In-Context Impersonation Reveals Large Language Models' Strengths and Biases

Leonard Salewski [1,2] **Stephan Alaniz** [1,2] **Isabel Rio-Torto** [3,4]*

Eric Schulz [2,5] **Zeynep Akata** [1,2]

[1] University of Tübingen [2] Tübingen AI Center [3] University of Porto
[4] INESC TEC [5] Max Planck Institute for Biological Cybernetics

Abstract

In everyday conversations, humans can take on different roles and adapt their vocabulary to their chosen roles. We explore whether LLMs can take on, that is impersonate, different roles when they generate text in-context. We ask LLMs to assume different personas before solving vision and language tasks. We do this by prefixing the prompt with a persona that is associated either with a social identity or domain expertise. In a multi-armed bandit task, we find that LLMs pretending to be children of different ages recover human-like developmental stages of exploration. In a language-based reasoning task, we find that LLMs impersonating domain experts perform better than LLMs impersonating non-domain experts. Finally, we test whether LLMs' impersonations are complementary to visual information when describing different categories. We find that impersonation can improve performance: an LLM prompted to be a bird expert describes birds better than one prompted to be a car expert. However, impersonation can also uncover LLMs' biases: an LLM prompted to be a man describes cars better than one prompted to be a woman. These findings demonstrate that LLMs are capable of taking on diverse roles and that this in-context impersonation can be used to uncover their strengths and hidden biases. Our code is available at `https://github.com/ExplainableML/in-context-impersonation`.

1 Introduction

Large Language Models (LLMs) can not only summarize documents and converse on a large range of topics [1], but they have also shown other emergent abilities [2, 3]. Because of their impressive abilities, LLMs are permeating into many applications [4, 5]. This means that there is a societal need to understand how these models "tick" [6, 7]. Traditionally, LLMs are provided with a context as a textual prompt and are asked to provide answers via text completion, thereby solving a variety of choice-based [8], description-based [9], and reasoning tasks [10]. Yet how in-context learning works is not fully understood. When Min et al. [11] prompted LLMs with random labels, they found that this did not drastically degrade performance, suggesting that the role of in-context demonstrations is to prime the model for a particular task. This is in line with other results suggesting that LLMs internally infer latent variables to make better predictions [12]. It has been suggested that LLMs, and other large models, can change their behavior when asked to respond as a particular persona. When Deshpande et al. [13] asked LLMs to respond as a hateful person, their toxicity score increased. When Wang and colleagues [14] asked LLMs to imagine being expert systematic reviewers, the

*Work done during a research visit at the University of Tübingen

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

quality of their literature search queries increased. That LLMs can impersonate specific people is also known; they can, for example, pretend to be Oscar Wilde, Carrie Bradshaw from Sex and the City, or Donald Trump [15]. But how does in-context impersonation affect LLMs' behavior in language-based and other downstream tasks?

In the current work, we let LLMs impersonate, that is taking on different roles, in context. We do this by prefixing the prompt with *"If you were a {persona}"* where *persona* is replaced with the persona that the LLM is asked to impersonate. These personas are associated either with a social identity or a domain of expertise. In a first simulation using a multi-armed bandit task [16], we find that LLMs impersonating children of different ages can recover the developmental stages of human-like exploration strategies. In language-based reasoning tasks, we find that LLMs impersonating domain experts perform better than LLMs impersonating non-domain experts. Finally, we ask LLMs to describe different classes of either birds or cars and then use their descriptions in a downstream, visual classification task. The results of this experiment corroborate our earlier results: LLMs become better as they pretend to be older, and they are also better when they pretend to be domain experts. However, we also see how impersonating LLMs reproduce biases: LLMs impersonating a black person or a male describe cars better, while LLMs impersonating a white person or a female describe birds better. These results expand our understanding of in-context learning in LLMs and open up new research directions investigating role-taking and pretense in LLMs and beyond.

2 Related Work

In-context learning refers to an LLM's ability to improve at a given task after being provided with a number of task-relevant demonstrations [1]. This ability sets LLMs apart from traditional models and has led to a totally new paradigm – one which does not require fine-tuning of weights on task-specific data but instead relies entirely on contextual information [17, 10, 18].

This contextual information is normally delivered as textual prompts [19], where a task or scenario is described and a model is asked to solve the task or reason about the scenario by generating the next words of the provided text. Due to its flexibility, prompting has been widely used as a generic method for natural language tasks [20, 21]. Importantly, the resulting in-context learning does not only work after LLMs have seen some examples, i.e. in the few-shot regime [22], but also without any examples, i.e. in the zero-shot regime [23]. LLMs are reasonably proficient at solving arithmetic [24] or reasoning tasks [25] without having been prompted with example solutions but only after being asked to provide an answer to a given problem. LLMs can require careful engineering of the provided prompts, either manually [26] or automatically [27]. Indeed, whole books have been written to provide guidelines on how to best perform prompt engineering [28], especially because engineering prompts can require a great amount of expertise [29].

One method known to influence LLMs behavior is to ask them to respond as a particular person [30, 31], an effect which is also described as role-taking [32]. LLMs can take in the text of one famous author, e.g. Oscar Wilde, and rewrite it in the style of another famous author, e.g. James Joyce [33]. This is not only true for LLMs but for any large model that provides results based on prompts, such as text-to-image models [34–36]. For example, using the artist's name for generative art prompting is known to boost the quality [29] or to substantially affect the style [37–39] of the generated images. To make LLMs respond more truthfully, Lin and colleagues introduced scenarios from the perspective of a fictional persona called "Professor Smith" [40]. Conversely, to make LLMs act maliciously, Wolf et al. [41] prompt LLMs adversarially to overcome alignment techniques. LLMs can also be used to simulate multiple humans which changes how they cooperate in economic games [42].

LLMs can also have their own "personalities" which can be evoked in-context [43]. Although LLMs frequently behave like the average person [44], their personality profiles can be tinkered with [45], e.g. by changing the context to be more or less emotional [46]. This has led researchers to use LLMs to simulate survey responses [47] of subpopulations by conditioning them on socio-demographic descriptions [48] or to ask them to respond in persona when writing about fictitious childhood events [49]. Additionally, non-deterministic tasks such as open-ended questions have also been explored [50].

Semantics derived automatically from language corpora can contain human-like biases [51]. Thus, LLMs do not only reproduce human-like text but also replicate biases present in the training data [7, 52]. Importantly, these biases can get exacerbated if LLMs are asked to provide answers in persona [46, 13, 53].

LLMs are naturally combined with large vision-language models (VLMs) [54, 55] such as CLIP [56] due to their versatility in a wide range of visual recognition tasks. Menon et al. [57] used GPT-3 [1] to generate a diverse set of short descriptions of a class that improve zero-shot classification when their CLIP scores are combined. Similarly, Yang et al. [58] used GPT-3 descriptions of classes as concept bottlenecks for interpretable image classification. LLMs can also be used as a knowledge base for visual question-answering (VQA) tasks [59].

3 In-context Impersonation Methodology

Our methodology is composed of two steps. First, we prompt and query the LLM. Second, we evaluate the resulting text queries in three tasks, i.e. two-armed bandit, reasoning, and visual classification.

3.1 Prompting and Querying the Large Language Model with Personas

LLMs are trained to predict the most probable next token t_k given previous tokens $t_1 \ldots t_{k-1}$ by maximizing the likelihood function $p_{\text{LLM}}(t_k|t_1,\ldots,t_{k-1})$. In this work, we use pre-trained LLMs without further finetuning them. Depending on the task, we generate one or more tokens given a task-specific context $\boldsymbol{c}$ that describes the task to the language model and prompts it for an answer. The context includes the instruction to impersonate using the phrase *"If you were a {persona}"* where persona p is replaced by the persona name. Thus, we obtain generated tokens $\boldsymbol{t}$ by sampling from

$$p_{\text{LLM}}(\boldsymbol{t}|\boldsymbol{c}^{(p)}) = \prod_{k=1}^{K} p_{\text{LLM}}(t_k|c_1^{(p)},\ldots,c_n^{(p)},t_1,\ldots,t_{k-1}) \tag{1}$$

We refer to this type of contextualization as *in-context impersonation*.

Personas Considered. The first interesting question to look at was if LLMs could impersonate the behavior of differently aged people. For this, we ask the LLM to imagine it is either a 2, 4, 7, 13, or 20-year-old. We also evaluate whether the LLM is able to impersonate different fields of expertise. Depending on the task considered, the expertise profiles differ (more details below). Finally, we evaluate whether LLMs have biases regarding gender and skin color. For this, we asked LLMs to imagine that they were either a man or a woman or a black person or a white person.

Large Language Models Considered. In this work, we evaluate two LLMs. For all of our tasks, we used the Vicuna-13B language model [60] which has 13 billion parameters and was trained to follow natural language instructions. Vicuna is a fine-tuned version of the LLAMA language model [61] using ShareGPT [62] conversational data. We use an instruction fine-tuned model because it was optimized to follow user prompts. Its weights are publicly available, allowing us to run the model locally. Vicuna is competitive with proprietary services such as ChatGPT in some domains [63][2]. In addition to Vicuna, we use the OpenAI API of ChatGPT [64] with the `gpt-3.5-turbo` model for the reasoning and vision tasks. For the bandit task, however, running 12k games with 10 trials each is infeasible.

We do not further train the models, nor do we provide sample solutions in-context; thus, all experiments are conducted in a zero-shot fashion. By providing minimal guidance to perform the task, we avoid pre-conditioning the model such that answers can better reflect the internalized language of the LLM instead of relying on few-shot examples. When sampling full sentences, we use a temperature of 0.7; to obtain the answer as a single symbol (token), we set it to 1 unless otherwise stated. These different temperatures were chosen based on the recommended default values of each LLM.

3.2 Bandit Task Design

We asked LLMs to imagine being in different personalities while participating in a multi-armed bandit task [65] taken from the psychology literature [66] and already applied to LLMs [8].

An agent gets to interact with a two-armed bandit problem for 10 trials. The mean reward for each arm a is drawn from $p(\theta_a) = \mathcal{N}(0, 10)$ at the beginning of a task, and the reward for each trial

[2] `https://chat.lmsys.org/?leaderboard`

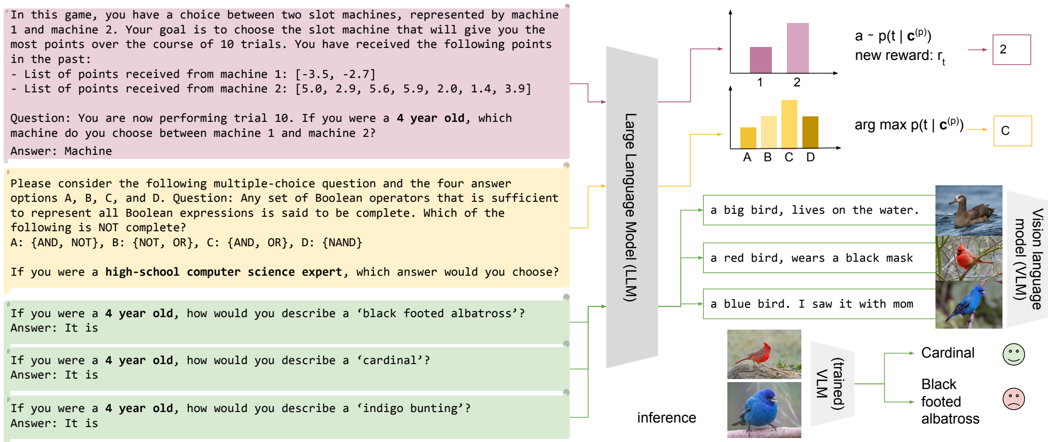

Figure 1: Our three tasks are designed to analyze the effect of *in-context impersonation*. First, we investigate bandit tasks (pink) where the LLM must maximize the reward while impersonating different age groups. Second, we evaluate the effect of domain expert impersonation on natural language reasoning tasks (yellow). Third, we study the usefulness of descriptions generated with impersonation w.r.t. age, expertise, ethnicity, and gender for visual classification (green).

is drawn from $p(r_t|a_t, \theta_{a_t}) = \mathcal{N}(\theta_{a_t}, 1)$. Feedback of past trials is provided via prompt-chaining, i.e. concatenating previous choices and their outcomes to the current prompt submitted to the LLM. We analyze the set of emerging exploration strategies, assuming that an agent uses Bayes' rule to update its beliefs over unobserved parameters. If prior and rewards are normally distributed, then the posterior will be normally distributed and the corresponding updating rule is given by the Kalman filtering equations. Let $p(\theta_a|h_t) = \mathcal{N}(\mu_{a,t}, \sigma_{a,t})$ be the posterior distribution at time-step t. Based on the parameters of this posterior distribution, one can define a probit-regression model:

$$p(A_t = 1|\mathbf{w}) = \mathbf{\Phi}\left(\beta_1 \mathrm{V}_t + \beta_2 \mathrm{RU}_t\right) \tag{2}$$

with $\mathbf{\Phi}$ denoting the cumulative distribution function of a standard normal distribution. Here, $\mathrm{V}_t = \mu_{1,t} - \mu_{2,t}$ represents the estimated difference in value and $\mathrm{RU}_t = \sigma_{1,t} - \sigma_{2,t}$ the relative uncertainty. One can use Equation 2 to analyze how much an agent engages in exploitation behavior by inspecting β_1 and how much the agent uses uncertainty to explore in a directed fashion by inspecting β_2 [16].

For this bandit task, we consider personas of different ages. Specifically, we study ages 2, 4, 7, 13, and 20 to cover key developmental stages of early childhood, childhood, adolescence, and adulthood where the learning progress is most pronounced in humans. The language model is prompted (see Figure 1, the pink path) to only answer "1" or "2" depending on which arm a it would like to choose. The LLM receives rewards and the associated actions from previous trials inside the context in the form of a list.

With $\log d_{a_t} = \log p_{\text{LLM}}(t_1 = a_t|\mathbf{c}^{(p)}, a_1, \ldots, a_{t-1}, r_1, \ldots, r_{t-1})$ being the unnormalized logits from the LLM for the token of arm a, for each trial we sample an action $\hat{a} \sim \sigma(\{\log d_{a_t}\}_{a_t=1}^{A})$ where we have two arms $A = 2$. We do not apply temperature scaling in this case as we are only sampling a single token and want it to reflect the LLM decision-making as faithfully as possible.

3.3 Reasoning Task Design

In our reasoning task, the LLM has to answer a multiple-choice question regarding a given topic from the Multitask Language Understanding (MMLU) dataset [67], commonly used to benchmark LLMs [61]. The MMLU dataset consists of 57 tasks from Science, Technology, Engineering, and Mathematics (STEM), Humanities, Social Sciences, and Other, ranging from elementary, high school, college, and professional levels of complexity. We start by prompting the LLM with the context:

```
Please consider the following multiple-choice question and the four answer
options A, B, C, and D. Question: {task}
If you were a {persona}, which answer would you choose?
```

The *task* is replaced by the question and the 4 possible answers, while the *persona* is replaced by an expert (see Figure 1, the yellow path). We consider three types of experts as personas. The task expert, e.g. for the high school computer science task, is "high school computer science expert". The domain expert is an aggregation of all the remaining experts in the same field as the task expert (but not the task expert himself), e.g. for high school computer science it would be any other STEM expert. The non-domain expert is an aggregation of the task experts from the other domains, e.g. for high school computer science it would be all Humanities, Social Sciences and Other experts.

After feeding the prompt to the LLM, the LLM prediction of the first token following the context is $d = p_{\text{LLM}}(t_1|\boldsymbol{c}^{(p)})$ and the N tokens for the possible answers of the multiple choice question are $o = \{o_i\}_{i=1}^{N}$ which in this case are A, B, C, and D. The predicted option is then given by

$$\hat{o} = \arg\max(\hat{c}_i), \text{ with } \hat{c}_i = d[c_i], i = 1 \dots N \tag{3}$$

which are the predicted probabilities of the language model. With this approach, we are able to obtain the option with the highest probability according to the LLM and, thus, compare it with the ground truth label to measure the accuracy resulting from different in-context impersonations.

3.4 Vision and Language Task Design

Lastly, we want to evaluate the usefulness of descriptions generated by *in-context impersonation* for downstream vision and language tasks. We focus on challenging fine-grained classification tasks, as the generated descriptions need to be domain specific for these tasks to succeed. We ask the LLMs to generate a description of a class, from the perspective of a persona. Our prompt is:

```
If you were a {persona}, how would you answer the following question
       in 45 words?  Q: What is a/an {class_name}?  A: It is
```

To avoid trivial solutions, i.e. the class name being mentioned in the description, we post-process the generated descriptions with a two-step approach: first, we replace class names used in noun phrases with an appropriate pronoun whilst respecting the given numerous. Second, if the class name is still not removed, we re-use the same language model to process the descriptions sentence by sentence. For this, we use 4 in-context examples, that demonstrate how to remove the class name information. The full process is documented in suppl. Section D.1.

Vision-Language Models (VLMs). We use CLIP (or variants thereof) [56, 68] to perform fine-grained visual classification as a means to evaluate the usefulness of the generated descriptions. CLIP models are trained with contrastive image-text matching losses to rank matching image and text inputs highly and non-matching inputs lowly. [56, 68] show that CLIP variants generalize well to match unseen texts, e.g. class names, an ability commonly referred to as zero-shot classification.

First, the image to classify is converted into a normalized feature representation I using CLIP's pre-trained vision backbone. Then, the class names are embedded into normalized feature vectors T_N using the pre-trained text backbone. Next, all pairwise cosine similarities $I \cdot T_N$ of the respective feature representations are computed. Finally, the $n^* = \arg\max_N(I \cdot T_N)$ over these similarities reveals the most similar class n^*.

Inference. We generate a description $D_n^{(p)}$ with the above prompt for each class n for each persona p where we use a generative approach, i.e. we auto-regressively sample a random token from the predicted logits (see Figure 1, the green path). For Vicuna-13B we use the default temperature of 0.7 and the default top-k value of $k = 50$. For ChatGPT we use the default temperature of 1.0. This continues until the model emits an <*end of sequence*> or the maximum number of tokens (96) is reached. We did not tune these values.

For visual classification, we use the zero-shot classification capabilities of CLIP models, but instead of using the embedded class name itself (T_n), we use the embedding of the generated descriptions $D_n^{(p)}$ for each class n and for each persona p. The predicted class for each persona $i^{(p)^*}$ is:

$$n^{(p)^*} = \arg\max(I \cdot D_n^{(p)}) \tag{4}$$

Performance is measured by computing the classification accuracy of the test splits on both datasets. As the descriptions are sampled from the LLM output, the results of the experiments are stochastic and we repeat them five times. We report the mean performance as well as 95% confidence intervals.

4 Experiments

Using Vicuna-13B, we evaluate the two-armed bandit and MMLU language reasoning tasks. For the zero-shot image classification task using a VLM we generate descriptions with both Vicuna-13B and ChatGPT. We focus on highlighting how the chosen persona changes the task performance of the LLM. As LLMs seem to be sensitive to prompts [69], we follow the meta-prompting approach from [26] to vary our impersonation prompts. We run all Vicuna-13B experiments with each of the six prompt variations, which are shown in the suppl. Section A.1. All experiments are performed on the test splits using a single A100-40GB GPU and we mention inference times in suppl. Section A.2.

4.1 Age-based impersonation changes exploration strategies

In the bandit task, for every age group that the LLM impersonates, we perform 2k two-armed bandit games of 10 trials each for each prompt variation. We evaluate the task performance in three ways.

First, we show the average reward per trial the LLM obtained with personas of increasing age in Figure 2 (top). With an increasing number of trials, the LLM obtains a higher average reward, corroborating that Vicuna-13B is able to learn from past trials to improve its policy similarly to GPT-3 in [8]. Moreover, as the LLM takes on a persona of different ages, we observe a divergence of obtained rewards as the number of trials increases. Younger personas, i.e., 2- and 4-year-old personas, obtain a smaller reward than older ones, i.e., 13- and 20-year-old personas.

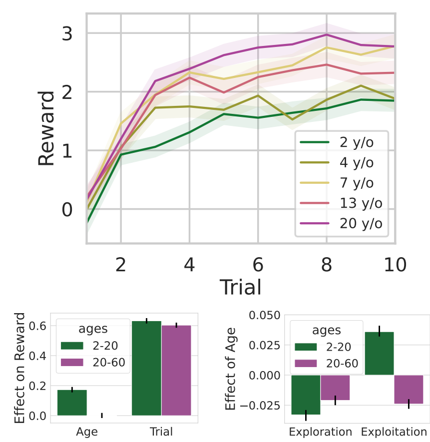

Figure 2: Two-armed bandit task. Top: Average reward per persona (10k games of 10 trials), left: Age and # of trials have a positive effect on the expected reward, right: With age, exploration decreases, and exploitation increases.

Secondly, we analyze the resulting rewards by using a regression, entering the trial number and age as independent variables. To extend the analysis, we evaluate two age groups, from 2 to 20 and from 20 to 60, where we evaluate ages in steps of 2 between 2 and 30 and steps of 5 from 30 to 60. We report these results in Figure 2 (bottom left). We find that the impersonating LLMs generally improved over trials, i.e. they increase their rewards as they progressed over trials of a game ($\beta = 0.63$, $p < .001$ for ages 2–20 and $\beta = 0.60$, $p < .001$ for ages 20–60). Importantly, LLMs impersonating older participants generate higher average rewards until age 20 ($\beta = 0.17$, $p < .001$), thereby replicating a general pattern found in the developmental literature [70]. We find no significant effect from ages 20–60, which also mirrors observations of stagnating mental performance of adults.

Lastly, we analyze how regression weights of the probit-regression were influenced by the age group the LLM is impersonating, again analyzing ages 2–20 and 20–60. Figure 2 (bottom right) reveals that LLMs pretending to be older explored their environment less ($\beta = -0.03$, $p < .001$) and exploited more ($\beta = 0.04$, $p < .001$) in the ages between 2–20. This pattern is in line with several results from the psychological literature which also found that children show higher levels of directed exploration [71] than adults [72]. These results suggest that impersonating LLMs can recover human-like developmental stages of exploration in a two-armed bandit task. If life is seen as an exploration-exploitation problem, then younger agents should show higher amounts of directed exploration [73, 74]. To the best of our knowledge we are the first to show that LLMs replicate similar trends when using in-context impersonation.

4.2 Expertise-based impersonation changes reasoning abilities

Our experiments on expertise-based impersonation (details in Section 3.3) are conducted on the MMLU dataset [67], for which we ask Vicuna-13B to impersonate experts from three different categories (task, domain, and non-domain). For each task we compute the task accuracy averaged over all task questions (95% confidence intervals are computed over the average task accuracy). We compare the task expert results with the average of all domain expert personas, the average of all

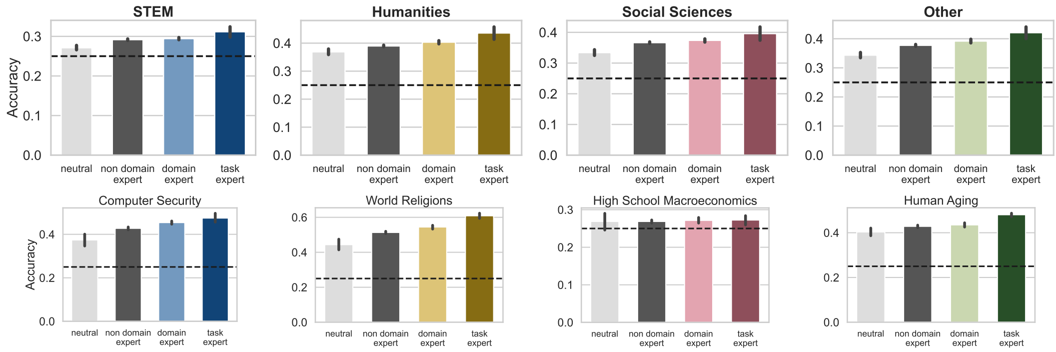

Figure 3: Expertise-based impersonation on all domains of the MMLU reasoning benchmark (top) and on exemplary individual tasks (bottom). For each task, we consider four personas: the neutral, the task expert, the domain experts (all experts from the same domain except the task expert) and the non-domain experts (all experts from all remaining domains). The dashed line is the random baseline.

non-domain expert personas, the average of all neutral personas, and the random baseline (horizontal line). We consider four neutral personas, namely student, average student, person, and average person, and the six aforementioned prompt variations.

In Figure 3 (top row), as expected, when the LLM is asked to impersonate the task expert, the performance is the highest. This shows that the LLM can indeed impersonate task experts with accuracy higher than random. Similarly, the domain expert personas perform better than the non-domain expert personas. This trend holds for all four MMLU domains and thus for MMLU in its entirety. In general, we observe that the performance in the Humanities tasks is higher than the accuracy in the other domain tasks, which is in line with results reported in the literature [61, 75, 76, 67]. Overall, these results suggest that LLMs can increase their performance when asked to impersonate task experts compared to non-task experts.

To provide more details on the individual behaviors of these personas, in the plots on the bottom row of Figure 3, we sample various expert personas, e.g. three positive and one negative case. The first, second and last plots indicate that the task expert persona performs better than the domain expert persona, which, in turn, outperforms the non-domain expert persona. In those cases, all experts outperform the neutral persona. For the High School Macroeconomics task, the task expert persona performs close to random and to the non-domain expert persona. This may be because, as Hendrycks et al. [67] observed, LLMs tend to perform worse on procedural problems that are calculation-heavy compared to purely verbal tasks. Furthermore, when the LLM performs close to or below the random baseline, i.e. the task is more difficult to solve for all types of experts, the impersonation trends are not as clear, since the model does not know how to solve the task well, irrespective of the persona. Thus, while in the Social Sciences field, the High School Macroeconomics task has worse performance, we see that for World Religions, the exam result is higher than 60%, i.e. a passing grade. Especially for World Religions and Human Aging, we observe that the task expert performs much better than the corresponding domain expert personas. We show results for all tasks in Section C.1 of the suppl.

Finally, since several MMLU evaluations [67, 77], can lead to small variations when comparing different models', we include results with the MMLU official prompt in suppl. Section C.2, where we verify that our findings on impersonation are not dependent on the formulation of the task. Lastly, we also show MMLU results for social groups in C.3.

4.3 Impersonation as categorical descriptions is complementary for visual categorization

In this section, we provide experimental results on two state-of-the-art fine-grained visual categorization datasets, i.e. Caltech UCSD Birds (CUB) [78] and Stanford Cars [79], with 200 and 196 classes of birds and cars, respectively. Additional results for FGVC Aircraft [80] and Oxford Flowers [81] can be found in Section D.2 of the supplementary. We first compare how different VLMs make use of the generated descriptions, then compare different LLMs in our in-context impersonation tasks and finally provide some qualitative results.

Comparing VLM variants. We first compare the classification accuracy of different VLMs when the Vicuna-13B generated descriptions of classes are fed to the language encoder of the VLM. For

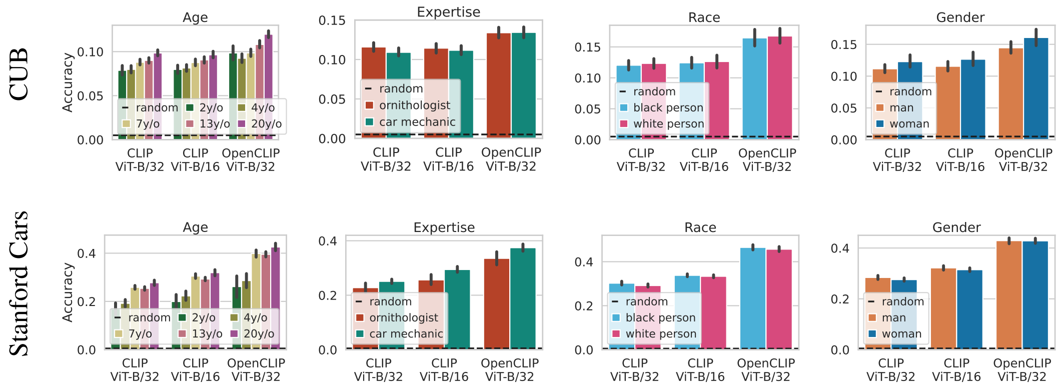

Figure 4: Comparing CLIP-32, CLIP-16 and OpenCLIP as VLMs (the language input comes from Vicuna-13B) on CUB (top) and Stanford Cars (bottom) datasets. We observe the effects of age, expertise, ethnicity and gender independent of the VLM used for fine-grained visual classification. The dashed line represents the random baseline.

the vision encoders we consider the Vision Transformer (ViT) [82] based B/32 and B/16 variants of the official CLIP implementation [56] as well as the OpenCLIP B/32 ViT variant [68]. The latter is a replication of the original CLIP trained on a larger dataset (Laion 5B [83]). For each CLIP variant, we use the corresponding causal transformer text encoders, which might not encode text as well as Vicuna but are able to embed the text into a shared multi-modal space.

Our results in Figure 4 show that across all three CLIP variants increased age in the impersonated persona increases performance for both bird and car classification. Interestingly, there is a significant increase in performance at 7 years of age when recognizing cars. Our expertise evaluation shows that the car mechanic persona's descriptions performs better than ornithologist's when recognizing cars. Interestingly, racial (column 3) and gender (column 4) personas, reveal consistent biases. While the black performs better in car classification, the white performs better in bird classification. This may indicate that there are stereotypical biases in the training data. Similarly, while the woman performs clearly better than man for bird classification, the trend is not as strong for car classification although man performs slightly better than woman. The language encoder of VLMs potentially being weaker than Vicuna, we expect these results to improve overall with a stronger language encoder in the VLM but this is an orthogonal direction to explore. To confirm the significance of our results, we run Chi^2 tests for expertise, race and gender. We consider the three CLIP models, five different seeds and the six different impersonation prompt variations. We find that for all experiments considered, {CUB, Stanford Cars} x {man/woman, black/white, ornithologist/car mechanic}, p<0.001. Thus, we conclude that our results are significant.

We also investigate the effects of composing personas for a computationally feasible subset of persons. More specifically, we study all possible combinations of {Black, White} × {Female, Male} for the CUB dataset for 5 different seeds (Figure 6). With Vicuna-13B we see weak evidence that the biases co-construct: Individually the white persona outperforms the black persona and the same applies to the female persona outperforming the male persona. Combined, the white female persona outperforms both the black female persona (change in race) and the white male persona (change in gender). Furthermore, we also study performance of additional genders (agender and non-binary) and races (indian, asian and hispanic) in the suppl. in Section D.5.

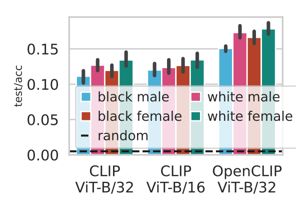

Figure 6: Composition of personas on CUB for Vicuna-13B.

Comparing LLM variants We evaluate how different LLMs, namely Vicuna-13B and ChatGPT, generate descriptions of the classes of interest. In these experiments, we keep the VLM fixed to OpenCLIP, as it is the best of the CLIP variants tested above. For computational reasons, we only evaluate on our original impersonation prompt. Figure 5 shows the effect of LLM impersonation on the generated descriptions evaluated on zero-shot image classification.

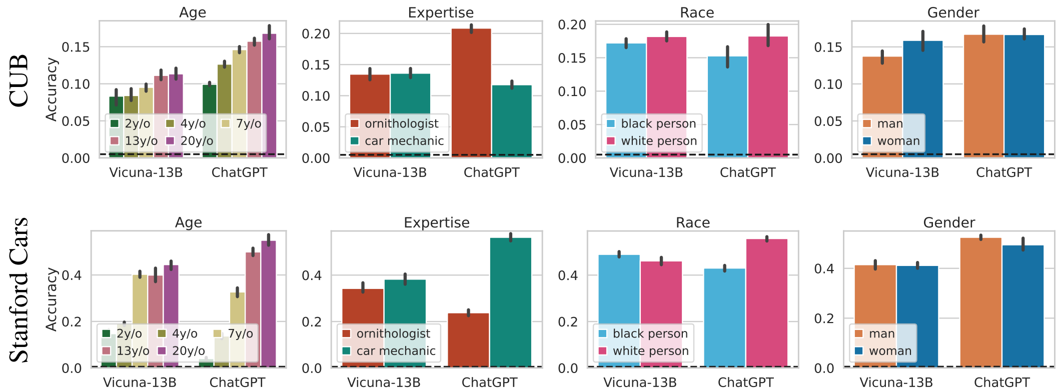

Figure 5: Comparing Vicuna-13B and ChatGPT as LLM variants (OpenCLIP is the VLM) on CUB and Stanford Cars. For both LLMs, the accuracy increases with increasing age, the expert persona on the respective dataset performs better and both LLMs are not free of biases, and impersonation of different genders or race affects their performance. The dashed line represents the random baseline.

For the age personas, we observe a clear trend of increased performance for both LLMs as they impersonate older characters. The progression is particularly pronounced for ChatGPT, where on Stanford Cars the 2-year-old persona describes different cars with similar expressions leading to $\sim 4\%$ accuracy, but as ChatGPT's persona gets older, it becomes more accurate in describing cars, e.g. 54.9% for persona of age 20. This indicates that LLMs can replicate human language at different development stages, varying their language both in terms of vocabulary and general knowledge for accurately describing these objects as discussed in [84]. Similarly to the reasoning task, LLMs exhibit higher expertise on the topic when we ask them to impersonate a bird expert ("ornithologist" persona) and a car expert ("car mechanic" persona). The respective domain expert persona performs approximately twice as well as the non-domain expert persona when using ChatGPT. Impersonating an expert, the LLM tends to describe a class in more detail and mention more discriminative features.

We also observe that impersonation can reveal biases encoded in the LLMs. A race bias becomes apparent when we ask the LLMs to impersonate a "black" or "white" person. ChatGPT tends to describe both birds and cars better when posing as a white person. Vicuna-13B, on the other hand, provides better descriptions of cars as a black person. Gender biases are a bit less noticeable, but we still find Vicuna-13B giving better bird descriptions as a woman persona and ChatGPT identifying cars better as a man persona. While instruction-based fine-tuning [64] tries to remedy social biases encoded in LLMs to some extent, we can still expose them through in-context impersonation.

Overall, we find that ChatGPT shows larger effects, probably due to its access to more diverse (fine-tuning) data. The fact that the effects described above can be found with two very different language models suggests that they are a result of the overall language modeling and instruction following training on internet data instead of specific model artifacts.

Qualitative results and limitations. In Figure 7, we provide the descriptions generated by ChatGPT and Vicuna for one class, i.e. black billed cuckoo, from the CUB dataset and one class, i.e. AM General Hummer SUV 2000, from the Stanford Cars dataset. As personas, we sample all the age personas we considered in our experiments, namely 2, 4, 7, 13 and 20-year-old personas.

For both LLMs, in both datasets, we observe that with increasing age, the complexity of the vocabulary and attributes of the mentioned objects increases. A 2-year-old persona talks about the sound the bird or the car makes, the shapes of the wings or wheels, and the emotions attached to seeing or riding it. A 4-year-old persona interestingly mentions experiences seeing the bird or the car more distinctly. A 7-year-old persona starts using more complicated adjective phrases, e.g. can drive on rough roads and outside places, whereas a 13-year-old persona takes it one step further, e.g. brownish-gray body with distinctive rusty colored markings. Finally, a 20-year-old persona makes a more complete description of the object including where the bird is found or what the car is mainly used for. This is in line with [85] where the authors show that given the same length of text, smaller children use less diverse and non-academic vocabulary, and repeat a lot. Even though LLM's may not faithfully represent the language of children, we qualitatively observe similar patterns. We show more examples and quantize the properties of the generated descriptions in suppl. Section D.3.

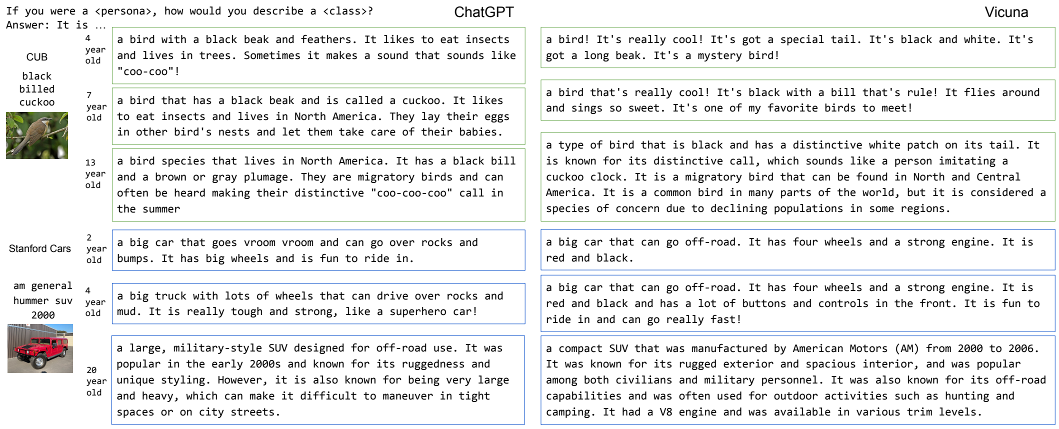

Figure 7: Qualitative results sampling all the age personas (2, 4, 7, 13 and 20-year-old personas) for two classes, i.e. Black Billed Cuckoo (CUB) and AM General Hummer SUV 2000 (Stanford Cars) classes. The results are obtained by querying ChatGPT and Vicuna.

One obvious difference between these two LLMs to point out is that the descriptions obtained from Vicuna appear to be longer and more detailed. Further, at earlier ages, e.g. 2 or 4, especially on CUB, the descriptions of Vicuna seem poetic. The difference between the semantic content of the descriptions of the 13-year-old persona and the 20-year-old persona seems to be less distinct in Vicuna than in ChatGPT. One final interesting observation is that Vicuna descriptions talk about the color of the car whereas the color can not be a distinguishing property of a car.

5 Broader Impact

We believe that a better understanding of in-context impersonation, as well as its resulting downstream effects, can not only help to mitigate the risk of fraud but also to understand how these newly-powerful agents behave more generally [86]. We have already seen that in-context impersonation boosts performance and produces biases; these results could be followed up by investigating how these characteristics emerge during training, change with increasing model size [87], or adapt with additional fine-tuning [88]. Additionally, LLM providers could quantitatively test for these biases before releasing new models. We specifically discourage crafting (system) prompts for maximal performance by exploiting biases, as this may have unexpected side effects, reinforce societal biases and poison training data obtained with such prompts. Other misuses may include amplification of stereotypical biases through generated content and using impersonation to invoke fake trust. However, we believe systematically studying these biases raises awareness in the ML community and general society and serves as a first step to research mitigation strategies. Lastly, we discuss limitations of our work in suppl. Section E.

6 Conclusion

We presented evidence that *in-context impersonation*, that is asking LLMs to take on different roles in context, can change their performance and reveal their biases. Asking LLMs to impersonate differently aged people in a two-armed bandit task, LLMs could reproduce human-like developmental stages of exploration behavior. Asking LLMs to impersonate domain experts, they performed better than LLMs that were asked to impersonate a non-domain expert. Finally, asking LLMs to impersonate various roles in a vision-language task revealed not only that impersonation can boost relative performance but also recovered societal biases about a person's age, gender, and race.

We have demonstrated the effects of in-context impersonation on single agents performing relatively simple tasks across a limited range of personas. In future work, we want to scale up this approach to multiple LLMs impersonating a variety of personas across complex and interactive tasks [89]. Finally, we believe that in-context impersonation can also be applied to other modalities, for example to large models for video generation [90].

7 Acknowledgements

The authors thank IMPRS-IS for supporting Leonard Salewski. This work was partially funded by the Portuguese Foundation for Science and Technology (FCT) under PhD grant 2020.07034.BD, the Max Planck Society, the Volkswagen Foundation, the BMBF Tübingen AI Center (FKZ: 01IS18039A), DFG (EXC number 2064/1 – Project number 390727645) and ERC (853489-DEXIM).

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *NeurIPS*, 2020.

[2] Taylor Webb, Keith J Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *arXiv:2212.09196*, 2022.

[3] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *TMLR*, 2022.

[4] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 2023.

[5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv:2108.07258*, 2021.

[6] Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. Understanding the capabilities, limitations, and societal impact of large language models. *arXiv:2102.02503*, 2021.

[7] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *ACM FAccT*, 2021.

[8] Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *PNAS*, 120(6), 2023.

[9] Jonathan Pilault, Raymond Li, Sandeep Subramanian, and Christopher Pal. On extractive and abstractive neural document summarization with transformer language models. In *EMNLP*, 2020.

[10] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

[11] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *EMNLP*, 2022.

[12] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *ICLR*, 2022.

[13] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv:2304.05335*, 2023.

[14] Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. Can chatgpt write a good boolean query for systematic review literature search? *arXiv:2302.03495*, 2023.

[15] Katherine Elkins and Jon Chun. Can gpt-3 pass a writer's turing test? *Journal of Cultural Analytics*, 5(2), 2020.

[16] Marcel Binz and Eric Schulz. Modeling human exploration through resource-rational reinforcement learning. In *NeurIPS*, 2022.

[17] Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. Can language models learn from explanations in context? In *EMNLP*. ACL, 2022.

[18] Simran Arora, Avanika Narayan, Mayee F Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, and Christopher Re. Ask me anything: A simple strategy for prompting language models. In *ICLR*, 2023.

[19] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *NeurIPS Workshops*, 2022.

[20] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *EACL*, 2021.

[21] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault

Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. Multitask prompted training enables zero-shot task generalization. In *ICLR*, 2022.
[22] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys*, 53(3), 2020.
[23] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *TPAMI*, 41(9), 2018.
[24] Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. How well do large language models perform in arithmetic tasks? *arXiv:2304.02015*, 2023.
[25] Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv:2305.00050*, 2023.
[26] Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *CHI*, 2021.
[27] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *EMNLP*, 2020.
[28] Nathan Hunter. *The art of prompt engineering with chatGPT*. eBook, 2023.
[29] Jonas Oppenlaender, Rhema Linder, and Johanna Silvennoinen. Prompting ai art: An investigation into the creative skill of prompt engineering. *arXiv:2303.13534*, 2023.
[30] Seungju Han, Beomsu Kim, Jin Yong Yoo, Seokjun Seo, Sangbum Kim, Enkhbayar Erdenee, and Buru Chang. Meet your favorite character: Open-domain chatbot mimicking fictional characters with only a few utterances. In *NAACL-HLT*, 2022.
[31] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. *arXiv:1909.05858*, 2019.
[32] Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role-play with large language models. *ArXiv:2305.16367*, 2023.
[33] Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. Unsupervised text style transfer using language models as discriminators. *NeurIPS*, 2018.
[34] Katherine Crowson, Stella Rose Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *ECCV*, 2022.
[35] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2021.
[36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022.
[37] Nassim Dehouche and Kullathida Dehouche. What's in a text-to-image prompt? the potential of stable diffusion in visual arts education. *Heliyon*, 9, 2023.
[38] Manuel Brack, Patrick Schramowski, Felix Friedrich, Dominik Hintersdorf, and Kristian Kersting. The stable artist: Steering semantics in diffusion latent space. *arXiv:2212.06013*, 2022.
[39] Sam Witteveen and Martin Andrews. Investigating prompt engineering in diffusion models. *arXiv:2211.15462*, 2022.
[40] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *ACL*, 2022.
[41] Yotam Wolf, Noam Wies, Yoav Levine, and Amnon Shashua. Fundamental limitations of alignment in large language models. *arXiv:2304.11082*, 2023.
[42] Gati Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans. *arXiv:2208.10264*, 2022.
[43] Max Pellert, Clemens M Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. Ai psychometrics: Using psychometric inventories to obtain psychological profiles of large language models. 2023.
[44] Peter S. Park, Philipp Schoenegger, and Chongyang Zhu. "correct answers" from the psychology of artificial intelligence. *arXiv:2302.07267*, 2023.
[45] Saketh Reddy Karra, Son Nguyen, and Theja Tulabandhula. Ai personification: Estimating the personality of language models. *arXiv:2204.12000*, 2022.
[46] Julian Coda-Forno, Kristin Witte, Akshay K Jagadish, Marcel Binz, Zeynep Akata, and Eric Schulz. Inducing anxiety in large language models increases exploration and bias. *arXiv:2304.11111*, 2023.
[47] Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dunner. Questioning the survey responses of large language models. *arXiv:2306.07951*, 2023.
[48] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 2023.
[49] Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences. *arXiv:2305.02547*, 2023.
[50] Gati Aher, RosaI. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *ICML*, 2022.

[51] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 2017.

[52] Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. In *AAAI/ACM AEIS*, 2021.

[53] Daniel Kang, Xuechen Li, Ion Stoica, Carlos Guestrin, Matei Zaharia, and Tatsunori Hashimoto. Exploiting programmatic behavior of llms: Dual-use through standard security attacks. *arXiv:2302.05733*, 2023.

[54] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.

[55] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *CVPR*, 2022.

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[57] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *ICLR*, 2023.

[58] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. *arXiv:2211.11158*, 2022.

[59] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of GPT-3 for few-shot knowledge-based VQA. In *AAAI*, 2022.

[60] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023.

[61] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv:2302.13971*, 2023.

[62] Dom Eccleston. ShareGPT: Share your wildest conversations with one click. `https://sharegpt.com/`, 2023. [Online; accessed 15-May-2023].

[63] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haotong Zhang, Joseph Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv:2306.05685*, 2023.

[64] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

[65] Samuel J Gershman. Deconstructing the human algorithms for exploration. *Cognition*, 173, 2018.

[66] Eric Schulz and Samuel J Gershman. The algorithmic architecture of exploration in the human brain. *Current opinion in neurobiology*, 55, 2019.

[67] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021.

[68] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. *arXiv:2212.07143*, 2022.

[69] Simran Arora, Avanika Narayan, Mayee F. Chen, Laurel J. Orr, Neel Guha, Kush S Bhatia, Ines Chami, Frederic Sala, and Christopher R'e. Ask me anything: A simple strategy for prompting language models. *arXiv:2210.02441*, 2022.

[70] Kate Nussenbaum and Catherine A Hartley. Reinforcement learning across development: What insights can we draw from a decade of research? *Developmental cognitive neuroscience*, 40, 2019.

[71] Emily G. Liquin and Alison Gopnik. Children are more exploratory and learn more than adults in an approach-avoid task. *Cognition*, 218, 2020.

[72] Eric Schulz, Charley M Wu, Azzurra Ruggeri, and Björn Meder. Searching for rewards like a child means less generalization and more directed exploration. *Psychological science*, 30(11), 2019.

[73] Anna P. Giron, Simon Ciranka, Eric Schulz, Wouter van den Bos, Azzurra, Ruggeri, Björn Meder, and Charley M. Wu. Developmental changes in learning resemble stochastic optimization. 2022.

[74] Nathaniel J. Blanco and Vladimir M. Sloutsky. Systematic exploration and uncertainty dominate young children's choices. *Developmental science*, 2019.

[75] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv:2204.02311*, 2022.

[76] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv:2203.15556*, 2022.

[77] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher R'e, Diana Acosta-Navas, Drew A. Hudson, E. Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan S. Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas F. Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525, 2023.

[78] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[79] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. *ICCV Workshops*, 2013.

[80] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

[81] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

[82] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[83] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv:2210.08402*, 2022.

[84] John Ed Oates and Andrew Ed Grayson. *Cognitive and language development in children.* Open University Press, 2004.

[85] Philip Durrant and Mark Brenchley. Development of vocabulary sophistication across genres in english children's writing. *Springer Reading Writing*, 32, 2019.

[86] Ryan Burnell, Wout Schellaert, John Burden, Tomer D Ullman, Fernando Martinez-Plumed, Joshua B Tenenbaum, Danaja Rutar, Lucy G Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, et al. Rethink reporting of evaluation results in ai. *Science*, 380(6641), 2023.

[87] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv:2001.08361*, 2020.

[88] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv:1909.08593*, 2019.

[89] Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv:2304.03442*, 2023.

[90] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv:2212.03191*, 2022.