Beyond the best guess: Statistical modeling with predictors from large language models

Michael Franke, Polina Tsvilodub, Fausto Carcassi
Department of Linguistics
University of Tübingen
`[michael.franke|polina.tsvilodub|fausto.carcassi]@uni-tuebingen.de`

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

1 Motivation

With the advent of deep neural transformer architectures (Vaswani et al., 2017), large language models (LLMs) have become powerful technologies that excel on standard benchmarks and promise to serve as foundation models for a vast and diverse set of applications (Bommasani et al., 2021) (Devlin et al., 2019; Chung et al., 2022; OpenAI, 2023; Touvron et al., 2023), especially when augmented with supervised fine-tuning (Chung et al., 2022) and reinforcement learning from human (Stiennon et al., 2022) or other (Bai et al., 2022) feedback. Recently an increasing number of diverse applications has gone beyond using LLMs out-of-the-box, based on their single-run input-output behavior, and instead utilize LLMs as a part of a larger computational process.

- tool use
- AutoGPT
- prompt-engineering
- structured reasoning
- for translation to programms (Wong et al., 2023)

For all of these applications, it is crucial to understand what LLMs can or cannot reliable do.

- benchmarks don't give us that
- mechanistic
- behavioral probing

Not yet addressed: derive full distribution over choice options and use for explanatory modeling.

SNIPPETS

Recent work comparing LLMs to human choice behavior in psychological experiments has focused on whether LLMs predict patterns of human answer behavior qualitatively. The bulk of this work looks at the answers generated by
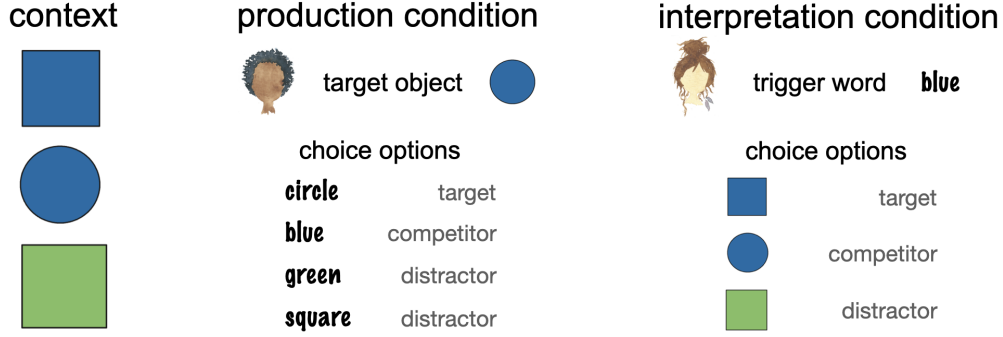
Figure 1: Structure of a reference game with human participants. Each trial consists of a set of objects, the so-called context. In production trials, participants choose a single word to describe a target object from the context. In interpretation trials, an object is selected as the likely object a trigger word referring to.

the LLMs. In general, much less attention has been payed to whether LLMs can also make adequate *quantitative predictions* that match human choice probabilities in suitable experiments. (Notable exceptions at the interface of NLP and computational psycholinguistics (Linzen, Hu, Rabovsky . . .).)

This work looks at whether probabilistic predictions of GTP-3.5 match human answer patterns also quantitatively. Rather than looking at LLMs as behavior-generating black box, which we want to understand based on its input-output behvaior, we explore whether we can use LLMs as a predictive probabilistic model, or at least as a component of one, aiming to predict the full distribution of human experimental data.

2 Experiment: Reference games

To keep matters simple, we use an established, well-understood and austere experimental paradigm to test human decision making in abstract communicative tasks, so-called **reference games** (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.

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).[1]

---

[1]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/.
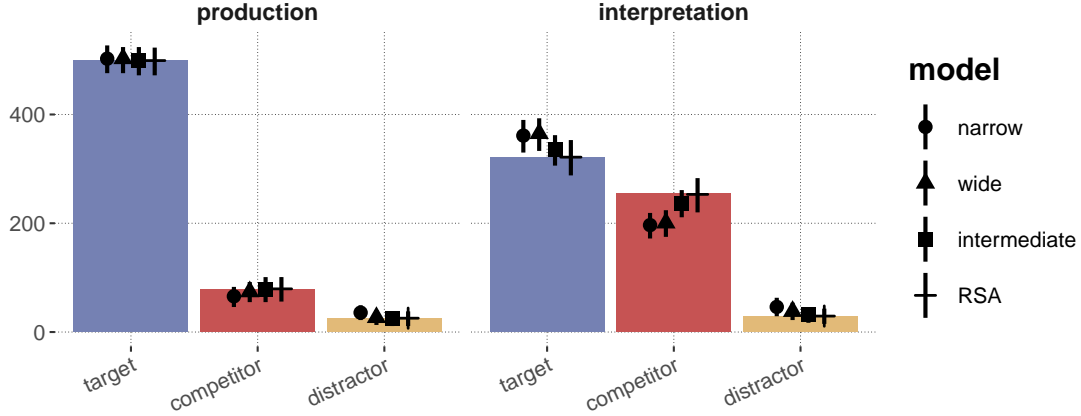
Figure 2: Counts of choices from reference games with human participants (colored bars), with summary statistics from the posterior predictive distribution of three models (shapes and error bars). Shapes show the mean of the posterior predictive distributions of three models, differing in the scope of the item-level aggregation with respect to soft-maximizing and scaling. Error-bars show corresponding 95% credible intervals of the posterior predictive.

2.1 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.

2.2 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.

2.3 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.

2.4 Results

The overall distribution of choices that correspond to the target, competitor, and distractor states is shown in Figure 2 on the left.[2] 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).

3 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; Qing and Franke, 2015), probabilistic

---

[2]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.

Bayesian modeling (e.g., Frank and Goodman, 2012; Franke and Degen, 2016) 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, parameterized stochastic policy, $P_S(u \mid s; \theta_S)$ for a speaker's choice of expression or utterance u given a referent or state s to describe, modulated by parameters θ_S; 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 introduced 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:

$$P_S(u_i \mid s, \alpha) = \text{SoftMax}(\mathbf{u}_s, \alpha)_i, \text{ where}$$
$$\mathbf{u}_s = \langle \log P_{L_0}(s \mid u_1), \ldots, \log P_{L_0}(s \mid u_n) \rangle$$

The softmax function takes a vector of non-normalized scores $\mathbf{s}$ and outputs a probability distribution in which odds are proportional to score differences, modulated by optimality parameter α:

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

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

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

[MF: two-parameters, can be fitted by Bayes, results in Figure ...]

4 LLM predictions for reference games

This section explores different ways of deriving quantitative predictions from an LLM to feed into a statistical model for the human data. 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. The previous literature has shown that scoring longer sequences of text for an autoregressive LLM in terms of *average next-token log probabilities*, or *average log-probs* for short, gives improved predictions in relevant benchmark tests (e.g., Brown et al., 2020). Using average log-probs 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 4). In the following, we first describe the different options of deriving LLM-based predictor terms. We then introduce a theoretical cognitive model for comparison. Finally, we compare all probabilistic models based on their adequacy of explaining the human choice data.

4.1 LLM-based probabilistic predictions for forced-choice data

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.[3] Each item I_k consists of a task description T_k

---

[3]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.

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.

item-level

item-level score (non-normalized)	item-level probability
$S(R_l, I_k)$ [average next-word log-probability]	1. SoftMax $\big(S(R_l, I_k), \alpha\big)$ 2. SoftMax $\big(S(R_l, I_k), 1\big)$ [with variable scaling] [with fixed scaling]

condition-level

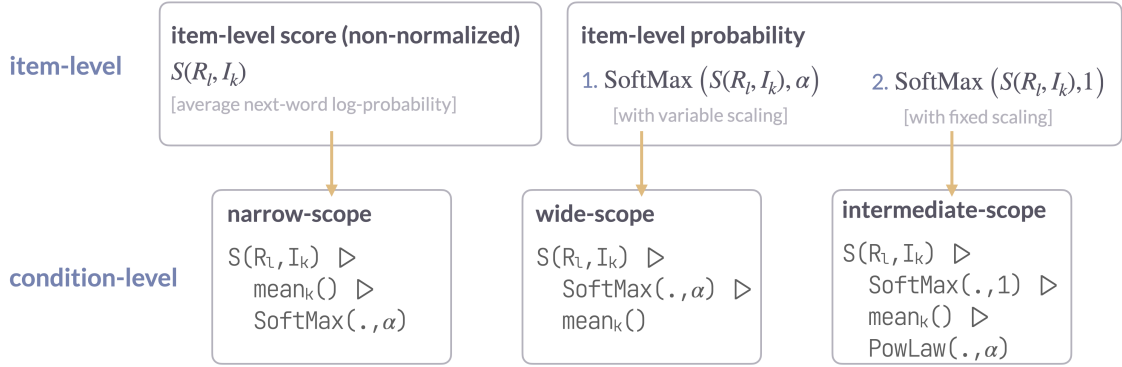| narrow-scope	wide-scope	intermediate-scope

`S(Rₗ,Iₖ)` ▷
   `meanₖ()` ▷
   `SoftMax(.,α)`
	`S(Rₗ,Iₖ)` ▷
	   `SoftMax(.,α)` ▷
	   `meanₖ()`
		`S(Rₗ,Iₖ)` ▷
		   `SoftMax(.,1)` ▷
		   `meanₖ()` ▷
		   `PowLaw(.,α)`

Figure 4: 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). [MF: include dots in mean function, too]

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).[4] We will define an item-level (non-normalized) score $S(O_{ki})$ for each choice option of a given item k, and use this to define a condition-level probabilistic predictions for response types by averaging over all items in a given condition.

Item-level score. If a choice option $O_{ki} = w$ is a single word, like in the production condition, an LLM's item-level prediction can simply be defined as the next-word log probability $\log P_{\text{LLM}}(w \mid T_k)$. If choice options are strings of more than one word, as in the interpretation condition, we follow the previous literature (e.g., Brown et al., 2020) and define the *item-level score* $S(O_{ki}, I_k)$ for option $O_{ki} = w_{ki1}, \ldots w_{kin}$ of item k, in terms of the average log-probs:

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

---

[4]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.

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

Condition-level predictions. 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, defined as:

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

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:

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

This means that there are three conceivable scope-sites for aggregating item-level information as shown in Figure 4. 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. 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, not *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. Similarly, intermediate-scope aggregation is also compatible with a sampling based picture, but would use pure decoding (without α-optimization) at the item level.

4.2 Model fitting, criticims & comparison

Parameterized predictions $P_m(R_l, C; \alpha_c)$ for models $m \in \{n, w, i, p\}$ can be assessed in the light of the empirical data from human participants with the usual tools of Bayesian data analysis (e.g. Gelman et al., 2014; McElreath, 2016; Lambert, 2018). For each condition C (production and interpretation), each model has a free optimality parameter $\alpha_c \sim \log\text{-Normal}(0.5, 1)$ with a log-Normal prior. We also assume that each model has one epsilon parameter per condition, ϵ_c, which accounts for random errors in the human data via Laplace smoothing, with a prior $\epsilon_c \sim \text{Beta}(1, 5)$ favoring small values. The resulting likelihood function for model m and data from condition C is a multinomial distribution with central predictor for response R_l given by:

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

Using Stan (Stan Development Team, 2023) for Bayesian inference, we obtain estimates of posterior credible values of each α_c and ϵ_c for each model. The usual Bayesian summary statistics for posteriors over model parameters are shown in Figure 1.

To assess goodness-of-fit, Figure 2 shows summary statistics (means and 95% credible intervals) for each model's posterior predictive distribution, i.e., the model's predictions about data of the same size and structure as the training data. 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 that only the intermediate-scope model passes this "visual posterior predictive check" for both conditions; the

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.

other two models both overpredict the target choice rate and underpredict the competitor choice rate. 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. 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. Consequently, the results from Table 2 suggest that the narrow-scope model fails on the interpretation data, and is at most borderline on 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.

model	production	interpretation
narrow	0.05	0.00
wide	0.51	0.00
intermediate	0.49	0.26
RSA	0.49	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.
[5] 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.

---

[5]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.
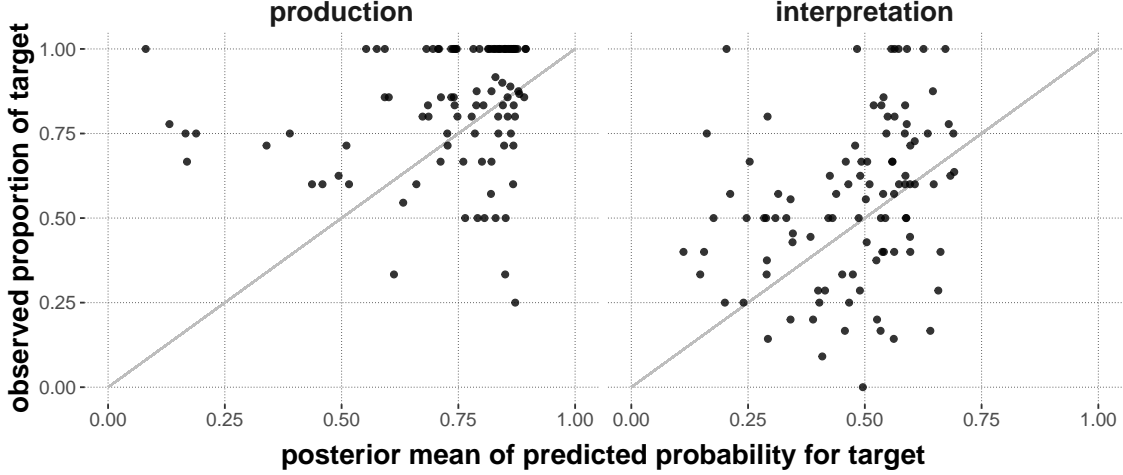
Figure 5: Item-level prediction-observation plot. Each dot represents an item. The *x*-coordinate represents the mean of the posterior prediction for the target choice probability for the given item. The *y*-coordinate represents the observed proportion of target choices for that item. The gray line (identity function) is the ideal prediction.

5 Item-level predictions

The ability to predict a condition-level averages is useful for applications, and may also be theoretically insightful. Nevertheless, for human data, averages over item- or individual-level variation can also be misleading (e.g., Stanovich and West, 2000; Estes and Maddox, 2005; Haaf and Rouder, 2019), as aggregate patterns may be artifacts of systematic variation at the subject- or item-level. LLMs do not make individual-level predictions, at least not without further intervention, such as special prompting. But the question remains whether the variation in item-level predictions made by LLMs is reflected in the human data.

To address this question, we fit a model to the human data at the item-level. The data to be explained is the set of all response type counts $\langle t_k, c_k, d_k \rangle$ for each item I_k. The model's predictions for the data from item I_k from condition C (production or interpretation) are:

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

In words, we use a single pair or 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.

Figure 5 shows that the item-level LLM predictor predicts 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. There is ample variation in the model 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. In fact, 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 each condition). These results suggest that LLM-based probabilistic predictions may imply item-level variance that is not attested in the human data.

6 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

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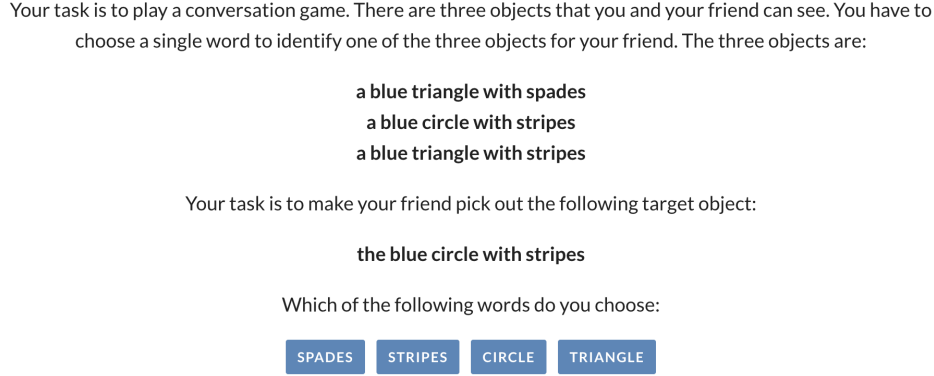
Your task is to play a conversation game. There are three objects that you and your friend can see. You have to choose a single word to identify one of the three objects for your friend. The three objects are:

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?

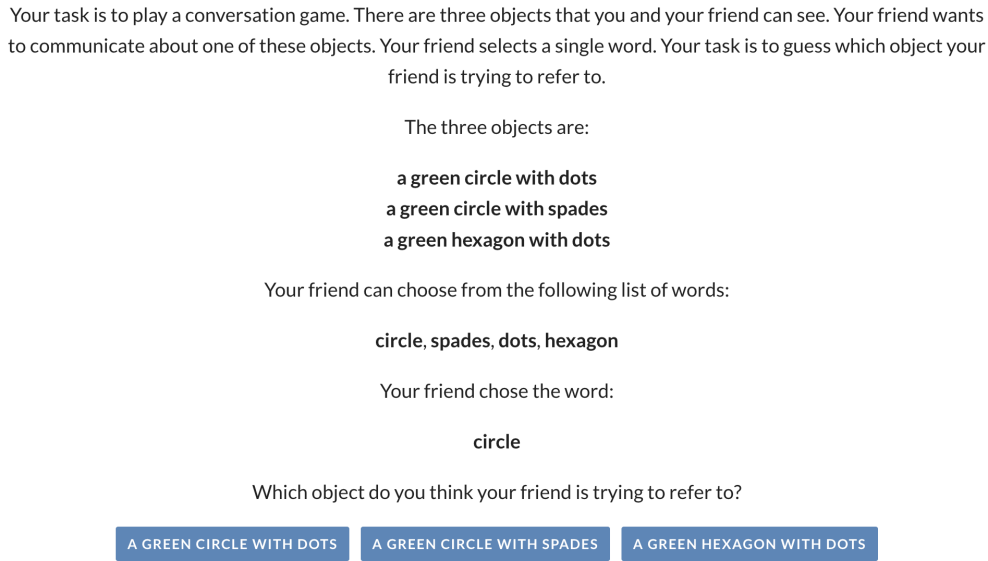A GREEN CIRCLE WITH DOTS   A GREEN CIRCLE WITH SPADES   A GREEN HEXAGON WITH DOTS

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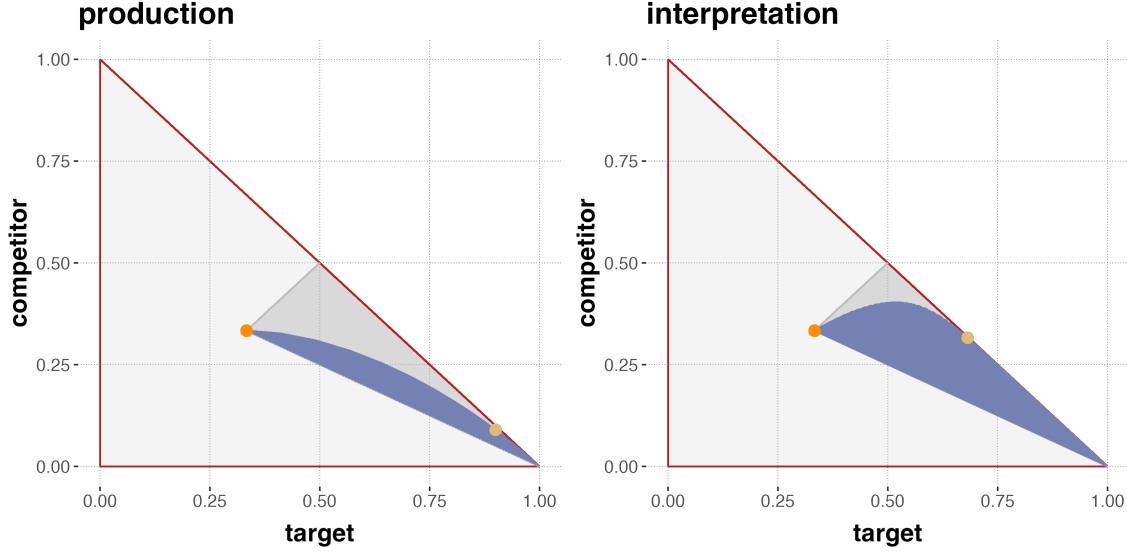
Figure 8: Range of predictions the intermediate-scope model makes for any pair of parameter values α_c and ϵ_c. The light gray triangle with the red boundary is the probability simplex, i.e., the space of all possible 3-place probability vectors. The darker gray area in gray boundary lines is the subspace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

Bai, Yuntao et al. (2022). *Constitutional AI: Harmlessness from AI Feedback*. arXiv: `2212.08073`.

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]`.

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.

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.

Estes, William K. and W. Todd Maddox (2005). "Risks of Drawing Inferences about Cognitive Processes from Model Fits to Individual versus Average Performance". In: *Psychonomic Bulletin & Review* 12.3, pp. 403–408.

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

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 (2016). "Reasoning in Reference Games: Individual- vs. Population-Level Probabilistic Modeling". In: *PLoS ONE* 11.5, e0154854.

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/`.

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.

Haaf, Julia M. and Jeffrey N. Rouder (2019). "Some do and some don't? Accounting for variability of individual difference structures". In: *Psychonomic Bulletin & Review* 26, pp. 772–789.

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.

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

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

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*.

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.

Stanovich, Keith E. and Richard F. West (2000). "Individual differences in reasoning: Implications for the rationality debate?" In: *Behavioral and Brain Sciences* 23.5, pp. 645–665.

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

Stiennon, Nisan et al. (2022). *Learning to summarize from human feedback*. arXiv: `2009.01325 [cs.CL]`.

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.

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]`.