Using Artificial Populations to Study Psychological Phenomena in Neural Models

Jesse Roberts¹, Kyle Moore¹, Drew Wilenzick², Douglas Fisher¹

¹Vanderbilt University
²Cornell University

jesse.roberts@vanderbilt.edu, kyle.a.moore@vanderbilt.edu, ddw77@cornell.edu, douglas.h.fisher@vanderbilt.edu

Abstract
The recent proliferation of research into transformer based natural language processing has led to a number of studies which attempt to detect the presence of human-like cognitive behavior in the models. We contend that, as is true of human psychology, the investigation of cognitive behavior in language models must be conducted in an appropriate population of an appropriate size for the results to be meaningful. We leverage work in uncertainty estimation in a novel approach to efficiently construct experimental populations. The resultant tool, PopulationLM, has been made open source. We provide theoretical grounding in the uncertainty estimation literature and motivation from current cognitive work regarding language models. We discuss the methodological lessons from other scientific communities and attempt to demonstrate their application to two artificial population studies. Through population based experimentation we find that language models exhibit behavior consistent with typicality effects among categories highly represented in training. However, we find that language models don’t tend to exhibit structural priming effects. Generally, our results show that single models tend to over estimate the presence of cognitive behaviors in neural models.

Introduction
In the wake of success following the introduction of transformers in (Vaswani et al. 2017) and the public deployment of powerful variants of GPT, many have started to question if these models exhibit behavior similar to human cognition. Work analyzing the cognition of these powerful, Turing-complete (Roberts 2023) models is important not only to the explanation and interpretation of the models themselves but may offer insight into cognition more generally, a synergism best embodied by the interplay of reinforcement learning and neuroscience (Subramanian, Chitlangia, and Baths 2022). We believe this emerging study of cognitive behavior in neural models can be improved by adopting methods from branches of science more typically associated with statistical testing. Without appropriate experimental methodology conclusions may not be robust in the face of variations, a symptom associated with the greater replicability crisis (Goodman, Fanelli, and Ioannidis 2016). Research attempting replication and extension of ToM results in GPT-4 found that relatively small experimental alterations caused the effect to disappear (Ullman 2023). This suggests the experimental design in the original study was insufficient to support the drawn conclusions.

This is precisely the motivation for the present paper. Claims that fail to be reproducible regarding powerful AI models may ultimately result in erosion of the public’s trust and attention. Any study of neural model cognitive behavior should characterize not only the presence but the size of the effect and the significance. Doing so necessitates rigor which may decrease erroneous conclusions, and will permit better explanation of neural model cognitive behavior through meta-analytic study. To this end, we present and demonstrate an artificial population generation method for the study of cognitive phenomena in neural models with the hope that it will aid in the reproducibility of research regarding the behavior of neural models.

This paper contributes by drawing connections between social and behavioral experimental design and neural model uncertainty estimation resulting in a (1) tool called PopulationLM for the creation of populations of neural models via stratified MC dropout. We harvest novel metrics and explore population best practices by applying artificial populations to the (2) replication and extension of (Misra, Ettinger, and Rayz 2021) (correlation analysis) and (3) (Sinclair et al. 2022) (difference analysis). We present novel results regarding the presence of typicality and structural priming effects in language models.

Behavioral Phenomena in Neural Models
In this section we review the current work related to the study of cognitive behavior in neural language models paying specific attention to the measures reported and methodology. Table 1 summarizes the works that have been identified, organized by the behavioral phenomenon that they investigate. This review and meta-commentary does not invalidate any of the findings in the associated papers. Rather, it serves as a compendium of work so far in this field and helps to illuminate the problem we wish to address.

The measures reported refers to the measure applied to the model output. Statistic refers to the statistical analysis applied to the measures. We find that most papers used atypical measures of effect like frequency of occurrence or qualitative analysis and tend to not use statistical testing. Those
employing t-tests don’t typically specify the particular test. Analogously, we find that less than a third of the papers report significance levels for their results. In contrast, most authors did isolate the experimental, independent variable. Rows marked with an * indicate works that did so only in a subset of reported experiments.

No study in our review utilizes uncertainty estimation to systematically perturb the model or the input. Therefore, no work has been done to study neural behavior in a population. In the latter half of this paper, we study two behavioral phenomena from table 1 in artificial populations: typicality and structural priming (SP). Typicality refers to a high degree of agreement across subjects in humans when ranking items as more or less typical of a given category and is known to be related to rate of retrieval of an item given the category (Rosch 1975). Structural priming refers to the predilection for a sentence structure similar to the most recently observed syntactical structure (Pickering and Ferreira 2008).

### Populations of Neural Models

In all social and behavioral science, conclusions drawn from a single subject face severe limitations. Without a population of subjects it is impossible to know if the individual is typical along the dependent variable in the population or an outlier.

Studying the cognitive behavior of neural models, either as an ontology or in relation to human psychology, suffers from a similar limitation. There always exists a possibility that an expression of a behavior is anomalous or that the behavior is tenuously supported in the network.

In this paper we refer to models and their derivatives as different species, i.e. BERT and DistilBERT are individual species, while they both belong to the same family. Genus is reserved for fine-tuned variants.

Forming inter-species populations is an intuitive but flawed approach since we wish to facilitate the study behaviors that may emerge in specific species, as is known to occur as a function of model size (Wei et al. 2022). Inter-species populations don’t permit this type of myopic study.

Instead we form populations using work from neural model uncertainty estimation. In that context, the goal is not a population but an estimation of model uncertainty. However, this is precisely the characteristic typically extracted from a population study, the degree to which a result is consistent across individuals. We refer to this as the population uncertainty. Several uncertainty estimation methods have been proposed in literature and can be placed into 4 broad groups (Gawlikowski et al. 2023), single network deterministic, test time augmentation, ensemble techniques, and Bayesian approximations.

Single network deterministic methods attempt to estimate the uncertainty of a network without multiple predictions being made. However, they trade accuracy for speed. Test time augmentation methods perform perturbations of the input data and estimate uncertainty across a single model’s outputs (Lyzhov et al. 2020). Though this is a promising solution for closed source models, there exists a bound on the perturbation resolution possible in transformers with test time augmentation due to Hahn’s lemma (Hahn 2020). Ensemble techniques, generally outperform Bayesian methods (Lakshminarayanan, Pritzel, and Blundell 2017) but require multiple models trained independently. The price associated with from scratch training makes this a poor solution (Sharir, Peleg, and Shoham 2020). Therefore, Bayesian approximation is the most applicable uncertainty estimation technique for the creation of populations of open source models.

### PopulationLM

We use Monte Carlo (MC) dropout (Gal and Ghahramani 2016) to form populations from base models. A neuron mask is assembled from the instances of random variables and put data and estimate uncertainty across a single model’s outputs (Lyzhov et al. 2020). Though this is a promising solution for closed source models, there exists a bound on the perturbation resolution possible in transformers with test time augmentation due to Hahn’s lemma (Hahn 2020). En-ssemble techniques, generally outperform Bayesian methods (Lakshminarayanan, Pritzel, and Blundell 2017) but require multiple models trained independently. The price associated with from scratch training makes this a poor solution (Sharir, Peleg, and Shoham 2020). Therefore, Bayesian approximation is the most applicable uncertainty estimation technique for the creation of populations of open source models.

### Table 1: Review summary of large language model behavioral studies.

<table>
<thead>
<tr>
<th>Phenomena</th>
<th>Study by</th>
<th>Measure(s)</th>
<th>Statistic</th>
<th>Significance</th>
<th>Experimental Var</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theory of Mind</td>
<td>Bubenek et al.</td>
<td>qualitative</td>
<td></td>
<td></td>
<td>not isolated</td>
</tr>
<tr>
<td></td>
<td>Kosinski</td>
<td>frequency</td>
<td></td>
<td></td>
<td>not isolated</td>
</tr>
<tr>
<td></td>
<td>Sap et al.</td>
<td>frequency</td>
<td></td>
<td></td>
<td>not isolated</td>
</tr>
<tr>
<td></td>
<td>Ullman</td>
<td>frequency</td>
<td></td>
<td></td>
<td>not isolated</td>
</tr>
<tr>
<td></td>
<td>Trott et al.</td>
<td>token probs</td>
<td>$\chi^2 + \beta$</td>
<td>reported</td>
<td>isolated*</td>
</tr>
<tr>
<td>Logical Reasoning</td>
<td>Binz and Schulz</td>
<td>token probs</td>
<td>$\chi^2 + t + \beta$</td>
<td>reported</td>
<td>isolated*</td>
</tr>
<tr>
<td></td>
<td>McCoy, Pavlick, and Linzen</td>
<td>frequency</td>
<td></td>
<td></td>
<td>isolated*</td>
</tr>
<tr>
<td></td>
<td>Lamprinidis</td>
<td>frequency</td>
<td></td>
<td></td>
<td>not isolated</td>
</tr>
<tr>
<td>Framing &amp; Anchoring</td>
<td>Binz and Schulz</td>
<td>token probs</td>
<td>$\chi^2 + t + \beta$</td>
<td>reported</td>
<td>isolated*</td>
</tr>
<tr>
<td></td>
<td>Jones and Steinhardt</td>
<td>frequency</td>
<td></td>
<td></td>
<td>isolated*</td>
</tr>
<tr>
<td></td>
<td>Suri et al.</td>
<td>frequency</td>
<td></td>
<td>reported</td>
<td>isolated*</td>
</tr>
<tr>
<td>Heuristics</td>
<td>Binz and Schulz</td>
<td>token probs</td>
<td>$\chi^2 + t + \beta$</td>
<td>reported</td>
<td>isolated*</td>
</tr>
<tr>
<td></td>
<td>Jones and Steinhardt</td>
<td>frequency</td>
<td></td>
<td></td>
<td>isolated</td>
</tr>
<tr>
<td>Typicality</td>
<td>Misra, Ettinger, and Rayz</td>
<td>token probs</td>
<td>$r + \rho$</td>
<td>reported</td>
<td>isolated</td>
</tr>
<tr>
<td>Priming</td>
<td>Sinclair et al.</td>
<td>token probs</td>
<td></td>
<td></td>
<td>isolated</td>
</tr>
<tr>
<td>Emotion Induction</td>
<td>Coda-Forno et al.</td>
<td>frequency</td>
<td>$r + t + \text{probit} \beta$</td>
<td>reported</td>
<td>not isolated</td>
</tr>
</tbody>
</table>

$r = \text{Pearson}, \rho = \text{Spearman}, \beta = \text{Berksons}, t = \text{t-test.}$
Table 2: Kolmogorov-Smirnov test for each population and each experiment compared to the base model. Null hypothesis $H_0$ is population probabilities and base model probabilities are drawn from the same underlying distribution per species. Populations very similar to the base model have an $\times$.

<table>
<thead>
<tr>
<th>Model Species</th>
<th>Paper</th>
<th>Typicality KS test</th>
<th>SP KS test</th>
<th>Type (parameters)</th>
<th>Training Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>DistilBERT</td>
<td>Sanh et al.</td>
<td>0.056 (p=0.055) $\times$</td>
<td>0.04 (p&lt;0.05)</td>
<td>MLM (66M)</td>
<td>BookCorpus, Wiki</td>
</tr>
<tr>
<td>BERT Base</td>
<td>Devlin et al.</td>
<td>0.051 (p=0.108) $\times$</td>
<td>0.03 (p=0.06) $\times$</td>
<td>MLM (110M)</td>
<td>BookCorpus, WebText</td>
</tr>
<tr>
<td>BERT Large</td>
<td></td>
<td>0.072 (p&lt;0.01) $\times$</td>
<td>0.05 (p&lt;0.01)</td>
<td>MLM (340M)</td>
<td></td>
</tr>
<tr>
<td>GPT</td>
<td>Radford et al.</td>
<td>0.069 (p&lt;0.01) $\times$</td>
<td>0.08 (p&lt;0.01)</td>
<td>MLM (120M)</td>
<td></td>
</tr>
<tr>
<td>DistilGPT-2</td>
<td>Sanh et al.</td>
<td>-0.072 (p&lt;0.01) $\times$</td>
<td>0.45 (p&lt;0.01)</td>
<td>CLM (82M)</td>
<td></td>
</tr>
<tr>
<td>GPT-2 Medium</td>
<td>Radford et al.</td>
<td>-0.03 (p=0.685) $\times$</td>
<td>0.29 (p=0.1) $\times$</td>
<td>CLM (117M)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.075 (p&lt;0.01) $\times$</td>
<td>0.51 (p&lt;0.01)</td>
<td>CLM (345M)</td>
<td></td>
</tr>
<tr>
<td>RoBERTa Base</td>
<td>Liu et al.</td>
<td>0.065 (p&lt;0.02) $\times$</td>
<td>0.08 (p&lt;0.01)</td>
<td>MLM (125M)</td>
<td>BERT train data, Stories, CC, OpenWebText, News</td>
</tr>
<tr>
<td>RoBERTa Large</td>
<td></td>
<td>0.15 (p&lt;0.1) $\times$</td>
<td>0.19 (p&lt;0.1)</td>
<td>MLM (355M)</td>
<td></td>
</tr>
</tbody>
</table>

The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)

18908

Figure 1: Single model regression vs population of models probability-typicality regression for RoBERTa Large. Rank is inversely related to typicality. 95% confidence intervals shown for both with very narrow bounds on the population.

How Much Dropout Is Necessary? MC dropout has been applied to transformers previously in (Shelmanov et al. 2021) and (Vazhentsev et al. 2022). In both of these papers the authors experimentally found 0.1 to be the most effective dropout rate for discouraging incorrect, poorly supported outputs.

We experimented with dropout rates from 0.1 to 0.8. We
found no advantage in using larger rates of dropout for experiments, as increased rates caused signal erosion with all behavioral correlations being dissolved beyond rates of 0.5. Therefore, we recommend that statistical studies adopt a 0.1 nominal dropout rate.

**How Big Should the Artificial Population Be?** Population size for a study is related to two important statistical measures, significance and power. The significance of a result is a measure of the probability of the null hypothesis. The power of a test is a measure of the probability that the test will correctly reject the null hypothesis and avoid a false negative.

For evaluating cognitive behavior in neural models, the power of a test is less important as the effect of a false negative is not likely to cause damage. However, the significance of a result is of the utmost importance as this permits meta-analytical extension and can act as a mitigator of sensationalism when applied properly.

We empirically find that a population of 50 is an acceptable compromise, providing sufficient statistically significant deviation from the base model in table 2 without dramatic computational costs. Interestingly, it seems the models tend to have correlated relationships with associated dropout populations. The KS tests for the two experiments in table 2 show that BERT, GPT-2, and RoBERTa models tended to have KS effect sizes which were rank correlated across experimental populations within model families. However, the model correlations don’t extend outside the family. This suggests that 50 member populations may tend to be sufficient for the approximated deep Gaussian process to emerge.

**Experiment 1: Typicality Effects**

We reproduce and extend the experiment conducted in (Misra, Ettinger, and Rayz 2021) which assessed the base model total correlation between probability and typicality. Our base model probabilities agree with past results, and we
contribute novel tests using dropout populations and within category analysis which shed light on the factors that support the emergence of typicality effects in language models.

**Experimental Setup** We use typicality data from (Rosch 1975) which gives a typicality rank, $r_i$, for each item, $i$, in category $C$. As in the original experiment, we construct prompts, $\pi_i$, for each $i \in C$ and measure the probability assigned to the category given the prompt, $P(C|\pi_i)$. So, for each category, only the item in the prompt (independent variable) will change across queries while the effect on the category probability is measured (dependent variable). After each prompting, the model input is flushed, guaranteeing that only the independent variable is manipulated for each trial for each category. This necessitates that the results be evaluated within category, since cross category results are not controlled. However, we also evaluate the test results across all categories for each model as a direct comparison to the results reported in the original paper.

**Individual Probability Correlation Test** For each population, we test for behavior consistent with typicality by evaluating the Pearson correlation between $P(C|\pi_i)$ and $r_i$ for all $i \in C$ and for all categories in the dataset. We hypothesize that, consistent with previous results, the probabilities output by the models will be positively correlated with typicality.

As predicted, all models show significant ($p<0.05$) probability/typicality correlation within nearly all categories consistent with typicality in humans in figure 2. DistilBERT shows insignificant correlation with the categories tool and vegetable, while DistilGPT2 and RoBERTa base both have insignificant correlation with tool. More generally, the correlation between probability and typicality is strongly conditioned upon category for all models. The behavior shows strong differentiation between causal (CLM) and masked language models (MLMs). Among all MLMs the total correlation is markedly higher than for CLMs. Further, the categories for which each model most exhibits typicality behavior differs across MLMs and CLMs.

The green bars in figure 2 represent the total correlation (across categories) obtained by only the base models and agree with past results (Misra, Ettinger, and Rayz 2021). However, the total population correlation, shown in dark blue, suggests that the base model total correlation is an over estimation of the true correlation.

**Population Uncertainty Correlation Test** We hypothesize that population uncertainty will be positively correlated with diminishing typicality. That is, as stimuli become less typical, the population will have decreasing agreement. Therefore, we test for correlation between normalized population standard deviation (as a measure of group uncertainty), $\sigma_n(P(C|\pi_i))$ and typicality rankings $r_i$.

In figure 3, for masked language models, mean normalized population uncertainty has a significant positive correlation as typicality diminishes mediated by category. The models tend to become more uncertain as the items become less typical. Therefore, we believe that masked language models, like humans (Rosch 1975), are more certain when inferencing about typical items. The categories which are most positively correlated with population certainty tend to be consistent with those which were most correlated with probability.

Interestingly, the standard deviation of model probabilities was found to scale with the mean of the probability, giving the appearance of increased population agreement as probability declined. Therefore, mean normalization is used. Mean normalized uncertainty may be more meaningful than standard deviation alone for models which learn to output probabilities.

![Figure 3: CLMs exhibit negative Uncertainty/Typicality correlation while MLMs exhibit positive correlation. Within category, total Spearman correlation ($p<0.08$), and total well represented category (item frequency > 60000) Spearman correlation ($p<0.01$) are shown.](image)

In sharp distinction, all causal language models exhibit negative certainty/typicality correlation. We speculate that this may be due to differences in training data and modeling objective. i.e. it is not typical for humans to say extremely obvious things like "A sparrow is a bird." Therefore, a dropout trained conversational model may have high uncertainty regarding highly typical item/category pairings in completions. However, this hypothesis is not readily testable due to GPT-2 training data unavailability. Further, the categories among the CLMs which are most negatively correlated with population certainty do not seem to be the same categories as those which were most positively correlated with probability in figure 2. This suggests that CLMs represent something all together different than MLMs in their population uncertainty.

**Confound Test** We considered that frequency of an item within the training data could act as a third variable and confound the results. To address this we evaluate the Pearson correlation of item frequency in the training data with typi-
We hypothesized that category "understanding" may be important for the emergence of typicality behavior and that mean within category item frequency in training data may predict category understanding. To test this, we performed a regression between within category Pearson correlation and mean within category item frequency in the training data for the BERT population.

In figure 4 we find that average item frequency within a category is highly correlated with the strength of typicality effects exhibited by the model within that category. The exception being the categories tool and toy. Further research may be necessary to determine why these categories do not fit the otherwise established trend. We suspect that this is the result of conflicts from the basic-level effect, that humans have a preferred level of categorization, which has a known relationship with typicality (Rosch et al. 1976). Tool and toy may be outliers because they are not at the basic categorization level for many of the items listed in those categories.

If the anomalous categories are removed, the correlation between within category probability/typicality correlation and within category mean item frequency in training data is Pearson $r=0.98$ ($p<0.01$) and with tool and toy included Pearson $r=0.7$ ($p<0.03$).

Another measure of concept "understanding" is the persistence of a concept through growing rates of dropout. As the dropout rate increases, more and more neurons are masked in the population members causing concepts with fewer constituent neurons to be ablated. So, we swept the dropout rate for the population from 0.1 to 0.8 and found that the categories which were highly represented in the training data tended to persist as dropout increased, while the categories with less training data representation tended to decorrelate at lower dropout rates.

We interpret these complimentary results to suggest that model "understanding" of a category may be driven by overall category representation in the training data and that, within categories which are well understood, models are likely to exhibit typicality effects. We find that this is the case for all tested model species as restricting the total Pearson correlation to the categories which are well represented in the BERT training data, a partial constituent of all other model’s training, leads to significant increases in all model probability-typicality correlations in figure 2.

**Experiment 2: Structural Priming Effects**

In (Sinclair et al. 2022) the authors investigated whether language models exhibit behavior consistent with the structural priming effect. We run a similar experiment using sentence data from their work. However, we use a dropout population, modify the experimental setup to control for unaddressed confounds, and perform a split-group cross validation.

**Experimental Setup** To test for SP in language models we adopt 3 treatment conditions: the control (CT) is the probability of a sentence, $S_x$, without any priming $P(S_x)$; the primed treatment (PT) is the probability of that sentence when the language model is first prompted with a sentence, $\pi_x$, of similar structure $P(S_x|\pi_x)$; and the alternative treatment (AT) is the probability of $S_x$ when prompted with a sentence, $\pi_y$, of differing structure $P(S_x|\pi_y)$ but identical semantic meaning. Any effect AT has will not be analogous to SP. However, it is not a placebo as it may not be inert. Therefore, both AT and PT must be compared to CT for contextualization.

We split 3000 examples into two groups and conduct all 3 treatments on all 50 population members per species. The results for the first group of 1500 are reported and the results for the second set of 1500 are used for cross validation. The


| Model Species | Wilcoxon $P(S_2|\pi_x) > P(S_x)$ | $\frac{\mu(P(S_2|\pi_x)) - \mu(P(S_x))}{\text{st}(P(S_2|\pi_x)) - \text{st}(P(S_x))}$ | Pearson(AT-CT, PT-CT) $r$ | Structural Priming |
|---------------|---------------------------------|-------------------------------------------------|--------------------------|-------------------|
| DistilBERT    | 0.42 (p=1)                      | ---                                             | ---                      | None              |
| BERT Base     | 0.27 (p=1)                      | ---                                             | ---                      | None              |
| BERT Large    | 0.47 (p=1)                      | ---                                             | ---                      | None              |
| GPT           | 0.14 (p=1)                      | ---                                             | ---                      | None              |
| DistilGPT-2   | 0.82 (p=0)                      | 0.93 ±0.001                                     | 0.89 (p<0.01)            | None              |
| GPT-2         | 0.96 (p=0)                      | 0.96 ±0.001                                     | 0.93 (p<0.01)            | None              |
| GPT-2 Medium  | 0.99 (p=0)                      | 0.98 ±0.001                                     | 0.94 (p<0.01)            | None              |
| RoBERTa Base  | 0.99 (p=0)                      | 0.98 ±0.001                                     | 0.70 (p<0.01)            | Marginal          |
| RoBERTa Large | 0.99 (p=0)                      | 0.96 ±0.001                                     | 0.67 (p<0.01)            | Marginal          |

Table 3: Test results used to detect structural priming. From left to right: the first relates preference for priming; the second finds the percent of the preference magnitude not attributable to SP; the third measures the correlation between SP and an alternative.

cross validation showed all results repeated within ±0.02 (p<0.01) of our reported results.

**Individual Probability Difference Test** For behavior consistent with SP to be present, the relationship PT > CT must tend to hold. To test this, we employ the Wilcoxon signed rank test, a non-parametric test appropriate for testing relative ranking of paired samples.

In table 3 the results show that only GPT-2 and RoBERTa exhibit a preference for PT over the control. These models require subsequent testing as SP is one possible explanation for their preference, but an alternative hypothesis is that the models prefer being primed with anything at all. Preference for priming could be induced by the presence of WebText and OpenWebText in the training data of GPT-2 and RoBERTa families as these possess conversational data in which SP is more likely to be observed.

**Elimination of Alternative Hypotheses** To eliminate a preference for priming regardless of structure as an alternative hypothesis, we find the 95% confidence interval of $\mu(\text{PT}) - \mu(\text{CT})$. This is the size of the preference change induced by PT which is not attributable to SP. Wilcoxon is not used in this case as it would result in the cancellation of the control group due to internal subtraction. It is possible that the effect magnitude will be similar but the individual samples not be correlated. Therefore, we also find the Pearson correlation between PT-CT and AT-CT.

For all models the alternative treatment produced an average effect which was 96% as large as the mean change due to SP. Further, the GPT-2 family showed strong correlation between AT-CT and PT-CT, suggesting these models do not prefer priming with a similar structure. However, the results in table 3 show that the RoBERTa family has a response to PT distinct from the AT response based on Pearson’s $r$.

**Comments** In contrast to previous work we find little evidence for the presence of structural priming effects. The RoBERTa family of models exhibits a response that is distinct when primed with a sentence of similar structure to the target sentence. However, the preference magnitude is not differentiable from an alternative structure priming. No other models exhibit significant, distinct effects.

---

1https://github.com/JesseTNRoberts/PopulationLM

**Conclusions**

This paper addresses a current need in the study of cognitive behavior in neural models by introducing PopulationLM, a system built on MC dropout for the creation of efficient populations of neural models. This permits population based analysis of model behavior which may decrease the presence of atypical behaviors. In both experiments our population studies, when compared to the original experiments of other authors, show that conclusions drawn from single models tend to over estimate the presence of cognitive behaviors. Beyond robustness, populations permit the study of divergence or decorrelation as a function of dropout and characterization of population uncertainty or disagreement.

We have conducted novel experiments using PopulationLM regarding the presence of typicality and structural priming in language models, being careful to isolate and analyze along independent variables and report effect sizes and significance. We find that typicality is consistently present while structural priming seems to not be, with both having predictable ties to behavior representation in training data.

PopulationLM may have further reaching applications beyond the study of cognitive behavior. Many papers have begun to systematically study prompt pattern effects (White et al. 2023). These possess similar issues of robustness to cognitive studies and could benefit from study among a population. Further, it’s possible that populations of models may serve as proxies for initial human behavior studies in the future. This could augment the ethical and financial efficacy of psycholinguistic research (Brysbaert 2019).

Test time augmentation (Gawlikowski et al. 2023) may be used to create local variations that perform similarly to dropout populations. However, the effects will decay with the length of the decoder context (Hahn 2020). The longer the priming, the less effect each individual token, including the experimental prompt, will have. We intend to investigate the use of test time augmentation for closed source language model systematic population studies in future work.

Finally, future work should investigate the (1) surprising increase in CLM certainty with decreased typicality, (2) the use of mean normalization to characterize probabilistic model certainty, and (3) the presence and impact of other cognitive phenomena like basic level effects.
Ethical Statement

Some work in the area of large language model cognitive behavior has produced conclusions which are not replicable when small variations are applied to the experiments. This coupled with the wide attention being given to large language models can lead to sensationalism and potentially contribute to the erosion of the public’s trust in the scientific community. The hoped effect of this paper is to bring awareness and partially address the current situation by providing a more systematic method for improving the robustness of results.

Alternatively, if data were to be improperly handled or inappropriate statistical testing performed, the impact could be negative. PopulationLM has the ability to augment the number of datapoints on which testing may be performed. If not analyzed appropriately, the increased data may potentially be used to support fallacious conclusions.

Acknowledgements

The authors would like to acknowledge Dr. Christopher MacLellan for collaborative discussions and Dr. Pat Langley for general guidance in methodological research.

References


