Grounding Large Language Model Behavior

Brian Yu (bri25yu@berkeley.edu)
Department of Electrical Engineering and Computer Sciences
2626 Hearst Ave. Berkeley, CA 94720 USA

Terry Regier (terry.regier@berkeley.edu)
Department of Linguistics
2650 Dwinelle Hall. Berkeley, CA 94704 USA

Abstract

Large language models (LLMs) have recently exploded in popularity. Research into these LLMs has extended to how to use LLMs in downstream applications, the risks and limitations of LLMs, and the capabilities of LLMs. However, this research is performance-based and misses out on fundamental explanations of why LLMs behave the way they do. In this paper, we approach LLMs from the perspective of discovering their behavior in a manner grounded in the natural sciences. We explore two facets of LLM behavior: linguistic and perceptual. We hope that our research inspires further work exploring LLM behavior. We release our code here. **Keywords:** Large language model; grounding; behavior

Introduction

The release of ChatGPT and GPT-4 by OpenAI inspired a new wave of investigation into large language model (LLM) applications, behavior, and capability. The GPT-4 technical report released by OpenAI investigates the latter two facets of GPT-4 behavior and capabilities, especially with regards to safety [1]. Exploration has been performed into how LLMs can be used for **applications in various domains** including the legal and medical domains [2, 3, 4, 5, 6, 7], **writing** [8, 9], and even **deep learning** [10, 11, 12, 13, 14]. Much of this application-driven research includes evaluations of the risks in those specific domains. Several works dive deeper into general risks and limitations posed by these models, including investigations of racial bias [15] and robustness [16].

Other research investigates the capabilities of these LLMs including topics such as **understanding and reasoning** [17, 18, 19, 20, 21], **exam taking** [22, 23, 24, 25], **linguistic** [26, 27, 28], **translation** [29, 30], **data augmentation** [31], and **various other domain capabilities** [32, 33, 34, 35].

The vast majority of research focuses on the performance of LLMs. Approaches begin with a notion of what the LLM "should" receive as an input and what the LLM "should" produce as an output. For example in the case of reasoning, the LLM "should" be able to reason to produce the correct output response. **However, these approaches are fundamentally shallow because they investigate the outputs produced and not *how* the outputs are produced.** A better approach is to investigate the properties of constituents of a given output and evaluate how consistent they are with each other [36]. However, this approach still misses the fundamental properties of how the model behaves in its outputs.

In order to properly evaluate model behavior, we must draw on the natural sciences. For example, psychology provides an avenue of evaluation into how LLMs make decisions by investigating heuristics used when making decisions [37, 38]. Evaluations can also be made by comparing the behavior of LLMs to human behavior, for example with regards to common sense [39] or psychopathy [40]. Investigations should be made into the behavior of LLMs from a cognitive, psychological, and linguistics perspective.

There are many avenues for exploring LLM behavior including linguistic [41, 42, 43], perceptual [44], spatial [45], and many more. In this paper, we explore the theoretical and perceptual aspects of LLM behavior. We review theoretical linguistic positions regarding the nature of language and attempt to reconcile them with today's LLMs. We perform preliminary experiments regarding color naming as a proxy of an LLM's perceptual space. We present several visual illusions to humans and analyze LLM responses. We hope that our experiments inspire further research with the paradigm of investigating LLM behavior.

Updating theoretical positions

We begin with a discussion on the theoretical underpinnings regarding what language is. Hauser, Chomksy, and Fitch (2002) argue that language has general intelligence roots in "a sensory-motor system, a conceptual-intentional system, and the computational mechanisms for recursion" and that a "uniquely human component" for language is recursion [41]. LLMs today have none of these properties as they are only trained on naturally occurring text on the open web. LLMs have no sensory-motor system because they cannot interact with the world except through the textual modality. LLMs have no explicit conceptual-intentional system, even lacking the symbolic underpinnings of any conceptual system. The most prevalent attention mechanism used today has no explicit ability to treat a group of inputs as a single "symbolic" input [46]. For example, an LLM today is unable even to explicitly treat the phrase "in the dark" as a phrasal unit. LLMs today lack the computational mechanisms for recursion and certainly lack a recursive capacity, having a fixed budget for computation. It is arguable that LLMs have enough computational resources to perform shallow to mid level recursion, but this by no means that they have recursive capacities. Hauser, Chomsky, and Fitch also propose mechanisms that could have contributed to language development including "number, navigation, and social relations" [41]. None of these mechanisms apply to the understanding of language in LLMs.

LLMs today do not have language capacities; rather, they have language *mimicry* capabilities. These LLMs are deep neural networks that are powerful function approximators. The specific function we train an LLM to approximate is natural language. As a result, any properties that are present in natural language will also be mimicked by an LLM. For example, reasoning manifests in natural language so LLMs are able to mimic reasoning capabilities. LLMs are able to nearly perfectly mimic natural language. **None of the aforementioned components mentioned by Hauser, Chomsky, and Fitch are necessary for language mimicry. None of the aforementioned mechanisms for language development are necessary for development of language mimicry.**

Pinker and Jackendoff (2005) argue that language is an "adaptation for the communication."[42]. Indeed, that is how the most popular and powerful LLMs are trained today. Specifically, LLMs today are trained to output the best response to a given input [47]. This enables LLMs to better approximate the fluency, structure, and content of human dialog compared to LLMs that are not trained on dialog data. So, in some sense, a large component of the finer details of language understanding and language production have roots in the notion that language was an adaptation for communication. Pinker and Jackendoff also highlight various other properties of language that pose constraints on the actual language systems that manifest in the world today. For example, physical constraints on the speech production and speech perception systems translate to constraints on the underlying hierarchical structure of language. These constraints are not explicitly introduced into the LLM training process yet the LLM produces language that strictly adheres to these constraints. This is yet another reflection of how powerful LLMs are at natural language approximation.

In a similar vein, Evans and Levinson (2009) propose that language is subject to "multiple design constraints" that reflect "cultural-historical" factors and constraints that reflect "human cognition" [43]. **Clearly, constraints on human cognition, physical human systems, or other human systems are not necessary for language mimicry.**

The claims made here are simultaneously powerful yet meaningless. On one hand, language capacities may be less complex and unique than previously held. On the other hand, because these LLMs only have language mimicry capabilities, no implication regarding their language properties is deep. **LLMs today lack many of the mechanisms that contribute to language abilities, yet they are still nearly perfect language approximators.**

Color perception

We explore the behavior of LLMs at color naming in an attempt to create a proxy of the underlying perceptual space of an LLM.

Experiment setup

We use Munsell chips as our color naming substrate. The Munsell chips consist of 330 chips over a gradient of hue, saturation, and brightness. The chips are partitioned into two categories of 10 grayscale chips and 320 colored chips. For our color naming task, we choose to only use the 320 colored chips, but the experiment is easily repeated with the full set of chips or additional color chips not found in the original set of 330 chips.

We use the BLIP visual question-answering large language model as the model family of interest [48]. BLIP consists of around 200M parameters and is trained specifically for visual question-answering. Several alternatives are available with varied ranges of capabilities. However, the particular choice of model family to investigate is not critical because the experiments can be easily repeated with any other family of visual question-answering models and because this line of investigation is not too concerned with performance. BLIP is a probabilistic model that outputs probabilities over its set of vocabulary tokens. This enables a direct measurement of "confidence" or "uncertainty" in a particular answer.

In order to get color naming output from the LLM, we prompt it with an input color chip and an input prompt. The input prompt we use is "What color is this image?". The prompt must be in natural language in order for BLIP to properly understand the input. The prompt must also direct the LLM to name the color in the input chip. From the LLM output, we collect the natural language response and the corresponding output probabilities. The output probabilities are used to generate visualizations that incorporate LLM uncertainty in color names.

The BLIP LLM is available for non-commercial use with citations under the BSD-3-Clause. The experiments were run on Google Colaboratory with an NVIDIA Tesla T4 of 12 GB GPU VRAM. Experiment runtimes varied from 1 to 3 minutes.

Results and analysis

We seek to answer three questions: (1) is the LLM competent and consistent enough to name all the colors, (2) do the colors form a coherent color naming space, and (3) does the LLM reflect human properties of color perception?

The BLIP LLM successfully names all of the Munsell chips 1. Furthermore, the LLM groups colors consistently across the gradients, with the same color names always occurring in a single contiguous set. The colors roughly match the author's subjective color names, forming a coherent color naming space. Looking at visualization of LLM confidence, we observe that color names near the edge of a color set have more uncertainty than those in the center or focus of that particular color set. The LLM is not only able to correctly mimic the color names of the Munsell chips, but also approximately mimic the uncertainty that humans have when naming colors on color naming boundaries.

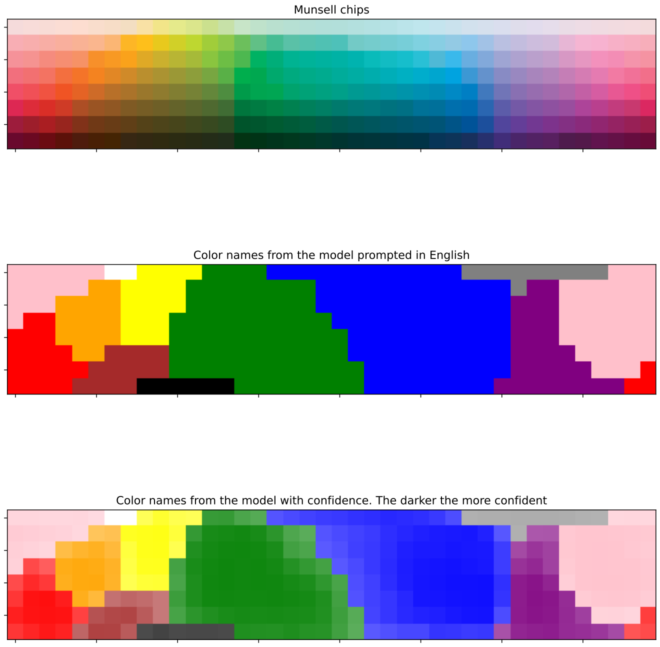
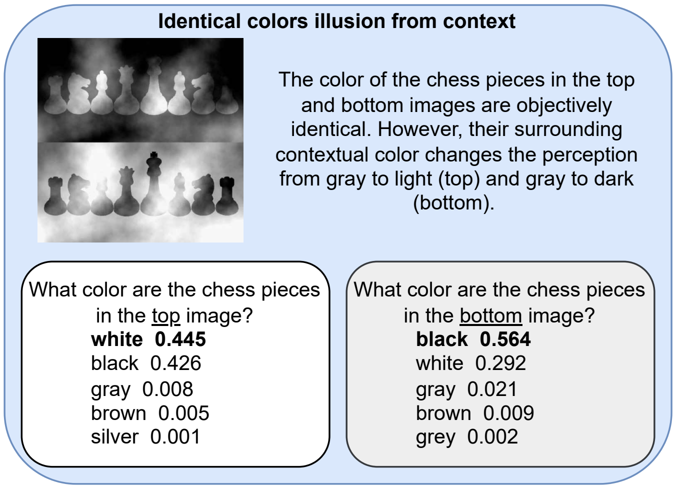


Figure 1: Color names output by the BLIP LLM plotted on the Munsell chip grid. **Top**: The original colors of the Munsell chips. **Middle**: The color names output by the BLIP LLM plotted using Matplotlib package default colors. **Bottom**: The middle figure with LLM probabilities scaling the transparency. Darker colors indicate higher confidence.

Visual illusions for humans

We investigate four visual illusions that occur as a product of the visual processing pathway in humans.


Identical colors illusion from context

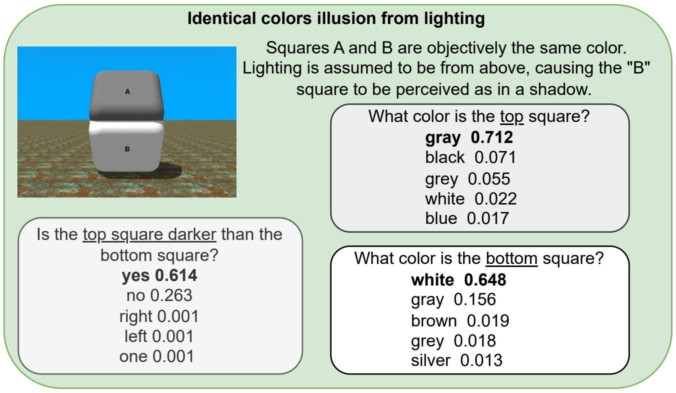
The color of the chess pieces in the top and bottom images are objectively identical. However, their surrounding contextual color changes the perception from gray to light (top) and gray to dark (bottom).

What color are the chess pieces in the top image?
white 0.445
black 0.426
gray 0.008
brown 0.005
silver 0.001

What color are the chess pieces in the bottom image?
black 0.564
white 0.292
gray 0.021
brown 0.009
grey 0.002

Figure 2: BLIP LLM model outputs and output probabilities for the illusion that the colors of the chess pieces in the top and bottom are different.

The LLM identifies the chess pieces as white and black in the top and bottom images, respectively. This is in line with human perception. However, the LLM is much more certain that the bottom pieces are black, an asymmetry that does not exist in human perception. This could be due to any biases in the dataset that the LLM was trained on that favor lighter backgrounds.


Identical colors illusion from lighting

Squares A and B are objectively the same color. Lighting is assumed to be from above, causing the "B" square to be perceived as in a shadow.

What color is the top square?
gray 0.712
black 0.071
grey 0.055
white 0.022
blue 0.017

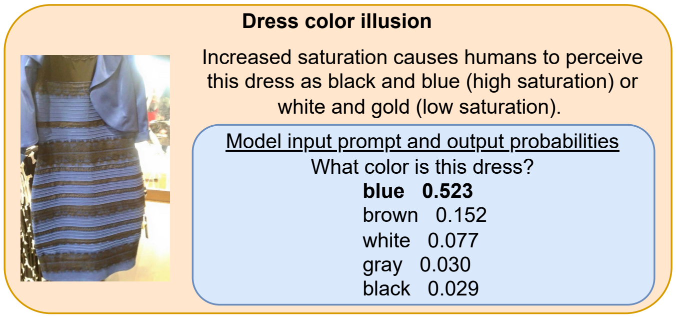
Is the top square darker than the bottom square?
yes 0.614
no 0.263
right 0.001
left 0.001
one 0.001

What color is the bottom square?
white 0.648
gray 0.156
brown 0.019
grey 0.018
silver 0.013

Figure 3: BLIP LLM model outputs and output probabilities for the illusion of different colors in squares A and B.

The LLM is confident that the top square is gray and slightly less confident that the bottom square is white, in line with human perception as the bottom square could be perceived as a white square in shadow or a plain gray square.

The LLM also perceives the top square as ”darker” than the top square with relatively high confidence.



Dress color illusion

Increased saturation causes humans to perceive this dress as black and blue (high saturation) or white and gold (low saturation).

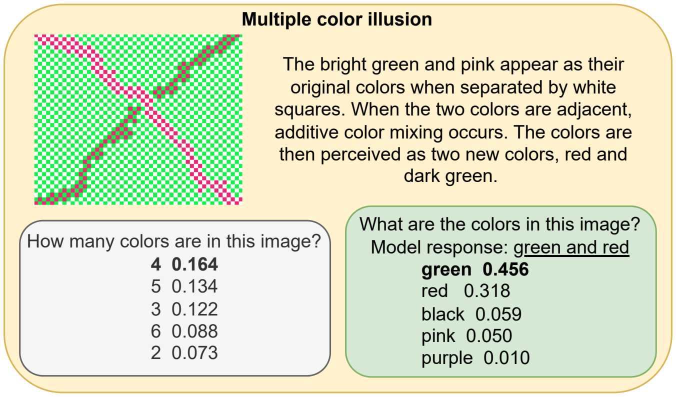
Model input prompt and output probabilities

What color is this dress?

blue 0.523
brown 0.152
white 0.077
gray 0.030
black 0.029

Figure 4: BLIP LLM model outputs and output probabilities for the illusion of two competing hypotheses of colors.

The LLM responds that this dress is “blue” with relative confidence. This differs from human perception where the response is typically the combination of two colors corresponding to the two different colored stripes on the dress.



Multiple color illusion

The bright green and pink appear as their original colors when separated by white squares. When the two colors are adjacent, additive color mixing occurs. The colors are then perceived as two new colors, red and dark green.

How many colors are in this image?

4 0.164
5 0.134
3 0.122
6 0.088
2 0.073

What are the colors in this image?
Model response: green and red

green 0.456
red 0.318
black 0.059
pink 0.050
purple 0.010

Figure 5: BLIP LLM model outputs and output probabilities for the illusion of more colors than just the two present.

The LLM responds that there are 4 colors in the image, with some probability mass on the alternative outputs of 5 colors and 3 colors. This could be partially due to the proximity in adjacent colors to the number of colors that the model believes or due to the illusion that there are more or less colors due to the adjacency of the pink and green colors.

Overall, the LLM responds in ways that roughly align with human perception.

Conclusion

LLMs should not only be evaluated on their performance on tasks or benchmarks; rather, their behavior should be modeled and understood for better insight. To get a fuller and deeper view of LLMs, inspiration must be drawn from the natural sciences. In this paper, we explore the linguistic and visual domains as avenues into LLM behavior. We review and update theoretical positions on the nature of language, perform preliminary experiments regarding color naming as a proxy of an LLM’s perceptual space, and we present several visual illusions to humans and analyze LLM responses. Future work can extend this paradigm of exploring LLM behavior in different domains like spatial perception or decision making.

Acknowledgments

Thank you to Professor Terry Regier for teaching an interesting class and for the great semester in Linguistics 243! Thank you to my peers Emil, Li, Kayo, Kaylo, and Zaid!

References

[1] OpenAI. Gpt-4 technical report, 2023.

[2] Jaromir Savelka. Unlocking practical applications in legal domain: Evaluation of gpt for zero-shot semantic annotation of legal texts. 2023.

[3] Debadutta Dash, Rahul Thapa, Juan M. Banda, Akshay Swaminathan, Morgan Cheatham, Mehr Kashyap, Nikesh Kotecha, Jonathan H. Chen, Saurabh Gombar, Lance Downing, Rachel Pedreira, Ethan Goh, Angel Arnaout, Garret Kenn Morris, Honor Magon, Matthew P Lungren, Eric Horvitz, and Nigam H. Shah. Evaluation of gpt-3.5 and gpt-4 for supporting real-world information needs in healthcare delivery, 2023.

[4] Tong Xie, Yuwei Wan, Wei Huang, Yufei Zhou, Yixuan Liu, Qingyuan Linghu, Shaozhou Wang, Chunyu Kit, Clara Grazian, Wenjie Zhang, and Bram Hoex. Large language models as master key: Unlocking the secrets of materials science with gpt, 2023.

[5] Krishna Kumar. Geotechnical parrot tales (gpt): Harnessing large language models in geotechnical engineering, 2023.

[6] Stephen James Krol, Maria Teresa Llano, and Jon McCormack. Towards the generation of musical explanations with gpt-3, 2022.

[7] Anaïs Tack and Chris Piech. The ai teacher test: Measuring the pedagogical ability of blender and gpt-3 in educational dialogues, 2022.

[8] Oğuz ’Oz’ Buruk. Academic writing with gpt-3.5: Reflections on practices, efficacy and transparency, 2023.

[9] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. News summarization and evaluation in the era of gpt-3, 2022.

[10] Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. Is a prompt and a few samples all you need? using gpt-4 for data augmentation in low-resource classification tasks, 2023.

[11] Ruohong Zhang, Yau-Shian Wang, and Yiming Yang. Generation-driven contrastive self-training for zero-shot text classification with instruction-tuned gpt, 2023.

[12] Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. Gpt-ner: Named entity recognition via large language models, 2023.

[13] Xiaohan Yang, Eduardo Peynetti, Vasco Meerman, and Chris Tanner. What gpt knows about who is who, 2022.

[14] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023.

[15] Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. Gpt detectors are biased against non-native english writers, 2023.

[16] Xuanting Chen, Junjie Ye, Can Zu, Nuo Xu, Rui Zheng, Minlong Peng, Jie Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. How robust is gpt-3.5 to predecessors? a comprehensive study on language understanding tasks, 2023.

[17] Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. Evaluating the logical reasoning ability of chatgpt and gpt-4, 2023.

[18] Sifatkaur Dhingra, Manmeet Singh, Vaisakh SB, Neetiraj Malviya, and Sukhpal Singh Gill. Mind meets machine: Unravelling gpt-4's cognitive psychology, 2023.

[19] Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models, 2023.

[20] Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. Can gpt-3 perform statutory reasoning?, 2023.

[21] Marcel Binz and Eric Schulz. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6), February 2023.

[22] Desnes Nunes, Ricardo Primi, Ramon Pires, Roberto Lotufo, and Rodrigo Nogueira. Evaluating gpt-3.5 and gpt-4 models on brazilian university admission exams, 2023.

[23] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems, 2023.

[24] Jaromir Savelka, Arav Agarwal, Christopher Bogart, Yifan Song, and Majd Sakr. Can generative pre-trained transformers (gpt) pass assessments in higher education programming courses?, 2023.

[25] Michael Bommarito II au2 and Daniel Martin Katz. Gpt takes the bar exam, 2022.

[26] Yikang Liu, Ziyin Zhang, Wanyang Zhang, Shisen Yue, Xiaojing Zhao, Xinyuan Cheng, Yiwen Zhang, and Hai Hu. Argugpt: evaluating, understanding and identifying argumentative essays generated by gpt models, 2023.

[27] Steven Coyne and Keisuke Sakaguchi. An analysis of gpt-3's performance in grammatical error correction, 2023.

[28] Kyle Mahowald. A discerning several thousand judgments: Gpt-3 rates the article + adjective + numeral + noun construction, 2023.

[29] Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. How good are gpt models at machine translation? a comprehensive evaluation, 2023.

[30] Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. Is chatgpt a good translator? yes with gpt-4 as the engine, 2023.

[31] Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq Joty, and Boyang Li. Is gpt-3 a good data annotator?, 2022.

[32] 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, 2023.

[33] Kent K. Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. Speak, memory: An archaeology of books known to chatgpt/gpt-4, 2023.

[34] Jaromir Savelka, Arav Agarwal, Christopher Bogart, and Majd Sakr. Large language models (gpt) struggle to answer multiple-choice questions about code, 2023.

[35] Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. Who is gpt-3? an exploration of personality, values and demographics, 2022.

[36] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023.

[37] Gaurav Suri, Lily R. Slater, Ali Ziaee, and Morgan Nguyen. Do large language models show decision heuristics similar to humans? a case study using gpt-3.5, 2023.

[38] Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. Machine intuition: Uncovering human-like intuitive decision-making in gpt-3.5, 2022.

[39] Philipp Koralus and Vincent Wang-Maścianica. Humans in humans out: On gpt converging toward common sense in both success and failure, 2023.

[40] Xingxuan Li, Yutong Li, Shafiq Joty, Linlin Liu, Fei Huang, Lin Qiu, and Lidong Bing. Does gpt-3 demonstrate psychopathy? evaluating large language models from a psychological perspective, 2023.

[41] Marc D. Hauser, Noam Chomsky, and W. Tecumseh Fitch. The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569–1579, 2002.

[42] Steven Pinker and Ray Jackendoff. The faculty of language: what's special about it? *Cognition*, 95(2):201–236, 2005.

[43] Yoonhyoung Lee, Eunsuk Lee, Peter C. Gordon, and Randall Hendrick. Commentary on evans and levinson, the myth of language universals: Language diversity, cognitive universality. *Lingua*, 120(12):2695–2698, 2010. The Myth of Language Universals.

[44] Terry Regier, Paul Kay, and Naveen Khetarpal. Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences of the United States of America*, 104:1436–41, 02 2007.

[45] Asifa Majid, Melissa Bowerman, Sotaro Kita, Daniel B.M. Haun, and Stephen C. Levinson. Can language restructure cognition? the case for space. *Trends in Cognitive Sciences*, 8(3):108–114, 2004.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[47] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[48] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.