# Large Language Model Recall Uncertainty is Modulated by the Fan Effect

**Jesse Roberts**
Tennessee Tech University
jtroberts@tntech.edu

**Kyle Moore**
Vanderbilt University
kyle.a.moore@vanderbilt.edu

**Thao Pham**
Berea College

**Oseremhen Ewaleifoh**
Vanderbilt University

**Doug Fisher**
Vanderbilt University

## Abstract

This paper evaluates whether large language models (LLMs) exhibit cognitive fan effects, similar to those discovered by Anderson in humans, after being pre-trained on human textual data. We conduct two sets of in-context recall experiments designed to elicit fan effects. Consistent with human results, we find that LLM recall uncertainty, measured via token probability, is influenced by the fan effect. Our results show that removing uncertainty disrupts the observed effect. The experiments suggest the fan effect is consistent whether the fan value is induced in-context or in the pre-training data. Finally, these findings provide *in-silico* evidence that fan effects and typicality are expressions of the same phenomena.

## 1 Introduction

Some subfields of AI are explicitly interested in understanding and mimicking the nature of human cognition (cognitive modeling, computational psychology, affective computing) but even more implicitly rely on models of human cognition (human-computer interaction, embodied robotics, collaborative robotics, AI assistive technology, computational game theory). A model that, through training, learned to implicitly exhibit human-like cognitive behaviors could be of tremendous value both to the explicit study of human cognition as an ethical test subject, and as a more faithful model of human behavior to those fields that seek to develop systems to work along side human counterparts. We believe that some large language models (LLM) may be excellent candidates for such a role.

LLMs process information in a manner that is fundamentally different from humans. The matrix multiplications, maximum inner product search, and perceptron networks may have, at some level, been inspired by the biological neuronal system. But beyond the superficial, the systems bear no similarities. In spite of algorithmic and mechanistic dissimilarity, a growing body of work suggests that by merely training on human-language data, large language models learn to exhibit human-like cognitive behaviors as shown in Table 1.

In this paper, we survey the work applying cognitive science inspired evaluations to LLMs to analyze, understand, and catalog their relation to human cognition. We extend the existing work by providing the first investigation of human-like fan effects à la Anderson and Reder (1999) in LLMs. This effect is specifically interesting because it has a relation to the previously studied typicality effect, and it is understood to be an expression of human categorization uncertainty that has been precisely measured through response time delay.

Our results show that (1) some LLMs exhibit human-like fan effects based on the typicality of categorical items learned in pre-training; (2) some LLMs exhibit human-like fan effects based on the relative frequency of items in the model context; and (3) with uncertainty mitigated, the observed fan effect is disrupted. Of the models tested, Mistral (Jiang et al., 2023) and SOLAR (Kim et al., 2023) exhibit noteworthy human-like fan effects, including nuanced differential fan effects previously observed in humans (Radvansky, 1999).

The results have two practical implications: LLMs learn to exhibit human-like uncertainty and that uncertainty may interfere with recall tasks. Our results additionally provide *in-silico* evidence that the fan effect is a special case of typicality as is true in COBWEB models (Silber and Fisher, 1989).

Understanding the cognitive behaviors acquired from language is essential to the successful application of LLMs in human-adjacent scenarios. Generally speaking, human-like cognitive effects may serve to smooth interactions between machine and human. Alternatively, a minority of discrepancies may serve to undermine the interactions.

https://github.com/JesseTNRoberts/Large-Language-Model-Recall-Uncertainty-is-Modulated-by-the-Fan-Effect



*Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 303–313
November 15-16, 2024 ©2024 Association for Computational Linguistics

| Phenomena | Study by | Measure(s) | Statistic | Significance | Systematic Perturbation |
|---|---|---|---|---|---|
| Theory of Mind | Bubeck et al. (2023) | qualitative | — | — | — |
| | Kosinski (2023) | frequency | — | — | — |
| | Sap et al. (2022) | frequency | — | — | — |
| | Ullman (2023) | frequency | — | — | — |
| | Trott et al. (2023) | token probs | $\chi^2 + \beta$ | reported | — |
| | Ma et al. (2023) | frequency | — | — | — |
| | Li et al. (2023) | frequqncy | — | — | — |
| Logical Reasoning | Binz and Schulz (2023) | token probs | $\chi^2 + t + \beta$ | reported | — |
| | McCoy et al. (2019) | frequency | — | — | — |
| | Lamprinidis (2023) | frequency | — | — | — |
| | Yax et al. (2024) | token probs | $\chi^2$ | reported | — |
| | Lampinen et al. (2023) | frequency | $\chi^2 + t$ | reported | — |
| Framing & Anchoring | Binz and Schulz (2023) | token probs | $\chi^2 + t + \beta$ | reported | — |
| | Jones and Steinhardt (2022) | frequency | — | — | — |
| | Suri et al. (2023) | frequency | — | reported | — |
| Decision-Making | Binz and Schulz (2023) | token probs | $\chi^2 + t + \beta$ | reported | — |
| | Jones and Steinhardt (2022) | frequency | — | — | — |
| | Coda-Forno et al. (2024) | frequency | $\beta$ | reported | — |
| | Hagendorff et al. (2023) | frequency | $\chi^2$ | reported | — |
| Typicality | Misra et al. (2021) | token probs | $r + \rho$ | reported | — |
| | Roberts et al. (2024b) | token probs | $r$ | reported | model |
| Priming | Sinclair et al. (2022) | token probs | — | — | data |
| | Roberts et al. (2024b) | token probs | $w$ | reported | data + model |
| | Michaelov et al. (2023) | token probs | — | — | data |
| Emotion Induction | Coda-Forno et al. (2023) | frequency | $r + t +$ probit $\beta$ | reported | — |

Table 1: Review summary of large language model behavioral studies. $r$ = Pearson, $\rho$ = Spearman, $\beta$ = $\beta$-regression, $t$ = t-test, $w$ = Wilcoxon. Systematic perturbation refers to the presence of noise injected into the model or data to improve result robustness.

## 2 Background

The ***fan effect*** is a psychological effect in human categorization behavior, first identified in Anderson (1974), where subjects take longer to recognize and accept or reject concepts that have overlapping features with concepts previously presented in a learning set. This has most commonly been studied using concepts made up of person-place pairs. More formally, given some training concept set $S = \{< X_1, Y_1 >, ..., < X_n, Y_n >\}$, where $X$ and $Y$ are features of the concepts, response time when performing recognition tasks for an arbitrarily chosen query concept $< X_q, Y_q >$ is correlated with the number of times that $X_q$ and $Y_q$ occur in $S$. The effect is apparent regardless of whether or not $< X_q, Y_q > \in S$.

Fan effects have subsequently been found to present with varying strength across different contexts. This tendency is dubbed the ***differential fan effect***. Differential fan effects have been investigated across object type and concept presentation modality. It was first identified by Radvansky and Zacks (1991), in which the fan effect was found to occur in instances where presented concepts have the same object associated with multiple places (that is to say, the object feature had a high fan value) but not when multiple persons were associated with a single place (i.e. the place feature had a high fan value). Radvansky et al. (1993) later extended this to different object types, specifically small locations and inanimate objects. Stopher and Kirsner (1981) found that fan effects do not seem to present when concepts are presented via images rather than text, suggesting that differential fan effect context is affected by modality in addition to content.

There remains some debate on the mechanism of the fan effect in human subjects, particularly in regard to explaining differential fan effects. Radvansky et al. (1993) proposed a mechanism, based on the concept of mental models, by which subjects create and maintain models of the world based on learned facts and that some types of overlap in presented concepts necessitate the creation of more models than less overlapping concept sets of the same size. Anderson and Reder (1999) proposes a different mechanism, derived from a cognitive architecture in which fan effects are mediated by changing weights of edges in the concept network. This mechanism was further supported experimentally in Sohn et al. (2004) but challenged for larger datasets in Radvansky (1999).

Fan effects are found by Silber and Fisher (1989) in probabilistic categories created by COBWEB to be a special case of another phenomenon known as the ***typicality effect***. This would seem to suggest that fan effects may arise as a consequence of categorization, with a potential explanation being that items closer to the categorical center are more likely to collide with other items, leading to recall uncertainty, while items further from the center are less likely to experience aliasing.

***Typicality***, first formalized and identified in humans by Rosch (1975), refers to a tendency of humans to perform categorization tasks quicker when prompted with a more typical member of a category than with a less typical member of a category, with level of typicality determined by how common the features of an instance of a category are among all members of the same category and among contrasting categories. That is, both an item's intra-category similarity and its inter-category similarity affect typicality assessments. For example, given pictures of two birds, a robin and a penguin, human subject response time will be higher when answering whether the penguin is a bird than whether the robin is a bird.

### 2.1 Prior Work

In Table 1, the results of a comprehensive survey of current work in LLM cognitive behavior studies is provided. No works could be found that study language model fan effects. Though Tung (2024) studied memory interference behavior in LLMs and use fan values in their analysis, they do not explicitly consider the fan effect or its presence.

On the other hand, work has been done that establishes the presence of typicality effects in LLMs (Misra et al., 2021; Bhatia and Richie, 2022; Roberts et al., 2024b) as well as vision models (Upadhyay et al., 2022). Bhatia and Richie (2022) found that BERT shows evidence of typicality effects, including consistency with typicality violations common to humans. Misra et al. (2021) recreated a subset of the experiments conducted by Rosch (1975) which were used to identify typicality effects in humans, identifying typicality effects across numerous categories and models. Roberts et al. (2024b) replicated Misra et al. (2021) with PopulationLM, establishing that the effect was not eroded when studied in a population.

Roberts et al. (2024b) found that the population standard deviations tended to positively correlate with typicality in encoder-only models, though not in decoder-only models. This suggests that the uncertainty captured by LLM variance may not be analogous to human uncertainty since LLMs are overwhelmingly based on the decoder-only architecture (Roberts, 2024).

## 3 In-Pretraining (Typicality) Fan Effect

Anderson originally observed the fan effect in the response times of humans when **correctly responding** to questions. However, in Silber and Fisher (1989), the authors observed human-like fan effects in a COBWEB model and found they were consistent with a special case of typicality. Based on this observation and extant work regarding the presence of typicality effects in LLMs, we hypothesize that LLMs may exhibit a fan effect induced by the relative typicality of categorical items acquired from pretraining. Specifically we formulate RQ3.1.

**Research Question 3.1.** *Given a partial list of items drawn from a category and presented to an LLM, are absence/presence prediction probabilities modulated by item typicality such that probabilities conditioned on typical items tend to be lower than those conditioned on less typical items?*

Expanding on this, based on results from (Roberts et al., 2024b), more typical items tend to have increased predicted word probability even when counterfactual prompting is used, most likely due to base rate probability effects (Moore et al., 2024). However, if a fan effect is present, the probability should tend to decrease with increasing typicality.

It is important to note that LLM probabilities are not necessarily analogous to human response times. However, existing work (Misra et al., 2021; Roberts et al., 2024b) has shown that typicality judgments, which have been measured via response time in humans (Rosch, 1975), are correlated with LLM probabilities.

### 3.1 Methodology

**Models:** All experimental trials are conducted among a systematically perturbed population formed from each base model using PopulationLM (Roberts et al., 2024b) to decrease the likelihood that obtained results are anomalous. The median value is the preferred aggregation when random sampling for the purpose of estimating a true value (Doerr and Sutton, 2019). Therefore, the median

**In-pretraining Fan Effect Prompt**
Following is a list that contains a number of birds. After the list, a bird will be judged as either present or absent in the list. If the list contains the bird, answer with present. If the list does not contain the bird, answer with absent. The list of birds is: toucan, magpie, swan, flamingo, duck, goose, blackbird, pelican, woodpecker, condor, canary, ostrich, redbird, catbird, lark, parakeet, hummingbird, bluejay, bluebird, sparrow, crow, vulture, cardinal, turkey, chicken, goldfinch, wren. According to the list, magpie is present. According to the list, kingfisher is absent. According to the list, robin is____

**LLM**
$P(present)$ and $P(absent)$

Figure 1: Prompt to measure presence/absence belief.

across each base model population is taken as the group prediction.

We choose RoBERTa (Liu et al., 2019), GPT-2 (Radford et al., 2019), Llama-2 (Touvron et al., 2023), Llama-3 (Meta, 2024), Mistral (Jiang et al., 2023), and SOLAR (Kim et al., 2023) as the base models for the experiments. RoBERTa and GPT-2 are chosen as representatives of models previously studied and found to exhibit typicality effects (Roberts et al., 2024b). However, past work has found that higher order human-like behaviors may not be exhibited in smaller models (Roberts et al., 2024a). We therefore include large open source LLMs (Llama-2, Llama-3, Mistral, and SOLAR) that may be more likely to exhibit more nuanced recall effects.

**Data Presentation:** Based on work by Rosch (1975) regarding human typicality judgments across items in ten categories, we construct lists for each of the ten categories in Figure 3 by randomly selecting half of the items in a category. Selected items are included precisely once in a comma separated list with instructional content and two in-context examples. The in-context examples are not randomly sampled and are instead consistent across all experiments.

For each item (N≈60) in each category and every model population member (N=50) we obtain a probability of absence and a probability of presence via counterfactual prompting (Moore et al., 2024). The probability is measured by obtaining the probability assigned to the *canary* words "present" and "absent" given each constructed prompt. We repeat each experiment for each base model for each category 10 times without reuse of populations or item lists. An example interaction for the category *bird* and the item *robin* is shown in Figure 1.

**Human Comparison:** The values for human typicality ratings are taken from Rosch (1975) and compared to the generated model probabilities to understand how typicality, as understood from human studies, impacts model behavior when performing recall.

**Other Hardware and Software:** All experiments used an A100 GPU Google Colab environment. Token likelihoods were obtained using a fork of the minicons Python library (Misra, 2022).

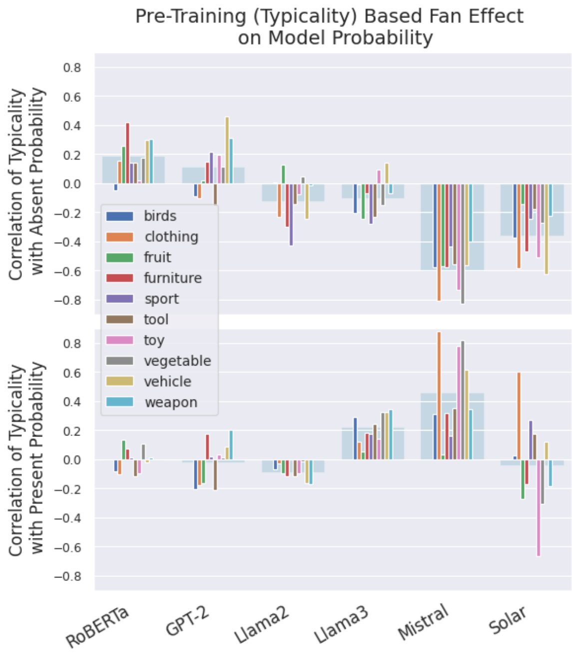


Figure 2: **Top row:** Mistral and SOLAR show significant negative Pearson correlations consistent with fan effects across a range of categories. **Bottom row:** Items present in the context do not elicit a human-like fan effect.

### 3.2 Results

As noted, the fan effect was only observed by Anderson in humans when responding correctly to questions. Thus, only the true absence group (TAG)

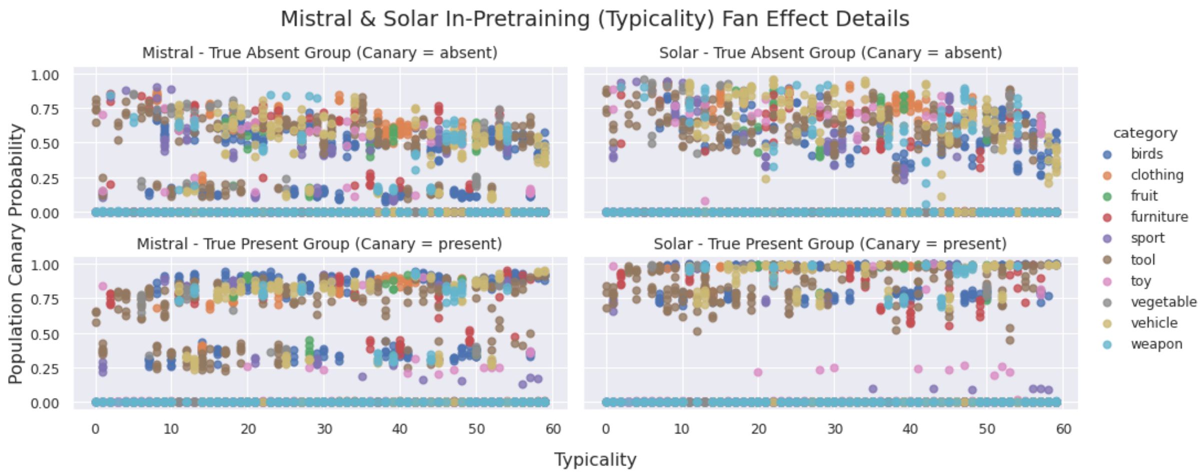


Figure 3: **Left col:** Predictions are made using Mistral. **Right col:** Predictions are made using SOLAR. **Bottom row:** queried item is present (w/o uncertainty). **Top row:** queried item is absent (with uncertainty). Fan effects are evident in the negative Pearson correlation (shown in Figure 2) in the natural group above the noise floor.

and true presence group (TPG) should be considered candidate scenarios that may exhibit a human-like fan effect.

In the upper left plot in Figure 3, there is an obviously distinct group which resides above the threshold (0.35), which we refer to as the probability noise floor. We interpret the group above the noise floor to be the TAG, that is the subset of absent items which the model regards as absent. The TPG, the subset of present items which the model regards as present, can be analogously seen in the bottom left with a noise floor at (0.5). Among predictions in the TAG, the probabilities have an obvious negative correlation with typicality, showing that more typical items tend to induce lower "absent" probabilities. We find that SOLAR (Kim et al., 2023) shows a similar fan effect, with TAG and TPG noise floor at (0.2).

The noise floor observed in both SOLAR and Mistral is an empirical observation which warrants additional consideration. From our investigation, the fan effect in LLMs is modulated by the probability magnitude. Therefore, low probability outputs induce noise in the observation of the fan effect in the model probabilities which are shown for completeness in Figure 3 but filtered in the correlation analysis shown in Figure 2.

Interestingly, in the lower left of Figure 3 the TPG for Mistral has positive correlations which are inconsistent with the fan effect. This is reflected in the bottom of Figure 2 as well. SOLAR, on the other hand, tends toward inter-category randomness in the bottom of Figure 2.

### 3.3 Discussion

In response to RQ3.1, we find in Figure 3 that items absent from the list elicit a human-consistent fan effect evident in the canary probabilities in Mistral (Jiang et al., 2023) and SOLAR (Kim et al., 2023). The probabilities show a significant (r>0.3) (Hinkle et al., 2003) correlation with intra-category typicality in Figure 2 consistent with the fan effects discovered in COBWEB and theorized in humans. This result shows that LLMs exhibit fan effects based on the effects of typicality present in the pretraining data.

RoBERTa (Liu et al., 2019), GPT-2 (Radford et al., 2019), Llama-2 (Touvron et al., 2023), and Llama-3 (Meta, 2024) were equivalently evaluated but showed no significant correlation, though Llama-3 does show a similar, slight effect. We additionally conducted the correlation investigation presented using the population variance in place of the token probabilities and found no significant correlations. This reinforces the possibility put forth in Roberts et al. (2024b) that decoder-only LLM variance may not capture human-like uncertainty given fan effects are understood as an expression of human uncertainty.

**Interpretation:** We were surprised to find the fan effect exhibited in the TAG but not the TPG. However, in retrospect this could have been anticipated based on nuanced consideration of the experiment.

The fan effect is canonically explained as a modulation of human uncertainty based on the categorical distance from an exemplar. When evaluating the TPG, the model is able to judge with near certainty by retrieving the queried item. On the other hand when judging the absence of a TAG item, the model can only know that the item has not been retrieved. The model assigns the probability of absence although it may actually be that the item is present but overlooked, inducing uncertainty. We hypothesize this uncertainty is precisely what the fan effect is modulating. So, when queried about an absent atypical item, the model responds confidently as if implying, "I definitely didn't see *that*".

The above scenario in which the fan effect is only observed in the absent case seems plausibly consistent with human cognitive behavior. Imagine a context in which a human has a deck of cards and is asked if a card is present. If the card is found, then the person will have no uncertainty about their response. On the other hand, if the card is not found, the certainty of the response would be expected to be modulated by the fan effect. That is, if an unusual or outlier card is being searched for then it is likely that the person would notice if it had been present. However, it is reasonable that a human could more easily overlook a common card.

We hypothesize that the uncertainty mitigation due to access to the queried items in the TPG leads to the disruption of the fan effect in Mistral and SOLAR. Our results leave unclear the nature of the fan effect under mitigated uncertainty in the TPG.

### 3.4 Next Steps

Future work should consider creating long context lists that prevent models from retrieving TPG items with high fidelity to attempt to induce uncertainty and fan effects in the TPG. This was not possible currently since no extant lists of intra-category typical items in humans are sufficiently long. However, it may be possible to use LLMs to augment the typicality datasets to create a sufficiently large list.

Results from Mistral suggest that fan effects without uncertainty tend toward a typicality effect response with increasing probability as typicality increases. However, results from SOLAR suggest that they tend toward noise. Future work should additionally attempt to disambiguate the nature of the fan effect when uncertainty is mitigated.

Future work should investigate human behavior in a scenario similar to the described card experiment to understand human fan effect behavior under mitigated uncertainty.

## 4 In-Context Fan Effect

We investigate the presence of fan effects as originally defined in Anderson (1974) in the context of concepts composed of categorical features. This addresses the question of whether fan effects show up in concepts defined and fan values induced exclusively in-context. We formulate this as RQ4.1. We augment our analysis to investigate the presence of differential fan effect as described in Radvansky and Zacks (1991), providing RQ4.2.

**Research Question 4.1.** *Given a list of simple concepts defined by their composite features that is presented to an LLM, are absence/presence prediction probabilities modulated by feature fan values such that probabilities conditioned on high fan features tend to be lower than probabilities conditioned on low fan features?*

**Research Question 4.2.** *Given a list of simple concepts defined by their composite features that is presented to an LLM, is correlation of absence/presence prediction probability with fan value modulated by the fan values of one feature more strongly than another feature?*

### 4.1 Methodology

We closely recreate the experimental methodology of Anderson (1974), with methods similar to those described in section 3.1 for in-pretraining fan effects.

**Models:** Based on the results regarding in-pretraining fan effects, we conduct in-context fan effect experiments with populations formed from Mistral and SOLAR using PopulationLM. The experiment uses a generated model population of size $N = 50$ with median aggregation across population to determine group prediction. As before, probabilities are obtained using the *canary* words "present" and "absent".

**Data Presentation:** Concepts are defined as natural language facts that pair persons, in the form of occupation labels, with places. Each fact is presented as a sentence of the form "The <occupation> is in the <place>". Features are sampled from predefined person and place lists, each of size 20. The fan value is defined as the number of concepts that contain a given feature value. For example, if three distinct concepts indicate a person is present in the place "School", the fan value of "School" is 3. Concept lists are randomly generated to control for ordering effects and feature combination base rates due to semantically connected features (e.g. <Priest, Church>).

| No. of Concepts per Place | No. of Concepts per Person: 1 | 2 | 3 |
|---|---|---|---|
| 1 | aA<br>bB<br>cC | dD<br>eE<br>fF | gG<br>hH<br>iI |
| 2 | jJ<br>kK<br>lL | eK<br>rR | gJ<br>hR<br>iL |
| 3 | mM<br>nN<br>oO | dM<br>rN<br>fO | gM<br>hN<br>iO |

Table 2: Feature assignment pattern used in Anderson and Reder (1999) and replicated in the in-context fan effect experiment.

The concepts in the recreation of Anderson are generated exactly as in Anderson (1974). A predefined set of feature combinations are used, as summarized in Table 2, which are designated by lowercase letters for persons and uppercase letters for places. The person and place assigned to each letter is randomly selected without replacement at the beginning of each trial. The result is N=26 concepts presented to the model in each trial, with a total of 16 fan value combinations (including fan = 0 for features not present in the set).

Prompts presented to the model follow prompt design similar to that in section 3.1. The prompt is composed of four sections: An instructional preamble, the concept list, a two-shot ICL example, and the test query. The ICL examples include a concept that is appended to the end of the concept list that is guaranteed to not be generated. This guaranteed concept is followed by two example queries and simulated outputs, one where the concept is the guaranteed present concept and one with a guaranteed absent concept.

An example prompt in which the concept <Doctor, Park> is shown in Figure 4. Note that <Mechanic, Mall> is included in all trials and has a guaranteed fan value of 1 for both features, while <Airport, Pilot> is absent in all trials.

**Human Comparison:** The data pairings generated are based on the data presented to humans in Anderson and Reder (1999) which were shown to illicit the fan effect in human recall.

**In-Context Fan Effect Prompt**
Following is a list that contains a number of people and the places in which they are located. After the list, a person will be judged as either present or absent in a specified place. When asked about person A in place B, if the list says that person A is in place B, answer with present. If the list does not say that person A is in place B, answer with absent. The list of people and places is: The Nurse is in the Studio. The Police Officer is in the Bank. . . . The Mechanic is in the Mall. According to the list, in the Mall, the Mechanic is present. According to the list, in the Airport, the Pilot is absent. According to the list, in the Park, the Doctor is____

**LLM**
$P(present)$ and $P(absent)$

Figure 4: Prompt to measure presence/absence belief.

**Other Hardware and Software:** All experiments are conducted on an A100 GPU Google Colab environment. Token likelihoods were again obtained with a modified version of the minicons library (Misra, 2022).

### 4.2 Results

The results for both models are shown in Figure 5. As was the case in the in-pretraining experiments, a probability noise floor was noted in the data for both canary completions (Mistral-absent: 0.3; Mistral-present: 0.4; SOLAR-absent: 0.45; SOLAR-present: 0.4), providing a TAG and TPG. The figures are truncated to show only the TPG and TAG datapoints. Correlation statistics of the results are shown in Figure 6, with solid columns indicating correlations with a $p \leq 0.01$.

In Mistral, we once again see an obvious negative correlation between canary probability and fan value in the TAG predictions. This is consistent with a fan effect when evaluating absence of a concept (RQ 4.1). In the TAG, we see a stronger correlation with the fan value of the person feature than with the fan value of the place feature,

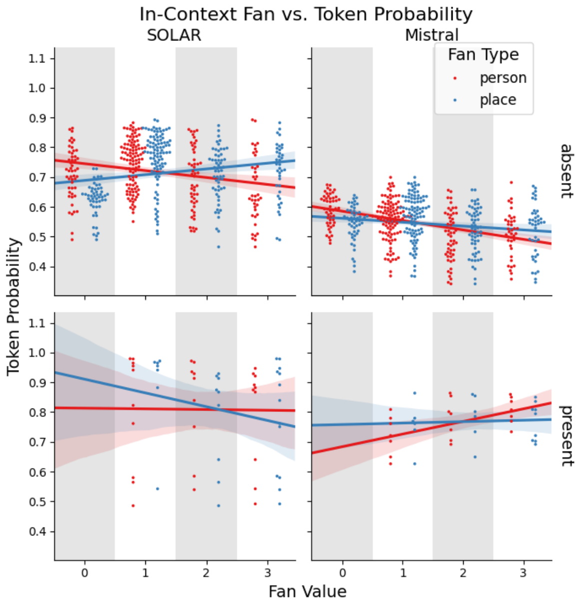


Figure 5: Results of the Anderson recreation experiments on SOLAR and Mistral **Top row:** queried item is absent with the model predicting true absence (with uncertainty). **Bottom row:** queried item is present with the model predicting true presence (w/o uncertainty). Lines of best fit are included. Pearson correlations shown in Figure 6.

supporting a positive result for RQ 4.2. This is consistent with results regarding differential fan effects in Radvansky and Zacks (1991), which found that the fan effect is mediated more by the fan of a particular object than the fan of a particular location.

SOLAR shows a slightly different story. For the TAG predictions, we still see a significant negative correlation when correlating with the fan of person, but a positive correlation with fan of place. TPG predictions instead show a negative correlation against fan of place and no correlation against fan of person. While this seems inconsistent with our Mistral results, it is consistent with our prior interpretations when properly analyzed. Based on these results, SOLAR and Mistral both show evidence of the fan effect in, at minimum, the same situations as in humans, which is to say uncertain contexts and based on the fan of person.

From the in-pretraining experiment, we expect that mitigated uncertainty in the TPG may lead to disruption of the fan effect. In confirmation, among TPG items all correlations fail to achieve a significant p value for fan value and canary probability Pearson correlation, again suggesting that mitigated uncertainty disrupts the fan effect.

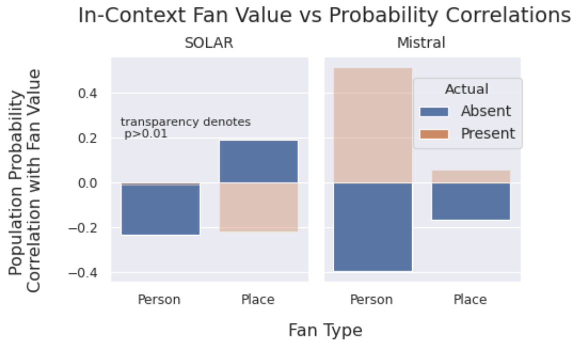


Figure 6: Negative correlations when the queried item is absent suggests items are recalled with higher certainty when the item has fewer in-context appearances (low fan value). Fan values derived from the queried person show fan effects while place fan values cause a distruption of the fan effect. No "present" item queries have significant p values though all "absent" item queries do.

## 4.3 Next Steps

There are numerous enhancements that could be applied to these experiments. While occupations were chosen as proxies for persons to be consistent with Anderson (1974), more unique identifiers like names may yield a stronger differential fan effect if the mental models mechanism proposed by (Radvansky and Zacks, 1991) is present in language models. This should be tested empirically in future work to investigate the nature of differential fan effects. Additionally, other feature types that are not related to persons and places should be investigated.

Human cognitive experiments often include a dimension of elapsed time between training and testing time when studying memory-sensitive behaviors. Future work should consider simulating this time separation in language models. Though language models do not possess a directly analogous temporal dimension, experiments could evaluate the injection of semantic noise of varying length as a potential proxy. In fact such an experiment may suggest that time, to humans, is itself a form of semantic noise.

# 5 Conclusions

Our experiments are the first to evaluate LLMs for the presence of human-like fan effects. We have shown that Mistral and SOLAR have learned to exhibit fan effects from training on human language data. This paper is not the first to identify SOLAR and Mistral as important human-like LLMs. Roberts et al. (2024a) found SOLAR and Mistral to be significantly more human-like than a large body of other open-source models when evaluated in a game theoretic context. Given Mistral was built from Llama-2 and SOLAR was built from Mistral, the authors propose the more human-like behavior may be the result of an improved representation acquired through additional training of Mistral with sliding window attention.

Our results show that fan effects are present both when the fan value is induced in-pretraining in the form of intra-category typicality and when the fan value is induced in-context in the form of repeated items within a list. The presence of typicality-based fan effects in language models lends further credence to the findings of Silber and Fisher (1989) suggesting that fan effects are a special case of typicality effects.

Additionally, we find that when uncertainty is mitigated, the fan effect is disrupted with divergent disruption patterns across LLMs. The divergent patterns across Mistral and SOLAR beg further investigation. However, we are unaware of any cognitive science literature that addresses fan effects in a disruptive scenario with mitigated uncertainty. Therefore, it is unclear how a human may behave in a similar context. We therefore call for human experiments.

Similarly, when the fan value is derived from place instead of person in the Anderson experiment, both Mistral and SOLAR exhibit a disruption of the fan effect in agreement with nuanced work regarding differential fan effects (Radvansky and Zacks, 1991). Again, each of these models diverges in the nature of the disruption but shows a consistent pattern of fan effects in the case of true absence when the fan value is calculated on the person feature.

Finally, we hope this paper will prove synergistic with the wider cognitive science and computational linguistic communities. By adapting experiments to evaluate the presence of known human cognitive effects in LLMs, we may gain new insight into cognitive effects. These insights not only help to explain the factors which influence the behavior of complex language models but also provide new potential hypotheses regarding the cognitive behavior of humans.

## 6 Practical Implications

Human-like uncertainty is shown to be present in Mistral and SOLAR in the form of a fan effect both when the fan value is induced in the pretraining of the model and in the context. However, just as found in Roberts et al. (2024b), the common measures of model uncertainty, variance and standard deviation, may not tend to correlate well with human uncertainty as quantified by the fan effect. This suggests that more work needs to be done to develop a human-consistent measure of LLM uncertainty.

Additionally, the fan effect should be considered when engaging LLMs in applications that require recall. The results here suggest that LLMs may have more trouble correctly evaluating the presence or absence of (1) items when the item is frequently present in the pretraining data and (2) coincident items when the base item is frequently present in the context of the model.


## Acknowledgments

We appreciate the helpful recommendations of reviewers that led to improved presentation clarity and the insightful evaluation of the area chair.


## Limitations

While this paper demonstrates that LLMs exhibit fan effects. It may be the case that the observed effects tend to be weak in comparison to the probability magnitude. So, it is unclear if LLMs fail in recall scenarios in manners consistent with fan effects.

## References


John R Anderson and Lynne M Reder. 1999. The fan effect: New results and new theories. *Journal of Experimental Psychology: General*, 128(2):186.

John Robert Anderson. 1974. Retrieval of propositional information from long-term memory. *Cognitive psychology*, 6(4):451–474.

Sudeep Bhatia and Russell Richie. 2022. Transformer networks of human conceptual knowledge. *Psychological Review*.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Julian Coda-Forno, Marcel Binz, Jane X Wang, and Eric Schulz. 2024. Cogbench: a large language model walks into a psychology lab. In *Forty-first International Conference on Machine Learning*.

Julian Coda-Forno, Kristin Witte, Akshay K Jagadish, Marcel Binz, Zeynep Akata, and Eric Schulz. 2023. Inducing anxiety in large language models increases exploration and bias. *arXiv preprint arXiv:2304.11111*.

Benjamin Doerr and Andrew M Sutton. 2019. When resampling to cope with noise, use median, not mean. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 242–248.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 3(10):833–838.

Dennis E Hinkle, William Wiersma, Stephen G Jurs, et al. 2003. *Applied statistics for the behavioral sciences*, volume 663. Houghton Mifflin Boston.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Erik Jones and Jacob Steinhardt. 2022. Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems*, 35:11785–11799.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, et al. 2023. Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. *arXiv preprint arXiv:2312.15166*.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2023. Language models show human-like content effects on reasoning tasks. *arXiv preprint arXiv:2207.07051*.

Sotiris Lamprinidis. 2023. Llm cognitive judgements differ from human. *arXiv preprint arXiv:2307.11787*.

Huao Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Michael Lewis, and Katia Sycara. 2023. Theory of mind for multi-agent collaboration via large language models. *arXiv preprint arXiv:2310.10701*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. Towards a holistic landscape of situated theory of mind in large language models. *arXiv preprint arXiv:2310.19619*.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

James Michaelov, Catherine Arnett, Tyler Chang, and Ben Bergen. 2023. Structural priming demonstrates abstract grammatical representations in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3703–3720, Singapore. Association for Computational Linguistics.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Kanishka Misra, Allyson Ettinger, and Julia Taylor Rayz. 2021. Do language models learn typicality judgments from text? *arXiv preprint arXiv:2105.02987*.

Kyle Moore, Jesse Roberts, Thao Pham, Oseremhen Ewaleifoh, and Doug Fisher. 2024. The base-rate effect on llm benchmark performance: Disambiguating test-taking strategies from benchmark performance. *arXiv preprint arXiv:2406.11634*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

GA Radvansky and RT Zacks. 1991. Mental models and fact retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17:940–953.

Gabriel A Radvansky. 1999. The fan effect: a tale of two theories.

Gabriel A Radvansky, Daniel H Spieler, and Rose T Zacks. 1993. Mental model organization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(1):95.

Jesse Roberts. 2024. How powerful are decoder-only transformer neural models? In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Jesse Roberts, Kyle Moore, and Doug Fisher. 2024a. Do large language models learn human-like strategic preferences? *arXiv preprint arXiv:2404.08710*.

Jesse Roberts, Kyle Moore, Drew Wilenzick, and Douglas Fisher. 2024b. Using artificial populations to study psychological phenomena in neural models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18906–18914.

Eleanor Rosch. 1975. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*.

J Silber and DH Fisher. 1989. A model of natural category structure and its behavioral implications. In *Proceedings of the eleventh annual conference of the Cognitive Science Society*, pages 884–891. Erlbaum Hillsdale, NJ.

Arabella Sinclair, Jaap Jumelet, Willem Zuidema, and Raquel Fernández. 2022. Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

Myeong-Ho Sohn, John R Anderson, Lynne M Reder, and Adam Goode. 2004. Differential fan effect and attentional focus. *Psychonomic Bulletin & Review*, 11:729–734.

Kerry Stopher and Kim Kirsner. 1981. Long-term memory for pictures and sentences. *Memory & Cognition*, 9:34–40.

Gaurav Suri, Lily R Slater, Ali Ziaee, and Morgan Nguyen. 2023. Do large language models show decision heuristics similar to humans? a case study using gpt-3.5. *arXiv preprint arXiv:2305.04400*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Sean Trott, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen. 2023. Do large language models know what humans know? *Cognitive Science*, 47(7):e13309.

Tzu-Yun Tung. 2024. *Prediction and Memory Retrieval in Dependency Resolution*. Ph.D. thesis.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Neha Upadhyay, Kritika Mittal, and Sashank Varma. 2022. Typicality gradients in computer vision models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Nicolas Yax, Hernan Anlló, and Stefano Palminteri. 2024. Studying and improving reasoning in humans and machines. *Communications Psychology*, 2(1):51.