LANGUAGE MODELS AS PROBABILISTIC COGNITIVE MODELS

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

State of the art large language models (LLMs) have shown impressive performance on a variety of benchmark tasks. But targeted assessment with methods inspired by experimental psychology highlights dissimilarities between human behavior and LLM predictions. Going even a step further, we here investigate the human-likeness of LLMs' predictions for multiple-choice decision tasks from the perspective of probabilistic cognitive modeling. We identify core differences, in terms of conceptual interpretation and usual practical evaluation, between LLMs and probabilistic cognitive models (PCMs). Concretely, we argue that LLMs are, first and foremost, predictors of item-level data, whereas PCMs primarily make predictions at a more abstract level (the condition-level). Using human data from a multiple-choice experiment on pragmatic language use, we find that LLMs fail to capture the variance in the human data at the item-level. We suggest different ways of deriving full distributional predictions from LLMs for condition-level data, thus building PCM-like predictions from LLM-derived information, and find that some, but not all ways of obtaining condition-level predictions yield adequate fits to our data. These results suggests that assessment of LLM performance depends strongly on seemingly subtle choices in methodology, and that LLMs are at best predictors of human behavior at an aggregate level, for which they are, however, not designed to make predictions in the first place.

1 Introduction

The invention of deep neural transformer architectures (Vaswani et al., 2017) has enabled a new generation of powerful large language models (Devlin et al., 2019; Chung et al., 2022; OpenAI, 2023; Touvron et al., 2023). State-of-the-art LLMs excel on many benchmark data sets (e.g., BIG-bench authors, 2023; Perez et al., 2023), and so 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 reliably 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 for engineering purposes, e.g., to systematically study of the effects of scale (e.g., BIG-bench authors, 2023), but is also used to ask to what extent LLM performance is "human-like." In more psychologically oriented work, LLM performance is therefore compared to human choice behavior in psychological experiments and investigates whether LLMs predict patterns of human answer behavior (e.g., Binz and Schulz, 2023; Hagendorff, 2023; Shiffrin and Mitchell, 2023). The main focus is often to compare qualitative patterns in LLM predictions and human data, but there is also work investigating whether LLMs can make adequate *quantitative predictions*. For example, work at the interface between NLP and computational psycholinguistics (Marvin

---

[*]Corresponding author.

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 the N400 component of event-related-potentials in EEG measurements (Lindborg and Rabovsky, 2021). The work presented in this paper seeks to extend the investigation of the human-likeness of LLMs by looking at empirical data from human multiple-choice behavior and comparing an established, theoretically motivated probabilistic cognitive model, which is commonly use to predict or explain this kind of data, to structurally similar quantitative predictions derived from LLMs. In short, while a lot of previous work has used targeted assessment of LLM capabilities from the point of view of the experimental psychologist, we here adopt the more specific perspective of the cognitive modeller.

For the purposes of this investigation, a probabilistic cognitive model (PCM) is characterized by providing a parameterized likelihood function $P_M(D_C \mid \theta, C)$, which assigns a probability to each possible data observation D_C for any experimental condition C and parameter values θ (e.g., Lewandowsky and Farrell, 2011; Lee and Wagenmakers, 2015). There are non-trivial, insightful differences between LLMs and PCMs, both at the conceptual level and in terms of common practices in evaluation, which we will explore in this paper in order to contribute to a better delineation of the potential and limitations of LLMs. For one, the "primitive predictions" that LLMs provide are not at the level of a condition C, but at a lower item-based level. For another, LLMs are often just assessed based on their ability to identify a target answer, not to predict the full distribution observed in the data.

This paper introduces and systematically compare different ways of carving numerical information obtained from an LLM into a structure similar to a parameterized PCM and address the empirical question of whether such an "LLM-based PCM" adequately captures the distribution of human forced-choice data at different levels of aggregation. To obtain data for this investigation, we subjected human participants and an LLM (a version of OpenAI's GPT-3.5; see Section 5) to the same forced-choice task, a simple reference game (Frank and Goodman, 2012). We introduce different ways of defining condition-level, distributional predictions based on LLM-provided numerical measures. We use standard tools of Bayesian data analysis (Gelman, Carlin, et al., 2014) to train, test and compare these "LLM-based PCMs" against each other, and against an established, theoretically informed PCM, namely the Rational Speech Act (RSA) model (Frank and Goodman, 2012). This is quite different, and arguably more stringent and insightful, than the standard practice of evaluating LLMs based on accuracy scores, determined by a winner-takes-all strategy to identify a single target response. We find that variability predicted by the LLM at the item-level is not borne out by the human data, and that some but not all ways of constructing condition-level predictions are equally good. These results suggest that it is not impossible to build predictive models for condition-level data, similar to PCMs, on top of LLM-supplied information, but that seemingly subtle details in the derivation of likelihoods matter. The conceptual comparison of LLMs and PCMs also clearly identifies a sense in which the latter are explanatory, while the former are not. [MF: double-check formulation of results]

The paper is structured as follows. Section 2 motivates the investigation in this paper, by enlarging on the importance of deriving distributional predictions for conditional-level forced-choice data, and further elaborating on the relevant differences between LLMs and PCMs. 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 commonly used, theoretically-motivate PCM, namely the Rational Speech Act (RSA) 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 the theoretically-motivated 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.

2 Motivation: Distributional predictions for condition-level forced-choice data

Psychological research into the workings of the human mind aims to find generalizable patterns in the way information is processed within or across different domains of cognition. Experimental work therefore often compares human performance in different *experimental conditions*, which reflect the general factors that are hypothesized to influence behavior. Yet a single experimental condition is often instantiated with different *experimental items*, which are not under scrutiny for any systematic, predictable effect on the observed measurements. For example, classical research on human memory (Atkinson and Shiffrin, 1968) investigated the effect of rehearsal on memory consolidation. Relevant experiments compare recall with and without rehearsal (experimental conditions), while using different items (words, numbers, etc., to be memorized) in each instantiation of the same memory task. Likewise, when studying how hearing a color word can facilitate a same-or-different judgement of color swatches (Rosch, 1975), the main experimental manipulation concerns the typicality of shown color swatches, while the variability between different color words like "blue" or "green", is less important to this research question and so treated as *item-level variation*. Consequently, a typical psychological experiment is mainly interested in assessing behavior at the level of the experimental condition, because that is where the distinctions relevant to the research question reside. Nevertheless, each experimental condition

can be, and often is, instantiated with different items, variation among which is deemed less relevant to the research question at hand.

Data from human participants for experiments of this kind usually shows some variability between items, and also variability between participants. This variability is commonly incorporated in statistical analyses as random stochastic variation, e.g., by using hierarchical (regression) models (Jaeger, 2008; Barr et al., 2013; Sorensen, Hohensteinb, and Vasishth, 2016). Still, the focus of interest usually remains at the condition-level effects, because it is this more abstract level of behavioral aggregation that is relevant for generalizable theory building. Similarly, when analyzing or explaining data from psychological experiments with a probabilistic cognitive model (PCM), the PCM will naturally be set-up to predict data at the level of the relevant experimental manipulation, i.e., at the condition-level. Even though item-level (and individual-level) variation can, and sometimes should, be incorporated into probabilistic cognitive modeling (e.g., Nilson, Rieskamp, and Wagenmakers, 2011; Lee, 2011; Scheibehenne, Rieskamp, and Wagenmakers, 2013), the focus of cognitive modeling is to explain systematic variation between experimental conditions. Therefore, a basic PCM is essentially characterized by a parameterized likelihood function $P_M(D_C \mid \theta, C)$, which assigns a probability to each logically possible data observation D_C for any experimental condition C and numerical parameters θ (Lewandowsky and Farrell, 2011; Lee and Wagenmakers, 2015). The difference between a theory-driven PCM and a standard, theory-neutral statistical model (like a regression model), which may also define a parameterized likelihood function of condition-level data, is that the PCM's likelihood function is defined in terms that are considered to be conceptually meaningful for the research domain at hand (see Section 4 for a concrete example).

PCMs, thus construed, differ markedly from LLMs. LLMs make predictions about the likelihood of words given a preceding or surrounding text. They do not, without further derivation, provide predictions at a conceptually meaningful level of abstraction, similar to experimental conditions. In this way, LLMs are not explanatory models in the same sense as PCMs might be: they simply do not operate at a level of generalizable abstraction which is arguably necessary for increasing human understanding of the phenomenon in question (Dellsén, 2020; Grimm, 2021). Nevertheless, we can derive predictions from LLM's word-level predictions, first at the item-level and also at the condition-level. Whether or not these predictions are deemed explanatory, the empirical question arises of whether any of these predictions are empirically adequate. This empirical question can be addressed by subjecting "LLM-based PCMs" to the same statistical model checking as the more usual theoretically motivated PCMs.

Standard benchmark-testing on multiple-choice tasks usually uses a winner-take-all (WTA) strategy (e.g., BIG-bench authors, 2023) to define accuracy scores. The WTA-strategy can also be used to derive probabilistic item-level and condition-level predictions from LLMs. Let $\{I_1, \ldots, I_m\}$ be m be instances of the same task, or items belonging to the same (logical) condition in a behavioral experiment. Each item $I_k = \langle x_k, \langle y_{k1}, \ldots, y_{kl} \rangle \rangle$ consists of an input prompt x_k, which is a string of text, and l choice options $\langle y_{k1}, \ldots, y_{kl} \rangle$, all of which are strings as well, possibly composed of $|y_{ki}|$ words, $y_{ki} = w_{ki1}, \ldots, w_{ki|y_{ki}|}$, and each of which . For simplicity of notation, we assume that the i-th choice option y_{ki} for each item k belongs to the same answer category R_i, e.g., that y_{k1} is always the designated *target option* that is assumed to be the "true" or to-be-selected goal answer, and that the other choice options are equally assignable into meaningful categories of answer types.

The most obvious *item-level score* an (autoregressive) LLM provides for each choice option y_{ki} is its log-probability:[2]

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

The WTA-based approach selects answers that maximize the item-level score (with random tie-breaking). Concretely, if $B_k = \arg\max_i S_{ki}$ is the set of all options that maximize the item-level score for item k, then the *item-level prediction* for item I_k of the WTA is:

$$P_{\text{item}}^{\text{WTA}} = \begin{cases} \frac{1}{|B_k|} & \text{if } y_{ki} \in B_k, \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Notice that the item-level WTA-prediction is usually a degenerate probability distribution placing all probability mass on a single choice option.

---

[2][MF: How to extend to masked language models?] 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.

The *condition-level prediction* of the WTA approach is obtained by averaging over the item-level predictions for each answer category:

$$P_{\text{cond}}^{\text{WTA}}(R_i) = \frac{1}{m} \sum_{k=1}^{m} P_{\text{item}}^{\text{WTA}}(y_{ki}) \ .$$

The *WTA-based accuracy* is the probability $P_{\text{cond}}^{\text{WTA}}(R_1)$ of selecting the target option over all instances of the task. This is the most prevalent measure of the quality of LLM predictions.

The WTA approach implicitly assumes that item-level choices are resolved by a greedy-like choice of the best alternatives. This procedure ignores information about relative differences between scores among choice options; information which may be conceptually and practically relevant. To generalize to less extreme sampling scenarios, the WTA approach can be generalized to softmax sampling, where the *item-level prediction* of choosing option y_{ki} is:

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

The *condition-level prediction* can be defined analogously to the above by averaging over items:

$$P_{\text{cond}}^{\text{SM}}(R_i) = \frac{1}{m} \sum_{k=1}^{m} P_{\text{item}}^{\text{SM}}(y_{ki}) \ .$$

As before, the *softmax-based accuracy*, $P_{\text{cond}}^{\text{SM}}(R_i)$, is the average probability of choosing the designated target option y_{k1}. For $\alpha \to \infty$, the softmax strategy converges to the WTA strategy, but for finite α the approaches differ. This has interesting entailments for accuracy assessment in more engineering-oriented contexts, as well as for questions of exploring the "human-likeness" of LLM predictions.

Looking at accuracy assessment, WTA-based accuracy measures can even differ qualitatively from accuracy measures obtained from the more general softmax-based strategy. Differences in accuracy measures depend on the variation in item-level scores, in particular the relation between score-ordering and score-differences. The example below shows that categorically different results are, in principle, possible when using WTA- or softmax based accuracy. The example 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.

> **Example:** Imagine that there are two options, and that the target option's score is a small ϵ higher in 80% of the task's items, and otherwise lower. 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 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]

The upshot of these considerations is that the standard practice of WTA-based performance assessment for LLMs gives false, or at least misleading or inaccurate results, whenever not all downstream applications use a greedy-like sampling strategy (which is almost certainly the case), and there is variability in item-level predictions.

When aspiring to predict experimental data from human participants with LLMs, the WTA-based strategy makes very strong predictions at the item-level, where it usually predicts a degenerate probability distribution according to which every data observation for a particular item should fall into the exact same response category. This may seem like an *a priori* unlikely prediction, but needs to be tested empirically, just as the item-level predictions by the softmax-based approach should (see Section 5). Whether item-level predictions by an LLM are empirically vindicated or not, is logically independent of whether these item-level predictions serve well as the backbone for deriving condition-level predictions. We therefore also need to test, against empirical data, whether different ways of deriving condition-level predictions are adequate (see Section 6).
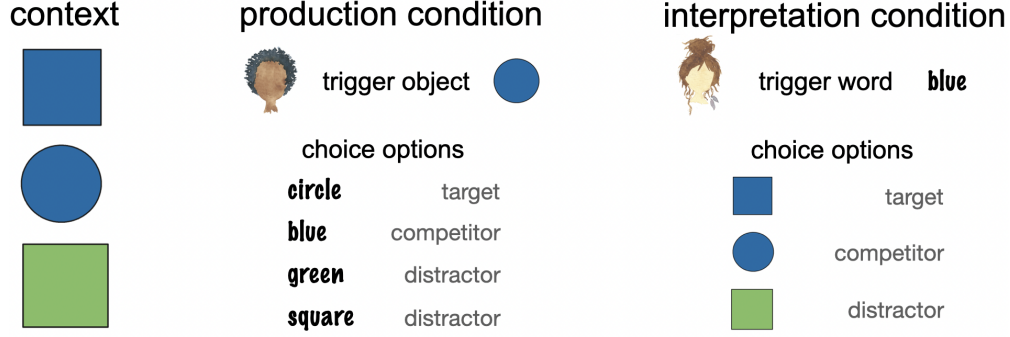
Figure 1: Structure of a reference game with human participants. Each trial consists of a set of objects, the so-called context. In production trials, participants choose a single word to describe a trigger object from the context. In interpretation trials, an object is selected as the likely object a trigger word referring to.

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 *trigger 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 trigger object; and, for the interpreter, to guess correctly which object the speaker had in mind when using the trigger word.

The example in Figure 1 is a standard case, which we will use throughout, where human choices are informative about the pragmatic reasoning that human 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 trigger 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 trigger object. The **distractor utterance** is the word that is true of the trigger object, but also true of another object. The other utterances, both of which are false of the trigger object 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 human data, when both LLMs and humans processed the same textual representation of the 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]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: $4/5 \; 1/2 + 1/5 \; 0 = 2/5$.

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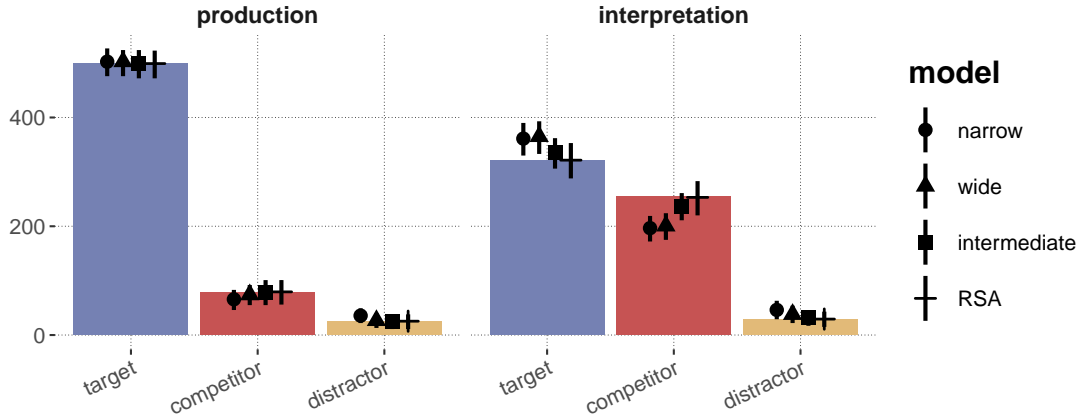
Figure 2: Counts of choices from reference games with human participants (colored bars), with summary statistics from the posterior predictive distribution of four models (shapes and error bars). Shapes show the mean of the posterior predictive distributions of the RSA model and three further models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

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. As in the original paper by Frank and Goodman (2012), 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 of target, competitor and distractor options: $\langle 0.882, 0.118, 0 \rangle$ 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 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

---

[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.

Semantic meaning

$\mathfrak{L}$	■ (blue sq)	● (blue circ)	■ (green sq)
"square"	1	0	1
"circle"	0	1	0
"green"	0	0	1
"blue"	1	1	0

Pragmatic speaker ($\alpha = 1$)

	"square"	"circle"	"green"	"blue"
■ (blue sq)	.5	0	0	.5
● (blue circ)	0	0.66	0	0.33
■ (green sq)	0.33	0	0.66	0

Literal listener

	■ (blue sq)	● (blue circ)	■ (green sq)
"square"	.5	0	.5
"circle"	0	1	0
"green"	0	0	1
"blue"	.5	.5	0

Pragmatic listener ($\alpha = 1$)

	■ (blue sq)	● (blue circ)	■ (green sq)
"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) for $\alpha = 1$ are shown as row-stochastic matrices.

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 policy is defined as soft-optimizing the choice of utterance to minimize the literal listener's surprisal for the state to be communicated, i.e., the trigger object:

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

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 a flat prior $P(s)$ 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 $\langle 3/5, 2/5, 0 \rangle$ 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 small error probability ϵ, with which a choice would be made at random (e.g., Lee and Wagenmakers, 2015): 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:

$$P_r(R_l, C; \alpha_c, \epsilon_c) = (1 - \epsilon)\, P_r(R_l, C; \alpha_c) + \epsilon_c/3,$$

where ϵ_c is a (condition-specific) parameter giving the probability that a choice was made by randomly guessing. [MF: FC objected to "condition-specificity" which is fair, but I leave it as it was; my reasoning: we use model checking to rule out models; in this setting the conservative strategy is to err on the side of making the models too powerful (even at the cost of conceptual plausibility)] The data D_C from condition C, see Figure 2, consists of counts for each response category. The parameterized likelihood function entailed by the RSA model for D_C is:

$$P_r(D_C \mid C, \alpha_C, \epsilon_C) = \text{Multinomial}(D_C, \langle P_r(R_l, C; \alpha_c, \epsilon_c) \rangle_{1 \leq l \leq 3}),.$$

The result is a four-parameter model, one pair of parameters per condition.

Parameterized predictions like $P_r(D_C \mid C, \alpha_C, \epsilon_C)$ can be assessed in the light of the empirical data with the usual tools of Bayesian data analysis (e.g. Gelman, Carlin, et al., 2014; McElreath, 2016; Lambert, 2018). Let $\alpha_c \sim \log\text{-Normal}(0.5, 1)$ have a reasonably wide log-Normal prior, and let $\epsilon_c \sim \text{Beta}(1, 5)$ 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).[6]

To assess goodness-of-fit, we use the *posterior predictive distribution*, i.e., the model's predictions about data of the same size and structure as the training data. Figure 2 shows summary statistics (means and 95% credible intervals) for the posterior predictive distribution of the RSA model (among other models, which are to be introduced later). We see that for both conditions the RSA model passes a "visual posterior predictive check" (Kruschke, 2015), namely that the distribution of posterior predictions includes the observed choice rates for each answer option. 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. 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. A very small value indicates that the model might be inadequate for reproducing the data it was trained on. This is clearly not the case for the RSA model with posterior predictive p-values close to 0.5. In sum, the condition-level predictions by the RSA model, a theoretically motivated PCM, are not discredited by the condition-level data.

We can also use the condition-level predictions for predicting the item-level data. If the whole data set for condition C is chunked into item-level data, $D_C = \left\langle D_C^1, \ldots, D_C^m \right\rangle$, the likelihood function for analysis at the item-level uses the same (condition-level) predictor to separately predict data from each item:

$$P_r^{\text{item}}(D_C \mid C, \alpha_C, \epsilon_C) = \prod_{k=1}^{m} \text{Multinomial}(D_C^k, \langle P_r(R_l, C; \alpha_c, \epsilon_c)\rangle_{1 \le l \le 3}), .$$

The following sections apply these tests 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 human choices for all of the 100 different items combined. An LLM, on the other hand, first and foremost makes predictions about each item. As mentioned previously, 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 irrelevant 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: @PT 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 option after processing the task description. Let C be the experimental condition of interest (here: the production or interpretation condition). Let $\{I_1, \ldots I_m\}$ be the multi-set of items that occurred in the human experiment for condition C.[7] Each item I_k consists of a task description T_k and l choice options $O_{k1}, \ldots, O_{kl}$, 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).[8] The *item-level score* $S(O_{ki}, I_k)$ for

---

[6]Posterior samples where generated for four chains with 2000 samples each, after a warm-up of 1000 samples. The "adapt-delta" value was set to 0.99. Converge was checked with $\hat{R}$-statistics (Gelman and Rubin, 1992).

[7]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.

[8]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.
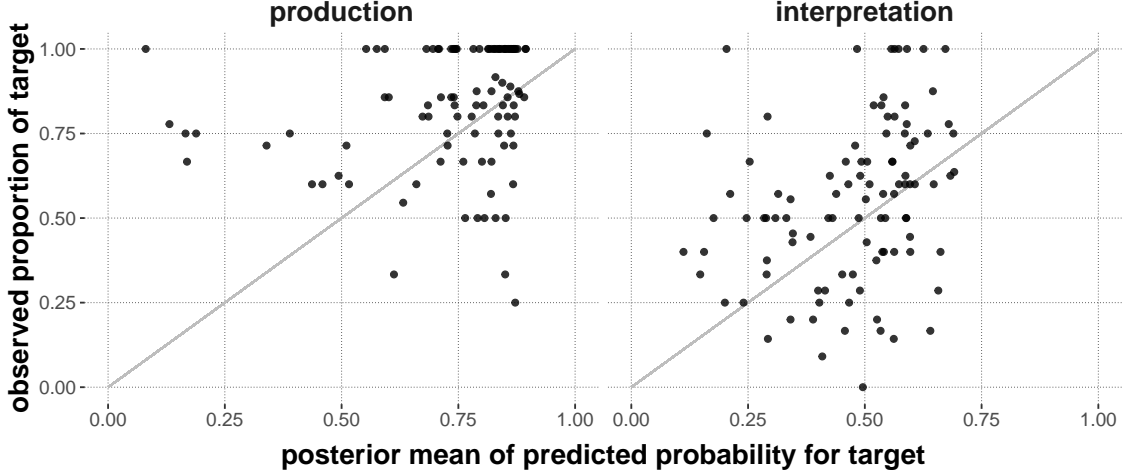
Figure 4: Item-level prediction-observation plot. Each dot represents an item. The *x*-coordinate represents the mean of the posterior prediction for the target choice probability for the given item. The *y*-coordinate represents the observed proportion of target choices for that item. The gray line (identity function) is the ideal prediction.

option $O_{ki} = w_{ki1}, \ldots w_{kin}$ of item I_k is defined as the log-probability of continuation O_{ki} after prompt T_k:[9]

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

[MF: check if formula is redundant; already introduced?] 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 [MF: which model exactly, when assessed, cite paper]. 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:

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

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.

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.

---

[9]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.
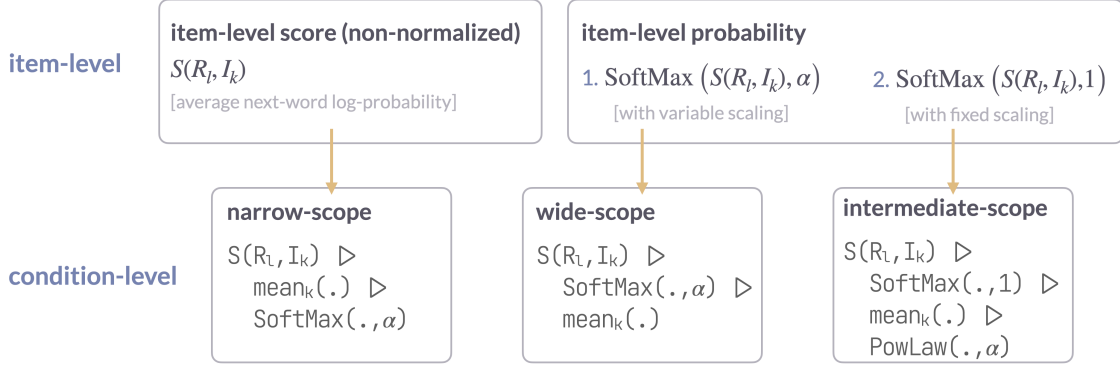
Figure 5: Schematic overview over three different approaches of obtaining condition-level predictions by aggregating over item-level scores. The basic item-level score is average next-word log probability. This can be taken as-is into averaging (narrow-scope aggregation), or first transformed from log- 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).

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 predictors. 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 $\mathbf{s} = \langle s_1, \ldots, s_l \rangle$ to probabilities $\mathbf{p} = \langle p_1, \ldots, p_l \rangle$ 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):

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

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.

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

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):

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

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.

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. Numbers indicate estimates for lower and upper bounds of 95% credible intervals and the posterior means. [MF: make wide format]

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 $\alpha \rightarrow \infty$. 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.

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

	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.

for optimization and ϵ for random error), the human choice probability can be reconstructed faithfully in its entirety.[10] 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

---

[10]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.
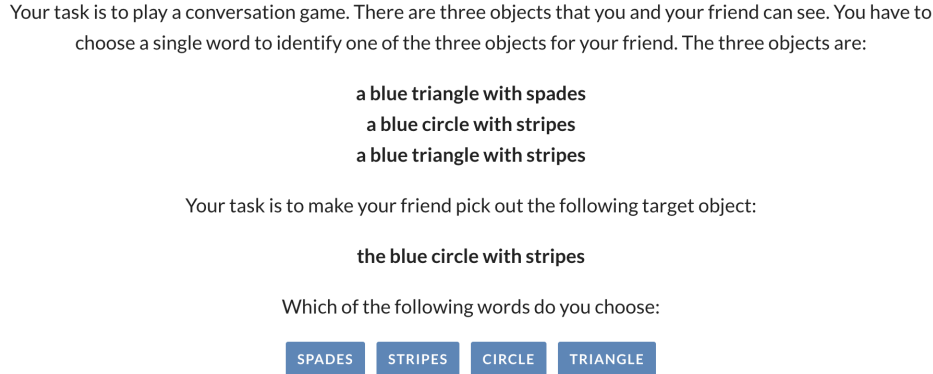
Your task is to play a conversation game. There are three objects that you and your friend can see. You have to choose a single word to identify one of the three objects for your friend. The three objects are:

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:

SPADES   **STRIPES**   **CIRCLE**   **TRIANGLE**

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?

A GREEN CIRCLE WITH DOTS   **A GREEN CIRCLE WITH SPADES**   **A GREEN HEXAGON WITH DOTS**
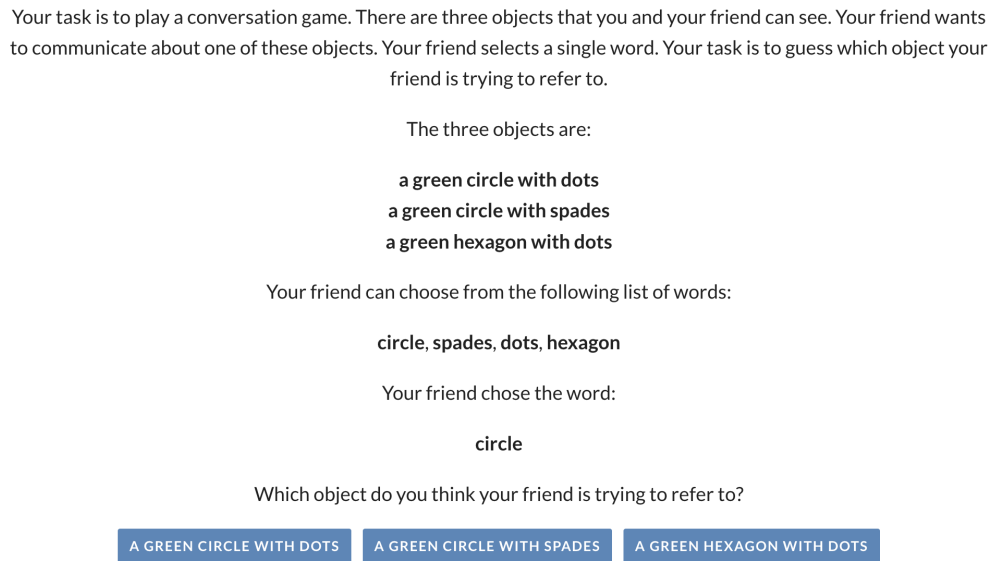
Figure 7: Screen shot from an interpretation trial of the online experiment.

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:
```
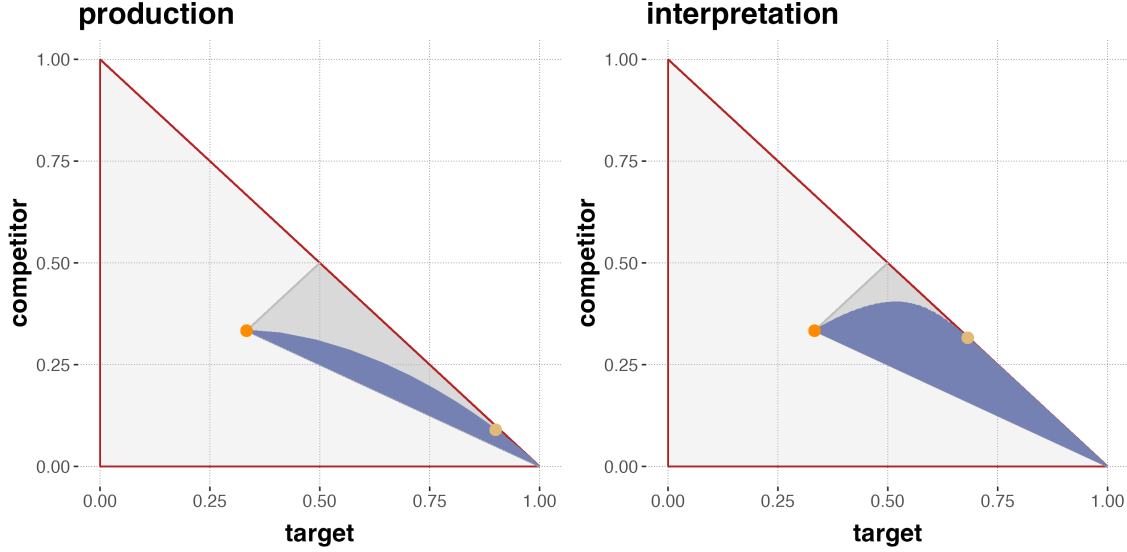
Figure 8: Range of predictions the intermediate-scope model makes for any pair of parameter values α_c and ϵ_c. The light gray triangle with the red boundary is the probability simplex, i.e., the space of all possible 3-place probability vectors. The darker gray area in gray boundary lines is the subspace 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 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

Atkinson, Richard and Richard Shiffrin (1968). "Human memory: A proposed system and its control processes". In: *The Psychology of Learning and Motivation*. New York: Academic Press.

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.

Dellsén, Finnur (2020). "Beyond Explanation: Understanding as Dependency Modelling". In: *The British Journal for the Philosophy of Science* 71.4, pp. 1261–1286.

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

Gelman, Andrew and Donald B. Rubin (1992). "Inference from Iterative Simulation Using Multiple Sequences (with discussion)". In: *Statistical Science* 7, pp. 457–472.

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

Grimm, Stephen (2021). "Understanding". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2021 Edition. Metaphysics Research Lab, Stanford University.

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. (2011). "How Cognitive Modeling Can Benefit From Hierarchical Bayesian Models". In: *Journal of Mathematical Psychology* 55, pp. 1–7.

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

Lewandowsky, Stephan and Simon Farrell (2011). *Computational Modelling in Cognition: Principles and Practice*. Sage Publications.

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.

Nilson, Hakan, Jörg Rieskamp, and Eric-Jan Wagenmakers (2011). "Hierarchical Bayesian Parameter Estimation for Cumulative Prospect Theory". In: *Journal of Mathematical Psychology* 55, pp. 84–93.

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]`.

Perez, Ethan et al. (July 2023). "Discovering Language Model Behaviors with Model-Written Evaluations". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 13387–13434.

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.

Rosch, Eleanor (1975). "The Nature of Mental Codes for Color Categories". In: *Journal of Experimental Psychology: Human Perception and Performance* 1.4, pp. 303–322.

Scheibehenne, Benjamin, Jörg Rieskamp, and Eric-Jan Wagenmakers (2013). "Testing the Adaptive Toolbox Models: A Bayesian Hierarchical Approach". In: *Philosophical Review* 120.1, pp. 39–64.

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.

Sorensen, Tanner, Sven Hohensteinb, and Shravan Vasishth (2016). "Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists". In: *The Quantitative Methods for Psychology*.

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]`.

Snippets

- for conclusion?
    - 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).

- PT: the high-level take-away that item-level LLM predictions don't fit human predictions well might, as always, be met with the comment that the task is rather artificial. maybe this should go into the discussion

- FC: The fact that the aggregate predictions track aggregate human behaviour means that the discrepancies are washing out. But there's also the interesting question of finding the systematic differences between LLMs and people. The task then would be to find sets of stimuli where the differences do *not* wash out and ask: What if anything do these stimuli share? This paper could also be a jumping board for this line of research.

- PT: maybe a conclusion in the spirit of "cogmod in the age of LLMs"/ "computational psycholinguistics & careful assessment of reasoning / linguistic abilities of LLMs against human performance" might be not entirely ludicrous too