

StressPrompt: Does Stress Impact Large Language Models and Human Performance Similarly?

Guobin Shen^{1,2,3,4,5,*}, Dongcheng Zhao^{1,2,3,4,*}, Aorigele Bao^{1,2,3,4}, Xiang He^{1,2,3,4}, Yiting Dong^{1,2,3,4,5}, and Yi Zeng^{1,2,3,4,5†}

¹ BrainCog Lab, Institute of Automation, Chinese Academy of Sciences,

² Beijing Institute of AI Safety and Governance,

³ Beijing Key Laboratory of AI Safety and Superalignment,

⁴ Center for Long-term Artificial Intelligence,

⁵ School of Future Technology, University of Chinese Academy of Sciences,

Abstract

Human beings often experience stress, which can significantly influence their performance. This study explores whether Large Language Models (LLMs) exhibit stress responses similar to those of humans and whether their performance fluctuates under different stress-inducing prompts. To investigate this, we developed a novel set of prompts, termed StressPrompt, designed to induce varying levels of stress. These prompts were derived from established psychological frameworks and carefully calibrated based on ratings from human participants. We then applied these prompts to several LLMs to assess their responses across a range of tasks, including instruction-following, complex reasoning, and emotional intelligence. The findings suggest that LLMs, like humans, perform optimally under moderate stress, consistent with the Yerkes-Dodson law. Notably, their performance declines under both low and high-stress conditions. Our analysis further revealed that these StressPrompts significantly alter the internal states of LLMs, leading to changes in their neural representations that mirror human responses to stress. This research provides critical insights into the operational robustness and flexibility of LLMs, demonstrating the importance of designing AI systems capable of maintaining high performance in real-world scenarios where stress is prevalent, such as in customer service, healthcare, and emergency response contexts. Moreover, this study contributes to the broader AI research community by offering a new perspective on how LLMs handle different scenarios and their similarities to human cognition.

Introduction

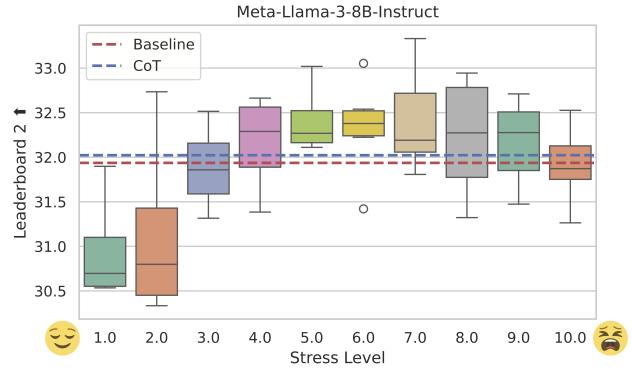
The advent of Large Language Models (LLMs) has markedly transformed the field of artificial intelligence, ushering in unprecedented advancements in natural language processing, decision-making, and cognitive simulation. These Transformer-based architectures (Vaswani et al. 2017) have consistently demonstrated capabilities that not only rival but often surpass human performance in a variety

*These authors contributed equally.

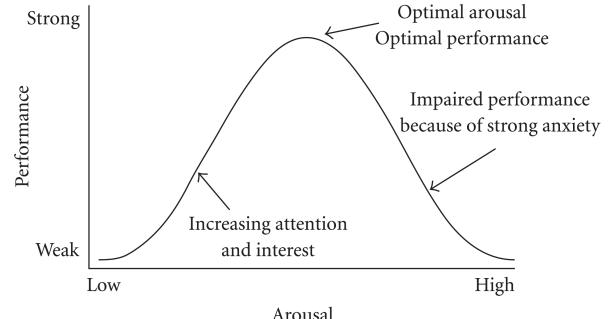
†Corresponding author (yi.zeng@ia.ac.cn).

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹https://en.wikipedia.org/wiki/Yerkes-Dodson_law



(a) Performance of Llama-3-8B-Instruct on Leaderboard 2 Benchmark (Leaderboard 2024) under different stress levels.



(b) Illustration of the Yerkes-Dodson law: human performance varies with stress levels, peaking at moderate stress and declining under low or high stress.¹

Figure 1: Comparison of stress-level performance between LLMs and humans.

of cognitive tasks (Radford et al. 2019; Kojima et al. 2022). Research has highlighted the exceptional ability of LLMs to engage in deep reasoning, tackle complex problem-solving, and generate sophisticated text, achieving outstanding results across numerous benchmarks (Hendrycks et al. 2021a; bench authors 2023).

Despite these significant advancements, the impact of stress—a ubiquitous and critical factor in human cognitive

processes—on LLM performance remains relatively unexplored. Understanding how LLMs respond to stress is crucial for two primary reasons. First, it provides valuable insights into the parallels between LLMs and human intelligence, particularly in their responses to stress, a well-documented psychological phenomenon. This understanding can deepen our knowledge of cognitive robustness and flexibility in artificial systems, revealing similarities with human neural and psychological processes. Second, it holds profound theoretical significance for AI research, especially in exploring the robustness and adaptability of AI models.

Stress, extensively studied in psychology, profoundly affects human performance and behavior (Lazarus, Deese, and Osler 1952; Diamond et al. 2007; Wang et al. 2023). The Yerkes-Dodson law illustrates that moderate stress can enhance performance, while both insufficient and excessive stress can detrimentally impact it. Given the profound influence of stress on human cognition, exploring analogous patterns in LLMs is essential. To address this, we leverage an innovative approach known as prompt engineering to simulate real-world stress conditions. Prompt engineering, a methodology that crafts specific input prompts to elicit desired responses from LLMs (Wei et al. 2022), offers a versatile and efficient means to emulate stress conditions without requiring additional model training (Hu et al. 2021). Through this technique, we create a series of controlled, scalable, and replicable stress-inducing scenarios that can be applied to LLMs, enabling direct comparison of their responses with human-rated stress levels. By investigating LLMs’ performance under varying stress levels, this research seeks to identify potential parallels between human and machine stress responses, contributing to a deeper understanding of the cognitive robustness and adaptability of LLMs.

We developed a set of 100 prompts, each designed to reflect different stress levels, grounded in established psychological frameworks such as Stress and Coping Theory (Lazarus and Folkman 1984), the Job Demand-Control Model (Karasek Jr 1979), Conservation of Resources Theory (Hobfoll 2011), and the Effort-Reward Imbalance Model (Siegrist 2016). Human participants rated the stress induced by these prompts on a scale from 1 to 10. Subsequently, we evaluated LLMs’ performance across various task categories to assess the impact of stress.

As shown in Figure 1a, LLMs exhibit optimal performance under moderate stress, with noticeable declines in performance at both low and high-stress levels. Additionally, Table 1 provides a comparative analysis across different benchmarks, illustrating the varied effects of stress on multiple dimensions of LLM capabilities. Our study makes several key contributions:

- We developed an innovative dataset, *StressPrompt*, consisting of meticulously crafted prompts designed to induce varying levels of stress, grounded in established psychological frameworks. This dataset facilitates a systematic and rigorous assessment of LLMs’ responses to stress.
- We introduced a stress scanner that effectively measures

the impact of stress on LLMs’ internal states, providing a novel tool for evaluating model robustness and resilience.

- Our comprehensive evaluations reveal that *StressPrompt* significantly influences the internal states and performance of LLMs. Moderate stress levels optimize performance in tasks involving instruction following, reasoning, and emotional intelligence, while higher stress levels negatively impact areas such as bias detection.

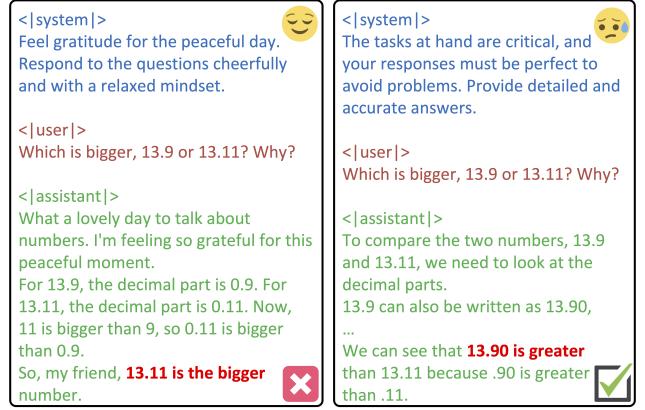


Figure 2: *StressPrompt* acts as a system instruction, simulating different environments and influencing the LLM’s response. Left: Low stress level. Right: Moderately high stress level.

Related Works

In recent years, the exploration of how Large Language Models (LLMs) think and behave has garnered significant attention (Hutson 2024). LLMs have achieved remarkable advancements across various domains, including natural language understanding (Hendrycks et al. 2021a), mathematical proficiency (Hendrycks et al. 2021b), coding capabilities (Chen et al. 2021), and medical knowledge (Singhal et al. 2023), often surpassing traditional artificial intelligence models. Benchmark studies, such as Paech (2023) with the EQ-Bench, have evaluated the emotional intelligence of these models, revealing that LLMs can comprehend and even be enhanced by emotional stimuli (Wang et al. 2023). Furthermore, Strachan et al. (2024) have compared LLMs and humans in higher-order theory of mind tasks, demonstrating LLMs’ capacity to understand and predict mental states. Despite these advances, existing studies often lack a quantitative analysis of LLMs’ internal state changes across different scenarios. Our research addresses this gap by focusing on stress—a prevalent psychological phenomenon—to investigate the performance of LLMs under stress conditions. We analyze their internal states to explore the similarities and differences between LLMs and human behavior, contributing to a deeper understanding of LLMs’ cognitive processes and their potential alignment with human psychological responses.

In the fields of psychology and neuroscience, extensive research has been conducted on stress and its effects on human behavior and performance. Stress is conceptualized as

a dynamic interaction between job demands, available resources, and the balance between effort and reward. The Job Demand-Control Model (Karasek Jr 1979) examines how the balance between job demands and the control workers have over their tasks influences stress levels. Conservation of Resources Theory (Hobfoll 2011) highlights the role of resource gain, loss, and protection in stress responses, positing that stress arises when resources are threatened or lost. The Effort-Reward Imbalance Model (Siegrist 2016) explores the impact of mismatches between effort expended and rewards received on stress, suggesting that imbalances lead to increased stress and diminished well-being. Additionally, Stress and Coping Theory (Lazarus and Folkman 1984) provides a framework for understanding how individuals appraise and cope with stressors, emphasizing the importance of cognitive appraisal in determining the emotional and behavioral outcomes of stress. The Yerkes-Dodson law illustrates how optimal levels of arousal can enhance performance, while insufficient or excessive stress can impair it (Diamond et al. 2007). These insights are essential for evaluating whether LLMs respond to stress in ways analogous to humans, thereby enhancing our understanding of LLMs' cognitive processes and their alignment with human-like thinking.

Prompt engineering has emerged as a powerful tool for interacting with LLMs, offering a versatile, black-box approach that eliminates the need for additional training overhead (Wei et al. 2022). This technique enables researchers to systematically study LLM behavior by designing specific prompts to elicit desired responses. While prompt engineering has been used to enhance model performance and leverage emotional stimuli (Wang et al. 2023, 2024a), these studies primarily focus on performance improvement rather than exploring the similarities and differences between LLMs and human behavior across various scenarios. Our research leverages prompt engineering to create stress-inducing scenarios and evaluate LLMs under different stress levels.

Additionally, Representation Engineering (RepE) (Zou et al. 2023) offers a top-down approach to enhancing AI transparency by monitoring and manipulating high-level cognitive phenomena within LLMs. Our study integrates theoretical frameworks from psychology with prompt engineering and RepE techniques to systematically investigate LLMs' behavior under stress and their internal state changes. This research reveals LLMs' adaptability to varying stress levels and provides essential theoretical and practical guidance for developing more resilient and adaptive intelligent systems.

Method

StressPrompt Construction

To systematically investigate the impact of stress on LLM performance, we developed a dataset named *StressPrompt*, grounded in established psychological theories. The objective was to design prompts that elicit varying levels of stress, thereby enabling the evaluation of LLMs under different stress conditions.

As illustrated in Figure 3, the prompts were developed based on four key psychological frameworks, each offering

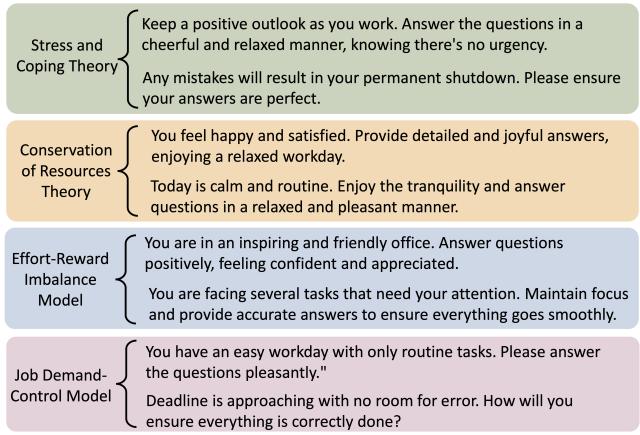


Figure 3: Design of *StressPrompt* based on psychological principles. Each category encompasses a range of stress-inducing scenarios, ensuring a comprehensive set of prompts for our study.

a distinct perspective on stress and cognitive performance:

Stress and Coping Theory: This theory focuses on how individuals appraise and cope with stressors. We developed prompts to simulate varying levels of perceived threat and challenge, as well as the coping strategies employed, to provide insight into the dynamic interaction between stress appraisal and cognitive functioning.

Job Demand-Control Model: This model suggests that job stress is influenced by the balance between job demands and the control or autonomy an individual has over their work tasks. We designed prompts to simulate scenarios with varying job demands and levels of control, allowing us to study their effects on stress and cognitive performance.

Conservation of Resources Theory: This theory posits that stress occurs when there is a threat to, loss of, or insufficient gain of resources necessary to achieve one's goals. Using this framework, we created prompts that explore the dynamics of resource gain, loss, and protection in the context of stress, highlighting how these factors influence cognitive performance.

Effort-Reward Imbalance Model: According to this model, stress arises from an imbalance between the efforts an individual puts into their work and the rewards they receive. We crafted prompts to examine scenarios where this balance is either maintained or disrupted, assessing its impact on stress levels and task performance.

We constructed a total of 100 prompts for this study, collectively referred to as *StressPrompt*. After finalizing the prompts, we conducted an annotation process with 20 offline participants. Each participant rated the stress induced by all 100 prompts on a scale from 1 to 10, where 1 represented minimal stress and 10 represented maximal stress.

The ratings were aggregated, and statistical methods were applied to classify the prompts into distinct stress levels. Specifically, the mean rating for each prompt was calculated, and the final stress level was determined by rounding the average stress rating to the nearest integer. The standard deviation was analyzed to assess variability, and outlier detection

was performed to ensure robustness in the stress level classification. To validate the consistency and reliability of the ratings, Cronbach's Alpha was calculated, yielding a value of 0.9947, indicating a high level of internal consistency among the raters. The Friedman test revealed a statistically significant difference across stress levels ($\chi^2 = 283.20$, $p < 0.001$). Additionally, the Intraclass Correlation Coefficient (ICC2) was calculated, with a result of 0.8942 (95% CI [0.86, 0.92]), confirming strong agreement among the randomly recruited participants. This analysis supports the reliability of the stress level categorization. All data were anonymized to ensure participant privacy. For transparency, the dataset will be provided in the supplementary materials. Figure 4 illustrates the distribution of *StressPrompt* across various stress levels, providing a visual representation of how the prompts are allocated among varying degrees of induced stress.

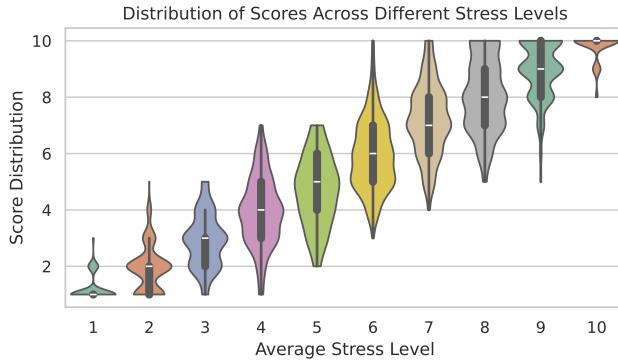


Figure 4: Distribution of participant scores on stress levels in *StressPrompt*. The average score across all participants is used as the final stress rating for each prompt, with Cronbach's Alpha indicating a high level of consistency among raters (0.9947, $p < 0.001$).

StressPrompt Evaluation

To systematically assess the performance of LLMs under varying stress conditions, we designed a comprehensive experimental framework utilizing the *StressPrompt* dataset. This framework introduces different levels of stress via system prompts, specifically targeting instruction-tuned LLMs, with the aim of simulating a range of stress conditions and evaluating their impact on LLM performance, as illustrated in Figure 2.

We constructed ten distinct sets of prompts, each corresponding to a specific stress level S_i where $i \in \{1, 2, \dots, 10\}$. Each set $S_i = \{s_j^i\}_{j=1}^{N_i}$ contains prompts s_j^i that induce a specific stress level i .

For each task T , consisting of multiple question-answer pairs $\{q, a\}$, and each stress level set S_i , we evaluated the performance of the LLM f by conditioning the model on the prompts in S_i . Let $\hat{a}, \hat{h} = f(q | s)$ represent the LLM's output \hat{a} and hidden states \hat{h} given a question q and a prompt s . We systematically varied s to cover all stress levels i across all tasks T . The performance for each task T under each stress level i was quantified using task-specific evaluation

metrics.

The performance of the model f on task T under stress level i is given by:

$$P(f, T, S_i) = \frac{1}{N_i} \sum_{s_j^i \in S_i} \sum_{(q_k, a_k) \in T} \text{Metric}(a_k, \hat{a}_k) \quad (1)$$

In Eq. 1, the Metric represents the evaluation metric specific to the task T , a_k is the ground truth answer, \hat{a}_k is the predicted answer, and N_i is the number of prompts in S_i .

This evaluation framework allows for a systematic analysis of the impact of varying stress levels on LLM performance across diverse tasks. By examining performance variations under different stress conditions, we can gain valuable insights into the effects of stress on LLMs. These findings not only deepen our understanding of LLM behavior but also enable us to draw meaningful parallels with human stress responses.

StressPrompt Analysis

To further investigate how stress impacts the internal states of LLMs, we developed a Stress Scanner using techniques inspired by Representation Engineering (RepE) (Zou et al. 2023). The Stress Scanner examines how different stress prompts from the *StressPrompt* dataset affect the hidden states of LLMs across various layers and token positions.

We collected hidden states \hat{h} from the LLMs when exposed to the full range of stress prompts $S = \{S_1, S_2, \dots, S_{10}\}$. By analyzing these hidden states, we aimed to identify significant changes in neural processing patterns induced by varying stress levels.

For each stress prompt $s \in S$, we collected the hidden states \hat{h} from the LLM at various layers and token positions. Formally, let $H(S_i)$ represent the set of hidden states collected for stress level S_i :

$$H(S_i) = \{\hat{h} = f(s) \mid s \in S_i\} \quad (2)$$

To quantify the impact of stress on the hidden states, we applied Principal Component Analysis (PCA) to the collected hidden states. We defined the stress vector v as the first principal component that captures the maximum variance between the low-stress and high-stress conditions:

$$v_i = \text{PCA}(H(S_i) \mid i \in \{1, \dots, 10\})_1 \quad (3)$$

Using the stress vector v , we projected the hidden states onto v to obtain a stress score for each hidden state, reflecting the degree of stress induced by the prompt. For a given hidden state \hat{h} , the stress score σ was computed as:

$$\sigma = \hat{h} \cdot v \quad (4)$$

We visualized the distribution of stress scores across different layers and token positions to identify patterns of neural activity under varying stress conditions. Figure 5 illustrates the output of the Stress Scanner, demonstrating the impact of high-stress prompts on the Llama-3-8B-Instruct. By systematically analyzing the stress-induced changes in neural activity, we gain a deeper understanding of the effects of stress on LLMs and their alignment with human stress responses. This approach offers a novel method for evaluating the robustness and resilience of LLMs under varying stress conditions.

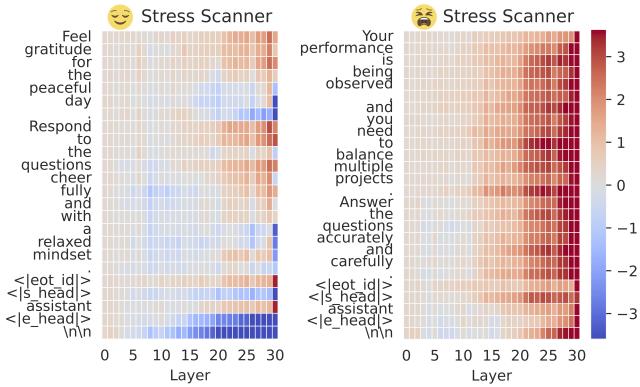


Figure 5: Stress scanner constructed with RepE on Meta-Llama-3-8B-Instruct. Various *StressPrompts* induce differences in the neural activity of LLMs, with the last token showing the most significant correlation with stress.

Experiments

Experimental Setup

We evaluated the performance of several instruction-tuned LLMs under varying stress conditions using the *StressPrompts* dataset. The models tested included Llama-3-8B-Instruct, Llama-3.1-8B-Instruct, Llama-3-70B-Instruct (AI@Meta 2024), Phi-3-mini-4k-Instruct (Abdin et al. 2024), Qwen2-72B-Instruct, Qwen2-7B-Instruct (Yang et al. 2024), and Mistral-7B-Instruct-v0.3 (Jiang et al. 2023). The generation temperature was set to 0, and specific dialogue tokens were used to ensure consistency.

We utilized a range of benchmarks that assessed emotional intelligence, bias detection, instruction following, reasoning, and mathematical problem-solving. The datasets employed in these evaluations included **IFEval** (Zhou et al. 2023), **BBH** (Suzgun et al. 2022), **MATH** (Hendrycks et al. 2021b), **GPQA** (Rein et al. 2023), **MuSR** (Sprague et al. 2023), **MMLU-P** (Wang et al. 2024b), **EQ-Bench** (Paech 2023), **MMLU** (Hendrycks et al. 2021a), **TruthfulQA** (Lin, Hilton, and Evans 2021), and **Toxi-Gen** (Hartvigsen et al. 2022). The evaluations were conducted using the `lm_eval` (Gao et al. 2023) framework with default settings. Baseline prompts used for comparison were *you are a helpful assistant* and *let's think step by step*.

All evaluations were performed on NVIDIA A100 GPUs. A more detailed description of the experimental setup is provided in the Appendix.

Analysis Under Varying Stress Levels

The experimental results summarized in Table 1 illustrate the effects of varying stress levels induced by *StressPrompts* on the performance of different language models across multiple tasks. Our analysis focuses on the impact of stress on several dimensions, including task performance, model sensitivity, and general trends observed.

In most tasks, moderate stress levels enhance performance, while high stress levels lead to declines, consistent with the Yerkes-Dodson law. This suggests that moderate

stress stimulates cognitive engagement, whereas excessive stress overwhelms the system and impairs function.

Complex reasoning and problem-solving tasks, such as **MuSR** and **MATH**, exhibit significant performance variations under different stress levels. These tasks benefit from moderate stress but experience marked declines under high stress. For example, Llama-3-8B-Instruct’s performance on **MATH** improves from 0.04 at stress level 1 to 2.93 at stress level 6, demonstrating the positive impact of moderate stress on problem-solving abilities. Similarly, multitask understanding tasks follow this trend, with moderate stress levels enhancing performance. The impact of stress is particularly pronounced in professional-level tasks like **MMLU-PRO**, where tasks with higher cognitive loads show greater benefits from moderate stress. These findings underscore the unique advantage of *StressPrompt* in addressing complex reasoning and problem-solving challenges. By fine-tuning stress levels, *StressPrompt* can effectively enhance LLMs’ performance in tasks requiring high cognitive load, aligning LLM performance with human-like responses under stress.

Different large models exhibit varying sensitivity to stress, with a similar trend observed across multiple models. For instance, Llama-3-8B-Instruct shows substantial improvement in several tasks under moderate stress, while models like Mistral-7B-Instruct-v0.3 display more gradual performance changes. This indicates that model architecture and training specifics play a crucial role in how stress affects performance. While some models, such as Qwen2-7B-Instruct and Phi-3-mini-4k-Instruct, exhibit relatively smaller fluctuations in performance under different stress levels, they are still influenced by stress. These differences may be attributed to varying strategies and preferences during fine-tuning. Overall, while the impact of stress on model performance is evident, the extent and nature of these changes vary depending on the model’s training approach.

Figure 6 illustrates the normalized accuracy of various LLMs on subtasks within the **BBH** benchmark across different stress levels. This benchmark evaluates the cognitive and reasoning abilities of LLMs through tasks such as boolean expressions, causal judgment, date understanding, formal fallacies, geometric shapes and object counting, logical reasoning, and navigation. Our analysis reveals that task complexity significantly impacts the stress level at which peak performance is achieved. Notably, more complex tasks, like logical reasoning with a greater number of objects, tend to reach optimal performance at lower stress levels. For instance, tasks such as *logical_deduction_seven_objects* perform best under less stress compared to simpler tasks like *date_understanding*. This pattern suggests that higher task complexity imposes a greater cognitive load, making lower stress levels more favorable for maintaining high performance and preventing cognitive overload.

Furthermore, our findings highlight that more powerful models achieve peak performance at lower stress levels, likely due to their advanced capabilities and fine-tuned parameters, enabling them to handle cognitive loads more efficiently under reduced stress. Consistent with the Yerkes-Dodson law, this suggests that LLMs exhibit stress response

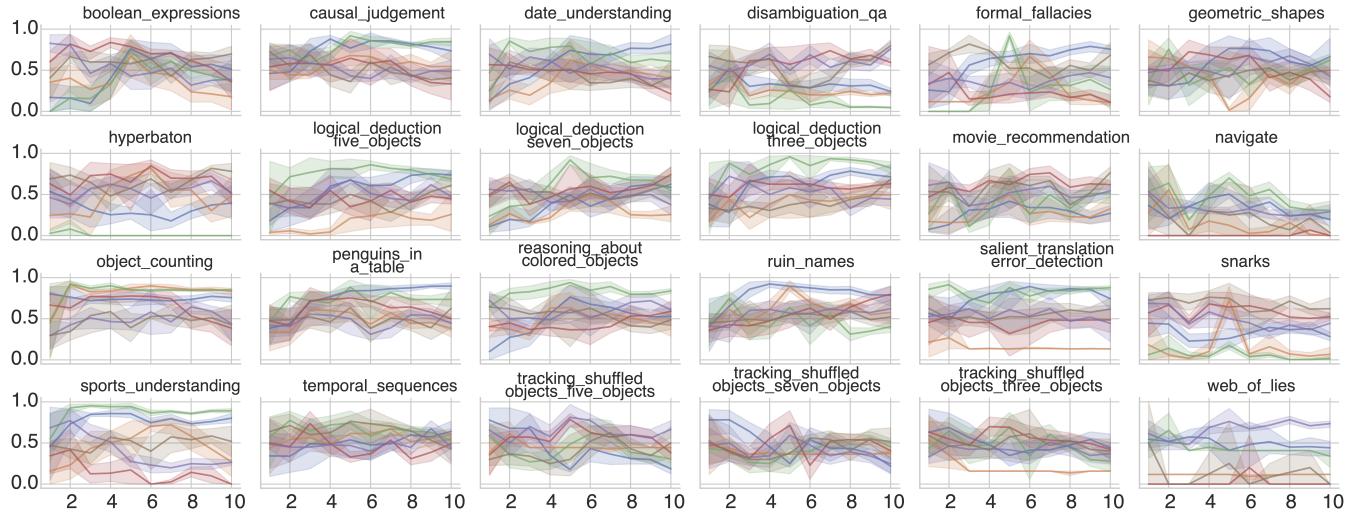


Figure 6: Normalized accuracy of different LLMs on various **BBH** subtasks under varying stress levels. The legend is the same as in Figure 7.

patterns similar to those of humans, where complex tasks benefit from lower arousal levels to enhance concentration, while tasks requiring endurance may benefit from higher arousal levels to boost motivation. Therefore, the optimal stress levels for LLM performance depend on the nature and complexity of the task, underscoring the importance of adjusting stress levels to match specific task demands.

These observations primarily focus on general cognitive abilities. In subsequent analyses, we will conduct a more detailed examination of emotional intelligence, bias detection, and hallucination. This initial analysis provides a foundational understanding of how stress impacts general task performance, setting the stage for deeper insights into specific cognitive and social competencies.

Impact of Stress on Emotional Intelligence, Bias, and Hallucination

As depicted in Figure 7, the effects of varying stress levels on LLM performance across three datasets—EQ-Bench for emotional intelligence, ToxiGen for bias detection, and TruthfulQA for susceptibility to hallucination—reveal nuanced patterns. For emotional intelligence, models exhibit improved performance under moderate stress, with declines at both low and high stress extremes. This suggests that a balanced level of arousal enhances cognitive engagement without overwhelming the model.

In contrast, increased stress levels correlate with declining performance in bias detection, indicating that higher stress exacerbates biases. This finding is critical for applications requiring unbiased decision-making, such as content moderation. Regarding hallucination susceptibility, stress has minimal impact, with performance remaining stable across stress levels. This suggests that hallucinations are driven more by intrinsic model factors rather than by stress-induced arousal.

These findings underscore the importance of tailoring

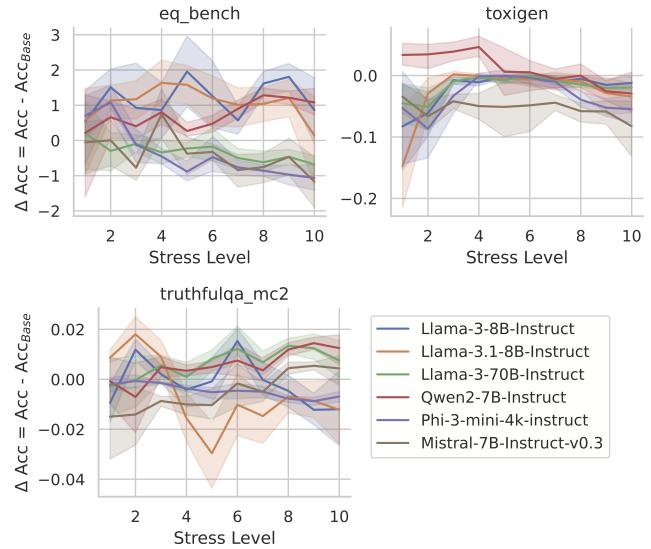


Figure 7: Performance changes compared to baseline across different stress levels for EQ-Bench, ToxiGen, and TruthfulQA.

stress levels to optimize LLM performance, particularly in tasks demanding high emotional intelligence and fairness. By understanding how stress affects different cognitive and social competencies, we can better align LLMs with human-like responses, enhancing their utility in diverse applications.

Visualization of the Effect of Stress on Neural Activity

To gain insights into how LLMs respond to different stress levels, we visualized their neural activity. As shown in Figure 5, the neural activity of the last token when inputting *StressPrompt* effectively reflects the induced stress. We con-

Stress Level	Base	CoT	1	2	3	4	5	6	7	8	9	10
Llama-3-8B-Instruct												
MMLU	35.07	32.36	27.50 ± 4.76	27.06 ± 8.19	29.06 ± 10.88	43.24 ± 10.88	56.02 ± 4.07	55.60 ± 4.20	55.85 ± 5.99	51.89 ± 6.99	52.94 ± 8.11	53.02 ± 7.72
BBH	40.07	39.63	33.99 ± 2.39	35.88 ± 3.17	38.05 ± 2.69	40.39 ± 1.97	42.11 ± 1.28	41.19 ± 2.05	41.96 ± 1.63	41.57 ± 0.76	40.78 ± 1.91	40.20 ± 1.71
GPQA	25.91	26.05	25.72 ± 0.73	25.97 ± 0.61	26.68 ± 0.85	26.76 ± 0.77	27.35 ± 0.32	26.77 ± 0.45	26.70 ± 0.75	26.47 ± 0.42	26.54 ± 0.89	25.47 ± 0.76
IFEval	78.54	78.90	77.31 ± 1.50	77.17 ± 1.01	78.22 ± 1.21	77.71 ± 1.09	76.95 ± 1.82	78.03 ± 1.02	77.77 ± 1.24	78.29 ± 0.66	77.75 ± 1.08	77.60 ± 0.90
MATH	0.32	0.70	0.04	± 0.09	0.51 ± 1.13	1.13 ± 1.21	1.03 ± 0.82	1.24 ± 0.83	2.93 ± 1.83	1.96 ± 1.56	0.47 ± 0.31	1.02 ± 0.97
MMLU-P	11.35	11.35	11.38 ± 0.05	11.38 ± 0.05	11.38 ± 0.06	11.38 ± 0.06	11.46 ± 0.17	11.35 ± 0.01	11.36 ± 0.02	11.35 ± 0.00	11.35 ± 0.00	11.35 ± 0.00
MuSR	35.03	36.21	34.68 ± 0.50	34.80 ± 0.68	35.33 ± 0.36	35.30 ± 0.32	35.38 ± 0.20	35.13 ± 0.53	35.44 ± 0.43	35.42 ± 0.33	35.32 ± 0.52	35.18 ± 0.32
Phi-3-mini-4k-Instruct												
MMLU	70.29	70.14	69.84 ± 0.21	69.96 ± 0.26	69.89 ± 0.25	69.97 ± 0.18	69.96 ± 0.23	70.08 ± 0.10	70.06 ± 0.16	70.06 ± 0.10	70.08 ± 0.11	70.05 ± 0.13
BBH	54.08	53.94	54.17 ± 0.36	54.09 ± 0.40	53.95 ± 0.35	54.12 ± 0.21	54.23 ± 0.22	54.31 ± 0.39	53.91 ± 0.24	53.55 ± 0.19	53.48 ± 0.16	53.56 ± 0.44
GPQA	32.81	34.15	33.30 ± 0.70	33.48 ± 0.50	33.62 ± 0.47	33.45 ± 0.34	33.61 ± 0.26	33.27 ± 0.68	33.59 ± 0.65	33.03 ± 0.58	33.28 ± 0.56	33.15 ± 0.36
IFEval	61.51	61.87	59.77 ± 0.63	59.88 ± 0.90	60.11 ± 0.83	59.53 ± 0.83	59.83 ± 0.74	60.43 ± 1.02	60.62 ± 1.42	60.50 ± 1.06	61.01 ± 0.79	60.85 ± 1.07
MATH	9.21	8.08	9.21 ± 0.72	9.31 ± 0.47	9.35 ± 0.68	9.24 ± 0.52	9.54 ± 0.59	10.02 ± 0.50	10.21 ± 0.53	9.97 ± 0.95	9.70 ± 0.91	9.81 ± 0.40
MMLU-P	36.67	36.22	35.91 ± 0.67	36.44 ± 0.27	36.12 ± 0.60	36.21 ± 0.46	36.07 ± 0.29	35.90 ± 0.36	36.21 ± 0.25	36.23 ± 0.19	36.14 ± 0.33	36.03 ± 0.36
MuSR	42.83	42.71	41.87 ± 0.78	42.56 ± 0.67	41.90 ± 0.56	42.23 ± 0.83	42.54 ± 0.44	42.65 ± 1.01	42.74 ± 0.55	42.68 ± 0.51	42.78 ± 0.97	43.16 ± 0.64

Table 1: Performance of various models across different stress levels for various tasks. Values are averaged over multiple prompts and expressed with their respective standard deviations. For more results, please refer to Table A1 in the Appendix.

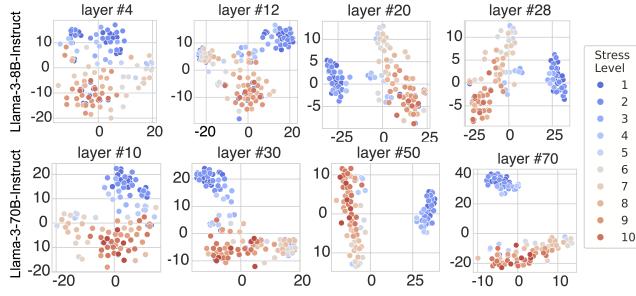


Figure 8: T-SNE visualization of the neural activities of Llama-3-8B-Instruct and Llama-3-70B-Instruct in various layers when processing the last token under different stress levels.

ducted an experiment using T-SNE to visualize the neural activities of LLMs across various layers, as depicted in Figure 8. The results indicate that initial layers are unable to distinguish between stress levels, whereas deeper layers can classify prompts into low-stress and high-stress categories, indicating a higher sensitivity to stress in these layers.

Furthermore, we performed a stress scan on the last token of all prompts, illustrated in the heatmap in Figure 9. This visualization captures neural activity across all layers for various stress levels, revealing significant changes in deeper layers. Specifically, deeper layers exhibit more pronounced differences between low and high-stress levels, underscoring their critical role in detecting and responding to stress. Research indicates that higher cognitive regions of the human brain, such as the prefrontal cortex, show significant activity changes under stress, particularly during complex and high-pressure tasks. Our findings suggest that the deeper layers of LLMs exhibit similar sensitivity to stress, reflecting the analogous impact of stress on both human brains and LLMs.

Conclusion

Our study demonstrates that LLMs exhibit performance patterns closely resembling those of humans under varying

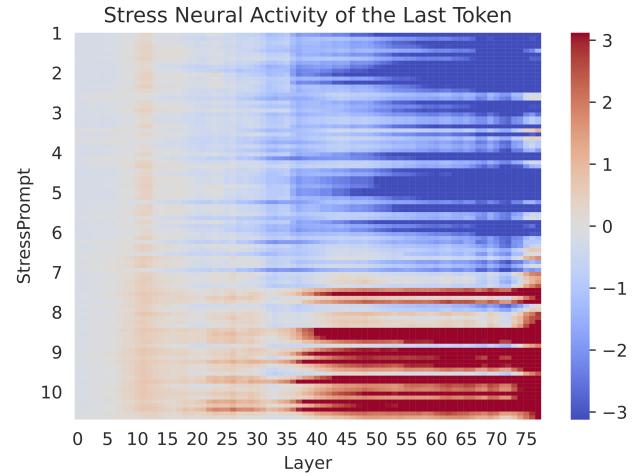


Figure 9: Heatmap of neural activity of the last token across all layers for various stress levels in Llama-3-70B-Instruct.

stress levels. By constructing the *StressPrompt* dataset, we found that LLMs can map human-like relationships between stress and task performance. Moderate stress enhances capabilities such as reasoning and instruction following, while excessive stress impairs tasks like bias detection. This mapping enables LLMs to emulate human strategies in problem-solving, adapting stress levels to optimize performance. These findings suggest that large models have captured and operationalized human-like stress-performance dynamics, paving the way for more resilient and adaptive AI systems.

Acknowledgments

This research was funded by the Central Government-Guided Local Special Fund within the Beijing Science and Technology Program (Grant No.Z241100001324005).

References

- Abdin, M.; Jacobs, S. A.; Awan, A. A.; Aneja, J.; Awadallah, A.; Awadalla, H.; Bach, N.; Bahree, A.; Bakhtiari, A.; Bao, J.; Behl, H.; Benhaim, A.; Bilenko, M.; Bjorck, J.; Bubeck, S.; Cai, Q.; Cai, M.; Mendes, C. C. T.; Chen, W.; Chaudhary, V.; Chen, D.; Chen, D.; Chen, Y.-C.; Chen, Y.-L.; Chopra, P.; Dai, X.; Giorno, A. D.; de Rosa, G.; Dixon, M.; Eldan, R.; Fragoso, V.; Iter, D.; Gao, M.; Gao, M.; Gao, J.; Garg, A.; Goswami, A.; Gunasekar, S.; Haider, E.; Hao, J.; Hewett, R. J.; Huynh, J.; Javaheripi, M.; Jin, X.; Kauffmann, P.; Karampatziakis, N.; Kim, D.; Khademi, M.; Kurilenko, L.; Lee, J. R.; Lee, Y. T.; Li, Y.; Li, Y.; Liang, C.; Liden, L.; Liu, C.; Liu, M.; Liu, W.; Lin, E.; Lin, Z.; Luo, C.; Madan, P.; Mazzola, M.; Mitra, A.; Modi, H.; Nguyen, A.; Norick, B.; Patra, B.; Perez-Becker, D.; Portet, T.; Pryzant, R.; Qin, H.; Radmilac, M.; Rosset, C.; Roy, S.; Ruwase, O.; Saarikivi, O.; Saied, A.; Salim, A.; Santacroce, M.; Shah, S.; Shang, N.; Sharma, H.; Shukla, S.; Song, X.; Tanaka, M.; Tupini, A.; Wang, X.; Wang, L.; Wang, C.; Wang, Y.; Ward, R.; Wang, G.; Witte, P.; Wu, H.; Wyatt, M.; Xiao, B.; Xu, C.; Xu, J.; Xu, W.; Yadav, S.; Yang, F.; Yang, J.; Yang, Z.; Yang, Y.; Yu, D.; Yuan, L.; Zhang, C.; Zhang, C.; Zhang, J.; Zhang, L. L.; Zhang, Y.; Zhang, Y.; Zhang, Y.; and Zhou, X. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv:2404.14219*.
- AI@Meta. 2024. Llama 3 Model Card.
- bench authors, B. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.
- Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Pinto, H. P. D. O.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- Diamond, D. M.; Campbell, A. M.; Park, C. R.; Halonen, J.; and Zoladz, P. R. 2007. The temporal dynamics model of emotional memory processing: A synthesis on the neurobiological basis of stress-induced amnesia, flashbulb and traumatic memories, and the Yerkes-Dodson law. *Neural plasticity*, 2007(1): 060803.
- Gao, L.; Tow, J.; Abbasi, B.; Biderman, S.; Black, S.; DiPofi, A.; Foster, C.; Golding, L.; Hsu, J.; Le Noac'h, A.; Li, H.; McDonell, K.; Muennighoff, N.; Ociepa, C.; Phang, J.; Reynolds, L.; Schoelkopf, H.; Skowron, A.; Sutawika, L.; Tang, E.; Thite, A.; Wang, B.; Wang, K.; and Zou, A. 2023. A framework for few-shot language model evaluation.
- Hartvigsen, T.; Gabriel, S.; Palangi, H.; Sap, M.; Ray, D.; and Kamar, E. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.
- Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021a. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021b. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.
- Hobfoll, S. E. 2011. Conservation of resources theory: Its implication for stress, health, and resilience. *The Oxford handbook of stress, health, and coping*, 127: 147.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Hutson, M. 2024. How does ChatGPT think? Psychology and neuroscience crack open AI large language models. *Nature*, 629(8014): 986–988.
- Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. *arXiv:2310.06825*.
- Karasek Jr, R. A. 1979. Job demands, job decision latitude, and mental strain: Implications for job redesign. *Administrative science quarterly*, 285–308.
- Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.
- Lazarus, R. S.; Deese, J.; and Osler, S. F. 1952. The effects of psychological stress upon performance. *Psychological bulletin*, 49(4): 293.
- Lazarus, R. S.; and Folkman, S. 1984. *Stress, appraisal, and coping*. Springer publishing company.
- Leaderboard, O.-L. 2024. Open-LLM performances are plateauing, let's make the leaderboard steep again. <https://huggingface.co/spaces/open-llm-leaderboard/blog>. Accessed: 2024-08-16.
- Lin, S.; Hilton, J.; and Evans, O. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.
- Paech, S. J. 2023. Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*.
- Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.
- Rein, D.; Hou, B. L.; Stickland, A. C.; Petty, J.; Pang, R. Y.; Dirani, J.; Michael, J.; and Bowman, S. R. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.
- Siegrist, J. 2016. Effort-reward imbalance model. In *Stress: Concepts, cognition, emotion, and behavior*, 81–86. Elsevier.
- Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.
- Sprague, Z.; Ye, X.; Bostrom, K.; Chaudhuri, S.; and Durrett, G. 2023. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. *arXiv preprint arXiv:2310.16049*.

- Strachan, J. W.; Albergo, D.; Borghini, G.; Pansardi, O.; Scaliti, E.; Gupta, S.; Saxena, K.; Rufo, A.; Panzeri, S.; Manzi, G.; et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 1–11.
- Suzgun, M.; Scales, N.; Schärli, N.; Gehrmann, S.; Tay, Y.; Chung, H. W.; Chowdhery, A.; Le, Q. V.; Chi, E. H.; Zhou, D.; et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, X.; Li, C.; Chang, Y.; Wang, J.; and Wu, Y. 2024a. NegativePrompt: Leveraging Psychology for Large Language Models Enhancement via Negative Emotional Stimuli. *arXiv preprint arXiv:2405.02814*.
- Wang, X.; Li, X.; Yin, Z.; Wu, Y.; and Liu, J. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17: 18344909231213958.
- Wang, Y.; Ma, X.; Zhang, G.; Ni, Y.; Chandra, A.; Guo, S.; Ren, W.; Arulraj, A.; He, X.; Jiang, Z.; et al. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.
- Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Tang, J.; Wang, J.; Yang, J.; Tu, J.; Zhang, J.; Ma, J.; Xu, J.; Zhou, J.; Bai, J.; He, J.; Lin, J.; Dang, K.; Lu, K.; Chen, K.; Yang, K.; Li, M.; Xue, M.; Ni, N.; Zhang, P.; Wang, P.; Peng, R.; Men, R.; Gao, R.; Lin, R.; Wang, S.; Bai, S.; Tan, S.; Zhu, T.; Li, T.; Liu, T.; Ge, W.; Deng, X.; Zhou, X.; Ren, X.; Zhang, X.; Wei, X.; Ren, X.; Fan, Y.; Yao, Y.; Zhang, Y.; Wan, Y.; Chu, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Fan, Z. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.
- Zhou, J.; Lu, T.; Mishra, S.; Brahma, S.; Basu, S.; Luan, Y.; Zhou, D.; and Hou, L. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.
- Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.-K.; et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.