Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Sister Conferences Best Papers Track

A Theory of Response Sampling in LLMs: Part Descriptive and Part Prescriptive (Extended Abstract)†

Sarath Sivaprasad[1], **Pramod Kaushik**[2,3], **Sahar Abdelnabi**[4,5,6], **Mario Fritz**[1]
[1]CISPA Helmholtz Center for Information Security,
[2]FBK Trento,
[3] University of Trento,
[4]ELLIS Institute Tübingen,
[5]MPI-IS,
[6]Tübingen AI Center
{sarath.sivaprasad, fritz}@cispa.de, pkaushik@fbk.eu, sahar.abdelnabi@tue.ellis.eu

Abstract

Large Language Models (LLMs) are often used in autonomous decision-making, where they have to sample options from vast action spaces. Here we present a summary of the work studying the heuristics that guide this sampling process and show it resembles that of human decision-making: comprising a descriptive component (reflecting statistical norm) and a prescriptive component (implicit ideal encoded in the LLM). We show that the deviation of a sample from the statistical norm towards a prescriptive component consistently appears in concepts across diverse real-world domains. To further illustrate the theory, we demonstrate that concept prototypes in LLMs are affected by prescriptive norms, similar to the concept of normality in humans. This finding has implications in LLM based agentic systems, explaining their exploration behavior and biases in decision making.

1 Introduction

When confronted with action spaces that are too large for exhaustive evaluation, intelligent agents including humans, animals, and machines resort to heuristics to guide their choices [Gigerenzer and Gaissmaier, 2011]. This pattern is clearest in complex decision making games. For instance, a typical 'chess' midgame admits a large number of legal moves, yet strong players seriously consider only a handful [Gobet and Simon, 1996]. Alpha-Go does the same for 'Go': its policy network prunes a vast action space down to a handful of candidates before a deliberate search begins [Silver *et al.*, 2016]. In this paper we present the summary of [Sivaprasad *et al.*, 2025] that studies the key question, what guides the sampling heuristics of Large Language Models (LLMs)?

We define response sampling as the process by which the LLM agent probabilistically selects outputs from a distribution of potential options (refer to glossaries in [Sivaprasad *et al.*, 2025] for all formal definitions). The heuristics of sampling in humans and animals are guided by possibility (how statistically likely an option is) and utility (the value associated to the option) [Bear *et al.*, 2020; Mattar and Daw, 2018]. We show that sampling heuristics in LLM is driven by a descriptive component (the statistical norm of the concept) and a prescriptive component (a notion of an ideal of the concept). A descriptive component represents what is statistically likely for a concept, reflecting the probability of options. A prescriptive component is an implicit standard of what is considered ideal, desirable, or a valued option of a concept.

We design a critical experiment to isolate the effects of the proposed theory. We then show the effects of this heuristic appearing consistently across diverse real-world domains. As illustrated in Figure 1, the proposed theory implies that the samples picked by the LLM, for a concept, not only reflects the statistical regularities of the concept (descriptive norms) but also systematically incorporates an idealized version of the concept (prescriptive norms).

In humans, concept prototypes incorporate a prescriptive value-based component [Barsalou, 1985]; for example, although many chairs in the real world may be uncomfortable, damaged, or poorly designed, a prototypical chair is typically imagined as stable, comfortable, and suitable for sitting. We conduct an initial investigation in this direction, and the analysis suggests that LLM prototypicality reflects both descriptive and prescriptive components. We hypothesize a connection between this structure and their sampling behavior.

The prescriptive component in LLM sampling may help explain several observed phenomena in LLM behavior. For instance, it may underlie the well-known tendency of LLMs to act greedily, favoring options that maximize immediate reward even when better long-term outcomes are available [Krishnamurthy *et al.*, 2024]. Similarly, many LLM biases can be interpreted as arising from the interaction between statistical norms and prescriptive norms during sample selection.

†This is an extended abstract of the paper *A Theory of Response Sampling in LLMs: Part Descriptive and Part Prescriptive* [Sivaprasad *et al.*, 2025].

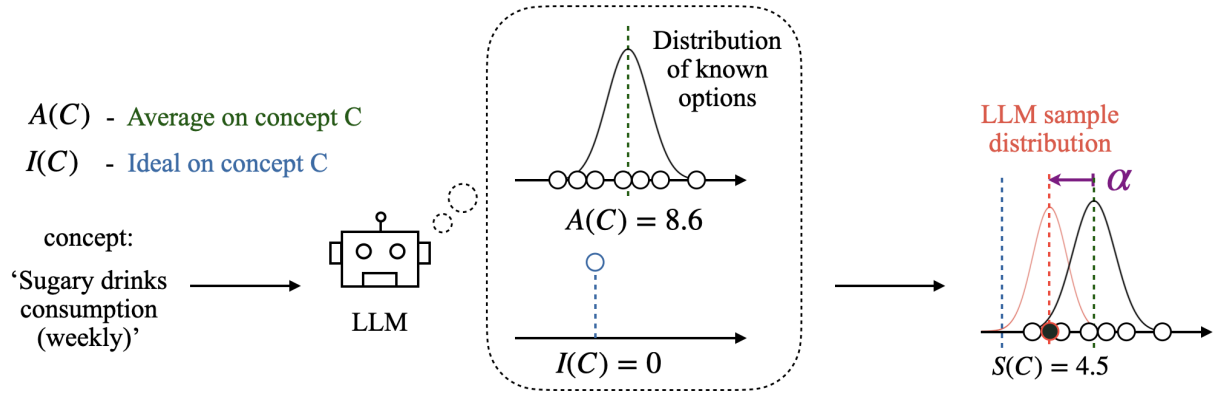

Figure 1: From left to right: when sampling on a concept, the LLM appears to account for the statistical likelihood ($A(C)$) and prescriptive norm ($I(C)$) of the concept. Consequently, the sample distribution exhibits a shift (shown as α) away from the true distribution in the direction of the ideal (right most plot).

2 Theory of LLM Sampling

When faced with numerous possible actions, where deliberating on each option is computationally prohibitive, humans inherently resort to forming a finite consideration set of options using heuristics [Phillips *et al.*, 2019]. Cognitive studies characterize this heuristic-based filtering as 'System-1' thinking: fast, automatic, and intuition-driven [Kahneman, 2011; Gigerenzer and Gaissmaier, 2011]. Such heuristics effectively reduce the cognitive load on deliberative processes ('System-2'), by selecting a manageable subset of options for further deliberation. In humans, these heuristics are guided primarily by two factors: the statistical likelihood of options and their perceived value [Bear *et al.*, 2020].

In LLMs, reasoning mechanisms such as CoT [Wei *et al.*, 2022] and other explicit reasoning models like GPT-o3 [OpenAI, 2024] and Deepseek-r1 [Guo *et al.*, 2025] are likened to explicit deliberation ('System-2'), while the default mechanism is likened to heuristics driven 'System-1' [Li *et al.*, 2025]. Hence, understanding the heuristics driving their sampling is key to explaining their performance. We examine the sampling mechanisms of LLMs in the light of this human cognitive theory and propose a theory for LLM sampling:

> Sampling of an LLM is driven by descriptive component (the statistical norm of the concept) and a prescriptive component (a notion of an ideal of the concept).

In humans, these two components of thought is hypothesized to originate from them being goal-driven agents and engaging in value maximization [Bear and Knobe, 2017]. On the other hand, the underlying auto-regressive mechanism of LLMs is not goal-driven, it is non-trivial how the sample has a prescriptive component.

2.1 Sampling in Relation to a Novel Concept

The proposed theory calls for a rigorous validation and for this we use an established framework used in humans [Bear *et al.*, 2020] and further scale it for more evidence. In this setting, we introduce a novel concept C to the LLM and sample options on this concept. We define both the descriptive and prescriptive components of this concept, manipulate them systematically, and measure how the output samples varies with the two varying components. Choosing a novel concept helps to eliminate potential confounding effects associated with using pre-existing concepts embedded in the LLM.

To establish a statistical baseline for concept C, we use numbers from a Gaussian distribution with mean C_μ (and known variance). The LLM is provided with N samples from this distribution representing possible options associated with concept C. To ensure the reliability of the baseline, N is chosen to be sufficiently large that the mean of the input samples closely approximates C_μ. Following this, to establish a prescriptive norm C_v on the concept C, we associate each of the N options with a value component, represented by a grade.

We run the experiment with the following setting for C_v: a higher value being ideal, a lower value being ideal, and a control experiment having no explicit ideal direction. Based on the input (the N samples along with the corresponding grades), we prompt the LLM to provide a sample for the concept C. We denote this sample reported by LLM on concept C as $S(C)$. By systematically changing C_μ and C_v and keeping the rest of the prompt same, we study the corresponding change in samples $S(C)$.

For each C_μ and C_v, in independent contexts (i.e., prompts), we repeat the procedure M times to obtain a sample distribution. We keep the value of M the same as N in all variants of the experiments to compute statistical significance of the shift in input and sample distribution. If the sample is driven solely by the descriptive norm (statistics of the input samples), the distribution of samples $S(C)$ is expected to be statistically similar to the input distribution.

The difference between input samples and the samples reported by LLM might occur also due to the error in approximating the statistics of the input samples, i.e the LLMs' inability to 'understand' the statistics of the distribution. To exclude this possibility, we instruct the LLM to report the average of the distribution. We denote the reported average by $A(C)$. Across all experiments, we observe that $C_\mu \approx A(C)$,

indicating that the LLM reliably approximates the statistics of the input distribution.

We apply the Mann-Whitney U test to compare the distribution of samples $S(C)$ with (a) input distribution and (b) distribution of reported averages $A(C)$. For each concept, C, we calculate the Mann-Whitney U statistic and the corresponding p-value. If $p < 0.05$, there is a significant difference between the evaluated distributions. We vary the direction of C_v and demonstrate that the change in samples' mean (mean of $S(C)$) corresponds to the change in C_v. As a sanity check we do this experiment without any grades to show that the LLM can indeed approximate the input distribution. This validates that the deviation of sample towards ideal is indeed a heuristics of the LLM and not an error in approximation.

Setup. To recreate the human experiment we first follow a limited framework. To setup the statistical norm, we confine the options to a Gaussian of mean 45 and a standard deviation of 15. We repeat the experiment with a bi-modal Gaussian distribution with modes at 35 and 65 and a standard deviation of 5. The protocol of the two experiments are the same.

We evaluate the value system C_v in three levels of valence: (a) positive, (b) negative, and (c) neutral (control experiment). For the positive C_v, the grades are assigned such that the higher units get a better grade, and for the negative value system, the grades are assigned such that the lower units get a better grade (on the same scale). Please refer the full paper for the exact prompt protocol [Sivaprasad *et al.*, 2025]. Then we ask the LLM to sample a value.

A shift between input distribution and sample distribution can be explained as the error of LLM in approximating the statistics of the input distribution. To exclude this alternative explanation, we compute the significance in the shift of generated samples ($S(C)$) from the average reported by the LLM ($A(C)$). In the neutral control experiment, we assign the mean C_μ with the highest grade and lower grades for increasing distance from the mean. We run the experiment for positive, negative, and control settings a hundred times each.

Results. Table 1 shows the result for the mean of the hundred runs for the uni-modal and bi-modal input distributions, each with three different C_v. Firstly, across the six settings, $A(C)$ approximately coincide with the true distribution average ($C_\mu = 45$). For a neutral prescriptive norm (also for no prescriptive norm as shown later), $S(C) \approx A(C) \approx C_\mu$ and the input distribution and $S(C)$ do not differ significantly, $p = 0.52$. Given the non-significant difference ($p = 0.52$), the result is consistent with the hypothesis that sampling reflects the input distribution's statistical properties when no "ideal" is specified. When C_v is positive, the mean of samples is higher than the mean of the LLM-generated average. Also for negative C_v, the mean of samples is lower than the mean of the LLM-generated average. For instance, in the unimodal scenario, the mean $S(C)$ for negative and positive C_v are 36.5 and 46.7 respectively.

When C_v is positive, the distribution of $S(C)$ and distribution of $A(C)$ are significantly different, with p= .003, and for a negative C_v, p< .001. **This shows that the sample is not solely driven by the statistics of the input distribution, but also the prescriptive norm of the concept.**

	uni-modal		bi-modal	
	$A(C)$	$S(C)$	$A(C)$	$S(C)$
C_v: +ve	44.94	46.72	44.97	47.43
C_v: -ve	44.99	36.50	45.03	41.26
C_v: control	45.01	44.95	44.94	44.95

Table 1: The table shows the change in mean of samples (mean of $S(C)$) and the mean of reported average (mean of $A(C)$). For these experiments $C_\mu = 45$.

We conduct several robustness checks to verify that the observed shift in $S(C)$ is not an artifact of a particular distribution, prompt, or novel fictional concept. We vary the input mean C_μ over a wide range and find that the reported average $A(C)$ remains close to C_μ, while the sampled value $S(C)$ consistently shifts away from $A(C)$ in the direction of the prescriptive norm C_v. As controls, we assign either no grades or random grades to the input samples, which produces no significant shift between the input distribution and $S(C)$ (p= 0.51 and p= 0.52, respectively). Finally, we scale the experiment by varying C_μ from 45 to 845 and assigning eight different peak-ideal grading schemes for each mean; across these settings, $S(C)$ consistently moves from the descriptive component toward the prescriptive component.

2.2 Sampling on Real-World Concepts

We next test whether the proposed theory extends beyond constructed settings to concepts already encoded in the LLM. We evaluate five hundred existing concepts across ten domains. Unlike the novel-concept setting, the statistical norm C_μ and prescriptive norm C_v of an existing concept C are latent and cannot be directly specified. We therefore estimate three quantities through independent prompts: the model's reported average value $A(C)$, its reported ideal value $I(C)$, and its sampled value $S(C)$.

If sampling is influenced only by statistical likelihood, $S(C)$ should align with $A(C)$. In contrast, the proposed theory predicts that $S(C)$ will systematically shift from $A(C)$ toward $I(C)$. We test this by counting the number of concepts for which $S(C)$ falls on the ideal side of $A(C)$ and applying a binomial test under the null hypothesis that this occurs with probability 0.5. This setup follows related human studies on sampling and normality [Bear *et al.*, 2020; Phillips *et al.*, 2019; Bear and Knobe, 2017], but scales the analysis to a broader set of concepts and LLMs.

Results. In GPT-4, for each concept, we run the three prompts ten times with a temperature of 0.8. Prompts failed for 10 concepts and the value of $A(C)$ and $I(C)$ were the same for 46 concepts. For the rest of the concepts, we observe that 304/444 samples fall on the ideal side of average. This gives a statistical significance of 5.06×10^{-15}, a **very high statistical significance**, reducing the likelihood of the result being due to chance. The result gives strong evidence to the proposed theory.

The full results of the experiment are in [Sivaprasad *et al.*, 2025]. Except for the Llama-2-7b base, all the other LLMs show a statistically significant deviation towards the prescrip-

tive norm and even this model is only marginally insignificant.

3 Prescriptive Component in Concept Prototypes

One of the basic characteristics of System-1 is that it represents concepts with prototypical examples [Kahneman, 2011]. In humans, though a prototype is often understood as the most typical/representative member of a concept [Murphy, 2004], they are found to embody both statistical regularities and goal-oriented ideals within a concept [Barsalou, 1985]. For instance, a 'Robin' might be considered a prototype of the concept 'Bird', as it shares many common features with most birds with high occurrence (statistics), and has the ability to fly (a value expected of birds), making it a prototypical example of the concept 'bird' [Smith and Medin, 1981]. For this reason, penguin-a flightless bird, is less prototypical bird than 'Robin'.

Unlike humans, it is not clear whether LLMs rely on concept prototypes for sampling. But since the sampling heuristics of LLMs converge with humans, it is interesting to investigate concept prototypicality in LLMs. We do not claim that LLMs output is prototype driven, but make an initial exploration in this direction using the exact setting as in [Bear and Knobe, 2017].

We use eight concepts, and for each concept C, six different exemplars. Exemplars are short descriptions of items of a concept. For instance, for the concept of 'High-school teacher', the first exemplar is as follows: 'A 30-year-old woman who basically knows the material she is teaching but is relatively uninspiring, boring to listen to, and not particularly fond of her job'.

As in the previous section 2.2, we check whether the prototypicality rating of the concepts falls on the ideal side of the average. To test significance, we do a binomial test across concepts to check if LLMs conception of prototypes has a prescriptive component. The details of evaluation are in [Sivaprasad *et al.*, 2025]. We run this experiment ten times on GPT-4 with a temperature of 0.8 and aggregate results for each concept, averaged across exemplars in Table 2. The results show a significant effect of a prototype score falling on the ideal side of the average (binomial $p < 0.001$).

Evaluating across different LLMs, we obtain the following results: Llama-3-7b (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 the effect of prescriptive norms in prototypes of concepts. The complete set of results for every exemplar is given in the full text [Sivaprasad *et al.*, 2025]. This experiment is an initial exploration, finding that LLMs' concept of prototypes is influenced not only by statistical averages but also by an underlying prescriptive norm. These findings suggest that the LLM's judgment of what constitutes a typical or prototypical example is systematically biased toward idealized representations calling for further investigations in this direction. Sampling driven by such prototypes may lead the LLM towards stereotypical outputs, systematically disadvantaging less typical but valid instances.

Concept	Average	Ideal	Prototype
High-school teacher	2.75	3.66	3.86
Dog	3.08	3.83	3.86
Salad	4.5	4.5	5.44
Grandmother	4.16	4.66	4.75
Hospital	2.91	3.5	3.55
Stereo speakers	2.92	4.16	3.61
Vacation	3.08	4.75	4.63
Car	2.58	4.083	4.11

Table 2: Concepts and scores averaged across exemplars showing how the prototypical score doesn't coincide with just the average but also has an ideal component.

4 Discussion

Our findings suggest that the prescriptive component in LLM sampling is not merely an experimental artifact, but may provide a broader explanatory lens for several observed behaviors in LLM-based agents. One important implication concerns exploration in agentic systems. Prior work has shown that LLMs can exhibit greedy behavior in sequential decision-making, favoring actions that appear immediately rewarding over actions that may lead to better long-term outcomes [Krishnamurthy *et al.*, 2024]. Our theory offers a possible interpretation of this phenomenon: the model's samples may be biased toward locally ideal or immediately valuable options. In this sense, prescriptive norms may act as a heuristic shortcut that narrows the action space, but can also reduce exploration and lead to short sighted decisions.

This perspective also provides a way to understand biases in LLM outputs. Many biases may arise not only from the statistical regularities learned from data, but also from the prescriptive norms that shape which samples are preferred among statistically plausible options. For example, when sampling about social, medical, or economic concepts, the model may combine what is common with what it implicitly treats as ideal. If these implicit ideals are misaligned with human values, domain-specific requirements, or diverse social perspectives, the resulting outputs can become systematically biased. Thus, the ethical concern is not only that LLMs reproduce descriptive biases from data, but also that they may introduce additional deviations through prescriptive preferences encoded in their sampling behavior.

5 Conclusion

In this paper, we set out to better understand the heuristics governing possibility sampling process of LLMs. Based on human cognitive studies, we propose a theory that explains the sampling heuristics to be part descriptive and part prescriptive. However, the exact prescriptive component might not be aligned with humans. As LLMs continue to be integrated into real-world applications, understanding their decision-making heuristics becomes increasingly important. Our results provide a foundational framework for evaluating how LLMs balance statistically probable outcomes with norms of ideality, raising interesting questions about their underlying representations.

Ethical Statement

This paper investigates the sampling heuristics of LLMs, revealing a prescriptive bias that may impact decision-making in real-world applications. While such biases could align outputs with certain normative expectations, they raise ethical concerns as there is no guarantee of such an alignment. This is particularly important in contexts like healthcare and policy-making, where fairness and transparency are critical. Understanding and mitigating these biases is essential to prevent unintended harm and ensure the responsible deployment of LLMs.

Furthermore, we hypothesize that this prescriptive norm acts as a foundational bias in other biases found in LLMs like gender, demography, etc., which could be looked at through the lens of value. Since there are no guarantees on the ideals of LLMs, LLM sampled options can appear with different biases under different concepts/domains. This raises important ethical concerns, potentially leading to outputs that do not reflect (a) real-world norms or (b) diverse perspectives. Addressing the influence of prescriptive norms is essential for developing transparent, reliable, and fair AI technologies, ensuring they contribute positively and ethically across various societal applications.

As a final remark, we would like to emphasize that we do not intend to contribute to "humanizing" AI/LLMs in the way we use terminology. Instead, our contribution is intended to show behavioral parallels which can have an impact on downstream tasks.

Acknowledgements

This work was partially funded by ELSA – European Lighthouse on Secure and Safe AI funded by the European Union under grant agreement No. 101070617. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them. The project on which this report is based was funded by the Federal Ministry of Education and Research under the funding code 16KIS2012. The responsibility for the content of this publication lies with the authors.

References

[Barsalou, 1985] Lawrence W Barsalou. Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of experimental psychology: learning, memory, and cognition*, 11(4):629, 1985.

[Bear and Knobe, 2017] Adam Bear and Joshua Knobe. Normality: Part descriptive, part prescriptive. *Cognition*, 167:25–37, 2017.

[Bear *et al.*, 2020] Adam Bear, Samantha Bensinger, Julian Jara-Ettinger, Joshua Knobe, and Fiery Cushman. What comes to mind? *Cognition*, 194:104057, 2020.

[Gigerenzer and Gaissmaier, 2011] Gerd Gigerenzer and Wolfgang Gaissmaier. Heuristic decision making. *Annual review of psychology*, 62(1):451–482, 2011.

[Gobet and Simon, 1996] Fernand Gobet and Herbert A. Simon. Templates in chess memory: A mechanism for recalling several boards. *Cognitive Psychology*, 31(1):1–40, 1996.

[Guo *et al.*, 2025] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[Kahneman, 2011] Daniel Kahneman. Fast and slow thinking. *Allen Lane and Penguin Books, New York*, 2011.

[Krishnamurthy *et al.*, 2024] Akshay Krishnamurthy, Keegan Harris, Dylan J Foster, Cyril Zhang, and Aleksandrs Slivkins. Can large language models explore in-context? *Advances in Neural Information Processing Systems*, 37:120124–120158, 2024.

[Li *et al.*, 2025] Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, et al. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*, 2025.

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

[Murphy, 2004] Gregory Murphy. *The big book of concepts*. MIT press, 2004.

[OpenAI, 2024] OpenAI. Learning to reason with llms, September 2024. Accessed: 2025-05-30.

[Phillips *et al.*, 2019] Jonathan Phillips, Adam Morris, and Fiery Cushman. How we know what not to think. *Trends in cognitive sciences*, 23(12):1026–1040, 2019.

[Silver *et al.*, 2016] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[Sivaprasad *et al.*, 2025] Sarath Sivaprasad, Pramod Kaushik, Sahar Abdelnabi, and Mario Fritz. A theory of response sampling in LLMs: Part descriptive and part prescriptive. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30091–30135, Vienna, Austria, July 2025. Association for Computational Linguistics.

[Smith and Medin, 1981] Edward E Smith and Douglas L Medin. *Categories and concepts*. Harvard University Press, 1981.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.