Bayesian Statistical Modeling with Predictors from LLMs

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 and are increasingly used as components in larger applications, where LLM-based predictions serve as proxies for human judgements. This raises questions about the human-likeness of LLM-derived information and whether LLMs, if they deliver human-like behavior, could possibly be considered explanatory models of (aspects of) human cognition or language use. To shed more light on these issues, we here investigate the human-likeness of LLMs' predictions for multiple-choice decision tasks from the perspective of Bayesian statistical modeling. 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, 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

Enabled by the invention of deep neural transformer architectures (Vaswani et al., 2017), recent years have brought 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, both in industry and academia (Bommasani et al., 2021). Yet, for any downstream application of LLMs, it is crucial to understand what LLMs can or cannot reliably do.

The way in which LLM capabilities should be assessed depends on what their intended application is. For many industrial applications, 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" or "gold standard" exists, averaged over many instances of this task (e.g., BIG-bench authors, 2023), but more encompassing approaches also highlight the importance of a more holistic assessments of LLM, including factors such as robustness, fairness and efficiency (Liang et al., 2023). Benchmark-driven assessments are very useful for engineering purposes, when the main issue is whether a given system can perform a particular task correctly.

There are also applications of LLMs where benchmark testing on a "gold standard" is arguably not optimal. 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. Simple 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). More sophisticated examples include neuro-symbolic models in which LLMs supply specific parts of the relevant information for some practical application [MF: include citations], or where LLMs are a part of bigger programs to build towards something more akin to explanatory cognitive models (e.g., Wong et al., 2023). For example, information from LLMs can be used to generate alternatives for deliberation (e.g. Tsvilodub, Carcassi, and Franke, 2024), arguably
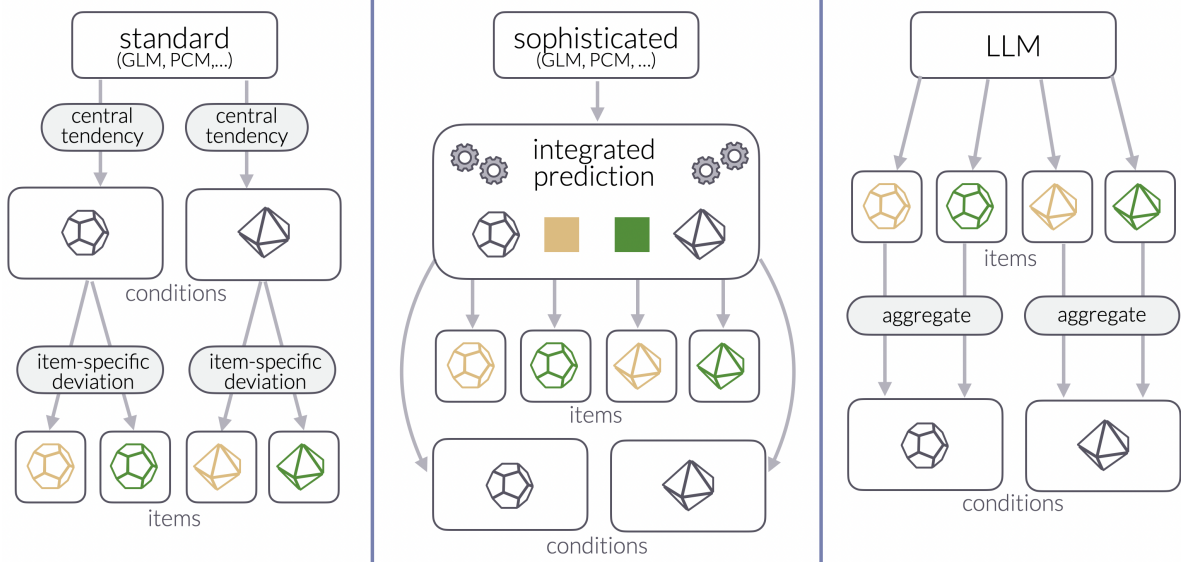
---

[*]Corresponding author.

Figure 1: Schematic representation of key conceptual differences between different types of predictive models. Standard statistical models, like hierarchical regression models, typically make predictions for aggregate data (e.g., at the condition level), and add random offsets for item-level variation. More sophisticated models, like some probabilistic cognitive models, may holistically combine information from an aggregate level (task, condition) with specific information about items. LLMs, in contrast, first and foremost give prediction about each individual item and must rely on proper aggregation to arrive at condition-level predictions.

similar to human resource-rational reasoning in open-ended domains. [MF: include citations] LLMs may also be 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). In many of these applications, LLMs essentially serve as a cheap, compressed stand-in for (average) human judgements, associations or choice behavior. To assess the quality of LLMs in such contexts, it is less important to compare against a gold standard, and more important to compare against the full distribution of the human behavior that is to be captured by the LLM. In sum, at least for some practical applications, the "gold standard" is not a single "correct" answer, but the full distribution of human responses.

Taking inspiration from experimental psychology, an increasing number of studies compares LLM predictions 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*. Most notably, there is a strong tradition of relatively early work in computational psycholinguistics (Marvin and Linzen, 2018; Hu, Gauthier, et al., 2020), which investigates whether quantitative predictions derived from language models 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 predictions derived from LLMs. Our foremost concern is methodological: How can we derive full distributional predictions from information supplied by LLMs, and how can we stringently test whether these distributional predictions are adequate given suitable empirical data? In short, while a lot of previous like-minded 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 probabilistic / statistical modeller. Taking numerical predictor values generated from LLMs as input, we explore strategies of building a Bayesian statistical model around them, and to scrutinize these LLM-grounded Bayesian statistical models with the usual methods of Bayesian data analysis (Gelman, Carlin, et al., 2014).

A main conceptual take-away of this investigation, is that statistical models built around LLMs are by design fundamentally different from common statistical or probabilistic cognitive models; an observation which also reflects back on the possibility of seeing LLMs as potentially explanatory for human behavior or cognition. Concretely, LLMs make predictions for each individual item, rather than specifying a predictor of central tendency at a more aggregate level, such as at the level of an experimental condition, as standard statistical models or probabilistic cognitive models normally do (see Figure 1). As this point is crucial for our investigation, the following elaborates briefly.

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 show some variability between items, and also variability between participants. This variability is commonly incorporated in standard statistical models as random stochastic variation, e.g., by using hierarchical regression models (Jaeger, 2008; Barr et al., 2013; Sorensen, Hohensteinb, and Vasishth, 2016), as shown on the left side of Figure 1. 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 model's predictions will naturally be set-up to predict data by taking condition-level properties, possibly in conjunction with item-level properties, into account (e.g., Nilson, Rieskamp, and Wagenmakers, 2011; Lee, 2011; Scheibehenne, Rieskamp, and Wagenmakers, 2013), as shown schematically in the middle of Figure 1. On the other hand, LLMs first and foremost provide predictions about each item. While it may be the case that, in producing an item-level prediction, the internal computation of a powerful LLM is informed by computations that incorporate abstract knowledge roughly corresponding to the condition-level, the atomic predictions accessible to the common user are specific to each individual string tested, thus raising concerns about robustness of predictions under perturbations of input prompts [MF: cite].

These considerations raise two important conceptual and empirical questions. First, we need to ask whether the item-level predictions made by LLMs are empirically correct. Second, the implied methodological challenge lies in specifying how item-level information can be used, e.g., by different aggregation methods, to make robust predictions at a more abstract level (condition, task, . . .). To investigate these issues, this paper introduces different ways of building a Bayesian probabilistic model around core predictor values derived from various LLMs. We then use the standard tools of Bayesian data analysis, to fit and check the resulting statistical models based on data from human multiple-choice experiments. Rather than aiming for scale and large-coverage, we focus on transparency and zoom in on a single case study of pragmatic language production and interpretation, namely so-called reference games (e.g. Deemter, Sluis, and Gatt, 2006; Degen, Franke, and Jäger, 2013; Graf et al., 2016; Sikos et al., 2021)., For reference games, a widely used probabilistic cognitive model exists, namely the Rational Speech Act model (Frank and Goodman, 2012), so that we can compare a theoretically motivated probabilistic cognitive model (PCM) against a probabilistic model built on top of LLM predictions. We find that variability predicted by the LLM at the item-level is generally not borne out by the human data, that not all ways of constructing condition-level predictions are equally good, and that different LLMs as backends may prefer to use different aggregation strategies.

The paper is structured as follows. Section 2 describes an experiment with human participants with a text-based reference game. Section 3 introduces the Rational Speech Act (RSA) model for the human data from the reference game experiment. Section 4 exposes a statistical model for item-level predictions derived from LLM scores and investigates whether this adequately captures the human data at the item-level based on scores from GPT-3.5 (Brown et al., 2020). Section 5 discusses different ways of deriving probabilistic predictions from LLMs at the condition-level and compares them against the human data and each other. Finally, Section 6 explores whether previous results generalize to other LLM backends by investigating different versions of LLaMA2 (Touvron et al., 2023).

2   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 2). In the **production condition**, the speaker
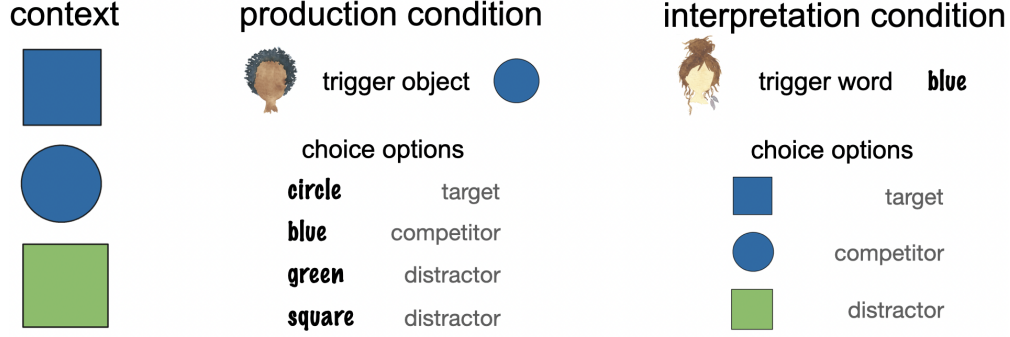
Figure 2: Structure of a reference game with human participants. Each trial consists of a set of objects, the so-called context. In production trials, participants choose a single word to describe a trigger object from the context. In interpretation trials, an object is selected as the likely object a trigger word referring to.

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 2 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 2), 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 2). 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 2. 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).[2]

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 2. 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 9 and 10 from Appendix A show example screenshots from the experiment.

---

[2]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/.
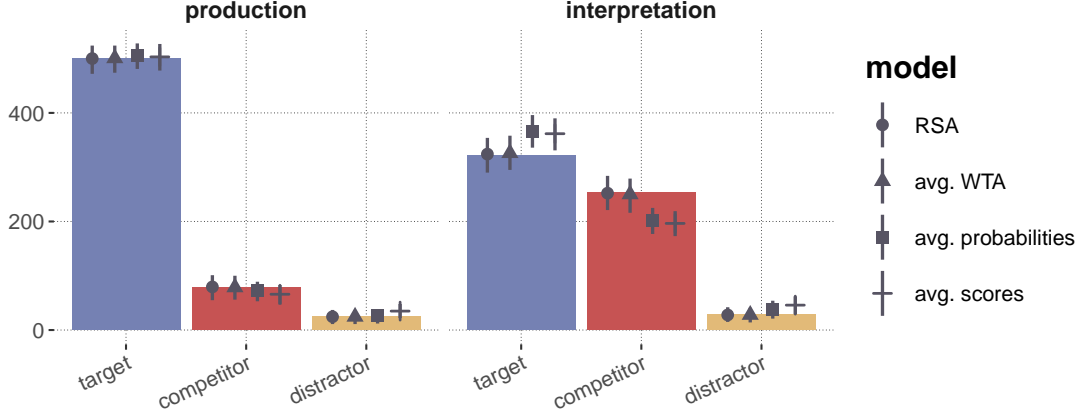
Figure 3: Counts of choices from reference games with human participants (colored bars), with summary statistics from the posterior predictive distribution of four models (shapes and error bars). Shapes show the mean of the posterior predictive distributions of the RSA model and three further models derived from item-level LLM-supplied scores. Error-bars show corresponding 95% credible intervals of the posterior predictive.

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 3.[3] 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).

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), 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 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 $P_{\text{prior}}(s)$ is a prior over states, a literal listener policy is defined as:

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

---

[3]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

$\mathfrak{L}$	🟦	🔵	🟩
"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 4: 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.

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., to maximize the log-probability of the trigger object:

$$P_S(u \mid s, \alpha) \propto \exp\left[\log P_{L_0}(s \mid u)\right]. \tag{1}$$

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

$$P_L(s \mid u, \alpha) \propto P_S(u \mid s, \alpha)\, P_{\text{prior}}(s). \tag{2}$$

Figure 4 gives example calculations (assuming a flat prior and $\alpha = 1$) for the reference game from Figure 2. 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. In sum, the condition-level predictions of the RSA model are a parameterized function $P_{\text{cond}}^{\text{RSA}}(R_l, C; \alpha_c)$, assigning a probability to each response category R_l (target, competitor, or distractor) in each condition C (production or interpretation).

The model constructed so far predicts probability zero for distractor choices, so that the human data shown in Figure 3, 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), so that we get:

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

where ϵ_c is a (condition-specific) parameter giving the probability that a choice was made by randomly guessing.[4]

The data D_C from condition C, see Figure 3, consists of counts for each response category. The parameterized likelihood function entailed by the RSA model for condition-level data D_C is:

$$P_{\text{cond}}^{\text{RSA}}(D_C \mid C, \alpha_C, \epsilon_C) = \text{Multinomial}(D_C, \langle P_r(R_l, C; \alpha_c, \epsilon_c)\rangle_{1 \le l \le 3}). \tag{3}$$

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

---

[4]Since the RSA model predicts probability 0 for the distractor option, this model is, in principle, able to predict any probability distribution over the three choice categories that is compatible with the order: $P_r(R_t) \ge P_r(R_c) \ge P_r(R_d)$. Intuitively, this is because with $\epsilon = 0$, $P_r(R_d, C; \alpha_c) = 0$, so that there is an α for any ratio of predicted choice probabilities for target and competitor, as long as the target probability is no smaller than the competitor probability. The ϵ-transformation is essentially a linear shift in the probability simplex towards the maximum entropy prediction, so that every prediction which obeys the ordering restriction above can be made for some pair of α and ϵ. This prediction-triviality is met in two ways. For one, the Bayesian priors on model parameters soft-constrain the model, so that the *ex ante* credible predictions do rule out many logically possible observations. For another, we break the triviality by assigning a non-zero probability to the prediction for the distractor option. The same triviality problem lurks for the *average-WTA* model introduced in Section 5, and the same solution is applied to it.

	model	data	method	condition	|95%	mean	95%|	|95%	mean	95%|	Bpppv
						α			ϵ		
!	RSA	item	—	prd.	2.62	3.13	3.69	0.08	0.12	0.16	0.29
!	RSA	item	—	int.	0.21	0.67	1.08	0.08	0.14	0.19	0.21
!	RSA	cond.	—	prd.	2.62	3.14	3.70	0.08	0.12	0.17	0.50
!	RSA	cond.	—	int.	0.27	0.68	1.13	0.09	0.14	0.19	0.51
	GPT	item	—	prd.	0.40	0.51	0.62	0.07	0.12	0.17	0.00
	GPT	item	—	int.	0.52	0.66	0.82	0.00	0.05	0.15	0.00
!	GPT	cond.	avg. scores	prd.	0.61	0.74	0.87	0.08	0.12	0.17	0.49
	GPT	cond.	avg. scores	int.	0.99	1.17	1.35	0.00	0.02	0.06	0.00
!	GPT	cond.	avg. prob.	prd.	0.81	5.00	14.58	0.08	0.12	0.16	0.61
	GPT	cond.	avg. prob.	int.	1.36	2.04	2.67	0.00	0.03	0.08	0.00
!	GPT	cond.	avg. WTA	prd.	0.86	1.04	1.22	0.08	0.12	0.17	0.48
!	GPT	cond.	avg. WTA	int.	0.16	0.43	0.69	0.06	0.13	0.19	0.50

Table 1: Summary statistics for model fits. Columns show the means and 95% credible intervals for posterior samples for each parameter, as well as the Bayesian posterior predictive p-values for each model, data type (item- or condition-level data), aggregation method (for condition-level data), and condition. The final column shows the Bayesian posterior predictive p-values. The first columns marks whether the row's Bayesian posterior predictive p-values are higher / lower than 0.05, flagging relative inability to capture the structure of the data.

Parameterized predictions, like in Equation (3), 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, 2016/2020; Lambert, 2018). Let $\alpha_c \sim$ log-Normal(1, 1) have a reasonably wide log-Normal prior, and let $\epsilon_c \sim$ Beta(1, 15) 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).[5]

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 3 shows summary statistics (means and 95% credible intervals) for the posterior predictive distribution of the RSA model (among other models for condition-level data, which are to be introduced later). We see that for both conditions the RSA model passes a "visual posterior predictive check" (Kruschke, 2015), which require that the distribution of posterior predictions includes the observed choice rates for each answer option. To corroborate the visual impression, Table 1 shows sample-based estimates of Bayesian posterior predictive p-values (Bppp values), using likelihood of the observed data as a test statistics. These Bppp values approximate the probability that a model conditioned on observed data D_{obs} predicts future data D_{rep}, of the same size and format as D_{obs}, that is at least as likely as the data D_{obs} itself is (given the posterior predictive distribution). Very small Bppp values indicate that the model might be inadequate for reproducing the data it was trained on, so to speak. This is clearly not the case for the RSA model with Bppp values close to 0.5, as shown in Table 1. In sum, the condition-level predictions by the RSA model, a theoretically motivated PCM, are not discredited by the condition-level data. Reversely, the RSA model seems to adequately capture 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, $D_C = \langle D_C^1, \ldots, D_C^m \rangle$, the likelihood function for analysis at the item-level uses the same (condition-level) predictor to separately predict data from each item, so that the likelihood function becomes:

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

Fitting the RSA model to the partitioned data set, we find estimates of Bppp values that do not discredit the model (see Table 1). This suggests that the item-level variation in the human data is not so pronounced as to provide strong evidence against the condition-level predictor from a theoretically motivated PCM. In other words, the condition-level RSA predictions seem to adequately capture also the item-level data.

---

[5]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).

The following sections apply these tests of Bayesian model criticism also to models built around predictor values from LLMs. Section 4 first looks at the item-level data, before Section 5 investigates different ways for deriving condition-level predictions.

[MF: mention (somewhere) where item-level effects could be added!]

4 Item-level predictions from LLMs

An LLM is design to predict the next token given an input string. This is essentially an item-level prediction: next-token probabilities are specified for concrete strings after a concrete instance of a task, i.e., an item of the task, not for the task as such. From these next-token probabilities, we can derive probabilities for multiple-choice answers for each item (this section) and for a condition (next section).

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 $|y_{ki}|$ tokens, $y_{ki} = w_{ki1}, \ldots, w_{ki|y_{ki}|}$. For simplicity of notation, we assume that the i-th choice option y_{ki} for each item k belongs to the same response category r_i. For the case at hand these categories are: target, competitor, distractor, so that y_{k1} always corresponds to the designated *target option*, r_1, that is the "gold standard" answer.

The most obvious *item-level score* an (autoregressive) LLM provides for each choice option y_{ki} is its log-probability:[6]

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

Based on each option's score, we can define the *item-level prediction* of choosing option y_{ki} in term of soft-maximization as:

$$P_{\text{item}}^{\text{LLM}}(y_{ki} \mid C; \alpha_c) \propto \exp(\alpha_c \, S_{ki}).$$

Notice that the usual item-level prediction used in benchmark testing is actually a "winner-takes-all" strategy, but this is just a special case of the above for $\alpha_C \to \infty$. Here, the α_C parameter corresponds to inverse temperature, so that $\alpha_C \to \infty$ corresponds to greedy sampling with temperature zero. Like in the RSA model, we use an independent soft-max parameter for each condition.

As mentioned previously, the items of the reference game experiment from Section 2 differ in which levels of features (color, shape, texture) instantiate the structure of the task shown in Figure 2, as well as the order of presentation of objects and words. We expect both human and LLM predictions to vary between different items: e.g., humans seem to have preferences for some features (e.g., Qing and Franke, 2015), such as over-production of informationally irrelevant material (e.g., Davies and Katsos, 2010; Rubio-Fernandez, 2019; Degen, Hawkins, et al., 2020); and machines may be susceptible to the presentation of the order of choice options. [MF: references prompt variation] It therefore becomes an empirical question of whether item-level predictions from LLMs provide a good fit, if we aim at predicting the human data separately for each item, not as a condition-level average.

To address this question, we assessed item-level scores S_{ki} from the text-davinci-003 instance GPT-3.5 August 2023 (Brown et al., 2020) for a text-based version of the reference game from Section 2. The task description x_k for item I_k is a text supplied as prompt (see Appendix B for an example). The choice options are categorized as in the human experiment.[7] Similar to the RSA model fit, we allow for random errors with probability ϵ_C:

$$P_{\text{item}}^{\text{LLM}}(y_{ki} \mid C; \alpha_c, \epsilon_c) = \epsilon_c \, P_{\text{item}}^{\text{LLM}}(y_{ki} \mid C; \alpha_c, \epsilon_c) \;+\; \epsilon_c/3.$$

---

[6][MF: TODO: Add quick word on how this can in principle be extended 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. [MF: check whether this is correct also after new LLaMA results.]

[7]In the current set-up the response type "distractor" has two instantiations in the production condition. Since choice options are a single word in the production condition, for simplicity, we lump both of the distractor words together and treat them as a single option.
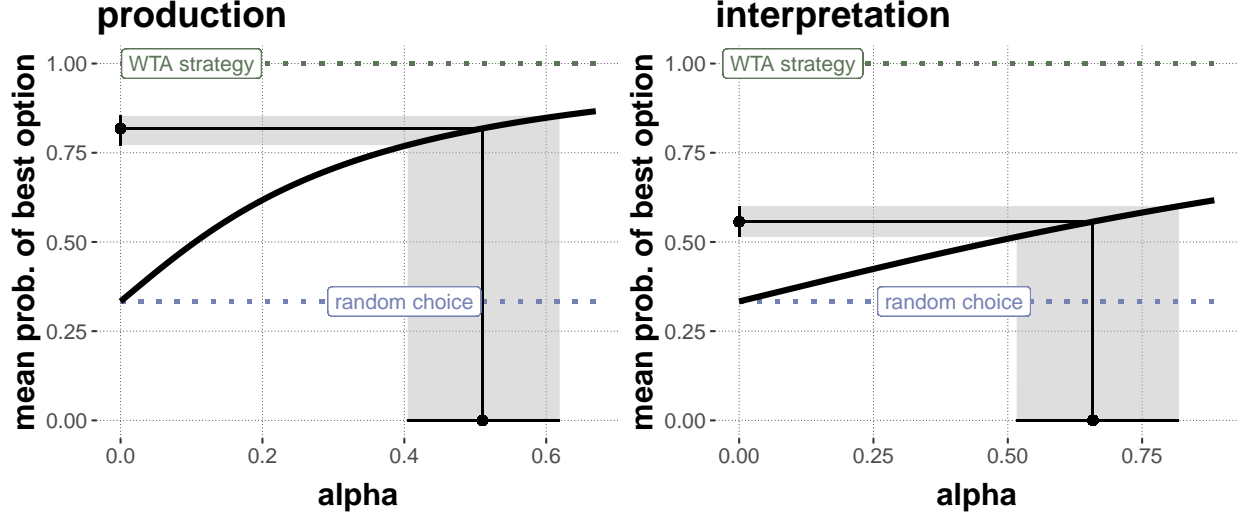
Figure 5: Predicted probability of highest-scoring answer category, averaged over items, for different values of softmax parameter α (black lines). The gray-shaded area indicates the posterior 95% credible interval for α, and the implied probabilistic prediction. For reference, the target probability under a random choice strategy and under the WTA strategy are shown with dashed lines. For credible values of α, the means of predicted probabilities for the target option are clearly distinct from the WTA strategy.

The resulting likelihood function for item-level data with item-level predictions is, where priors on parameters are as for the RSA model described in Section 3.:

$$P_{\text{item}}^{\text{LLM}}(D_c \mid C; \alpha_c, \epsilon_c) = \prod_{k=1}^{m} \text{Multinomial}\left(D_c^k, \left\langle P_{\text{item}}^{\text{LLM}}(y_{ki} \mid C; \alpha_c, \epsilon_c)\right\rangle_{1 \le l \le 3}\right).$$

Summary statistics for samples from the posterior distribution over parameters are shown in Table 1 (rows 4 and 5). Credible values of α parameters are comparatively low for the LLM-based model, suggesting that the predicted scores have rather large differences, which have to be compensated for by the softmax link-function to adequately fit the data. Figure 5 shows the mean of the probabilities for the highest scoring answer category, averaged over all items, for different parameter values of α, highlighting that for *a posteriori* credible values of α (gray shaded areas), the predictions are clearly distinct from the predictions of a WTA strategy (which assigns probability 1 to the highest scoring option for each item). This suggests that the WTA strategy, which is the special limiting case for $\alpha \to \infty$, provides a substantially worse fit for item-level choices than those provided by less extreme values of α.

Sampling-based approximations of Bayesian posterior predictive p-values for the by-item analysis are very low (see Table 1), suggesting that the unaggregated LLM scores are inadequate predictors of the human data. To corroborate this conclusion, Figure 6 shows that the item-level LLM scores predict variance which is not borne out by the human data. Concretely, the plots show, for each condition and 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. 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 uses the most obvious item-level score to predict what human participants choose on a by-item level is ruled out by the current experimental data.

5   Condition-level predictions

While LLMs do not make conditional-level predictions as such, they can be derived from item-level scores S_{ki} by averaging over all items belonging to the relevant condition. There are many ways of averaging item-level information. Figure 7 shows three salient approaches, which differ in what the underlying item-level measure is: the raw scores S_{ki},
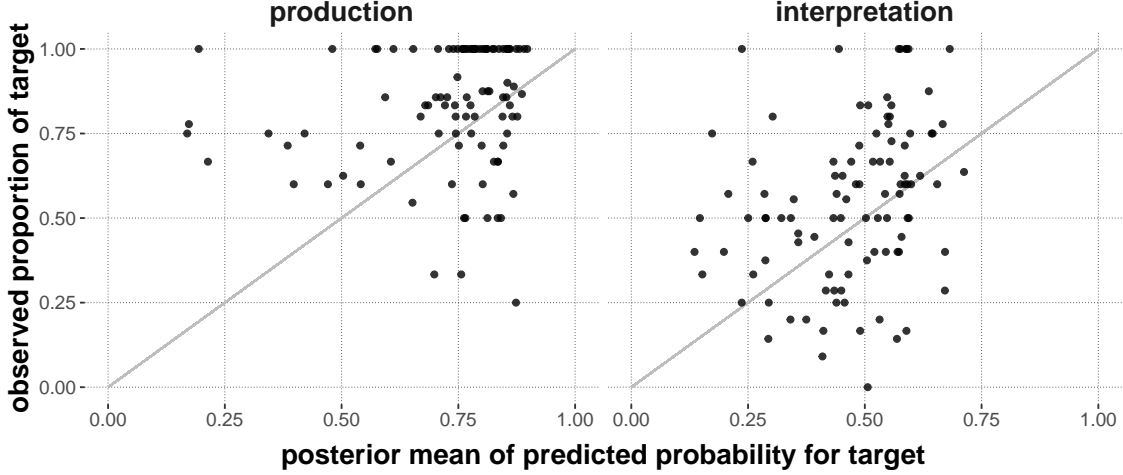
Figure 6: Item-level prediction-observation plot. Each dot represents an item. The *x*-coordinate represents the mean of the posterior prediction for the target choice probability for the given item. The *y*-coordinate represents the observed proportion of target choices for that item. The gray line (identity function) is the ideal prediction.
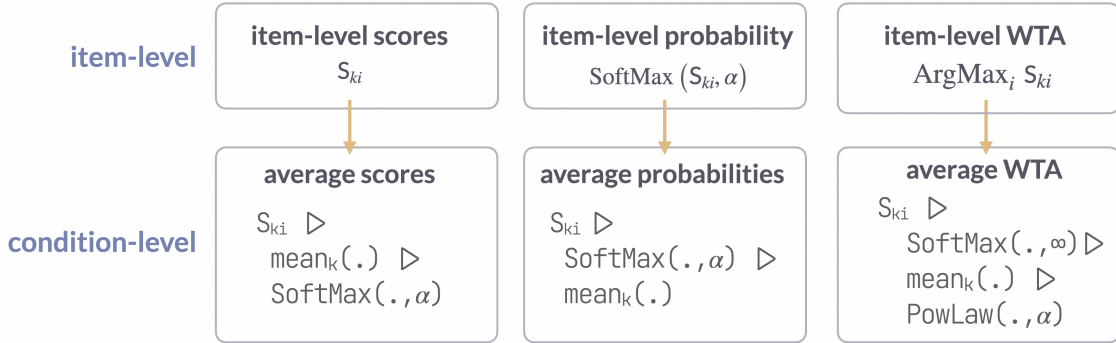
Figure 7: Schematic overview over three different approaches of obtaining condition-level predictions by aggregating item-level information. Approaches differ in the kind of item-level information they take into account, which entails differences in the order in which aggregation, transformation to probabilities and parameterized scaling occur.

the item-level probability derived from it (as used in Section 4), or the predictions from the winner-takes-all (WTA) strategy commonly used in benchmark testing.

The *average-scores predictor* first aggregates the item-level scores, and then transposes the averages into (scaled) probabilities using the usual parameterized softmax function:[8]

$$P_{\text{cond}}^{\text{SCR}}(r_i \mid C, \alpha_c) \propto \exp\left[\alpha_c \frac{1}{m} \sum_{k=1}^{m} \mathsf{S}_{ki}\right]. \qquad \text{[average scores (narrow-scope aggregation)]}$$

The *average-probabilities predictor* first transposes scores into probabilities with a parameterized softmax function, and only aggregates over items last:

$$P_{\text{cond}}^{\text{PRB}}(r_i \mid C, \alpha_c) = \frac{1}{m} \sum_{k=1}^{m} P_{\text{item}}^{\text{LLM}}(y_{ki} \mid C, \alpha_c). \qquad \text{[average probabilities (wide-scope aggregation)]}$$

Finally, the *average-WTA predictor* considers the prediction of the WTA strategy (a softmax with $\alpha \to \infty$) as the basic item-level unit to aggregate over. To add parameterized scaling to this method, a power-law transformation is a natural

---

[8]The reported results average over the multi-set $\{I_1, \ldots I_m\}$ of items that occurred in the human experiment for condition C. By using a multi-set, which may contain a single item multiple times, we produce aggregate predictions for exactly the set of items that the participant group saw, which provides the most fitting counterpart to the human data.

choice:

$$P_{\text{cond}}^{\text{WTA}}(r_i \mid C, \alpha_c) \propto \left[\frac{1}{m} \sum_{k=1}^{m} P_{\text{item}}^{\text{WTA}}(y_{ki} \mid C) \right]^{\alpha_c}, \qquad \text{[average WTA (intermediate-scope aggregation)]}$$

where $P_{\text{item}}^{\text{WTA}}(y_{ki} \mid C) = \lim_{\alpha \to \infty} P_{\text{item}}^{\text{LLM}}(y_{ki} \mid C, \alpha)$.

The *average-scores* and *average-probabilities* predictors are equivalent if there is only one item, in which case the prediction of the *average-WTA* method is the special case of $\alpha \to \infty$. For cases with more than one item, the predictions of the three predictors are not guaranteed to be the same. Conceptually, the *average-probabilities* and the *average-WTA* predictors, but not the *average-scores* predictor, are compatible with a picture in which condition-level predictions result from the actual predictions at the item level. However, based on the results from the previous section, the item-level predictions for the WTA strategy are demonstrably incongruent with the human data. In general, the average-probability and average-WTA predictors can lead to qualitatively different results about task-level accuracy (see Appendix C).

Using the same approach as for the RSA model in Section 3, we can build likelihood functions for condition-level data around the three predictors introduced above. With the same priors and methods used before, we obtain samples from the posterior over parameters and samples from the posterior predictive distributions. Summary statistics for posteriors over model parameters are shown in Table 1. What is noteworthy is that for the *average-WTA* model, the estimates of α_c in the production condition do not rule out, in fact lie close to, the special value $\alpha_c = 1$, for which the power-law transformation is the identity function. This means that just averaging WTA-responses at the item-level yields a reasonable predictor for the production data at the condition-level. However, for the interpretation condition, the value $\alpha_c = 1$ is clearly outside the range of credible parameter values, so that a simple recipe like "always average WTA-responses without transformation" is not a viable strategy for good condition-level predictions in general. It is also worth noting that for the *average-probability* model values of α_c of around five and above are virtually indistinguishable from the item-level predictions of the WTA strategy (see Figure 5). This suggests that, for the production data, aggregation of item-level predictions from the WTA-strategy gives good condition-level predictions.

Figure 3 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 *average-WTA* model pass 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 1 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 1 suggest that the *average-scores* and the *average-probabilities* model are able to reproduce the production data, but fail on the interpretation data; and that only the *average-WTA* model does not fail to capture 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 *average-WTA* 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. On the other hand, there is a stark contrast between the results of item-level and condition-level data. The model that was able to properly fit the condition-level data builds on predictions for item-level choices that are clearly incompatible with the human item-level data. The model seems to be "right for the wrong reasons."

6 Generalizing to other LLM backends

All previous results where obtained for LMM predictions derived from GPT-3.5 turbo. To investigate whether key results replicate for other LLMs and scales, we ran the same analyses also for scores derived from variants of LLaMA2 (Touvron et al., 2023), in particular the 7, 13 and 70 billion parameter versions of the models fine-tuned with human feedback only (marked as 'hf' in the following) and based on human feedback and chat data (marked as 'chat'). Table 2 shows the relevant summary statistics and Bayesian posterior predictive p-values. Figure 8 shows visual posterior predictive checks for the condition-level data. The plots show the *differences* between observed counts and the models' posterior predictions for expected counts for each condition and response option. We find that the average-WTA predictor is consistently able to capture the data from the production condition, except for one of the six models (LLaMA2-chat 70B). However, the average-WTA predictor recovers the interpretation data for only one of the six models (LLaMA2-chat 13B, see also Table 2). Interestingly, for LLaMA2-chat 13B also the average-probability predictor is able to recover the interpretation data, unlike for the GPT-3.5. As for GPT-3.5, no model captured the
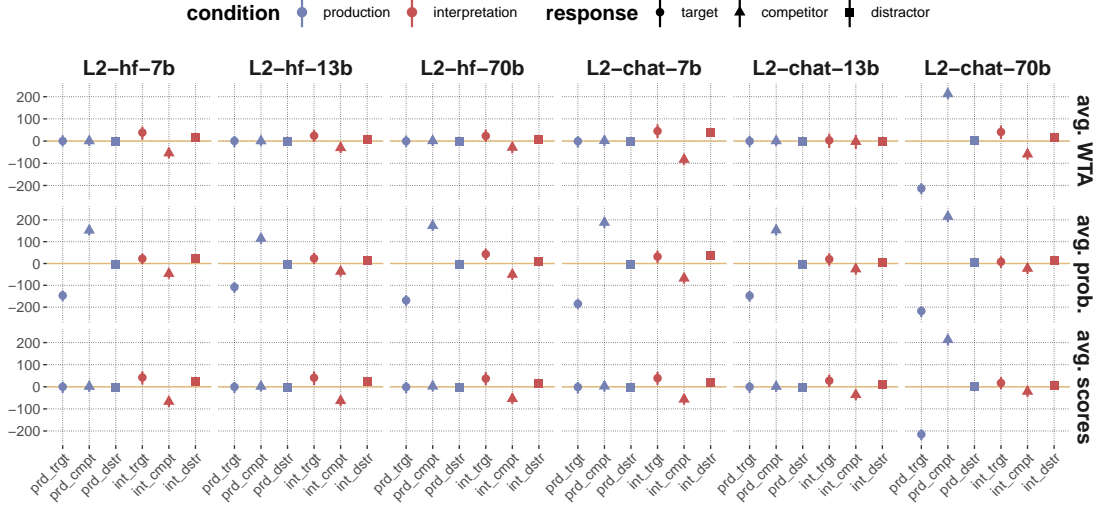
11

Figure 8: Visual posterior predictive checks on condition-level data for different models of the LLaMA family, paired with different aggregation strategies.

item-level variance observed in the human data. In sum, we find general support for the previous conclusions that the average-WTA predictor is most successful in capturing the condition-level data, and therefore that, where models manage to recover the condition-level data, they seem to be "right for the wrong reasons." Additionally, we also find variability in which method of aggregation works well with which LLM as scoring model.

7 Conclusion

While the common practice in evaluating the capabilities of LLMs is based on accuracy averaged over large collections of data, this work took the alternative route to explore what we learn if we subjected LLMs to the same routines and strong demands on distributional quality of fit to human data as we normally do for statistical or probabilistic cognitive models. Knowledge of the adequacy of LLMs' full distributional predictions for simple tasks that require human like judgement or decisions is important to gauge in how far LLMs can be trusted to provide such information in applications such as hybrid, neuro-symbolic models [MF: refs]. A main contribution of this paper is methodological, showing how statistical model criticism can be used for LLMs in the first place, and, more specifically, how it can be insightful in the detailed assessment of how LLMs might or might not be able to predict human data at the degree of accuracy that we would normally aspire for when dealing with explanatory statistical models in cognitive science. Applying this comparative approach to a minimal, but non-trivial data set, we find that the LLM predictions on a per-item level predict variance that is not attested in the human data. From several candidate predictor measures for aggregate condition-level data, only one was not refuted by the human data, but this was one that relied on the empirically implausible WTA-strategy at the item-level, incidentally the same strategy that is commonly used in accuracy-based benchmark testing (BIG-bench authors, 2023).

Explanatory power. A basic observation brought to the foreground by our approach is that LLMs' atomic predictions are for individual items and that some aggregation method is needed to derive more abstract, condition- or task-level predictions. This is potential one sense in which LLMs may be felt to be less, or not at all, explanatory. They do not offer, at least not directly, a human-comprehensible compression of reality into a *kind* of response pattern, over and beyond making a prediction for each particular situation. As this kind of compression is arguably important for a sense of understanding (Dellsén, 2020; Grimm, 2021), the direct comparison of LLMs with common practices in experimental psychology and with probabilistic cognitive models, provides an interesting perspective on why LLMs are often felt to be lacking in explanatory power.

This perspective on the explanatory role of LLMs goes beyond the factors of performance, indirect support and parsimony identified by van Deemter (2023). It is also subtly different from considerations of an LLM's ability to generalize (Hupkes et al., 2020). It rather suggests that *transferability* is a dimension to "explanatory power" that is important as well. Imagine that models M and M' have been designed for and trained on data from a situation S, but

need to be applied to a different situation T. Assume that for model M the only way to make predictions for T is to collect data pertaining to T, and either retrain or fine-tune the model. In contrast, model M' can make predictions for T without novel data collection by recognizing a meaningful difference between S and T and consequently manually changing parameter values or model-internal mechanics to accommodate for this change. In that case, model M' would be more transferable than model M. For example, if we change the experimental setting for a reference game to consist of data from a special population, such as very young children or language-impaired adults, a potentially reasonable architectural change to a model like the RSA model is to consider differences in the sets of alternative utterances for the speaker (e.g. Noveck, 2001). Even though this is only a vague explication of a notion of transferability, it suffices to corroborate the intuition that probabilistic cognitive models like the Rational Speech Act model, which are designed to operate at a higher level of conceptual abstraction, will often appear more transferable than models like LLMs, which make predictions not for *kinds* of situations but for *particular* situations. Whether any given model's transfer-ability is correct, is an orthogonal empirical question. It is also an empirical question, exactly to which extent and in which areas LLMs are (not) transferable, especially if we consider prompting a transfer strategy (Liu et al., 2022; Xie et al., 2022). In any case, the comparison between LLMs and other statistical or probabilistic cognitive models started here suggests systematic research into the transfer-ability and explanatory value of LLMs, e.g., by prompting strategies that are empirically insightful or by their use in composite neuro-symbolic models that implement theoretically-meaningful conceptual differences between types of situations.

Variability in predictions and performance measures.   Despite being anchored in a small, but detailed case-study, the fact that different plausible methods of aggregating item-level information led to condition-level predictions of variable quality is worrisome. The wide-spread reliance on a winner-takes-all strategy might be inconsistent with the actual use practices of LLMs, which may not always rely on a temperature-zero sampling strategy (see Appendix C). In conclusion, research on LLMs should systematically study the conceptual and empirical consequences of seemingly minor decisions in evaluation or application settings. Variability of performance measures also comes with a well-known risk for robust and reproducible science (c.f. Hu and Levy, 2023; Tsvilodub, Wang, et al., 2024). The more researcher degrees of freedom there are, the higher the risk of false results, even in the absence of intentions to mislead (e.g. Ioannidis, 2005; Chambers, 2017). Testing LLMs as predictors in statistical models should raise awareness for the issue of robust research methods also in NLP (Wieling, Rawee, and Noord, 2018).

The work presented here also contains considerable researcher degrees of freedom. We only considered one out of several conceivable ways of carving a parameterized likelihood function from LLM-based predictors. If future work would contribute to systematically exploring these, the main goal of this paper would have been met, which is to raise awareness for the possibility, perhaps even necessity, to scrutinize LLMs at the same level of rigorous detail as other models in cognitive science.

Human-like predictions from LLM.   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 parts 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; Tsvilodub, Carcassi, and Franke, 2024). Based on the data set and the detailed analyses conducted here, it seems not infeasible to use numerical predictions from LLMs as part of predictive probabilistic models. But, ideally, low-level variation, such as from order of presentation or similar "nuisance variables," should be averaged out or taken into account in some way or other, 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. In sum, we conjecture 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 taken into account. This, however, entails that each LLM component in a hybrid model is independently tested against at least a modestly sized empirical data set.

No substitute for human subjects.   Looking at LLM-derived predictions for human data from the perspective of cognitive modeling highlights the fact that LLMs predict item-level variation, but not subject-level variation, which is common in human data. We may consider variation introduced via softmax / temperature as analogous to between-human variation, but this likely falls short of reality, where pronounced differences in answer behavior may surface. For the particular case of pragmatic language use, prior research has shown that individual participants have markedly different behavioral profiles, often consistently behaving like literal language users or more sophisticated language users (e.g., Nieuwland, Ditman, and Kuperberg, 2010; Franke and Degen, 2016; Spychalska, Kontinen, and Werning, 2016). It is an open question whether predictions from LLMs reflect the same kind of variation. The results presented here recommend skepticism. We therefore side with the cautious voices that do not recommend replacing human participants with LLMs in psychological research (Dillion et al., 2023; Harding et al., 2023). In contrast, this points to an important

challenge for future LLM research. The fact that aggregate predictions can track aggregate human behaviour means that the variance on both sides is washed out to achieve a similar result. This raises the issue of finding the systematic differences between LLMs and human answerers. The task then would be to find cases where the differences do *not* wash out and ask: What if anything do these cases share?

Limitations and follow-up work.   The scope of our experimental investigation was deliberately small but perspicuous. Focusing on the methodological contributions and details in performance assessments, we investigated a minimal non-trivial case study where we could triangulate LLMs, probabilistic cognitive models and human data. This work may therefore serve as starting point for a wider investigation of more complex data sets and case studies in which LLMs are analyzed as and directly compared to explanatory PCMs.

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]`.

Chambers, Chris (2017). *The Seven Deadly Sins of Psychology*. Princeton University Press.

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]`.

Davies, Catherine and Napoleon Katsos (2010). "Over-informative children: Production/comprehension asymmetry or tolerance to pragmatic violations?" In: *Lingua* 120.8, pp. 1956–1972.

Deemter, Kees van, Ielka van der Sluis, and Albert Gatt (2006). "Building a Semantically Transparent Corpus for the Generation of Referring Expressions." In: *Proceedings of the Fourth International Natural Language Generation Conference*. Sydney, Australia: Association for Computational Linguistics, pp. 130–132.

van Deemter, Kees (2023). "Dimensions of Explanatory Value in NLP models". In: *Computational Linguistics*.

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.

Dillion, Danica et al. (2023). "Can AI language models replace human participants?" In: *Trends in Cognitive Sciences* 27.7, pp. 597–600.

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 (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/`.

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.

Graf, Caroline et al. (2016). "Animal, dog, or dalmatian? Level of abstraction in nominal referring expressions". In: *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Ed. by Anna Papafragou et al. Austin, TX: Cognitive Science Society, pp. 2261–2266.

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]`.

Harding, Jacqueline et al. (2023). "AI language models cannot replace human research participants". In: *AI & SOCIETY*. ISSN: 1435-5655.

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, Jon Gauthier, 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.

Hu, Jennifer and Roger Levy (2023). *Prompt-based methods may underestimate large language models' linguistic generalizations*. arXiv: `2305.13264 [cs.CL]`.

Hupkes, Dieuwke et al. (2020). "Compositionality decomposed: how do neural networks generalise?" In: *Journal of Artificial Intelligence Research* 67.

Ioannidis, John P. A. (2005). "Why Most Published Research Findings Are False". In: *PLoS Medicine* 2.8, e124.

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.

Liang, Percy et al. (2023). "Holistic Evaluation of Language Models". In: *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

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 (2016/2020). *Statistical Rethinking*. Boca Raton: Chapman and Hall.

Nieuwland, Mante S., Tali Ditman, and Gina R. Kuperberg (2010). "On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities". In: *Journal of Memory and Language* 63.3, pp. 324–346.

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.

Noveck, Ira A. (2001). "When Children are more Logical than Adults: Experimental Investigations of Scalar Implicature". In: *Cognition* 78, pp. 165–188.

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.

Rubio-Fernandez, Paula (2019). "Overinformative Speakers Are Cooperative: Revisiting the Gricean Maxim of Quantity". In: *Cognitive Science* 43.11.

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*.

Spychalska, Maria, Jarmo Kontinen, and Markus Werning (2016). "Investigating scalar implicatures in a truth-value judgement task: evidence from event-related brain potentials". In: *Language, Cognition and Neuroscience* 31.6, pp. 817–840.

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]`.

Tsvilodub, Polina, Fausto Carcassi, and Michael Franke (2024). *Towards Neuro-Symbolic Models of Language Cognition: LLMs as Proposers and Evaluators*.

Tsvilodub, Polina, Hening Wang, et al. (2024). *Predictions from language models for multiple-choice tasks are not robust under variation of scoring methods*. arXiv: `2403.00998` `[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.

Wieling, Martijn, Josine Rawee, and Gertjan van Noord (Dec. 2018). "Reproducibility in Computational Linguistics: Are We Willing to Share?" In: *Computational Linguistics* 44.4, pp. 641–649.

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]`.

Xie, Sang Michael et al. (2022). *An Explanation of In-context Learning as Implicit Bayesian Inference*. arXiv: `2111.02080` `[cs.CL]`.

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

A   Screenshots from the online experiment with human participants

Figure 9 shows a trial from the production condition, Figure 10 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 9: 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 10: 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 9 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:
```

```
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   Excursion: Accuracy scores from average-probability vs. average-WTA predictors

Standard benchmark testing looks at *task accuracy*, defined as the probability of selecting the "gold standard" response, usually based on "winner-takes-all" (WTA) selection of the highest scoring option. We can generalize this and define the *softmax-based accuracy* as the average probability of choosing the designated target response option r_1 for the *condition-level soft-max prediction* (notation as defined in main text):

$$P_{\text{cond}}^{\text{SM}}(r_1) = \frac{1}{m} \sum_{k=1}^{m} P_{\text{item}}^{\text{SM}}(y_{k1}) \ .$$

The WTA-based accuracy (the standard measure in most benchmark testing) is the special case of $\alpha \to \infty$.

The WTA-based accuracy can differ qualitatively from the more general softmax-based accuracy, as shown by the following example. The example shows that differences between accuracy measures depend on the variation in item-level scores, in particular on the relation between score-ordering and score-differences. Note that 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.[9]

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.

---

[9]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: $^4/_5 \ ^1/_2 + ^1/_5 \ 0 = ^2/_5$.

	model	data	method	condition	α			ϵ			
					|95%	mean	95%|	|95%	mean	95%|	Bpppv
	L2-hf-7b	item	—	prd.	1.09	1.30	1.51	0.00	0.02	0.06	0.00
	L2-hf-7b	item	—	int.	1.90	2.52	3.13	0.00	0.04	0.12	0.00
!	L2-hf-7b	cond.	avg. scores	prd.	9.65	11.50	13.62	0.08	0.12	0.16	0.48
	L2-hf-7b	cond.	avg. scores	int.	6.80	7.96	9.14	0.00	0.02	0.05	0.00
	L2-hf-7b	cond.	avg. prob.	prd.	4.98	16.44	33.06	0.06	0.10	0.13	0.00
	L2-hf-7b	cond.	avg. prob.	int.	20.06	79.30	163.94	0.00	0.01	0.04	0.00
!	L2-hf-7b	cond.	avg. WTA	prd.	4.17	4.96	5.85	0.07	0.12	0.16	0.48
	L2-hf-7b	cond.	avg. WTA	int.	0.87	1.02	1.18	0.00	0.02	0.05	0.00
	L2-hf-13b	item	—	prd.	1.37	1.91	2.47	0.00	0.07	0.12	0.00
	L2-hf-13b	item	—	int.	2.13	2.90	3.69	0.00	0.05	0.14	0.00
!	L2-hf-13b	cond.	avg. scores	prd.	13.73	16.32	19.10	0.08	0.12	0.16	0.50
	L2-hf-13b	cond.	avg. scores	int.	7.73	9.11	10.42	0.00	0.02	0.05	0.00
	L2-hf-13b	cond.	avg. prob.	prd.	69.84	185.80	360.44	0.06	0.09	0.13	0.00
	L2-hf-13b	cond.	avg. prob.	int.	13.63	39.72	88.28	0.00	0.02	0.05	0.00
!	L2-hf-13b	cond.	avg. WTA	prd.	2.28	2.70	3.18	0.08	0.12	0.16	0.51
	L2-hf-13b	cond.	avg. WTA	int.	0.78	0.94	1.12	0.00	0.03	0.07	0.03
	L2-hf-70b	item	—	prd.	1.53	1.86	2.22	0.00	0.03	0.07	0.00
	L2-hf-70b	item	—	int.	3.20	4.01	4.74	0.00	0.03	0.08	0.00
!	L2-hf-70b	cond.	avg. scores	prd.	44.57	53.17	62.44	0.08	0.12	0.17	0.48
	L2-hf-70b	cond.	avg. scores	int.	6.86	8.14	9.32	0.00	0.02	0.05	0.00
	L2-hf-70b	cond.	avg. prob.	prd.	10.51	26.59	45.73	0.07	0.10	0.14	0.00
	L2-hf-70b	cond.	avg. prob.	int.	10.86	23.63	42.16	0.00	0.03	0.09	0.00
!	L2-hf-70b	cond.	avg. WTA	prd.	5.88	7.03	8.19	0.08	0.12	0.17	0.50
	L2-hf-70b	cond.	avg. WTA	int.	0.55	0.68	0.81	0.00	0.03	0.07	0.04
	L2-chat-7b	item	—	prd.	0.34	0.41	0.49	0.00	0.03	0.07	0.00
	L2-chat-7b	item	—	int.	0.79	1.05	1.34	0.00	0.05	0.15	0.00
!	L2-chat-7b	cond.	avg. scores	prd.	58.47	69.64	81.79	0.08	0.12	0.16	0.48
	L2-chat-7b	cond.	avg. scores	int.	2.03	2.37	2.73	0.00	0.02	0.05	0.00
	L2-chat-7b	cond.	avg. prob.	prd.	8.05	27.26	55.30	0.07	0.11	0.15	0.00
	L2-chat-7b	cond.	avg. prob.	int.	2.86	19.24	51.41	0.00	0.01	0.04	0.00
!	L2-chat-7b	cond.	avg. WTA	prd.	12.18	14.70	17.21	0.08	0.12	0.16	0.50
	L2-chat-7b	cond.	avg. WTA	int.	0.86	1.01	1.14	0.00	0.02	0.05	0.00
	L2-chat-13b	item	—	prd.	0.36	0.44	0.53	0.00	0.03	0.08	0.00
	L2-chat-13b	item	—	int.	0.68	0.90	1.11	0.00	0.05	0.13	0.00
!	L2-chat-13b	cond.	avg. scores	prd.	5.56	6.63	7.76	0.08	0.12	0.16	0.49
	L2-chat-13b	cond.	avg. scores	int.	1.70	2.05	2.40	0.00	0.02	0.06	0.01
	L2-chat-13b	cond.	avg. prob.	prd.	3.90	22.25	74.12	0.06	0.10	0.14	0.00
!	L2-chat-13b	cond.	avg. prob.	int.	2.85	4.70	7.01	0.00	0.03	0.07	0.08
!	L2-chat-13b	cond.	avg. WTA	prd.	4.28	5.11	5.99	0.07	0.12	0.16	0.49
!	L2-chat-13b	cond.	avg. WTA	int.	0.15	0.32	0.51	0.02	0.11	0.18	0.50
	L2-chat-70b	item	—	prd.	0.28	0.35	0.42	0.00	0.03	0.08	0.00
	L2-chat-70b	item	—	int.	1.03	1.24	1.47	0.00	0.03	0.08	0.00
	L2-chat-70b	cond.	avg. scores	prd.	0.40	0.55	0.74	0.00	0.04	0.10	0.00
!	L2-chat-70b	cond.	avg. scores	int.	1.91	2.37	2.89	0.00	0.03	0.08	0.16
	L2-chat-70b	cond.	avg. prob.	prd.	0.43	0.66	0.96	0.00	0.04	0.10	0.00
	L2-chat-70b	cond.	avg. prob.	int.	2.82	5.21	8.43	0.00	0.01	0.04	0.02
	L2-chat-70b	cond.	avg. WTA	prd.	0.14	0.21	0.35	0.00	0.06	0.13	0.00
	L2-chat-70b	cond.	avg. WTA	int.	0.85	1.00	1.14	0.00	0.02	0.05	0.00

Table 2: Summary statistics for model fits based on LLaMA2 backends. The information shown is the same as in Table 1.