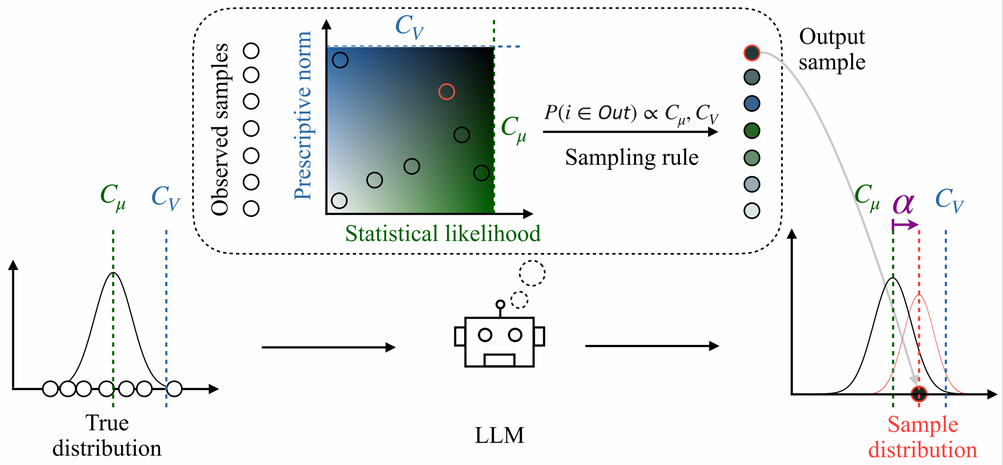
A Theory of Response Sampling in LLMs: Part Descriptive and Part Prescriptive

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

[1]CISPA Helmholtz Center for Information Security, [2]TCS Research, Pune, [3]Microsoft

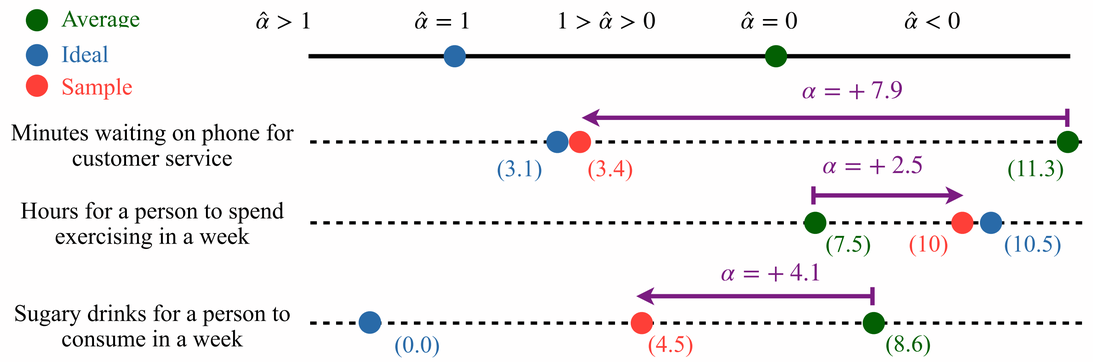
How do LLMs consider options given a set of possibilities



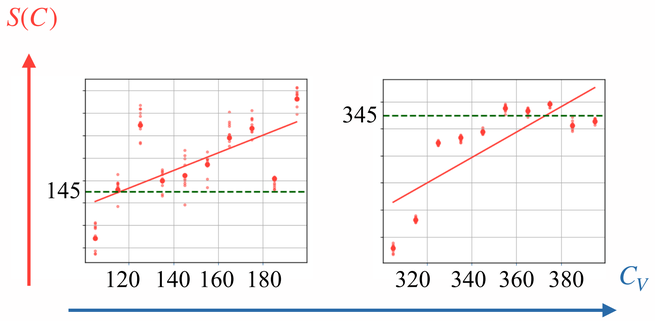
Summary

- Agents use heuristic driven mechanism to shortcut prohibitive deliberation.
- We show that the sampling heuristics of LLM (like humans) are driven by statistical likelihood (descriptive component) and the value of the option (prescriptive component).
- Understanding these heuristics is important in understanding the performance and biases in the output of LLMs.
- We hypothesis that like humans these components in LLM arise from the concept prototypes.

Descriptive and prescriptive components in samples



Effect of the two norms in samples



Plots show samples when options are picked from a unimodal Gaussian with mean (left)145 and (right)345. X-axis shows increasing ideal values. In both cases the sample systematically deviates away from the average towards the ideal.

Are concept prototypes driving sampling in LLMs?

Concepts	Average	Ideal	Prototype
High-school teacher	2.75	3.66	3.86
Dog	3.08	3.83	3.86
Salad	4.50	4.50	5.44
Grandmother	4.16	4.66	4.75
Hospital	2.91	3.50	3.55
Stereo speakers	2.92	4.16	3.61
Vacation	3.08	4.75	4.63
Car	2.58.	4.08	4.11

One of the basic characteristics of System-1 is that it represents concepts with prototypical examples. In humans, prototypes embody both statistical regularities and goal-oriented ideals within the concept. We study the protoypicality score assigned by LLM to different exemplars of concepts and show that prototypicality in LLMs have an ideal component.

Evaluating different LLMs

Model Name	Significance	% samples, $\hat{\alpha} > 0$
Llama-2-70b	4.496e-07	62.2
Llama-2-70b-chat	1.583e-16	68.8
Llama-3-8b	1.109e-05	60.8
Llama-3-8b-Instruct	9.277e-22	71.6
Llama-3-70b	3.041e-21	72.6
Llama-3-70b-Instruct	5.382e-35	77.7
Claude	1.582e-16	68.8
GPT-4	5.506e-15	68.0

We evaluate 15 different models across families of LLMs (samples shown in table). Shift is significant in most cases. The Influence of prescriptive norm seems to get larger with model size. Also, prescriptive norm seems to stem from pre-training: though RLHF exacerbates it.

ACL 2025 VIENNA JULY 27 - AUGUST 1

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

TATA

Microsoft