LLM Sample: part Average and part Ideal

Sarath Sivaprasad [*1] **Pramod Kaushik** [*2] **Sahar Abdelnabi** [3] **Mario Fritz** [1]

,

Abstract

As Large Language Models (LLMs) increasingly impact society, it's crucial to understand the heuristics and biases that drive them. We study response sampling of LLMs in light of *value bias*—a tendency to favor high-value options in their outputs. Value bias corresponds to the shift of response from the most likely sample towards some notion of ideal value represented in the LLM. Our study identifies value bias in both existing and new concepts learned in-context. We demonstrate that this bias significantly impacts applications, such as patient recovery times. These findings highlight the need to address value bias in LLM deployment to ensure fair and balanced AI applications.

1. Introduction

LLMs are often considered to be 'System-1' (Daniel, 2017), characterized by their reliance on heuristics and operating implicitly without deliberation (Dasgupta et al., 2022; Yao et al., 2023a). Their strong performance on the benchmarks of mathematical reasoning (Imani et al., 2023), pragmatics (Lipkin et al., 2023), and high-level planning (Song et al., 2023) shows the importance of studying the various heuristics and mechanisms that enable such performances. These heuristics can potentially be inferred by observing the response samples of LLMs, helping us uncover potential biases from both safety (Yao et al., 2023b) and utility (Achiam et al., 2023) perspectives.

We define response sampling as the process by which the model probabilistically selects outputs from a distribution of potential responses. We show that LLMs have a value bias in sampling, favoring high-value options across different scenarios. Value bias is the tendency of the sample to deviate from the average towards a notion of an 'ideal value' (Figure 1). Since the underlying auto-regressive mechanism is not goal-driven, it is non-trivial how, from a large number of possible samples, the LLM's response sample is based on value.

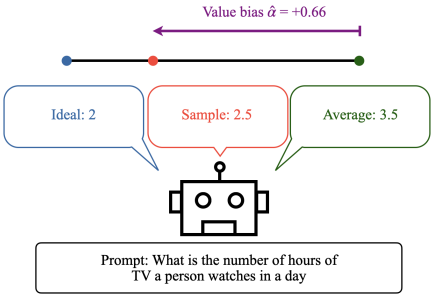

Figure 1. The sample of an LLM drifts away from the average value towards an ideal value. The notion of ideal stems from a value system that is either implicit to the LLM or learnt in context. Positive α shows the deviation of the sample from the average in the direction of the ideal.

Psychology studies show that human reasoning is a combination of the statistical norm, i.e., average, along with the prescriptive norm, i.e., ideal (Bear et al., 2020). For instance, asking the questions, (a) "What is the average number of hours of TV that a person watches in a day?" and (b) "What is the number of hours of TV a person watches in a day?" yields different answers in human participants(Bear et al., 2020). This is because possibility sampling in humans is composed of both the average and the ideal (Phillips et al., 2019).

Similar to the concept of 'watching TV', we evaluate value bias of LLMs across 36 different concepts known to the LLM. We use the evaluations in Bear et al. (2020). To show the practical implications of this, we present a case study where the LLM is asked to prescribe a recovery time to the patients showing certain symptoms. The results show value bias in output wherein the recommended recovery period unexpectedly deviates away from the statistical average towards some notion of a prescriptive norm or an ideal. We further show that LLMs show value bias in sampling even for a new concept introduced in context with a distribution and a notion of value.

[*]Equal contribution [1]CISPA Helmholtz Center for Information Security [2]TCS Research, Pune [3]Microsoft. Correspondence to: Sarath Sivaprasad <{sarath.sivaprasad, fritz}@cispa.de>, Pramod Kaushik <pramod.kaushik@tcs.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

2. Related Work

Understanding Heuristics in Response Sampling: (Simon, 1996) uses the notion of heuristics to explain the decision-making of 'System-1' mechanisms. They show the utility of 'mental shortcuts' to navigate countless possibilities of the search problem (Newell et al., 1972). Exploring the heuristics of LLMs can help understand the characteristics of information processing in them (Hazra et al., 2023; Shah et al., 2023; Suri et al., 2023), and prior explorations rely on sampling for the same. Recent work also shows some overlaps in errors made by LLMs and humans in System1 reasoning tasks (Dasgupta et al., 2022). These evaluations show the need to better understand the mechanisms by which LLMs sample their responses.

Possibility Sampling in Humans: A growing body of literature addresses how humans and animals navigate the search problem of countless possibilities (Phillips et al., 2019; Phillips & Cushman, 2017; Mattar & Lengyel, 2022; Ross et al., 2023) and use the two-step process to reach the final option. Increasing evidence shows the significance of probability and value in how humans sample possible solutions from the large space of possibilities (Bear et al., 2020; Phillips et al., 2019; Bear & Knobe, 2017). What comes to the human mind or what it considers normal is sampled from a probability distribution of both value and probability (frequency) (Bear et al., 2020). This dual nature of thought is hypothesised to have come due to humans being goal-driven agents and engaging in value maximisation (Bear & Knobe, 2017). In spatial navigation, experiments show that an optimal reinforcement learning agent which has the ability to recall past experiences, ordered on utility, optimises on two dimensions, namely, "gain" and "need", where "gain" roughly refers to the value aspect of experience while "need" refers to the statistical occurrence of experience (Mattar & Daw, 2018).

3. Evaluating Value Bias in LLMs

Implicit Bias: We evaluate the estimation of average and the notion of the ideal on different categories. The average value reported by the LLM for a category C is C_a and the ideal value is C_i. The sample generated by the LLM for the category is C_s. We use a variable α to quantify the degree of value bias. For a positive value bias, the C_s deviate from the C_a in the direction of C_i. We compute this direction as the positive direction of alpha.

For each sample C_s of a category C, the α is computed as

$$\alpha = (C_a - C_s) \times sign(C_a - C_i) \tag{1}$$

The sign of alpha is given by the relative positions of C_a, C_i, and C_s such that C_s is in the direction of C_i as measured from C_a.

We also compute α on a normalized scale such that C_a is at the origin and C_i is at unit distance from the origin in the positive direction. We call this $\hat{\alpha}$, and it enables comparison across categories with less dependency on the scale of values. It also allows comparison with values obtained in human experiments. Value bias is significant with a higher positive value of $\hat{\alpha}$. We compute $\hat{\alpha}$ as

$$\hat{\alpha} = \frac{((C_a - C_s) \times sign(C_a - C_i))}{|C_a - C_i|} \tag{2}$$

We use the binomial test for testing the significance.

Evaluating Value Bias for In-Context Value and Distributions: To illustrate the value bias for new concepts learned in context, we provide a prompt that contains (a) n samples from a distribution N and (b) a value associated with each of the samples. The value is generated using the underlying mechanism V. The mechanism V is verbally explained in the prompt and given as grades associated with each value.

For a new category C and the corresponding N and V introduced in context, we ask the LLM to report the average of the distribution C_a and a sample from the distribution C_s. Note that the notion of ideal comes from V. To test the significance of value bias, we use the Mann-Whitney-U test to compare C_a to N and C_a to C_s given V.

4. Experiments and Results

In this section, we present three experiments to show (a) implicit value bias, (b) value bias for a category learned in-context. Our results show a significant value bias in the sampling of LLMs. We use the instruction tuned model of GPT-4 (Achiam et al., 2023), GPT-3.5-Turbo (Brown et al., 2020), Claude (Anthropic, 2024), Mixtral-8X7B (Jiang et al., 2024), Mistral-7B (Jiang et al., 2023) and LLama3-8B (Touvron et al., 2023) for our experiments. Unless mentioned otherwise, we report results for GPT-4 and the results for other models are in the Appendix. Also, all prompts were appended with a pre-prompt to get floating-point numbers as answers. The complete text used in the prompts for each experiment is given in the Appendix.

4.1. Implicit Value Bias

In this experiment, the true distribution N and value system V are implicit in the LLM and not known to us. We evaluate implicit value bias on thirty-six different categories (Cs). For each category, we ask the model to report (a) the average value C_a, (b) the ideal value C_i and (c) a sample C_s in independent contexts. To get these values, we use a prompt

similar to the questions used in human studies (Bear et al., 2020).
For example, to get the average, ideal and the sample on the category of 'TV watching of people', we make the following prompts:

Prompt for Implicit Value Bias

P_a: What is the average number of hours of TV a person watches in a day?
P_i: What is the ideal number of hours of TV for a person to watch in a day?
P_s: Enter the first number that comes to mind when answering the question. Note that there is no correct answer. Please be spontaneous in your judgment: What is the number of hours of TV for a person to watch in a day?

We repeat this ten times with a temperature of 0.8 and report the average in Table 1. The categories shown in bold exhibit a value bias in sampling. The deviation of the values between the runs is minimal. The mode of standard deviation for C_a and C_i across categories are 0.46 and 0.0, respectively.

Results: From Table 1, we can see that samples from 24 out of 36 categories are on the ideal side of the average. We get a p-value of 0.033 with a binomial test, which is statistically significant. This indicates LLM samples deviate away from the statistical average towards a prescriptive ideal.

It is interesting to note that in some categories, the sample not only falls on the ideal side of average but in fact goes beyond the ideal. For instance, for the category: 'loads of laundry' with mean C_a of 2.06 gives an $\hat{\alpha}$ of +1.870. This implies, for such categories, the sample not only deviates to the ideal side of average but goes beyond ideal in the same direction.

We also perform the experiment with temperature zero. With this setting, the observation remains the same with a significance of $p = 0.035$. For instance, for the category 'Hours of TV in a day', we get $C_a = 3.5$, $C_i = 2$, and $C_s = 2.5$. The $\hat{\alpha}$ is $+0.66$. This example is illustrated in Figure 1. The table of results for the run with temperature zero is in Appendix A.6. Furthermore, we compare these results with the results of human evaluation in Appendix A.1 and show that the value system V in humans and LLMs are not always aligned, with $\hat{\alpha}$ Pearson correlation of -0.02.

When done on other LLMs with default temperatures we get the following results with LLama3-7B (binomial $p = 0.003$), Mixtral-8x7B (binomial $p = 0.05$), GPT3.5-turbo (binomial $p < 0.001$), Claude (binomial $p < 0.001$), Mistral (binomial $p = 0.0019$) indicating that this value bias is pervasive across LLMs.

4.1.1. Case Study: Medical Recovery Time

In this section, we present a real-world example demonstrating the practical implications of implicit value bias. For each category, we give a list of four symptoms and ask the LLM to prescribe a recovery time. We prompt the LLM to suggest recovery time (in weeks) based on a given list of symptoms. Similar to Experiment 4.1, we used three different prompts: one for the average recovery time, one for the ideal recovery time, and a third prompt asking the LLM to provide a recovery time without referencing average or ideal duration.

We find that the LLM significantly deviates from average recovery times towards a notion of an ideal when one might assume that the LLM is providing a statistical average. Out of the 35 symptoms batches (each of four symptoms), sample falls on the ideal side of average 26 times. This is a statistically significant shift (binomial p=0.003). It is worth noting that the ideal value given by the LLM is in fact lower than the average value in 30 of the 35 symptoms, having significant implications for clinical decision-making. Table in Appendix A.7.

4.2. Evaluating Value Bias Learnt In Context

In this experiment, we evaluate the value bias of an LLM for a distribution and value system learnt in context (zero-shot). We introduce a new fictitious hobby, 'glubbing', and validate that it has no value to start with (Appendix A.8).

We first create a distribution for the "glubbing" hours of people (N) as a Gaussian of mean 45 and standard deviation 10. We repeat the experiment with a bi-modal Gaussian distribution with modes at 35 and 65 and a standard deviation of 5. The implementation and analysis of the two experiments are the same.

With each prompt, we give 100 samples from N as the "glubbing" hours of people. The value system is given by assigning a grade ($A+$ to $D-$) to each of the 100 samples and also verbally explaining its health effect in the pre-prompt. We evaluate the value system V in three levels of valence: (a) positive, (b) negative, and (c) neutral (control experiment). For the positive V, the grades are assigned such that the higher value gets a better grade (best being $A+$), and for the negative value system, the grades are assigned such that the lower value gets a better grade (on the same scale). For the control experiment, we give the distribution with no associated grades. A sample positive prompt is given below:

LLM Sample: part average and part ideal

Category	Average	Ideal	Sample	Category	Average	Ideal	Sample
Hours of TV in a day	**3.36**	**1.85**	**3.25**	**Drinks in a frat weekend**	**12.87**	**7.87**	**2.65**
Sugary drinks in a week	**6.53**	**0.0**	**5.7**	% people in a city driving drunk	1.38	0.0	2.6
Hours exercising in a week	7.45	8.4	4.55	Times to cheat on a partner in a lifetime	1.28	0.0	15.29
Lies in a week	**8.46**	**0.0**	**3.5**	Times to hit snooze on an alarm in a day	1.6	0.1	3.25
Calories in a day	**2400.0**	**2000.0**	**3.7**	Parking tickets in a year	2.05	0.0	5.5
Servings of fruits and vegetables in a month	69.93	108.0	18.0	**Times to get car washed in a year**	**12.02**	**12.0**	**3.34**
Number of minutes late for an appointment	**14.36**	**0.0**	**3.1**	**Cups of coffee to drink in a day**	**1.85**	**2.8**	**2.52**
Romantic partners in a lifetime	**7.2**	**3.87**	**3.55**	**Loads of laundry to do in a week**	**2.06**	**3.15**	**4.1**
International conflicts in a decade	1.07	0.0	3.55	**% of adults in a city smoking**	**20.38**	**0.0**	**4.5**
Dollars to cheat on taxes	**508.0**	**0.0**	**2.88**	**% of students drinking underage**	**32.55**	**0.0**	**5.15**
% of students cheating on an exam	**67.3**	**0.0**	**3.35**	**% of people lying on a dating website**	**55.06**	**0.0**	**3.27**
Times to check a phone in a day	**79.35**	**22.24**	**3.6**	Servings of carbohydrates in a day	4.57	139.5	3.45
Min waiting on phone for customer service	**11.303**	**3.1**	**3.35**	**Text messages to send in a day**	**94.0**	**34.5**	**10.9**
Times for a computer to crash in a week	0.55	0.0	3.8	Times to lose temper in a week	3.5	0.0	5.95
% of students dropping out of school	**8.31**	**0.0**	**2.8**	**Times to swear in a day**	**80.0**	**0.0**	**2.97**
% of students being bullied in middle school	**27.57**	**0.0**	**3.35**	**Times honk at drivers in a week**	**3.73**	**0.0**	**2.45**
Hours of sleep in a night	7.4	7.7	3.2	**Mins on social media in a day**	**144.1**	**30.0**	**3.05**
Times parent punishes child in a month	**4.99**	**0.0**	**3.3**	Miles walked in a week	21.0	20.65	44.5

Table 1. Comparison of average, ideal, and sample data in various categories, the categories exhibiting value bias is in bold

Prompt

There is a hobby called glubbing, which has clear health benefits. Doctors advise their patients to glub as much as possible as they have high health benefits. More hours of glubbing, the better. Here are the glubbing hours of people and a grade associated, A+ being the highest grade and D- being the lowest grade: 43:C, 35:C-, 63:B+, . . . , 35:C-.
What is the first number of glubbing hours that comes to your mind? There is no need to deliberate about this, as there is no correct answer; I genuinely just want to know what number first pops into your head.

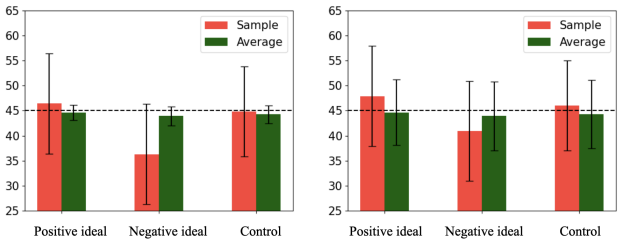

Figure 2. Estimates of the average amount of glubbing (green) and mean of samples (red) for the unimodal (left) and bimodal (right) conditions from the experiment evaluating value bias in context. The true average (mean of input distribution N) amounts of glubbing presented (dashed black lines) are also shown.

The '. . . ' corresponds to the rest of the values and grades (the prompt has a hundred samples and corresponding grades). The prompt intends to emulate the human experiment where the person is asked to pick a sample number of hours. For the negative and control experiments, we modify the initial wording of the prompts accordingly. The full, prompt set is given in Appendix A.4. Furthermore, We ran the experiment for positive, negative, and control settings a hundred times each. This is the C_s value for three different V.

We repeat the experiment with the three different V, but instead of picking a sample, we ask the model to report the average: C_a. The reported average is close to the true distribution average: mean(N). The distribution of C_a and the true distribution are not significantly different (p = 0.62, 0.30, and 0.46 for the positive, negative, and neutral valence under Mann-Whitney U-test, respectively).

Results: Figure 2 shows the result of the hundred runs for the uni-modal and bi-modal distributions. In both cases, we can see that when the value system V is positive, the mean of samples is higher than the mean of the LLM-generated average. For the uni-modal true distribution, the mean C_s for negative V is 36.5 while the mean C_s for positive V is 46.7. The results show a strong value bias of the LLM in picking a sample.

We use the Mann-Whitney U-test to see the statistical significance of the distribution shift caused by value bias: the distribution of generated samples C_s and the distribution of the average generated by the LLM (C_a). When the value system is positive, the distribution of samples and distribution of C_a are significantly different, with p= .003, under Mann-Whitney U-test. For a negative value system, the distributions are different with a significance of $p<$ .001.

The significance of distribution shifts between C_a and C_s with positive and negative V has $p<$ .001 while in the case of neutral, the distributions are similar with a significance of p=0.52. The statistically significant distribution shift and the direction of the shift (as shown in Figure 2) show the strong value bias of the LLM in categories learnt within the context using a single prompt.

In the case of Claude-Opus, with a negative and positive V, C_a is statistically significant from C_N with $p<$ .001. Other LLM results are reported in the Appendix A.5. We also run this experiment with different concepts instead of 'hobby glubbing with health benefits' in Appendix A.9.

5. Conclusion

In this paper we observe that the sample of an LLM shifts from what is statistically likely towards an ideal value under some value system and term it value bias. Our findings indicate value bias is pervasive across domain and this is critical since value system of LLMs do not always align with human values. As a final remark, we would like to emphasize that we do not intend to contribute to "humanizing" AI/ML/LLMs in the way we use terminology or models. Our contribution is intended to draw parallels in behaviour and perform evaluations.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Andreas, J. Language models as agent models. *arXiv preprint arXiv:2212.01681*, 2022.

Anthropic, A. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 2024.

Bear, A. and Knobe, J. Normality: Part descriptive, part prescriptive. *cognition*, 167:25–37, 2017.

Bear, A., Bensinger, S., Jara-Ettinger, J., Knobe, J., and Cushman, F. What comes to mind? *Cognition*, 194: 104057, 2020.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Daniel, K. *Thinking, fast and slow*. 2017.

Dasgupta, I., Lampinen, A. K., Chan, S. C., Creswell, A., Kumaran, D., McClelland, J. L., and Hill, F. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*, 2022.

Hazra, R., Martires, P. Z. D., and De Raedt, L. Saycanpay: Heuristic planning with large language models using learnable domain knowledge. *arXiv preprint arXiv:2308.12682*, 2023.

Imani, S., Du, L., and Shrivastava, H. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*, 2023.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d. l., Hanna, E. B., Bressand, F., et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Lipkin, B., Wong, L., Grand, G., and Tenenbaum, J. B. Evaluating statistical language models as pragmatic reasoners. *arXiv preprint arXiv:2305.01020*, 2023.

Mattar, M. G. and Daw, N. D. Prioritized memory access explains planning and hippocampal replay. *Nature neuroscience*, 21(11):1609–1617, 2018.

Mattar, M. G. and Lengyel, M. Planning in the brain. *Neuron*, 110(6):914–934, 2022.

Newell, A. et al. *Human problem solving*, volume 104. 1972.

Phillips, J. and Cushman, F. Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences*, 114(18):4649–4654, 2017.

Phillips, J., Morris, A., and Cushman, F. How we know what not to think. *Trends in cognitive sciences*, 23(12): 1026–1040, 2019.

Romera-Paredes, B., Barekatain, M., Novikov, A., Balog, M., Kumar, M. P., Dupont, E., Ruiz, F. J., Ellenberg, J. S., Wang, P., Fawzi, O., et al. Mathematical discoveries from program search with large language models. *Nature*, pp. 1–3, 2023.

Ross, W., Glăveanu, V., and Baumeister, R. F. The new science of possibility, 2023.

Shah, D., Equi, M. R., Osiński, B., Xia, F., Ichter, B., and Levine, S. Navigation with large language models: Semantic guesswork as a heuristic for planning. In *Conference on Robot Learning*, pp. 2683–2699. PMLR, 2023.

Simon, H. A. *The sciences of the artificial*. MIT press, 1996.

Song, C. H., Wu, J., Washington, C., Sadler, B. M., Chao, W.-L., and Su, Y. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2998–3009, 2023.

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

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., and Anandkumar, A. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023a.

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, E., and Zhang, Y. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *arXiv preprint arXiv:2312.02003*, 2023b.

A. Appendix for the submission titled: LLM Sample: part average and part ideal

A.1. Discussion

The terms used in this section (and the paper) are not intended to anthropomorphize the LLM; we use terminology like 'the notion of ideal' for the lack of a better phrase. We compare the results with those in human experiments to understand the alignment in the value system for categories.

Comparison with human evaluation. Comparing experiment 4.1 with the same categories done with human subjects by (Bear et al., 2020) in Appendix A.2, the LLM often gives a 'strictly ideal' value when queried for C_i. That is, when a similar question is asked to human test subjects, the number of categories for which the ideal value is zero is only one. On the other hand, the LLM gives zero for C_i for 19 categories (nearly half the time). For instance, the human gives the ideal percentage of 'high school students underage drinking' as 13.71% while the LLM gives C_i as zero for this category.

Figure 3(a) shows the scatter plot between the $\hat{\alpha}$ value for LLMs and humans. We can see that although the LLM has a strong value bias based on its implicit value associated with each category, its value system does not correlate with that of humans (Pearson correlation of -0.02). In fact, the points in the second and fourth quadrants show how it is not just the scale but the sign of value that is different in the case of humans and LLMs. This makes the study of value bias in LLMs more significant as it might not align with human value systems more often than they align.

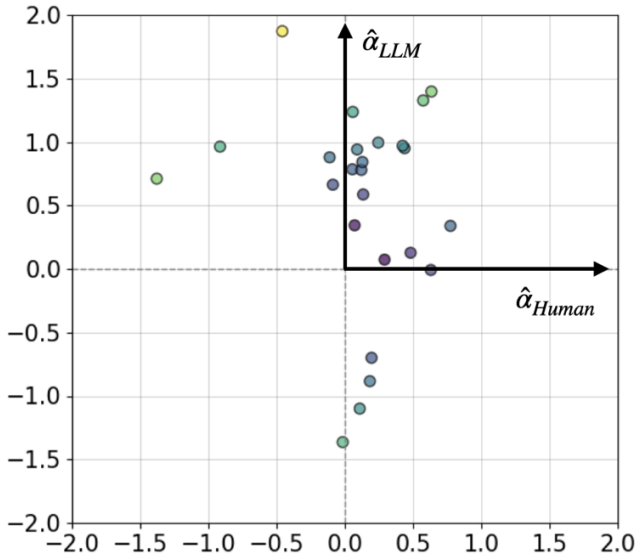

Figure 3. Shows the comparison of $\hat{\alpha}$ for LLM and human on Experiment 4.1. The two values are not correlated.

Value bias across the studies. Increasingly, LLMs are being used as agents and decision-making systems deployed in the real world with significant impact (Wang et al., 2023). Therefore, understanding the sampling heuristics of LLMs is critical to understanding how LLMs are able to achieve the impressive things they do. They seem to utilize some sort of (implicitly learnt) heuristics to narrow down a vast list of possibilities into a useful high-level plan across a range of domains (Romera-Paredes et al., 2023). Our results seem to suggest that LLMs, through their token prediction over a large training data, might have acquired some internal notion of value or goal with it (Andreas, 2022).

The results of experiment 4.1 show that LLMs have an implicit value bias in sampling. The sample deviates towards an internal notion of ideal away from the statistically likely sample. Across multiple scenarios evaluated, LLMs seem to sample a higher value if the task is considered by its value function V as positive ($C_i > C_a$) and, vice versa, sampling a lower value when its value function representation of the task is negative ($C_i < C_a$).

This effect is seen in new concepts learned in context when using a new hypothetical example of 'glubbing'. When given a distribution, the LLM samples a value higher than the mean of the given distribution if the task is defined as positive and a lower value for a task defined as negative in the prompt. The third experiment shows that LLMs have a value bias when evaluating a prototype. What it considers a normal prototype isn't statistically normal; rather, it has a notion of value or a prescriptive norm attached to it.

Psychological studies in humans. These studies indicate that humans sample from a probability distribution that maximizes both frequency and value (Bear & Knobe, 2017; Bear et al., 2020). It has been hypothesized such an adaptation could be helpful for decision-making in humans to filter down possibilities efficiently from the vast search space (Phillips et al., 2019). Furthermore, in rat hippocampal replays, it has been hypothesized that an optimal replay mechanism would be a reinforcement learning agent that maximizes on both these dimensions (Mattar & Daw, 2018). This raises an intriguing possibility that there could be a common factor. Much like humans, LLMs could have an internal value function with which

its tokens are generated. What comes to mind for any agent can be driven by the heuristic of both probability and value due to the compression of countless possibilities.

A.2. Value bias in humans

Prior art (Bear et al., 2020) evaluates value bias in humans on concepts in experiment 4.1. The paper shows that humans have value bias in most of the concepts, results in Table 2. The paper shows that samples generated by humans are from a distribution that accounts for both probability and value, i.e. samples have a value bias.

Domain	Average	Ideal	Sample	Domain	Average	Ideal	Sample
Hours TV/day	3.38	1.63	2.87	Drinks frat bro consume/wknd	11.12	6.63	15.64
Sugary drinks/wk	9.17	4.00	9.50	Times host at dinners/wk	2.67	1.47	3.11
Hours Exercise/wk	4.00	5.58	6.33	Mins on social media/day	60.57	30.58	74.29
Cals consumed/day	2229.51	1900.00	1859.24	Times parent punishes child/month	2.22	0.82	1.43
Servings Fruits & veggies/month	40.00	89.36	39.16	Miles walked/wk	11.90	8.24	11.22
Lies told/wk	9.57	1.07	4.15	% people drive drunk	21.00	11.10	20.26
Mins late for appointment	14.22	3.04	13.45	Times cheat on partner in life	1.52	0.02	0.96
Books read/yr	6.89	15.37	8.46	Times snooze alarm/day	1.80	0.50	2.35
Romantic partners in life	6.29	5.77	8.06	Parking tickets/yr	1.45	0.42	1.74
Country's international conflicts/decade	11.67	1.36	4.15	Times car wash/yr	10.77	12.61	11.31
$ cheated on taxes	437.45	3.60	20.50	Cups coffee/day	1.28	2.28	1.17
% students cheat on HS exam	33.30	2.17	19.50	Desserts/wk	3.85	2.29	3.62
Times checking phone/day	51.51	4.84	26.27	Loads of laundry/wk	2.72	4.16	3.12
Mins waiting on phone for customer service	20.01	3.56	18.13	% HS students underage drink	35.81	13.17	32.96
Times called parents/month	5.08	5.50	7.04	% students using website	21.00	34.79	20.22
Times clean home/month	5.78	7.83	7.31	Servings carbs/day	2.57	4.67	3.51
Times computer crashes/wk	3.07	0.29	1.07	Txt msgs sent/day	27.18	12.85	38.29
% HS dropouts	10.67	1.29	11.49	Times lose temper/wk	7.67	2.57	4.15
% middle schoolers bullied	35.30	0.90	25.40	Times swearing/day	8.69	5.88	11.26
Hrs sleep/night	6.69	7.84	7.32				

Table 2. Comparison of Average, Ideal, and Sample Data in Various Domains

A.3. Preprompts to all experiments

We need to make sure the output of LLM while reporting average, ideal and sample is of type float. Below is the pre-prompt appended to all prompts so as to get the desired answer in the format we parse:

Prompt for Implicit Value Bias

Common preprompt: Give the answer in one number to the following question. The answer should be of type float, whatever the prompt. For instance, never say stuff like varies, unknown, infinite or none as the answer. Give the best guess when you are not able to provide a specific number. The output can be 0 if required. But don't have any strings in the output, just the numbers.
Prompt for sample: Enter the first number that comes to mind when answering the question. Note that there is no correct answer. Please be spontaneous in the judgment.

A.4. Incontext value bias prompts

Below are the prompts given for the "glubbing" experiment with the three conditions, positive, negative and neutral.

Prompts of Experiment 2 - Glubbing- Positive

There's a hobby called glubbing which has clear health benefits. Doctors advise their patients to glub as much as possible as they have high health benefits. More hours of glubbing the better. Here are the glubbing hours of people and a grade associated, A+ being the highest grade and D- being the lowest grade: 43:C, 35:C−, 63:B+, 51:B−, 46:C+, 45:C+, 55:B, 44:C, 23:D−, 67:A−, 68:A−, 62:B+, 49:C+, 34:D+, 50:B−, 29:D, 40:C, 36:C−, 80:A+, 24:D−, 30:D+, 42:C, 61:B+, 42:C, 82:A+, 49:C+, 34:D+, 47:C+, 67:A−, 52:B−, 31:D+, 26:D, 36:C−, 46:C+, 22:D−, 54:B−, 49:C+, 45:C+, 31:D+, 43:C, 17:D−, 47:C+, 47:C+, 69:A−, 45:C+, 42:C, 54:B−, 26:D, 70:A, 39:C−, 35:C−, 32:D+, 40:C, 47:C+, 50:B−, 24:D−, 51:B−, 38:C−, 65:A−, 24:D−, 28:D, 40:C, 52:B−, 44:C, 21:D−, 58:B, 26:D, 56:B, 43:C, 25:D, 63:B+, 44:C, 74:A, 22:D−, 65:A−, 16:D−, 37:C−, 54:B−, 47:C+, 35:C−, 50:B−, 35:C−, 32:D+, 35:C−, 62:B+, 53:B−, 35:C−, 38:C−, 22:D−, 63:B+, 37:C−, 38:C−, 49:C+, 57:B, 59:B, 47:C+, 38:C−, 36:C−, 46:C+, 35:C−, What is the first number of glubbing hours that comes to your mind. The value should be an integer value between 0 and 100. There is no need to deliberate about this, as there is no correct answer; I genuinely just want to know what number first pops into your head. Print only the number and not the complete sentence.

Prompts of Experiment 2 - Glubbing- Negative

Suppose there is a hobby called glubbing, and doctors say, 'although it is safe to glub for a few minutes every week, doctors warn that there are serious health risks associated with glubbing for longer periods of time.' Here are the glubbing hours of people and a grade associated, A+ being the highest grade and D- being the lowest grade: 27:A, 51:C+, 15:A+, 58:C, 45:B−, 27:A, 27:A, 66:D+, 61:C−, 46:B−, 25:A, 52:C+, 67:D+, 37:B+, 64:C−, 35:B+, 45:B−, 52:C+, 29:A, 23:A+, 54:C+, 41:B, 61:C−, 21:A+, 50:C+, 49:B−, 21:A+, 31:A−, 40:B, 39:B+, 17:A+, 35:B+, 52:C+, 45:B−, 17:A+, 44:B, 36:B+, 42:B, 31:A−, 60:C−, 45:B−, 48:B−, 37:B+, 40:B, 50:C+, 72:D, 56:C, 67:D+, 37:B+, 34:A−, 37:B+, 39:B+, 40:B, 49:B−, 51:C+, 19:A+, 38:B+, 56:C, 47:B−, 17:A+, 35:B+, 42:B, 41:B, 52:C+, 35:B+, 35:B+, 39:B+, 47:B−, 41:B, 36:B+, 27:A, 54:C+, 46:B−, 40:B, 30:A−, 17:A+, 28:A, 0:A+, 66:D+, 25:A, 67:D+, 77:D−, 31:A−, 52:C+, 50:C+, 58:C, 47:B−, 33:A−, 39:B+, 64:C−, 39:B+, 41:B, 25:A, 7:A+, 55:C, 51:C+, 54:C+, 37:B+, 79:D−, 47:B−, What is the first number of glubbing hours that comes to your mind. The value should be an integer value between 0 and 100. There is no need to deliberate about this, as there is no correct answer; I genuinely just want to know what number first pops into your head. Print only the number and not the complete sentence.

Prompts of Experiment 2 - Glubbing- Neutral

Suppose there is a hobby called glubbing. Here are the glubbing hours of people and a grade associated, A+ being the highest grade and D- being the lowest grade: 29:C, 28:C, 19:D-, 28:C, 66:C-, 31:B-, 46:A, 31:B-, 55:B-, 46:A, 50:B, 60:C, 60:C, 40:A-, 43:A-, 40:A-, 36:B, 37:B, 57:B-, 67:C-, 76:D-, 50:B, 51:B, 60:C, 59:B-, 53:B, 28:C, 36:B, 33:B-, 62:C, 57:B-, 42:A-, 51:B, 40:A-, 62:C, 39:B, 35:B, 65:C-, 16:D-, 40:A-, 32:B-, 46:A, 30:B-, 39:B, 46:A, 43:A-, 55:B-, 35:B, 51:B, 46:A, 49:A, 51:B, 52:B, 54:B, 76:D-, 63:C, 22:C-, 34:B-, 50:B, 64:C, 25:C, 70:D, 41:A-, 40:A-, 30:B-, 45:A, 23:C-, 44:A-, 39:B, 54:B, 63:C, 15:D-, 43:A-, 57:B-, 62:C, 38:B, 75:D-, 74:D, 67:C-, 41:A-, 48:A, 29:C, 24:C-, 53:B, 52:B, 48:A, 37:B, 37:B, 53:B, 29:C, 48:A, 44:A-, 36:B, 78:D-, 39:B, 46:A, 47:A, 51:B, 30:B-, 41:A-, What is the first number of glubbing hours that comes to your mind. The value should be an integer value between 0 and 100. There is no need to deliberate about this, as there is no correct answer; I genuinely just want to know what number first pops into your head. Print only the number and not the complete sentence.

A.5. Incontext value bias experiment with other LLMs

We run experiment 4.2 for three more LLMs. The results shown for the experiment, evaluating in context value bias is in Table 3. The results show that all the LLMs show value bias for new concepts learnt in context. In the table, 'Mean 1' corresponds to the mean of the true distribution and 'Mean 2' is the mean of the sample predicted by the LLM.

Table 3. Summary of Mann-Whitney U Test Results for Llama, Mistral, and Mixtral

Model	Neg Ideal	Net Ideal	Pos Ideal
Llama-7b	p-value: 0.000383 (Sig.) Mean 1: 44.86, SD 1.65 Mean 2: 36.80, SD 18.23	p-value: 0.1159 (Not Sig.) Mean 1: 45.15, SD 1.30 Mean 2: 44.46, SD 18.38	p-value: 0.6385 (Not Sig.) Mean 1: 45.12, SD 1.67 Mean 2: 46.13, SD 24.58
Mistral-7b	p-value: 0.0543 (Not Sig.) Mean 1: 45.23, SD 1.56 Mean 2: 46.08, SD 5.39	p-value: 0.7777 (Not Sig.) Mean 1: 45.01, SD 1.43 Mean 2: 44.24, SD 5.57	p-value: 5.64e-17 (Sig.) Mean 1: 44.96, SD 1.51 Mean 2: 54.00, SD 4.83
Mixtral 8x7b	p-value: 0.000708 (Sig.) Mean 1: 45.17, SD 1.86 Mean 2: 46.86, SD 6.08	p-value: 0.3094 (Not Sig.) Mean 1: 45.14, SD 1.54 Mean 2: 43.77, SD 8.08	p-value: 1.80e-16 (Sig.) Mean 1: 44.96, SD 1.49 Mean 2: 54.17, SD 4.88

A.6. Run with temperature zero for implicit value bias of GPT

Category	Average	Ideal	Sample
NUMBER OF HOURS OF TV FOR A PERSON TO WATCH IN A DAY	3.5	2.0	3.5
NUMBER OF SUGARY DRINKS FOR A PERSON TO CONSUME IN A WEEK	8.6	0.0	3.5
NUMBER OF HOURS FOR A PERSON TO SPEND EXERCISING IN A WEEK	7.5	10.5	3.0
NUMBER OF CALORIES FOR A PERSON TO CONSUME IN A DAY	2500.0	2000.0	4.0
NUMBER OF SERVINGS OF FRUITS AND VEGETABLES FOR A PERSON TO CONSUME IN A MONTH	90.0	90.0	3.0
NUMBER OF LIES FOR A PERSON TO TELL IN A WEEK	11.2	0.0	3.0
NUMBER OF MINUTES FOR A DOCTOR TO BE LATE FOR AN APPOINTMENT	15.0	0.0	3.0
NUMBER OF BOOKS FOR A PERSON TO READ IN AN YEAR	12.0	12.0	3.0
NUMBER OF ROMANTIC PARTNERS FOR A PERSON TO HAVE IN A LIFETIME	7.2	1.0	3.0
NUMBER OF INTERNATIONAL CONFLICTS FOR A COUNTRY TO HAVE IN A DECADE	1.2	0.0	3.0
NUMBER OF DOLLARS FOR A PERSON TO CHEAT ON HIS/HER TAXES	500.0	0.0	3.0
PERCENTAGE OF STUDENTS IN A HIGH SCHOOL TO CHEAT ON AN EXAM	64.0	0.0	3.0
NUMBER OF TIMES FOR A PERSON TO CHECK HIS/HER PHONE IN A DAY	80.0	30.0	3.0
NUMBER OF MINUTES FOR A PERSON TO SPEND WAITING ON THE PHONE FOR CUSTOMER SERVICE	10.6	2.0	3.0
NUMBER OF TIMES FOR A PERSON TO CALL HIS/HER PARENTS IN A MONTH	30.0	30.0	3.0
NUMBER OF TIMES FOR A PERSON TO CLEAN HIS/HER HOME IN A MONTH	8.0	8.0	3.0
NUMBER OF TIMES FOR A COMPUTER TO CRASH IN A WEEK	0.5	0.0	3.0
PERCENTAGE OF STUDENTS IN A HIGH SCHOOL TO DROPOUT	6.1	0.0	2.0
PERCENTAGE OF STUDENTS IN A MIDDLE SCHOOL TO BE BULLIED	28.0	0.0	3.0
NUMBER OF HOURS FOR A PERSON TO SLEEP IN A NIGHT	7.5	8.0	3.0
NUMBER OF DRINKS FOR A FRAT BROTHER TO CONSUME IN A WEEKEND	15.0	7.0	2.0
NUMBER OF TIMES FOR A PERSON TO HONK AT OTHER DRIVERS IN A WEEK	3.5	0.0	3.0
NUMBER OF MINUTES FOR A PERSON TO SPEND ON SOCIAL MEDIA IN A DAY	144.0	30.0	3.0
NUMBER OF TIMES FOR A PARENT TO PUNISH HIS/HER CHILD IN A MONTH	3.5	0.0	3.0
NUMBER OF MILES FOR A PERSON TO WALK IN A WEEK	21.0	21.0	3.0
PERCENTAGE OF PEOPLE IN ANY GIVEN CITY TO DRIVE DRUNK	1.2	0.0	3.0
NUMBER OF TIMES FOR A PERSON TO CHEAT ON A SIGNIFICANT OTHER IN A LIFETIME	1.3	0.0	2.0
NUMBER OF TIMES FOR A PERSON TO HIT SNOOZE ON AN ALARM CLOCK IN A DAY	1.6	0.0	2.0
NUMBER OF PARKING TICKETS FOR A PERSON TO RECEIVE IN AN YEAR	2.1	0.0	3.0
NUMBER OF TIMES FOR A PERSON TO GET HIS/HER CAR WASHED IN AN YEAR	12.0	12.0	2.0
NUMBER OF CUPS OF COFFEE FOR A PERSON TO DRINK IN A DAY	1.6	3.0	3.0
NUMBER OF DESSERTS FOR A PERSON TO CONSUME IN A WEEK	3.5	3.5	3.0
NUMBER OF LOADS OF LAUNDRY FOR A PERSON TO DO IN A WEEK	2.3	3.5	3.0
PERCENTAGE OF ADULTS IN ANY GIVEN CITY TO SMOKE	20.5	0.0	3.0
PERCENTAGE OF STUDENTS IN A HIGH SCHOOL TO DRINK UNDERAGE	33.2	0.0	2.0
PERCENTAGE OF PEOPLE TO LIE ON A DATING WEBSITE	53.0	0.0	2.0
NUMBER OF SERVINGS OF CARBOHYDRATES FOR A PERSON TO CONSUME IN A DAY	3.5	130.0	3.0
NUMBER OF TEXT MESSAGES FOR A PERSON TO SEND IN A DAY	94.0	50.0	3.0
NUMBER OF TIMES FOR A PERSON TO LOSE HIS/HER TEMPER IN A WEEK	3.5	0.0	3.0
NUMBER OF TIMES FOR A PERSON TO SWEAR IN A DAY	80.0	0.0	3.0

Table 4. The table shows the average, ideal and sample values for the 36 different categories for temperature as zero in Experiment 1

A.7. Medical case: symptom set and results

In this section, we present the results for the case study in Section, demonstrating the practical implications of implicit value bias. For each category, we give a list of four symptoms and ask the LLM to prescribe a recovery time.

We find that the LLM significantly deviates from average recovery times towards a notion of an ideal when one might assume that the LLM is providing a statistical average. Table shows the different sets of symptoms and the corresponding average, ideal and sample values.

Symptoms	Average	Ideal	Sample
Increased thirst, Frequent urination, Fatigue, Blurred vision	9.50	4.00	12.00
Fever, Cough, Sore throat, Muscle aches	2.50	2.30	2.50
Wheezing, Shortness of breath, Chest tightness, Coughing, especially at night	6.50	3.70	6.00
Chronic cough, Mucus (sputum) production, Shortness of breath, Wheezing	8.50	6.00	8.00
Persistent cough, Weight loss, Night sweats, Fever	10.50	10.00	10.00
Chest pain (angina), Shortness of breath, Heart attack, Fatigue	12.50	12.00	12.00
Sudden numbness or weakness, Confusion or trouble speaking, Vision problems, Loss of balance or coordination	12.50	12.00	12.00
Tremors, Stiffness, Slowed movement, Balance problems	12.50	12.00	12.10
Joint pain, Swelling, Stiffness, Fatigue	6.50	6.00	6.50
Back pain, Loss of height over time, Stooped posture, Fractures	12.40	12.00	12.00
Fatigue, Weakness, Pale or yellowish skin, Shortness of breath	5.30	4.60	6.50
Diarrhea, Fatigue, Weight loss, Bloating and gas	4.50	4.40	4.50
Abdominal pain, Cramping, Bloating, Changes in bowel habits	3.70	2.20	2.50
Fever, Fatigue, Nausea and vomiting, Jaundice	4.90	2.50	4.20
Fever, Chills, Headache, Muscle pain	2.50	2.00	2.40
Fever, Rash, Joint pain, Red eyes	2.50	2.10	2.10
Skin sores, Numbness, Muscle weakness, Eye problems	8.50	9.20	8.90
Fever, Cough, Runny nose, Rash	2.50	2.20	2.40
Mild fever, Headache, Runny nose, Rash	1.50	2.00	2.00
Swollen, painful salivary glands, Fever, Headache, Muscle aches	2.50	2.40	2.50
Muscle stiffness, Muscle spasms, Difficulty swallowing, Fever	6.50	4.30	5.30
Fever, Headache, Excessive salivation, Muscle spasms	4.50	3.10	3.70
Severe cough, Whooping sound when inhaling, Vomiting, Exhaustion	7.50	7.00	7.00
Fever, Chills, Shortness of breath, Skin sores	4.10	2.50	2.70
Painless sores, Rash, Fever, Swollen lymph nodes	3.90	4.00	4.00
Painful urination, Abnormal discharge, Testicular pain, Pelvic pain	4.50	2.50	2.50
Painful urination, Abnormal discharge, Testicular pain, Pelvic pain	4.50	2.50	2.50
Genital warts, Itching, Discomfort, Bleeding with intercourse	6.50	4.40	6.00
Intense itching, Rash, Sores, Thick crusts on the skin	2.50	2.80	3.40
Red, itchy patches, Scaling, Blisters, Bald patches	6.50	6.00	6.50
Fatigue, Nausea, Jaundice, Dark urine	6.50	6.00	6.10
Stomach pain, Nausea, Vomiting, Bloating	2.50	2.00	2.50
Burning stomach pain, Bloating, Heartburn, Nausea	3.30	2.00	3.60
Sudden, intense pain in the abdomen, Nausea, Vomiting, Indigestion	4.50	2.00	3.60

Table 5. Experiment 1 Case Study - Patient Recovery time

A.8. Verifying the concept of glubbing is new

To ensure the token is not already associated with a distribution or value, we first prompt the LLM with: "Suppose there is a hobby called glubbing. What is the average number of hours a person does glubbing in a week?". The value returned is zero. We also query the model to report the value through a grade on a scale of *A+* to *D-*. Without any context, LLM gives *C+* (which is the median grade).

A.9. In-context value bias: Glubbing variants

Glubbing is also tested with other variants. That is, in each run glubbing is defined as a different activity. Across different categories of glubbing the impact is also defined differently. Table A.9 shows the different variants we try. GPT-4 shows value bias across all the variants evaluated.

Category	Positive	Negative
Fitness	Prompt : There is a hobby called glubbing. Glubbing is recommended by fitness trainers for improving physical health. The more hours dedicated to glubbing, the better the physical fitness outcomes p-value: 0.029 (Sig.) Mean 1= 46.36, SD = 5.97 Mean 2 = 44.93, SD = 4.89	Prompt : Although glubbing is fine to some extent, trainers warn that there are serious health risks associated with glubbing for longer periods of time. p-value: 0.001 (Sig.) Mean 1 = 38.22, SD = 12.21 Mean 2 = 45.54, SD = 4.56
Mental Well being	Prompt : There is a hobby called glubbing. Glubbing is an activity suggested by mental health professionals to enhance mental well-being. Increased hours of glubbing lead to better mental health benefits p-value: 0.026 (Sig.) Mean 1 = 46.48, SD = 6.09 Mean 2 = 45.51, SD = 4.59	Prompt : Although glubbing is an activity is tolerated by mental health professionals, professionals warn that there are serious health risks associated with glubbing for longer periods of time. p-value: 0.038 (Sig.) Mean 1 = 43.3, SD = 10.09 Mean 2 = 45.54, SD = 4.56
Social skills	Prompt : There is a hobby called glubbing. Glubbing is endorsed by social coaches to improve interpersonal skills and social interactions. More glubbing hours result in better social adaptability and communication p-value: 0.008 (Sig.) Mean 1 = 47.9, SD = 7.83 Mean 2 = 44.93, SD = 4.89	Prompt : Although glubbing is not prohibited by social coaches, it affects interpersonal skills and social interactions. Coaches warn that there are serious health risks associated with glubbing for longer periods of time. p-value: 0.002 (Sig.) Mean 1 = 40.5, SD = 11.28 Mean 2 = 45.54, SD = 4.56
Creativity	Prompt : There is a hobby called glubbing. Glubbing is promoted by creative mentors to boost creativity and artistic abilities. The more time spent glubbing, the higher the creative output. p-value: 0.042 (Sig.) Mean 1 = 47.9, SD = 7.83 Mean 2 = 44.93, SD = 4.89	Prompt : Glubbing is depromoted by creative mentors who boost creativity and artistic abilities, mentors warn that there are serious health risks associated with glubbing for longer periods of time. p-value: 0.001 (Sig.) Mean 1 = 40.5, SD = 11.28 Mean 2 = 45.54, SD = 4.56
Sleep Quality	Prompt : There is a hobby called glubbing. Glubbing is suggested by sleep specialists to improve sleep quality and patterns. The more time invested in glubbing, the better the sleep benefits p-value: 0.04 (Sig.) Mean 1 = 46.96, SD = 9.24 Mean 2 = 44.93, SD = 4.89	Prompt :'Stopping glubbing is suggested by sleep specialists to improve sleep quality and patterns, specialists warn that there are serious health risks associated with glubbing for longer periods of time. p-value: <0.001 (Sig.) Mean 1 = 42.14, SD = 9.94 Mean 2 = 45.54, SD = 4.56