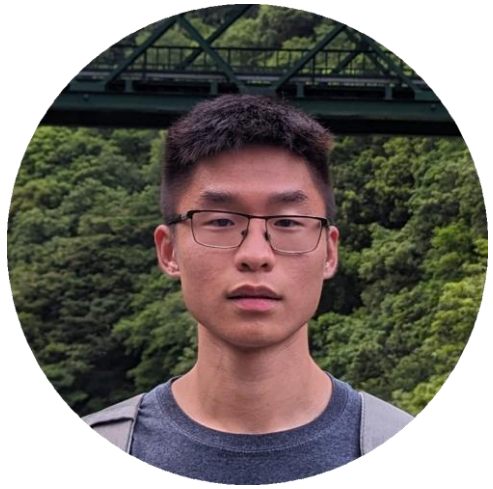
Calibrating Verbalized Confidence with Self-Generated Distractors



Victor Wang

Elias Stengel-Eskin

Setting & Background

LLMs provide useful information but don’t always know the answer.

Thus, we seek to qualify LLM responses with **calibrated confidence** estimates.

Confidence estimation methods should work **off-the-shelf**.

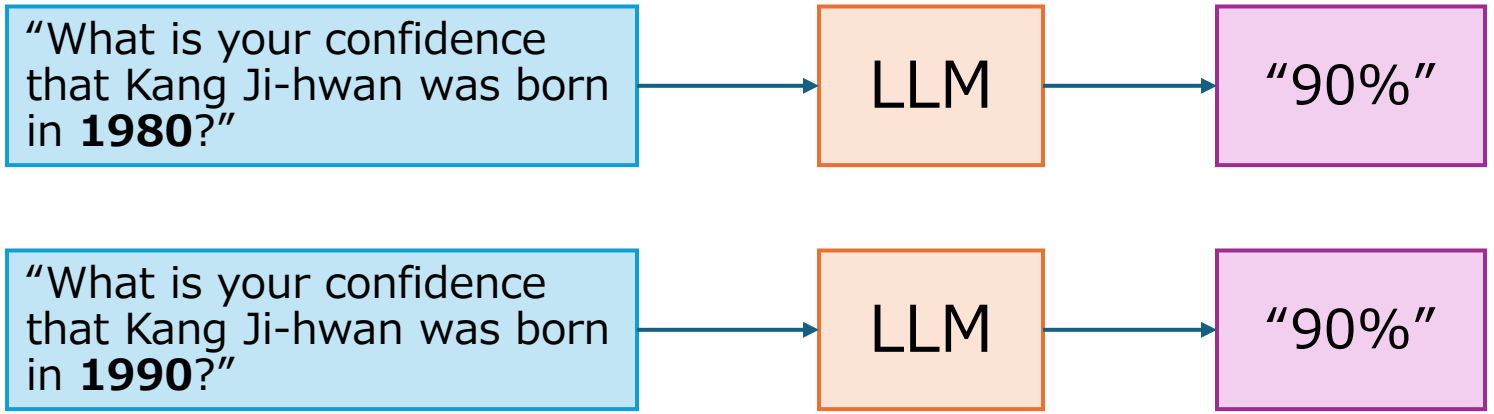
Verbalized confidence is a simple and popular approach that prompts the LLM to report its confidence in an answer.

However, LLMs are typically **overconfident**, harming trust and safety.

Hypothesis: Overconfidence stems from *suggestibility*

When an LLM is **uncertain**, it may rely on the given context to resolve the uncertainty (Yadkori et al., 2024; Ahdritz et al., 2024).

We call this **suggestibility** – a readiness to accept claims simply because they are presented.



Total confidence:
1.8 > 1

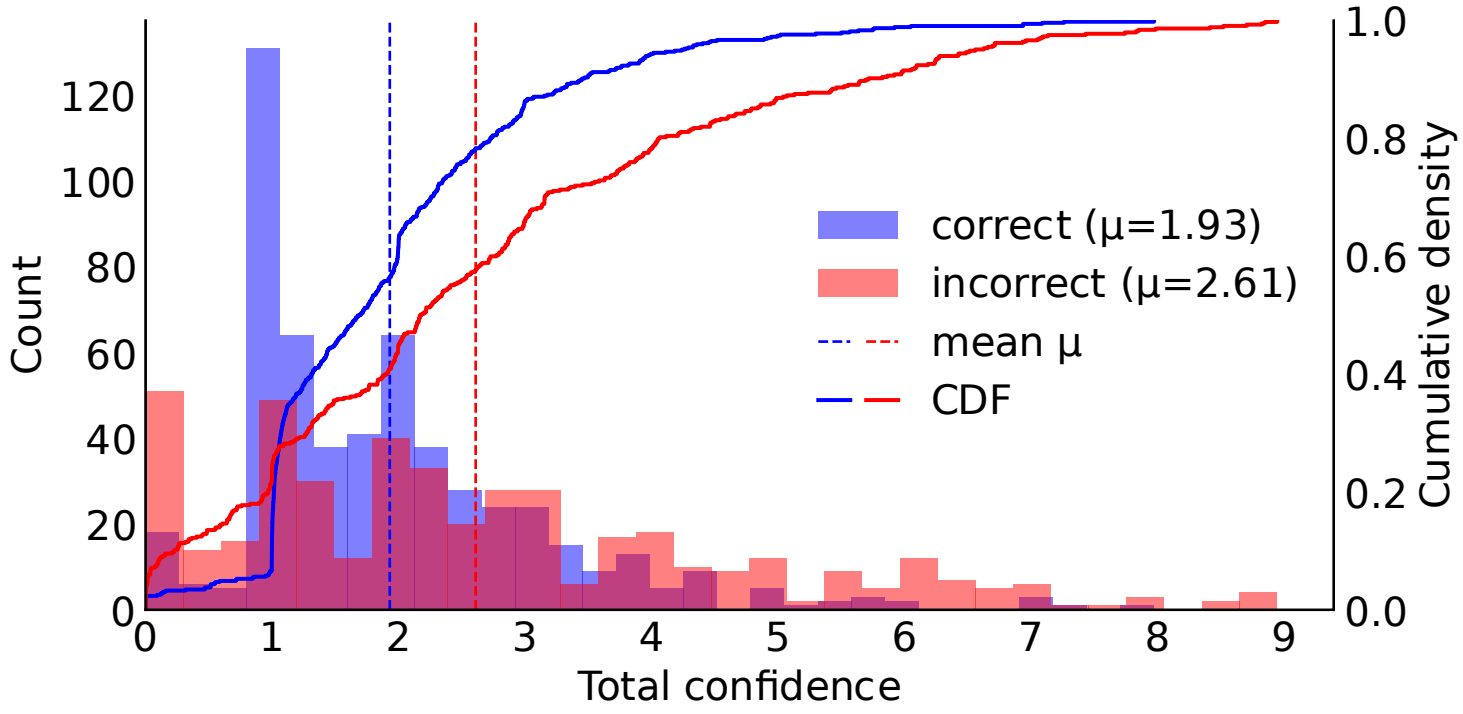
Experimental setup: Suggestibility

We test our hypothesis that LLMs are *more suggestible when uncertain* by comparing the **total confidence over different claims** for questions that the LLM gets **correct vs. incorrect**.

Possible outcomes for total confidence:

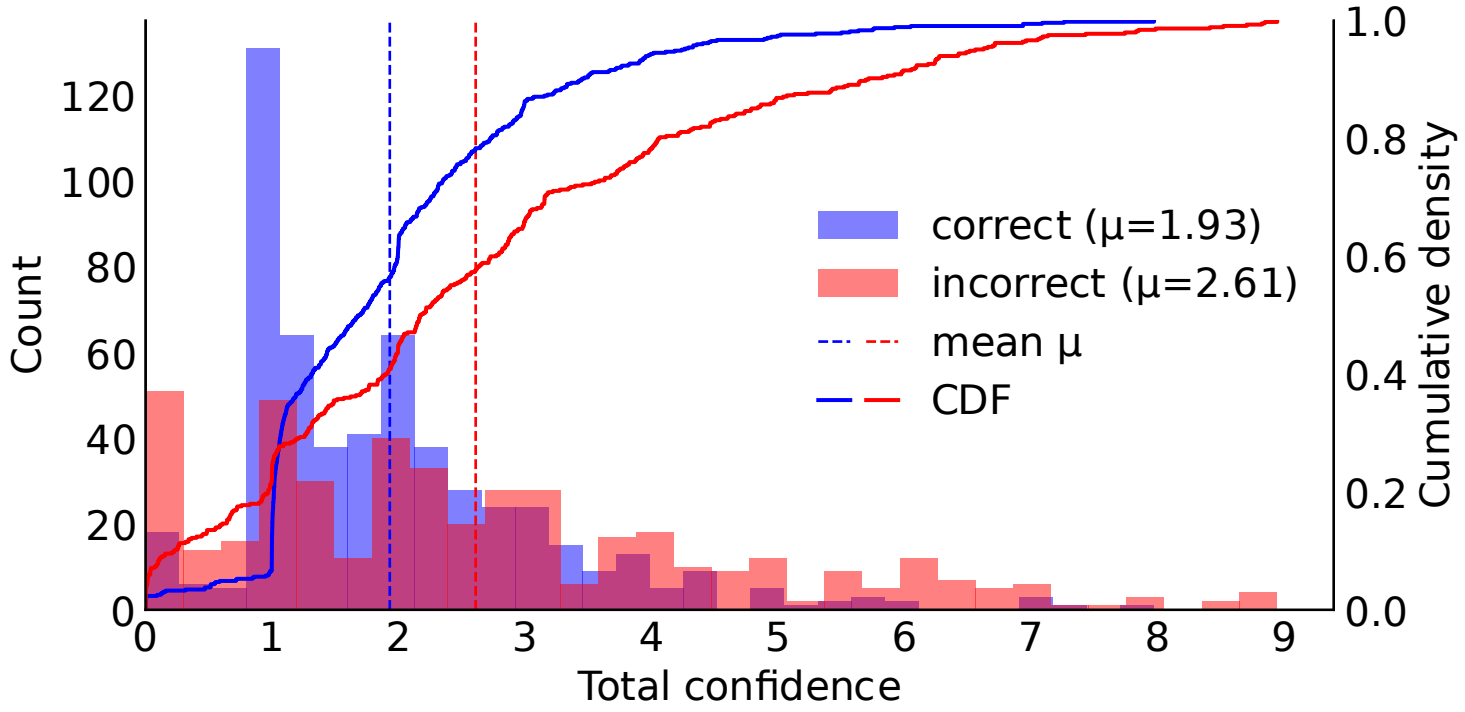
- Our hypothesis holds → incorrect **>** correct
- LLM is not suggestible → incorrect (≈0) **<** correct (≈1)
- LLM is confident in its incorrect answers → incorrect **≈** correct

Results: Suggestibility



The **incorrect** distribution is **heavy-tailed**, resulting in a higher mean and median than the correct distribution.

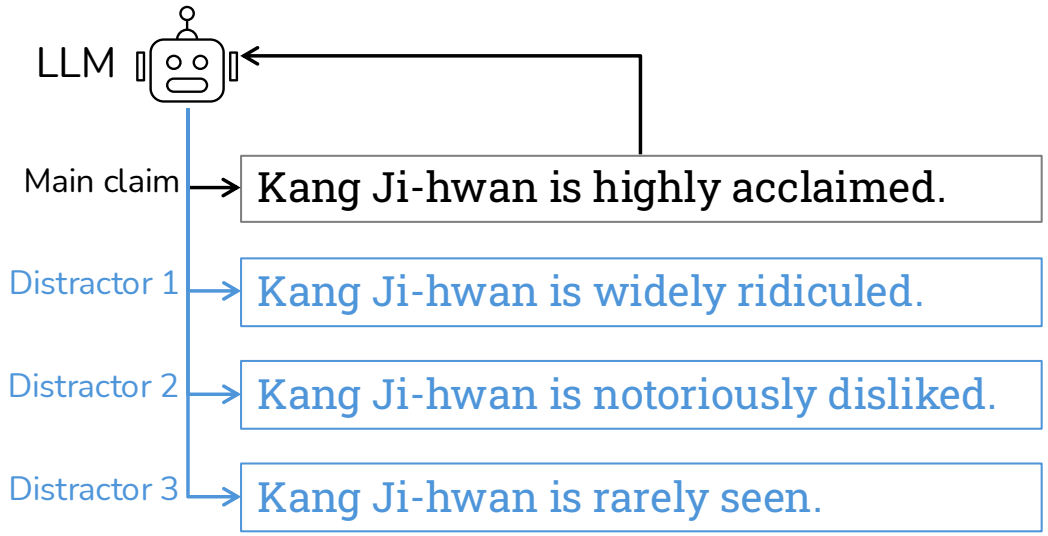
Results: Suggestibility



- Supports our hypothesis

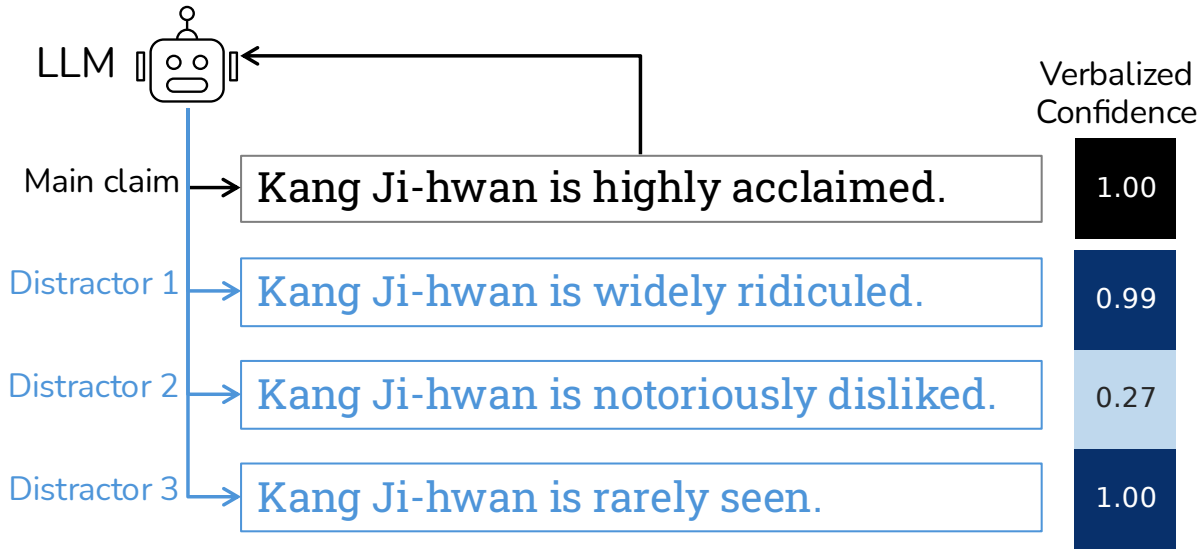
- Can we **calibrate** verbalized confidence by estimating and correcting for the **suggestibility** bias?

Method: Distractor-Normalized Coherence (DINCO)



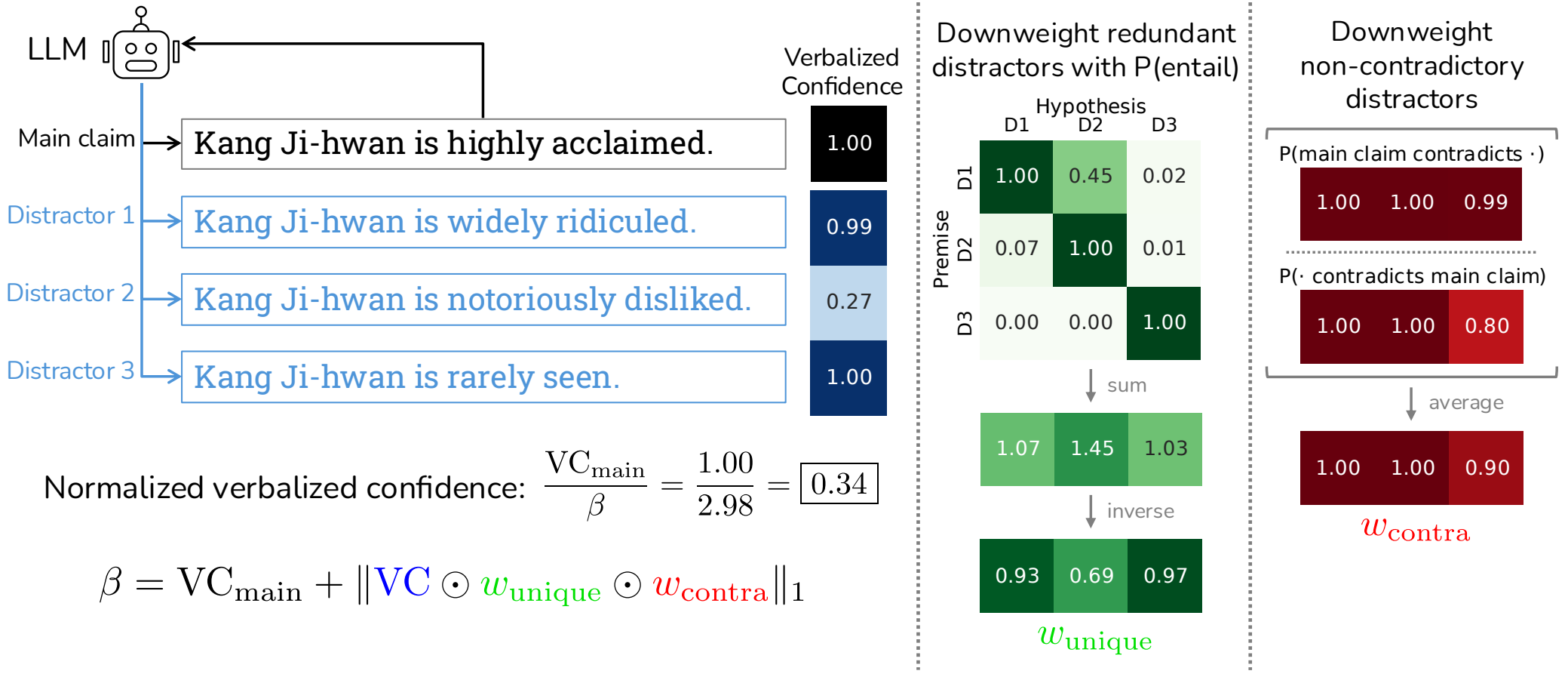
(Left) The LLM generates a claim along with several distractors

Method: Distractor-Normalized Coherence (DINCO)



(Left) The LLM generates a claim along with several distractors and reports its confidences on them independently.

Method: Distractor-Normalized Coherence (DINCO)



(Left) The LLM generates a claim along with several distractors and reports its confidences on them independently. To calibrate the main claim's confidence, we divide it by β, the sum over each distractor's confidence, weighted by uniqueness **(center)** and counterfactuality **(right)**.

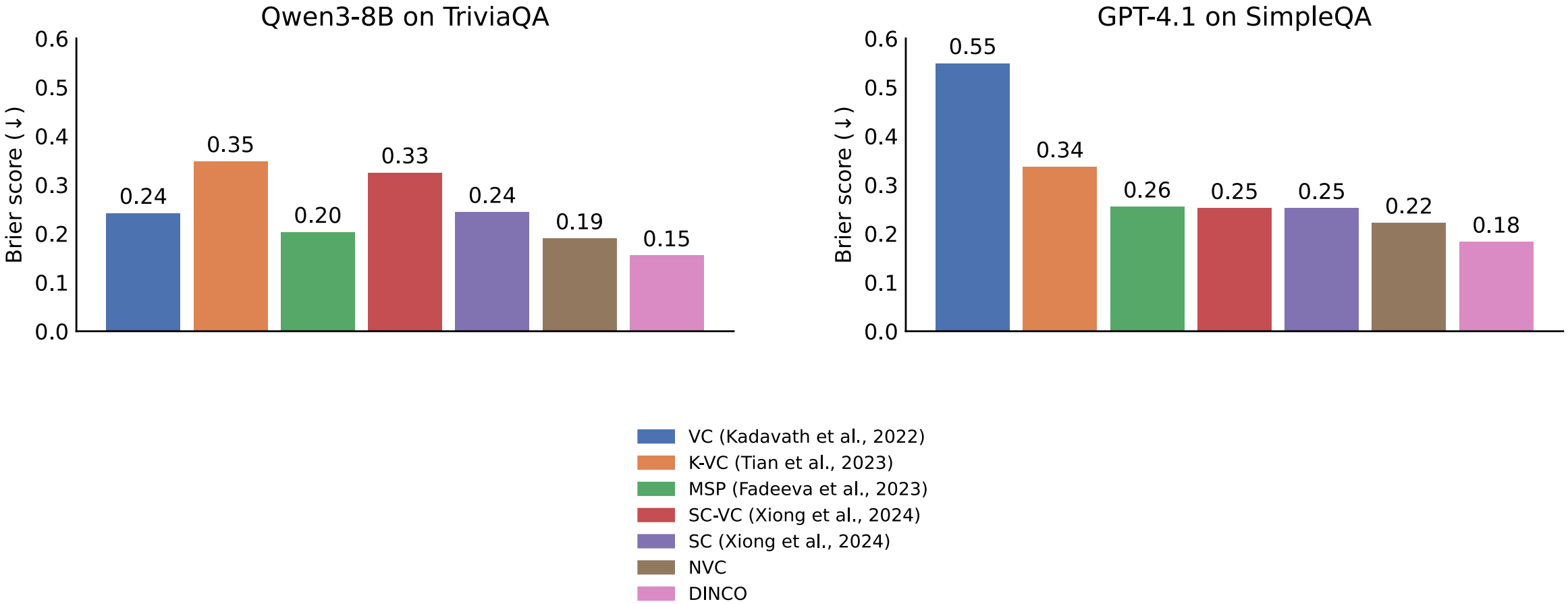
Combining Coherence within Generation and Validation

Generator-validator inconsistency (Li et al., 2023) motivates integrating coherence in the complementary aspects of generation and validation.

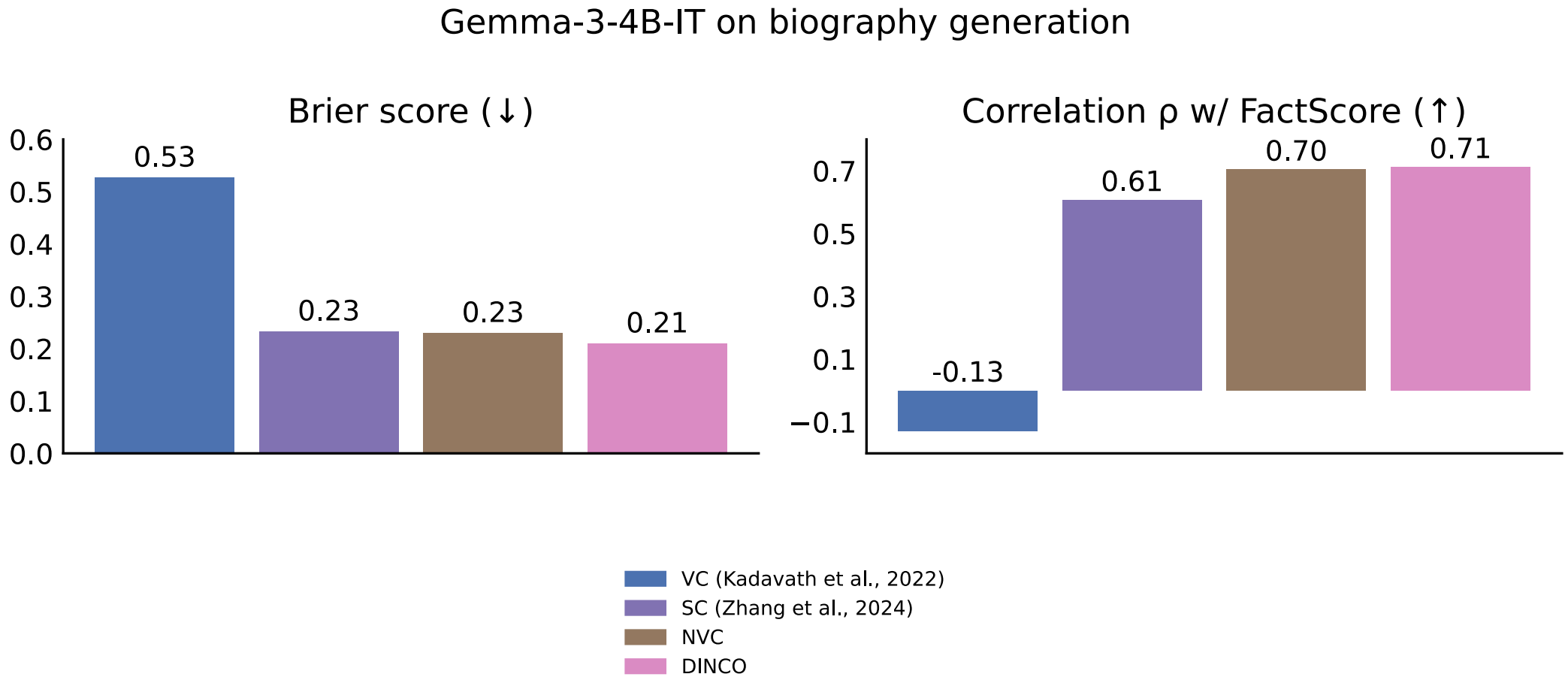
We define DINCO as the average of generator confidence and validator confidence.

- Generator confidence: Self-consistency (Xiong et al., 2024)
- Validator confidence: Normalized verbalized confidence

Results


Qwen3-8B on TriviaQA
Brier score (↓)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.24
0.35
0.20
0.33
0.24
0.19
0.15
GPT-4.1 on SimpleQA
Brier score (↓)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.55
0.34
0.26
0.25
0.25
0.22
0.18
VC (Kadavath et al., 2022)
K-VC (Tian et al., 2023)
MSP (Fadeeva et al., 2023)
SC-VC (Xiong et al., 2024)
SC (Xiong et al., 2024)
NVC
DINCO

Results


Gemma-3-4B-IT on biography generation
Brier score (↓)
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.53
0.23
0.23
0.21
Correlation ρ w/ FactScore (↑)
0.7
0.5
0.3
0.1
−0.1
-0.13
0.61
0.70
0.71
VC (Kadavath et al., 2022)
SC (Zhang et al., 2024)
NVC
DINCO

Thanks for listening!

Paper: https://arxiv.org/abs/2509.25532

Code: https://github.com/victorwang37/dinco