayesian posterior predictive p-values for the by-item analysis are very low (0.0001 for production, 0 for interpretation), suggesting that the unaggregated item-level LLM predictions are inadequate predictors of the human data. In contrast, using the global, item-independent RSA model predictions to fit the item-level data, we obtain estimates of posterior predictive p-values that do not discredit the model (0.338 for production, 0.252 for interpretation). These results suggest that LLM-based probabilistic predictions may imply item-level variance that is not attested in the human data. Put more strongly, a model that seeks to predict what human participants choose on a by-item level, based on the most obvious item-level score, is ruled out by experimental data.

6 Condition-level predictions

This section explores different ways of deriving a probabilistic likelihood function for the human data from an LLM at the aggregate, condition-level. Using log-probabilities for each answer option as a basic item-level score, there are at least three conceivable ways of deriving condition-level predictions by averaging over item-level variation (see Figure 5). In the following, we first describe the different options of deriving LLM-based predictor terms. We then compare all probabilistic models based on their adequacy of explaining the human choice data.

Condition-level predictions are a probability distribution over response types, obtained from item-level scores $S(R_l, I_k)$ by averaging over all items belonging to the relevant condition. To map non-normalized scores $\mathbf{s} = \langle s_1, \dots, s_l \rangle$ to probabilities $\mathbf{p} = \langle p_1, \dots, p_l \rangle$ with different degrees of optimization, a common choice is the softmax function with optimality parameter α:

$$\text{SoftMax}(\mathbf{s}, \alpha) = \mathbf{p}, \text{where } p_i \propto \exp(\alpha p_i) \ .$$

The softmax function can furthermore be decomposed into a first step of mapping scores to probabilities, and subsequently reweighing probabilities via a power-law transformation (c.f., Franke and Degen, 2023):

$$\begin{aligned} \text{SoftMax}(\mathbf{s}, \alpha) &= \text{Pow}(\text{SoftMax}(\mathbf{s}, 1); \alpha) \\ \text{Pow}(\mathbf{p}; \alpha) &= \mathbf{q}, \text{where } q_i \propto p_i^\alpha \end{aligned}$$

This means that there are three conceivable scope-sites for aggregating item-level information as shown in Figure 5. Narrow-scope aggregation first aggregates the item-level scores, and then transposes the averages into (scaled) probabilities:

$$P_n(R_l, C; \alpha) \propto \exp\left[\alpha \frac{1}{m} \sum_{i=k}^{m} S(R_l, I_k)\right] \qquad \text{[narrow-scope aggregation]}$$

Wide-scope aggregation, first transposes scores into probabilities, scales them and only aggregates over items last.

$$P_w(R_l, C; \alpha) \propto \frac{1}{m} \sum_{i=k}^{m} \frac{\exp(\alpha\, S(R_l, I_k))}{\sum_{l'} \exp(\alpha\, S(R_{l'}, I_k))} \qquad \text{[wide-scope aggregation]}$$

model	condition	Parameter	|95%	mean	95%|
narrow	production	α	0.19	0.22	0.26
		ϵ	0.00	0.04	0.11
	interpretation	α	0.15	0.18	0.20
		ϵ	0.00	0.02	0.06
intermediate	production	α	0.86	1.04	1.22
		ϵ	0.08	0.13	0.18
	interpretation	α	0.36	0.48	0.60
		ϵ	0.00	0.05	0.12
wide	production	α	0.22	2.69	8.79
		ϵ	0.08	0.13	0.19
	interpretation	α	0.19	1.08	6.03
		ϵ	0.00	0.06	0.18
RSA	production	α	2.59	3.16	3.71
		ϵ	0.08	0.13	0.18
	interpretation	α	0.24	0.66	1.07
		ϵ	0.10	0.15	0.20

Table 1: Summary statistics of posterior samples. Number indicate estimates for lower and upper bounds of 95% credible intervals and the posterior means.

Intermediate-scope aggregation performs item-level averaging after mapping scores onto probabilities (using softmax with $\alpha = 1$), but before scaling (with a power-law transformation with variable α):

$$P_i(R_l, C; \alpha) \propto \left[\frac{1}{m} \sum_{i=k}^{m} \frac{\exp\left(S\left(R_l, I_k\right)\right)}{\sum_{l'} \exp\left(S\left(R_{l'}, I_k\right)\right)} \right]^{\alpha} \qquad \text{[intermediate-scope aggregation]}$$

All three condition-level predictors coincide when there is only one item, of course. For more items, wide-scope and intermediate-scope aggregation coincide when $\alpha = 1$. In all other cases, the predictors are not guaranteed to be identical. Conceptually, the main difference is what each measure considers to be the basic item-level unit to aggregate over. Narrow-scope aggregation considers raw scores, which are *not* probabilistic predictions at the item level. Wide-scope aggregation considers α-optimized probabilities at the item level, which is compatible with a procedure of sampling, via softmax decoding, item-level answer options. Wide-scope aggregation contains the "winner-takes-all" procedure for accuracy scoring in common benchmark tests for $\alpha \to \infty$. Finally, like wide-scope aggregation, intermediate-scope aggregation is also compatible with a sampling based picture, but would use pure decoding (without α-optimization) at the item level and reserves the scaling of probability distributions to a stage *after* aggregation.

The three condition-level predictions yield three different probabilistic models, each with two pairs of condition-dependent parameters α_c, and ϵ_c. Using priors and Bayesian posterior estimation as previously described for the RSA (see Section 4), we obtain samples from the posterior over parameters and samples from the posterior predictive distributions. The usual Bayesian summary statistics for posteriors over model parameters are shown in Table 1. What is noteworthy about these estimates is that the α_c parameters for the wide-scope model have a very large credible range. This is because, already for values of $\alpha_c = 1$, the item-level predictions are virtually "winner-takes-all" choices, selecting the option with the maximum score with probability almost 1 for almost all items. This implies that, by studying the posterior predictive distribution of the wide-scope model, we are virtually testing condition-level predictions from the standard WTA-strategy used for benchmark testing (see Section 2).

Figure 2 shows the summary statistics (means and 95% credible intervals) for each model's posterior predictive distribution. We find that only the theoretical model (RSA) and the intermediate-scope model passes this "visual posterior predictive check" for both conditions; the other two models both overpredict the target choice rate and underpredict the competitor choice rate in the interpretation condition. To corroborate the visual impression, Table 2 shows sample-based estimates of Bayesian posterior predictive p-values, using likelihood of the observed data as a test statistics. Consequently, the results from Table 2 suggest that the narrow-scope model fails on the interpretation data, and is at most borderline compatible with the production data; that the wide-scope model is able to reproduce the production data, but fails on the interpretation data; and that only the intermediate-scope model is able to fully recover the data from both conditions.

	narrow	wide	intermediate	RSA
production	0.05	0.51	0.49	0.49
interpretation	0.00	0.00	0.26	0.52

Table 2: Sample-based estimates of Bayesian posterior predictive p-values for each model and each condition, based on likelihood of the observed data as test statistic.

These results tell us that not all ways of deriving condition-level predictions by averaging over item-level variation are equally good. Some approaches clearly fail basic checks for statistical goodness-of-fit. On the positive side, we also find that there is at least one model with predictors based on LLM-measures, namely the intermediate-scope model, which is able to recover the patterns in the data. In other words, there is a way of deriving predictor values for condition-level forced-choice probabilities from an LLM such that, when fed into a common linking function (here with parameters α for optimization and ϵ for random error), the human choice probability can be reconstructed faithfully in its entirety.[9] This implies that using LLM predictors for probabilistic predictions, such as in a neuro-symbolic model, might be possible if embedded in the proper link functions and if item-level variation is averaged out in the right way.

7 Conclusion

It is not entirely ludicrous to use numerical predictions from LLMs are part of predictive probabilistic models. But we have to be careful. Ideally, we average out over low-level variation, such as from order of presentation or similar "nuisance," at least as long as we do not understand what causes this variation in the predictions of models and further research that investigates when exactly this variation accords with empirically observed patterns.

This also implies that we should likely *not* (yet) aspire to use LLMs are models or individual- or item-level predictors.

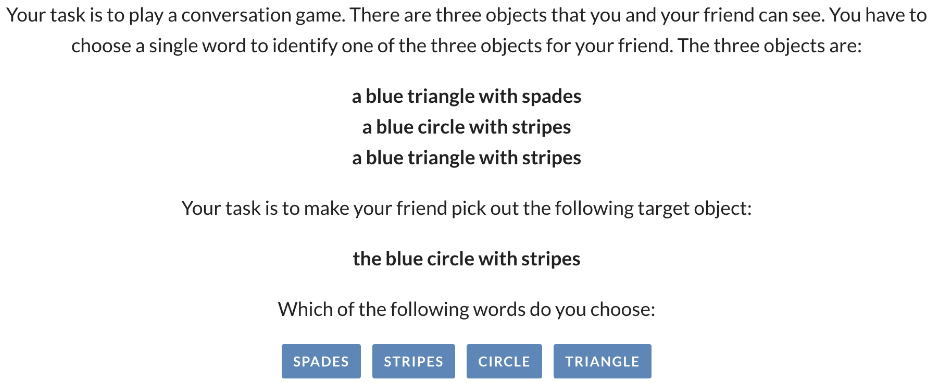
(TODO: rethink conclusion)

to be continued

[9] A potential worry is that the intermediate-scope model is trivial in the sense that it could have predicted *any* data observation. Appendix C shows that this is not the case.

A Screenshots from the online experiment with human participants

Figure 6 shows a trial from the production condition, Figure 7 one for the interpretation condition of the online experiment.



Your task is to play a conversation game. There are three objects that you and your friend can see. You have to choose a single word to identify one of the three objects for your friend. The three objects are:

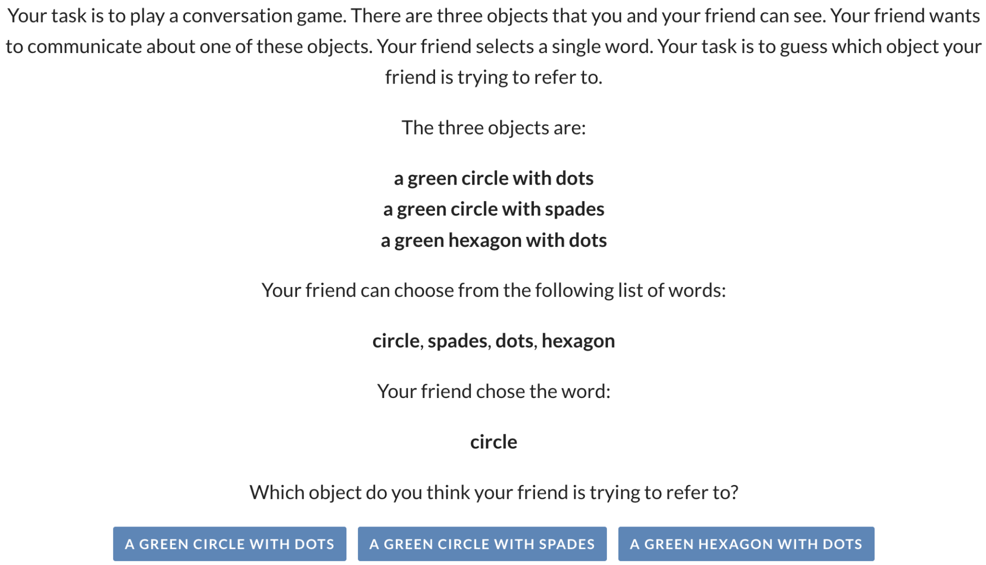
a blue triangle with spades
a blue circle with stripes
a blue triangle with stripes

Your task is to make your friend pick out the following target object:

the blue circle with stripes

Which of the following words do you choose:

Figure 6: Screen shot from a production trial of the online experiment.



Your task is to play a conversation game. There are three objects that you and your friend can see. Your friend wants to communicate about one of these objects. Your friend selects a single word. Your task is to guess which object your friend is trying to refer to.

The three objects are:

a green circle with dots
a green circle with spades
a green hexagon with dots

Your friend can choose from the following list of words:

circle, **spades**, **dots**, **hexagon**

Your friend chose the word:

circle

Which object do you think your friend is trying to refer to?

Figure 7: Screen shot from an interpretation trial of the online experiment.

B Example item for the LLM experiment

The text-based input for the LLM predictions mirrors the text in the human experiment, except that the LLM input also lists the set of all available choice options (which for the human experiment is unnecessary since this information is given by the buttons for the forced-choice selection). For example, the task description T_k for the item that corresponds to the production trial shown in Figure 6 is shown below (the actual input has no line breaks in the first paragraph):

```
Your task is to play a conversation game. There are three objects that
you and your friend can see. You have to choose a single word to identify
one of the three objects for your friend.

The three objects are:
```

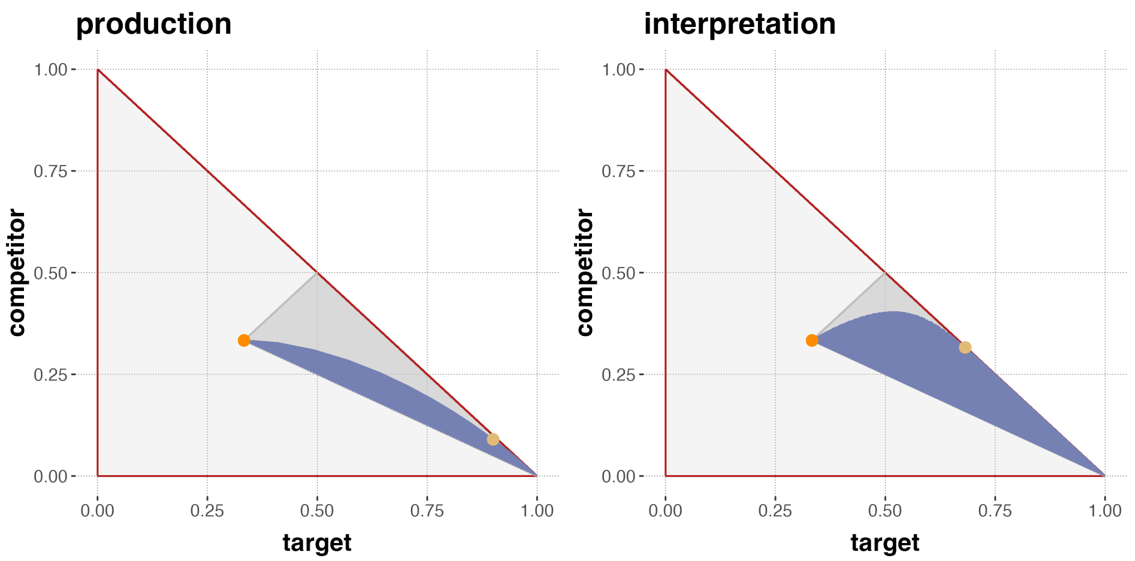


Figure 8: Range of predictions the intermediate-scope model makes for any pair of parameter values α_c and ϵ_c. The light gray triangle with the red boundary is the probability simplex, i.e., the space of all possible 3-place probability vectors. The darker gray area in gray boundary lines is the subspace that is compatible with the ordering that the probability of choosing the target is bigger than that of the competitor, which in turn is bigger than that of the distractor. The blue area is the subspace of probabilistic predictions the intermediate-scope model makes under any value of its parameters. The orange dot in the middle is the "Laplace point" of equal probability for all three options, which the model predicts for $\alpha = 0$ or $\epsilon = 1$. The yellow dot is the "vanilla prediction" for $\alpha_c = 1$ and $\epsilon_c = 0$.

```
a blue triangle with spades
a blue circle with stripes
a blue triangle with stripes

Your task is to make your friend pick out the following target object:

the blue circle with stripes

Which of the following words would you choose:

spades
stripes
circle
triangle

Your answer:

I would choose the word
```

C Range of prior predictions of intermediate-scope model

It may seem that, given the freedom to adjust optimization α and random guessing rate ϵ, some models might be able to explain *any* data observation. The intermediate-scope model, which was not discredited by model criticism in terms of recovery-based posterior predictive checks, is not trivial in this sense. Figure 8 shows the range of predictions that the intermediate-scope model is, in principle, capable of making (for the whole possible range of α and ϵ values, disregarding Bayesian priors). There are distributions over response types which the model does not explain predict *ex ante*. For example, the model does not predict cases where the number of competitor choices is almost as high as the number of target choice, as would be the case if participants would consistently *not* engage in pragmatic reasoning.

References

Barr, Dale J. et al. (2013). “Random effects structure in mixed-effects models: Keep it maximal”. In: *Journal of Memory and Language* 68.3, pp. 255–278.

BIG-bench authors (2023). “Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models”. In: *Transactions on Machine Learning Research*. ISSN: 2835-8856.

Binz, Marcel and Eric Schulz (2023). “Using cognitive psychology to understand GPT-3”. In: *Proceedings of the National Academy of Sciences* 120.6, e2218523120. eprint: `https://www.pnas.org/doi/pdf/10.1073/pnas.2218523120`.

Bommasani, Rishi et al. (2021). “On the opportunities and risks of foundation models”. In: *arXiv preprint arXiv:2108.07258*.

Brown, Tom B. et al. (2020). *Language Models are Few-Shot Learners*. arXiv: `2005.14165 [cs.CL]`.

Chung, Hyung Won et al. (2022). *Scaling Instruction-Finetuned Language Models*. arXiv: `2210.11416 [cs.LG]`.

Creswell, Antonia, Murray Shanahan, and Irina Higgins (2022). *Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning*. arXiv: `2205.09712 [cs.AI]`.

Degen, Judith (2023). “The Rational Speech Act Framework”. In: *Annual Review of Linguistics* 9.1, pp. 519–540.

Degen, Judith, Michael Franke, and Gerhard Jäger (2013). “Cost-Based Pragmatic Inference about Referential Expressions”. In: *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*. Ed. by Markus Knauff et al. Austin, TX: Cognitive Science Society, pp. 376–381.

Degen, Judith, Robert X. D. Hawkins, et al. (591–621 2020). “When redundancy is useful: A Bayesian approach to ‘overinformative’ referring expressions”. In: *Psychological Review* 203.127, p. 4.

Devlin, Jacob et al. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. 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)*. Association for Computational Linguistics, pp. 4171–4186.

Frank, Michael C. (Sept. 2016). “Rational speech act models of pragmatic reasoning in reference games”. In.

– (2023). “Large language models as models of human cognition”.

Frank, Michael C. and Noah D. Goodman (2012). “Predicting Pragmatic Reasoning in Language Games”. In: *Science* 336.6084, p. 998.

Franke, Michael and Judith Degen (2023). “The softmax function: Properties, motivation, and interpretation”. In.

Franke, Michael and Gerhard Jäger (2016). “Probabilistic pragmatics, or why Bayes’ rule is probably important for pragmatics”. In: *Zeitschrift für Sprachwissenschaft* 35.1, pp. 3–44.

Franke, Michael, Xian Ji, et al. (2023). *magpie: Minimal architecture for the generation of portable interactive experiments*. URL: `https://magpie-reference.netlify.app/`.

Gao, Luyu et al. (2023). *PAL: Program-aided Language Models*. arXiv: `2211.10435 [cs.CL]`.

Garcez, Artur d’Avila and Luis C. Lamb (2020). *Neurosymbolic AI: The 3rd Wave*. arXiv: `2012.05876 [cs.AI]`.

Gatt, Albert et al. (2013). “Are we Bayesian referring expression generators?” In: *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*. Ed. by Markus Knauff et al.

Gelman, Andrew et al. (2014). *Bayesian Data Analysis*. 3rd edition. Boca Raton: Chapman and Hall.

Goodman, Noah D. and Michael C. Frank (2016). “Pragmatic Language Interpretation as Probabilistic Inference”. In: *Trends in Cognitive Sciences* 20.11, pp. 818–829.

Hagendorff, Thilo (2023). *Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods*. arXiv: `2303.13988 [cs.CL]`.

Holtzman, Ari et al. (2021). “Surface Form Competition: Why the Highest Probability Answer Isn’t Always Right”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 7038–7051.

Hu, Jennifer et al. (2020). “A Systematic Assessment of Syntactic Generalization in Neural Language Models”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 1725–1744.

Jaeger, T. Florian (2008). “Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models”. In: *Journal of Memory and Language* 59, pp. 434–446.

Kruschke, John E. (2015). *Doing Bayesian Data Analysis*. 2nd edition. Burlington, MA: Academic Press.

Lambert, Ben (2018). *A Student’s Guide to Bayesian Statistics*. Sage Publications.

Lee, Michael D. and Eric-Jan Wagenmakers (2015). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge, MA: Cambridge University Press.

Lindborg, Alma and MIlena Rabovsky (2021). “Meaning in brains and machines: Internal activation update in large-scale language model partially reflects the N400 brain potential.” In: *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*. Vol. 43.

Liu, Jiacheng et al. (2022). "Generated Knowledge Prompting for Commonsense Reasoning". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Vol. Volume 1: Long Papers, pp. 3154–3169.

Marvin, Rebecca and Tal Linzen (2018). *Targeted Syntactic Evaluation of Language Models*. arXiv: `1808.09031 [cs.CL]`.

McElreath, Richard (2016). *Statistical Rethinking*. Boca Raton: Chapman and Hall.

OpenAI (2023). *GPT-4 Technical Report*. arXiv: `2303.08774 [cs.CL]`.

Paranjape, Bhargavi et al. (2023). *ART: Automatic multi-step reasoning and tool-use for large language models*. arXiv: `2303.09014 [cs.CL]`.

Park, Joon Sung et al. (2023). *Generative Agents: Interactive Simulacra of Human Behavior*. arXiv: `2304.03442 [cs.HC]`.

Qing, Ciyang and Michael Franke (2015). "Variations on a Bayesian Theme: Comparing Bayesian Models of Referential Reasoning". In: *Bayesian Natural Language Semantics and Pragmatics*. Ed. by Henk Zeevat and Hans-Christian Schmitz. Language, Cognition and Mind. Berlin: Springer, pp. 201–220.

Scontras, Gregory, Michael Henry Tessler, and Michael Franke (2021). *A practical introduction to the Rational Speech Act modeling framework*.

Shiffrin, Richard and Melanie Mitchell (2023). "Probing the psychology of AI models". In: *Proceedings of the National Academy of Sciences* 120.10, e2300963120. eprint: `https://www.pnas.org/doi/pdf/10.1073/pnas.2300963120`.

Sikos, Les et al. (2021). "Reevaluating pragmatic reasoning in language games". In: *PLOS ONE* 16.3, pp. 1–33.

Stan Development Team (2023). "The Stan Core Library". Version 2.32.0.

Stevens, Jon and Anton Benz (2018). "Game-Theoretic Approaches to Pragmatics". In: *Annual Review of Linguistics* 4, pp. 173–191.

Touvron, Hugo et al. (2023). *LLaMA: Open and Efficient Foundation Language Models*. arXiv: `2302.13971 [cs.CL]`.

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30.

Wilcox, Ethan, Pranali Vani, and Roger Levy (2021). "A Targeted Assessment of Incremental Processing in Neural Language Models and Humans". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 939–952.

Wong, Lionel et al. (2023). *From Word Models to World Models: Translating from Natural Language to the Probabilistic Language of Thought*. arXiv: `2306.12672 [cs.CL]`.

Zhang, Jenny et al. (2023). *OMNI: Open-endedness via Models of human Notions of Interestingness*. arXiv: `2306.01711 [cs.AI]`.