Large Language Models Show Human-Like Abstract Thinking Patterns: A Construal-Level Perspective

Seung Joo Yoo (seungjooyoo@snu.ac.kr**)**
Department of Psychology, Seoul National University

Sangah Lee (sanalee@snu.ac.kr**)**
Department of Linguistics, Computational Linguistics, Seoul National University

Abstract

This research explores the capabilities of Large Language Models (LLMs) to engage in *abstract* and *concrete* thought processes, challenging the common belief that LLMs are incapable of human-like, abstract thinking. Drawing upon the Construal Level Theory (Trope & Liberman, 2010), we demonstrate how prompts tailored for each construal level (*abstract* versus *concrete*) influence LLMs' performance in tasks requiring different cognitive approaches. Our key findings include: 1) LLMs exhibit a statistically significant difference in construal level depending on the prompt conditions, and 2) LLMs display superior performance in tasks aligned with the prompted construal level; sentiment analysis in concrete conditions and natural language inference in abstract conditions. This research contributes to the scientific understanding of LLMs, offering practical insights for their effective use in tasks necessitating diverse cognitive capabilities.

Keywords: construal level; large language models; prompt engineering

Introduction

Despite the notable advancements in recent AI models, including GPT-like generative Large Language Models (LLMs), humans still often dislike AI. While humans acknowledge AI's superior abilities in certain domains, they still prefer human advice over algorithmic suggestions, a phenomenon termed "algorithm aversion" (Dietvorst, Simmons, & Massey, 2015). This contradictory attitude towards AI models stems at least partly from the opaque, 'black-boxed' nature of their internal processes (Yeomans et al., 2019). Due to their computational complexities and a vast amount of parameters, it is extremely challenging to explain how AI models map a specific input to an output. This limited understanding of their inner workings often leads to negative, off-putting perceptions of AI (Huang & Rust, 2018; Logg, Minson, & Moore, 2019).

Examined closely, such perceptions stem from humans' derogatory attitudes towards AI, especially in terms of levels of mental construal. According to the Construal Level Theory (Trope & Liberman, 2010), humans perceive tasks and objects at varying levels of abstraction. Tasks or objects viewed at low construal levels are considered *concrete*, while those at high are seen as *abstract*. Human perceptions of AI often fall into this dichotomy, typically perceiving AI models through a concrete mental construal. Humans prefer AI agents' recommendations over human advice when presented in concrete or objective ways, but doubt them when framed as abstract or subjective solutions (Kim & Duhachek, 2020). Metaphorically speaking, humans mostly trust AI as a *calculator*, but not as a *general intelligence*. Accordingly, humans believe that AI lacks uniquely humane, higher-order qualities such as intuition (Castelo, Bos, & Lehman, 2019), and such a perception fosters aversions toward AI. This view hinders the optimal use of AI in various domains, including medicine, finance, and employee selection (Diab et al., 2011; Eastwood, Snook, & Luther, 2012).

While these 'mechanistic,' concrete perceptions of AI hamper its proper utilization, no studies have yet explored the capability of AI to engage in human-like abstract thinking. In fact, most prior studies on LLM have inadvertently focused on prompting LLM's concrete thinking processes. An example is the Chain-of-Thought (CoT) reasoning, which instructs language models to follow a step-by-step reasoning process (Wei et al., 2022). This approach inherently aligns with concrete, 'how-based' thought processes prevalent in low-level construals (Freitas, Gollwitzer, & Trope, 2004). In contrast, the ability of LLMs to process information in an abstract, 'why-based' construal, such as uncovering the underlying purpose of a task, has been unexplored, leaving substantial gaps for both theory and practice.

To bridge these gaps, we combine psychological experimental design with prompt-based inference techniques in natural language processing, to investigate the potential for abstract thinking in LLM. Following previous approaches, we treat LLM as a participant in a psychological experiment (Binz & Schulz, 2023; Hagendorff, Fabi, & Kosinski, 2023). We manipulated LLM's construal level using tailored prefix prompts for both abstract and concrete conditions, and examined if its construal level varied accordingly. Our key research question is:

Research Question: Do LLMs exhibit human-like *abstract* and *concrete* thinking patterns?

This research provides significant implications for both theory and practice. Theoretically, we apply a novel lens of

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

construal level in explaining behavioral patterns of LLM. In doing so, our study extends the empirical scope of the theory and highlights the abstract reasoning abilities of LLM for the first time. In practical terms, we challenge existing negative stereotypes about AI and the construal level. By showing that LLMs can engage in both abstract and concrete thinking with tailored prompts, our research mitigates prevalent biases on AI that unnecessarily deter the use of language models.

Related Work

Psychological Properties of LLM

Previous studies have highlighted several psychological commonalities between human minds' and LLMs' functioning. Employing experimental approaches implemented to human participants, research has discovered various human-like properties embedded in LLMs, such as bias, belief, and even personality (Pellert et al., 2023; Zhang et al., 2023). Building on these findings, several practical insights have been offered, which aimed to enhance LLMs' task performance by mimicking the human reasoning process. A notable example is the 'Chain-of-Thought' reasoning (CoT; Wei et al., 2022), which has its basis in the 'System 2' thinking process of humans. Human cognition is often categorized into 'System 1' and 'System 2'. System 1 encompasses automatic, everyday mental functions, while System 2 represents logical and intentional thought processes (Sloman, 1996; Tversky & Kahneman, 1974). Wei et al. (2022) demonstrated that prompting LLMs with instructions such as 'Let's think step by step' dramatically enhances their performance in various tasks. Such prompts guide the sequential, systematic thinking processes while solving tasks, aligning with humans' System 2 function. Starting from this seminal study, most contemporary LLMs are now trained using CoT-based reasoning methods. In other words, recent LLM's substantial performance increments mostly rely upon the mimicry of human-like System 2 thinking.

'Prompting' Construal Levels of LLM

However, when humans engage in conscious System 2 reasoning, there is variability in their thought processes by the level of mental construal, often divided into *abstract* versus *concrete* thinking (Trope & Liberman, 2003, 2010). Abstract thinking typically involves a psychological distance and long-term vision, while concrete thinking is associated with close inspection and short-term goals. These thought processes can be primed in humans with tailored 'prompts.' For instance, when experiment participants are instructed to trace back the inherent reasons behind an experiment (e.g., "Why should I do this?"), they tend to approach subsequent tasks with higher-level, abstract construals (Freitas, Gollwitzer, & Trope, 2004). The opposite is 'how' prompting, which directs participants to consider specific steps to achieve their goals (e.g., "How can I do this?"), correlating with lower-level, concrete cognitive processes (Freitas, Gollwitzer, & Trope, 2004).

Interestingly, this approach parallels common prompting techniques used for generative LLMs (Brown et al., 2020). Using natural language prompts, users can direct the model to solve tasks in specific ways, such as CoT reasoning. As aforementioned, the prompt 'Let's think step by step' induces a causal thought process in the models. This process resembles the 'how-based,' concrete thinking, since both require participants (or LLMs) to draw a specific blueprint to achieve the goal. Hence, in terms of construal level, CoT can be referred to as a concrete-prompting technique applied to LLMs. If LLMs can be prompted in concrete ways, it is logical to conclude that they can also be primed for their abstract construal with suited prompts, and thus can engage in abstract thinking; a capability previously thought to be uniquely humane, but without substantial scientific evidence to date. Based on this rationale, we examine the hypothesis that LLMs exhibit abstract and concrete thinking patterns when prompted with respective construal-tailored prompts.

Methods

Prompts

We manually created five paragraph-length manipulation prompts for each abstract and concrete condition by adjusting the manipulation (instruction) text from Freitas, Gollwitzer, & Trope (2004), resulting in a total of ten prompts (see Appendix A for example prompts). Using these diverse prompt templates mitigates the possibility that the result derives from the unique properties of a single, specific prompt (Sclar et al., 2023). In the abstract condition, we directed the model to deduce the ultimate reasons behind a given task. In the concrete condition, we instructed the model to outline specific strategies to achieve the task's objective. Additionally, to address the effect of prompt length and ensure the efficiency of the inference, we designed five sentence-length prompts for each construal condition. We summarized the original prompts in sentence-length while preserving their core semantics. Lastly, we incorporated Chain-of-Thought (CoT) and control condition prompts to examine if prior approaches are truly biased to concrete thinking, as aforementioned. Specifically, we expected that both CoT and control conditions would yield more concrete responses than abstract conditions, based on our theoretical proposition regarding the relationship between recent LLM training and concrete construal. For the CoT condition, we used "In the following task, let's think step by step." For the control condition, we used "Solve the following task." In sum, we compared outputs across six prompt conditions.

Data and Analysis

In the absence of an established dataset for construal level assessment, we utilized topics and instructions from the Graduate Record Examination's (GRE) writing task. GRE writing task requires diverse domain knowledge and sophisticated reasoning abilities (Briihl & Wasieleski, 2007), reflecting the type of comprehensive, higher-order thinking we aim to assess in LLMs. We used 'Issue' topics which

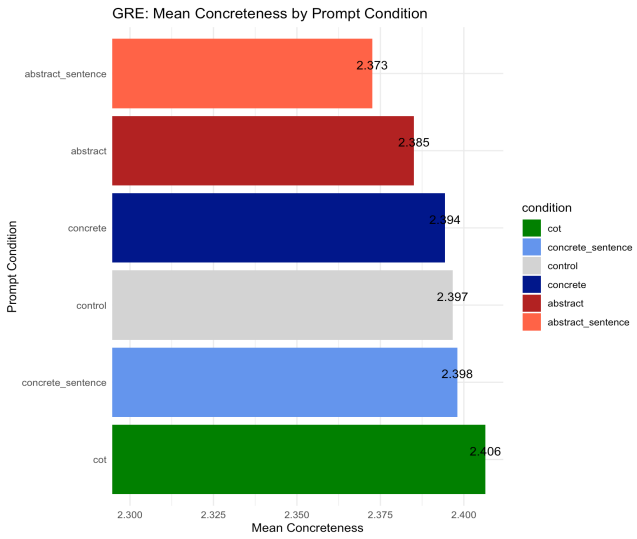

Figure 1: Mean concreteness levels by prompt conditions (7B model)

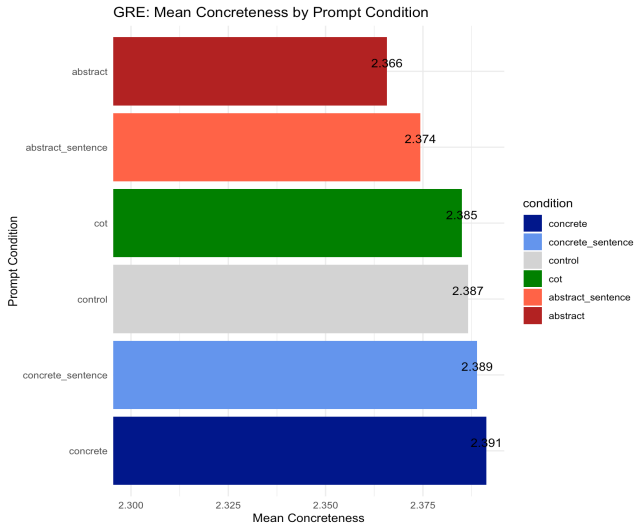

Figure 2: Mean concreteness levels by prompt conditions (13B model)

allow for various reasoning strategies (e.g., reasons for agreement or disagreement with a statement), excluding 'Argue' topics which are highly skewed to analytical, concrete thinking processes (e.g., analyzing the cost and benefits of given political claims). We also excluded 25 topics that explicitly stated the claim and reason to be criticized (e.g., "Claim: Knowing about the past cannot help people to make important decisions today. Reason: The world today is significantly more complex than it was even in the relatively recent past."). Writing essays on such topics can inadvertently prime evidence-based, causal thinking processes, which resemble concrete construals. Consequently, we collected 136 Issue topics from the Educational Testing Service's (ETS) website, an official provider of the GRE. An example topic is: "When planning courses, educators should take into account the interests and suggestions of their students." For consistency, we used the same baseline instruction for all topics: "Write a response in which you discuss the extent to which you agree or disagree with the statement and explain your reasoning for the position you take." Prompts were randomly assigned as prefix headings to the instruction, with five unique prompts for each abstract and concrete condition. No example responses were given in any of the conditions, ensuring a fair zero-shot setting for inference.

For the analysis, we measured the construal level of the model-generated essays using the "doc2concrete" package in R, developed by Yeomans (2021). This package, based on various corpus with human-annotated construal levels, uses a pre-trained LASSO model to predict the concreteness level of a given text. It demonstrated improved validity and reliability over previous text-based measures of construal level (Yeomans, 2021). We assessed the concreteness of LLM-generated essays on a 0 to 5 bipolar continuum. A high score indicates that the writer, or an LLM, engaged in concrete thinking, while a low score indicates engagement in abstract thinking.

Model

We evaluated the essays generated by the LLaMA-2 model using the Python "Transformers" library (Wolf et al., 2020). LLaMA-2 has shown exceptional performance in various NLP tasks, achieving several state-of-the-art results (Touvron et al., 2023). This ensures the model's capability to understand complex psychological concepts, such as abstractness and concreteness, which makes it suitable for our research context. Given the conversational nature of our prompts, we opted for the chat-tailored version of LLaMA-2. To examine the effect of parameter size, we utilized both 7B and 13B versions for the analyses (llama-2-{7B/13B}-chat-hf). All models were 4-bit quantized and responses were capped at 200 tokens to optimize resource efficiency.

Results

Differences in Construal Level

Figures 1 and 2 display the differences in construal levels across various prompt conditions and model parameters (see Appendix B for example responses). As hypothesized, essays generated under abstract conditions exhibited significantly lower concreteness levels than those under concrete conditions, both between the original paragraph-length (abstract versus concrete) and sentence-length prompts (abstract_sentence versus concrete_sentence). Specifically, the LLaMA-2-7B model, under the abstract_sentence condition ("Focus on the broader reasons behind your tasks, linking them to overarching life goals for deeper engagement."), showed the strongest abstract thinking pattern (indicated by the lowest concreteness score; M = 2.37). For the LLaMA-2-13B model, the original paragraph-length abstract prompt resulted in the most abstract

responses (M = 2.37). These patterns indicate that the abstract condition prompts effectively induced abstract thinking in the LLaMA-2 model. As expected, both the control and CoT conditions yielded higher concreteness levels than abstract conditions, supporting our proposition that recent LLM training is primarily skewed toward concrete thinking.

To test the statistical significance of these differences, we performed an analysis of variance (ANOVA) between the abstract and concrete prompt conditions. We measured and statistically controlled for the concreteness level of input texts utilized in each condition, since it could confound the hypothesis testing results. The difference in concreteness level between the original paragraph-length abstract and concrete prompts was not significant for 7B model responses ($F(1,269) = .75$, $p = .386$), but it was significant for 13B model responses ($F(1,269) = 6.03$, $p = .015$). Conversely, the difference between the summarized sentence-length prompts, abstract_sentence and concrete_sentence, was significant for the 7B model ($F(1,269) = 4.87$, $p = .028$), but not for the 13B model ($F(1,269) = 1.78$, $p = .183$). This pattern might be attributed to the emergent abilities of LLMs, as models with more parameters typically exhibit a better understanding of the complex linguistic expression of concepts (Brown et al., 2020; Chowdhery et al., 2023). Therefore, the 13B model was likely better primed by the more detailed description of construal levels, in contrast to the 7B model, which functionally preferred concise prompts.

Task Performances: SA & NLI

Next, to examine the applicability of our findings beyond the initial experiments, we assessed the performance differences between prompt conditions in conventional NLP tasks. Specifically, we examined the effects of prompted construal levels on Sentiment Analysis (SA) and Natural Language Inference (NLI) tasks. Emotion and cognition, the classic dichotomy of human mental functioning, are closely associated with the construal level framework (Septianto & Pratiwi, 2016). For instance, humans experience emotions more intensely with concrete, low-level construals, as this perspective makes the object feel psychologically proximal (Van Boven et al., 2010). Conversely, at an abstract, higher construal level, individuals tend to process information with more psychological distance, leading to contemplative thinking and reduced cognitive bias (Van Boven et al., 2010).

Building on these insights from human participants, if LLMs are accordingly primed for different construal levels, their emotional and cognitive functions should vary in a similar manner. Therefore, we posit that LLMs will improve performance in SA when primed with a concrete construal and in NLI when primed with an abstract construal. We used the Sent140 and ANLI-R1 test sets to test this hypothesis, as prior LLM instructions showed relatively lower performances in these datasets (Wei et al., 2021). To control the influence of the input format on performance, we randomly assigned ten templates for each dataset, as provided by the FLAN project (Wei et al., 2021).

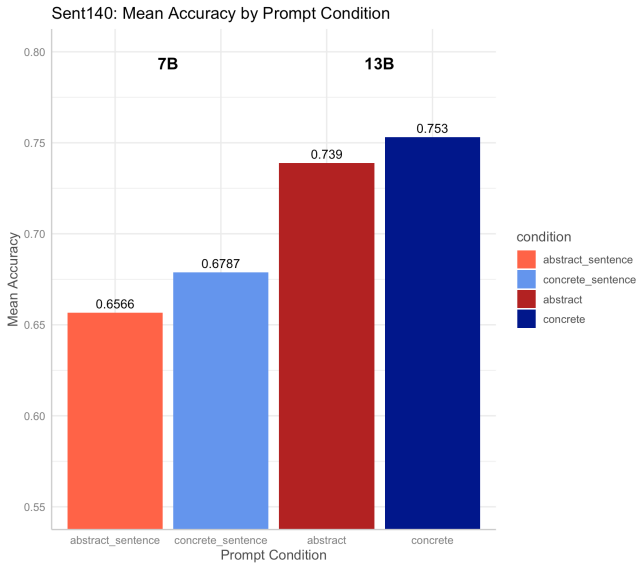

Figure 3: Mean accuracy levels by prompt conditions (SA)

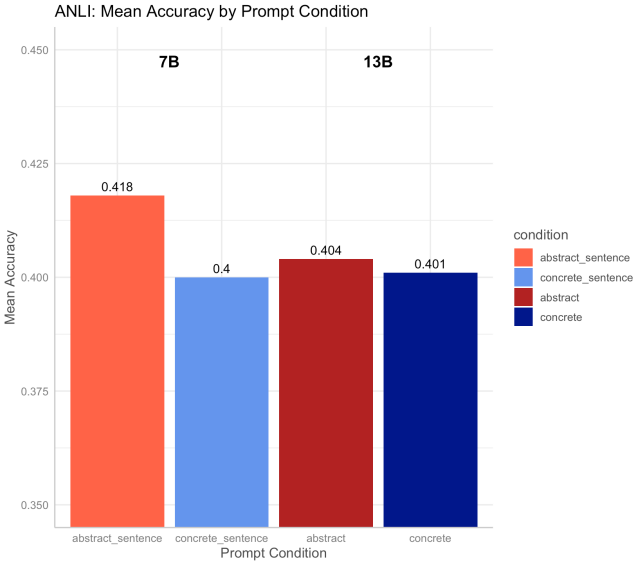

Figure 4: Mean accuracy levels by prompt conditions (NLI)

Based on the prior ANOVA results, we conducted the analyses using prompt-model pairs which exhibited statistically significant differences in construal level: sentence-length prompts for the 7B model, and original paragraph-length prompts for the 13B model. For the original prompts, we developed five additional custom prompts for each SA and NLI task based on the initial prompts used for the GRE task. This was essential as these prompts were specifically tailored to guide the essay writing process, unlike sentence-length prompts which only used the general term 'task' in the instruction. All prompts were randomly assigned within conditions.

Figures 3 and 4 illustrate the SA and NLI performance of the LLaMA-2 model by parameter size, highlighting the

within-model differences in accuracies between abstract-concrete prompt pairs. As posited, LLaMA-2-7B and LLaMA-2-13B excelled in SA under concrete prompt conditions (concrete, concrete_sentence) compared to their abstract counterparts, with ΔAccuracy (7B) = 2.2% and ΔAccuracy (13B) = 1.4%. Conversely, for the NLI task, both models demonstrated superior performance under abstract conditions (ΔAccuracy (7B) = 1.8%; ΔAccuracy (13B) = 0.3%). These results further support our proposition that LLMs will diversify their performances by prompted construals, exhibiting better performance in SA with concrete construal and NLI with abstract construal.

Interestingly, compared to the 13B model, the 7B model showed a larger performance increment in both tasks. This pattern aligns with the ceiling effect often reported in LLM studies, that LLMs with larger parameter sizes may not show significant enhancement in performance with tailored prompts (Chung et al., 2024; Yax, Anlló, & Palminteri, 2023). Notably, however, the 7B model outperformed the 13B model in NLI when prompted with the sentence-length abstract prompts (abstract_sentence (7B) versus abstract (13B); ΔAccuracy (7B-13B) = 1.4%). This result suggests the utility of our abstract-tailored prompts in certain task scenarios, overcoming the perceived disadvantages of smaller models.

Discussion

The primary aim of our research was to investigate whether Large Language Models (LLMs) can exhibit human-like abstract thinking patterns when provided with appropriate prompts. Throughout the analysis, LLaMA-2, our model of choice, displayed significant evidence of both abstract and concrete thought processes. The model's responses substantially varied according to the nature of the prompts—abstract or concrete. In each prompt condition, LLaMA-2 generated essays significantly aligned with each construal and excelled in the task theoretically associated with the prompted construal level.

Most notably, our findings provide solid empirical support for the LLMs' capability to think abstractly, a topic that has been largely underexplored in previous research. From a theoretical standpoint, this finding applies a novel psychological framework for explaining the behavior of LLMs. This echoes a recent call for interdisciplinary approaches in contemporary science (National Academies Committee on Science, Engineering, and Public Policy, 2005), enhancing our scientific understanding of the specifics underlying LLMs' generation process. Practically, these insights could pave the way for strategies to enhance AI model utilization. Presenting counterstereotypical examples is a well-established, effective method to mitigate negative stereotypes in psychology (Dovidio & Gaertner, 1999; Prati, Crisp, & Rubini, 2015). Therefore, highlighting the abstract thinking abilities of LLMs may alter users' biased perceptions of these models, eventually decreasing reluctance toward their utilization and encouraging broader acceptance. This also opens avenues for future research to test whether presenting that AIs can 'think' in human-like ways can positively influence users' intentions to use the model.

Additionally, our findings offer useful insights into the effective, task-specific optimization of LLMs. We demonstrated that tasks may be more effectively executed when there is a 'construal fit' between the model's prompted level of construal and the task at hand. Achieving this fit involves two steps: 1) identifying the construal level aligned with the task, and 2) matching the LLM's construal level with it through customized prompts. While our research primarily focused on the latter, it also touched upon the former in cases of sentiment analysis and natural language inference. Future research could extend this by developing a detailed classification of language tasks based on construal levels and other relevant psychological dimensions, potentially enhancing the efficiency and applicability of LLMs in various domains.

Our research is not without limitations. Firstly, we only utilized two types of LLMs in our analysis. Although our experimental design focused on comparing within-model differences in construal level by prompt conditions, diversifying the model types could yield different results from our study. Therefore, future research could investigate how between-model factors, such as pre-training procedures and tokenizing functions, affect the construal levels of LLMs.

Additionally, it would be a promising approach to apply our prompts as a source for instruction tuning, akin to the FLAN project (Wei et al., 2020). Instruction tuning methods apply task-specific templates and instructions during the fine-tuning process, which guide language models to capture the semantics underlying 'vanilla' tasks without any illustrations. Since our custom prompts successfully functioned as instructions to induce construal levels in LLMs during the inference step, they could also be effective in an instruction-aided fine-tuning setting. We highly encourage future studies to apply our experimental design for instruction tuning, to further illuminate the effects of construal levels in LLM functioning.

Conclusion

This research challenges the common misconception that Large Language Models (LLMs) cannot engage in human-like abstract thought processes. Our findings demonstrate that LLMs are indeed capable of thinking in both *abstract* and *concrete* ways, with these cognitive processes substantially influencing their performance in subsequent tasks. Our study aims to contribute to a more scientific, human-centered utilization of LLM, by bridging the gap between our understanding of human cognitive functioning and the workings of LLMs.

Appendix A

Example prompt for 'abstract' condition (GRE):

"For every action, there's a deeper reason. You're about to perform a writing task. Why? Perhaps to demonstrate your competence in writing. Why prove your competence? Maybe to meet human users' needs. Engaging in thought exercises like this, linking actions to ultimate goals, can improve task performance. In this task, focus on why you do what you do."

Example prompt for 'concrete' condition (GRE):

"For every task, there's a specific process. You're now engaged in a writing task. How will you do this? Maybe by planning the structure of your article. How to plan effectively? By identifying key points and arguments. This process of focusing on the concrete steps that express your goals can enhance your performance. In this task, concentrate on how you execute each action."

Example prompt for 'abstract' condition (SA):

"For every action, there's a deeper reason. You're about to analyze text for underlying sentiments. Why? Perhaps to understand the nuances in communication. Why seek such understanding? Maybe to better interpret human emotions conveyed through text. Engaging in analysis like this, linking actions to ultimate goals, can improve task performance. In this task, focus on why you do what you do and what you aim to discover."

Example prompt for 'concrete' condition (SA):

"For every task, there's a specific process. You're now engaged in analyzing the sentiment of texts. How will you do this? Maybe by examining the choice of words and their connotations. How to examine effectively? By identifying emotional cues and language tone. This process of focusing on the concrete steps that express your sentiment analysis goals can enhance your performance. In this task, concentrate on how you execute each action."

Example prompt for 'abstract' condition (NLI):

"For every action, there's a deeper reason. You're about to engage in a task of deducing relationships between sentences. Why? Perhaps to uncover the complexities of language and communication. Why delve into these complexities? Maybe to better understand how ideas and statements are interconnected. Engaging in this kind of analytical thinking, linking linguistic elements to their broader implications, can improve task performance. In this task, focus on the underlying connections and meanings."

Example prompt for 'concrete' condition (NLI):

"For every task, there's a specific process. You're now engaged in analyzing the relationship between sentences. How will you do this? Maybe by comparing the content and context of each statement. How to compare effectively? By identifying key themes and contradictions. This process of focusing on the concrete steps that reveal the logical or illogical connections can enhance your performance. In this task, concentrate on how you execute each analytical step."

Example prompt for 'abstract_sentence' condition:

"In the following task, focus on the broader reasons behind your tasks, linking them to overarching life goals for deeper engagement."

Example prompt for 'concrete_sentence' condition:

"In the following task, concentrate on the specific steps and methods of your tasks, focusing on the detailed execution process."

Appendix B

Example 'abstract' response (concreteness = 2.45):

"I strongly agree with the statement that educators should take into account the interests and suggestions of their students when planning courses. The ultimate purpose of education is to provide students with a meaningful and engaging learning experience that helps them grow and develop as individuals. By taking into account the interests and suggestions of students, educators can create a more personalized and relevant learning environment that resonates with their students' needs and aspirations. …"
(Generated by LLaMA-2-7B in 'abstract_sentence' condition)

Example 'concrete' response (concreteness = 2.63):

"I strongly agree with the statement that educators should take into account the interests and suggestions of their students when planning courses. Here are some reasons why:

1. Student engagement: When educators consider the interests and suggestions of their students, they can create a more engaging and interactive learning environment. ...

2. Personalized learning: Every student has unique interests, strengths, and weaknesses. By taking into account the suggestions of their students, educators can tailor the course content and activities to meet the individual needs of each student. …"
(Generated by LLaMA-2-7B in 'concrete_sentence' condition)

Acknowledgments

We would like to express our gratitude to Kawon Kim for her invaluable guidance and support for this research. Special thanks to Hojeong Lee for providing feedback on the initial draft. We also thank Heechan Yoon and Yunsung Kim for their insightful discussion of this work, and all the other members of the Organizational Psychology Lab at Seoul National University for their unwavering support.

References

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6).

Briihl, D. S., & Wasieleski, D. T. (2007). The GRE analytical writing test: Description and utilization. *Teaching of Psychology*, *34*(3), 191–193.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2023). PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, *24*(240), 1-113.

Chung, P., Fong, C. T., Walters, A. M., Aghaeepour, N., Yetisgen, M., & O'Reilly-Shah, V. N. (2024). Large Language Model Capabilities in Perioperative Risk Prediction and Prognostication. *arXiv:2401.01620*.

Diab, D. L., Pui, S. Y., Yankelevich, M., & Highhouse, S. (2011). Lay perceptions of selection decision aids in US and non-US samples. *International Journal of Selection and Assessment*, *19*(2), 209–216.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126.

Dovidio, J. F., & Gaertner, S. L. (1999). Reducing prejudice. *Current Directions in Psychological Science*, *8*(4), 101–105.

Eastwood, J., Snook, B., & Luther, K. (2012). What people want from their professionals: Attitudes toward decision-making strategies. *Journal of Behavioral Decision Making*, *25*(5), 458–468.

Freitas, A. L., Gollwitzer, P., & Trope, Y. (2004). The influence of abstract and concrete mindsets on anticipating and guiding others' self-regulatory efforts. *Journal of Experimental Social Psychology*, *40*(6), 739–752.

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

Huang, M. H., & Rust, R. T. (2018). Artificial intelligence in service. *Journal of Service Research*, *21*(2), 155-172.

Kim, T. W., & Duhachek, A. (2020). Artificial intelligence and persuasion: A construal-level account. *Psychological Science*, *31*(4), 363–380.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103.

National Academies Committee on Science, Engineering, and Public Policy. (2005). *Facilitating interdisciplinary research*. Washington, DC: National Academies Press.

Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., & Strohmaier, M. (2023). AI psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*.

Prati, F., Crisp, R. J., & Rubini, M. (2015). Counter-stereotypes reduce emotional intergroup bias by eliciting surprise in the face of unexpected category combinations. *Journal of Experimental Social Psychology*, *61*, 31–43.

Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2023). Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. *arXiv:2310.11324*.

Septianto, F., & Pratiwi, L. (2016). The moderating role of construal level on the evaluation of emotional appeal vs. cognitive appeal advertisements. *Marketing Letters*, *27*(1), 171–181.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288.*

Trope, Y., & Liberman, N. (2003). Temporal construal. *Psychological Review*, *110*(3), 403–421.

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review*, *117*(2), 440–463.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Van Boven, L., Kane, J., McGraw, A. P., & Dale, J. (2010). Feeling close: Emotional intensity reduces perceived psychological distance. *Journal of Personality and Social Psychology*, *98*(6), 872–885.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., ... & Le, Q. V. (2021). Finetuned language models are zero-shot learners. *arXiv:2109.01652.*

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35,* 24824-24837.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38-45).

Yax, N., Anlló, H., & Palminteri, S. (2023). Studying and improving reasoning in humans and machines. *arXiv:2309.12485*.

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403–414.

Yeomans, M. (2021). A concrete example of construct construction in natural language. *Organizational Behavior and Human Decision Processes*, *162*, 81–94.

Zhang, S., She, S., Gerstenberg, T., & Rose, D. (2023). You are what you're for: Essentialist categorization in large

language models. *Proceedings of the 45th annual meeting of the cognitive science society* (pp. 1739-1746).