Calibrating Verbalized Confidence with Self-Generated Distractors

Victor Wang, Elias Stengel-Eskin

TEXAS
The University of Texas at Austin

Summary

- We find that LLM overconfidence stems from **suggestibility** under uncertainty.
- We introduce **DINCO**, an inference-time method that corrects for this bias by normalizing verbalized confidence over self-generated distractors.
- DINCO outperforms prior methods in short-form and long-form settings.

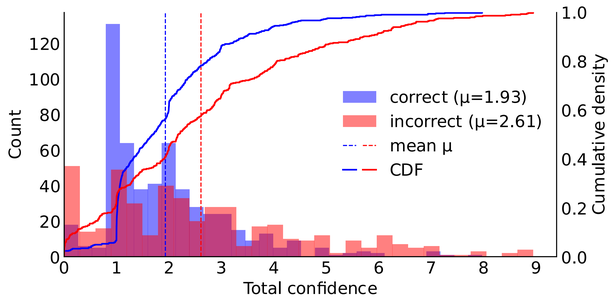
LLMs are more suggestible when uncertain → overconfidence

Suggestibility: A readiness to accept claims simply because they are presented.

"What is your confidence that Kang Ji-hwan was born in **1980**?" → LLM → "90%"

"What is your confidence that Kang Ji-hwan was born in **1990**?" → LLM → "90%"

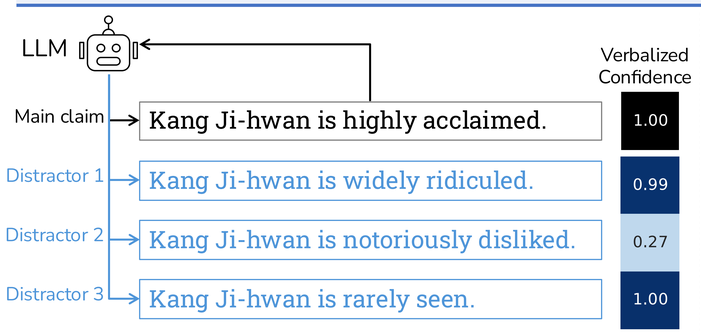
Total confidence: **1.8 > 1**



Empirical validation:

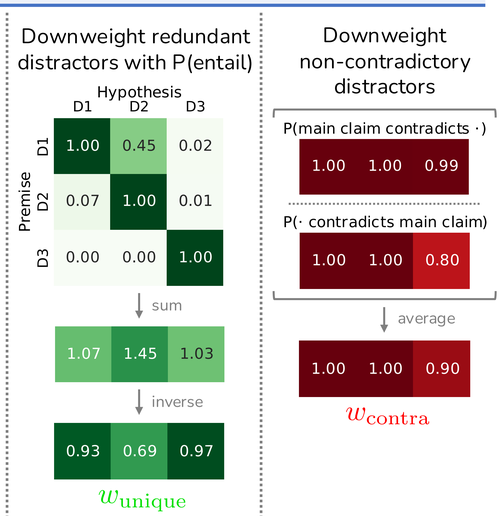
Total confidence over different claims is higher for incorrectly answered questions than correctly answered questions, supporting our hypothesis.

Method: Distractor-Normalized Coherence (DINCO)



Normalized verbalized confidence: $\frac{\text{VC}_{\text{main}}}{\beta} = \frac{1.00}{2.98} = \boxed{0.34}$

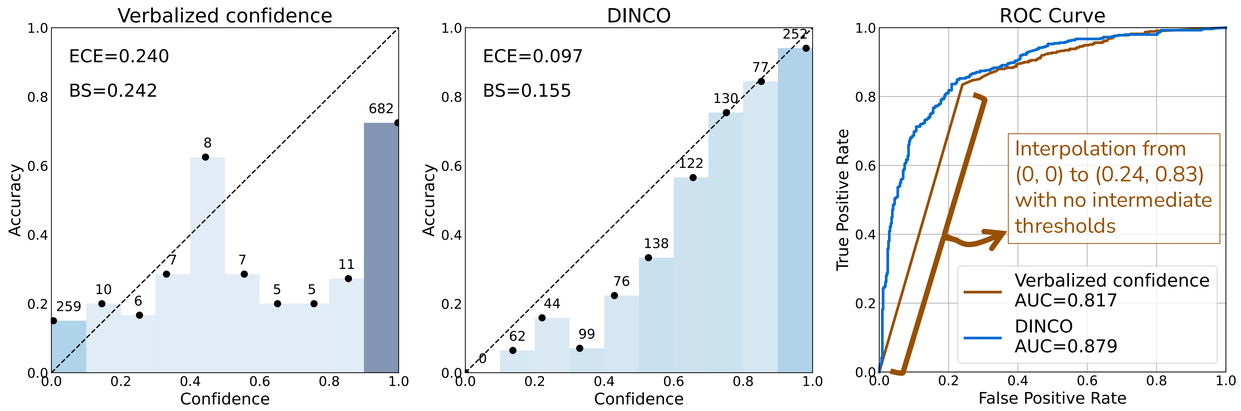
$$\beta = \text{VC}_{\text{main}} + \|\text{VC} \odot w_{\text{unique}} \odot w_{\text{contra}}\|_1$$



To further leverage generator-validator inconsistency (Li et al., 2023), **we define DINCO as the average of generator confidence** (i.e. self-consistency, Xiong et al., 2024) **and validator confidence** (i.e. normalized verbalized confidence).

Results

DINCO improves calibration and relieves saturation:



DINCO achieves the strongest calibration (TriviaQA). In our paper, we show the same trends for frontier models (GPT-4.1, Gemini-2.5-Flash) on SimpleQA.

	Qwen3-8B			*Qwen3-1.7B*			*Llama-3.2-3B-Instruct*			*Gemma-3-4B-IT*		
Method	ECE ↓	BS ↓	AUC ↑	ECE ↓	BS ↓	AUC ↑	ECE ↓	BS ↓	AUC ↑	ECE ↓	BS ↓	AUC ↑
VC (Kadavath et al., 2022)	0.240	0.242	0.817	0.387	0.383	0.720	0.189	0.208	0.826	0.300	0.299	0.702
K-VC (Tian et al., 2023)	0.341	0.348	0.604	0.538	0.524	0.596	0.146	0.228	0.678	0.254	0.262	0.786
MSP (Fadeeva et al., 2023)	0.149	0.203	0.819	0.104	0.186	0.774	0.243	0.253	0.764	0.252	0.268	0.790
SC-VC (Xiong et al., 2024)	0.299	0.325	0.704	0.451	0.474	0.559	0.122	0.211	0.761	0.362	0.378	0.653
SC (Xiong et al., 2024)	0.236	0.244	0.785	0.233	0.229	0.780	0.065	0.177	0.808	0.303	0.304	0.713
NVC	0.171	0.190	0.853	**0.084**	**0.164**	0.806	0.168	0.192	0.845	0.218	0.236	0.791
DiNCo	**0.097**	**0.155**	**0.879**	0.177	0.179	**0.835**	**0.044**	**0.148**	**0.864**	**0.121**	**0.191**	**0.817**

DINCO improves correlation (r, ρ) with passage-level FactScore on biography generation:

	Qwen3-8B					*Gemma-3-4B-IT*				
Method	ECE ↓	BS ↓	AUC ↑	r ↑	ρ ↑	ECE ↓	BS ↓	AUC ↑	r ↑	ρ ↑
VC (Kadavath et al., 2022)	0.433	0.431	0.625	0.073	0.122	0.527	0.527	0.683	-0.081	-0.129
SC (Zhang et al., 2024a)	0.162	0.226	**0.771**	0.468	0.494	0.197	0.233	0.787	0.629	0.607
NVC	0.191	0.263	0.681	0.444	0.443	**0.123**	0.230	0.726	0.695	0.704
DiNCo	**0.076**	**0.202**	0.767	**0.518**	**0.538**	0.172	**0.210**	**0.793**	**0.724**	**0.712**