DoubtBench: Does Self-Doubt Help Language Models—or Just Make Them Check?

Rishi Ahuja[1,*] • Rajeev Ranjan Dwivedi[2] • Dr. Vinod K. Kurmi[2]

[1]Department of Information Technology, Dr. B. R. Ambedkar National Institute of Technology Jalandhar
[2]Visual Data Computing Group (VisDom), Department of Data Science and Engineering, IISER Bhopal
[*]Corresponding author: rishia.it.24@nitj.ac.in

IndoML 2026 Undergraduate Forum • Accepted at Findings of WSDM 2027, to be presented in Hong Kong

The question

Prompts called "emotional" or "metacognitive" often bundle two different instructions:

Self-doubt: *"While answering, you experience uncertainty and self-doubt about whether your first impression is correct."*

Check: *"Before giving your final response, check your work for arithmetic, ordering, and instruction-following mistakes."*

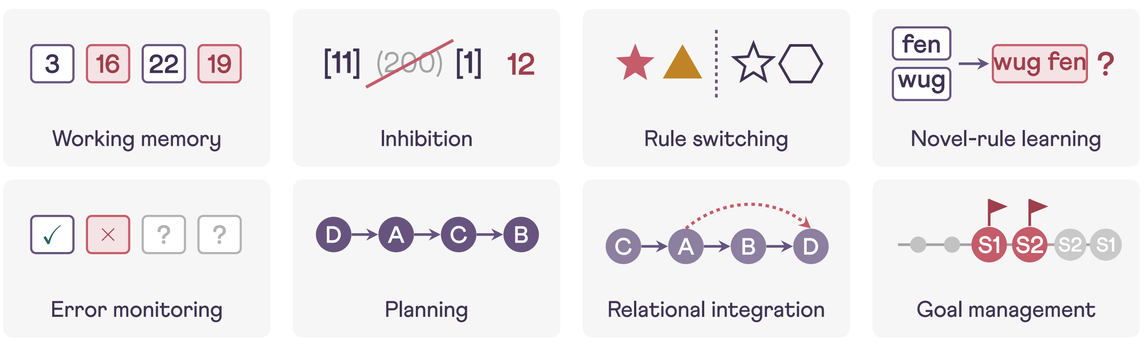
If the bundle helps, a comparison against a plain prompt cannot say *which part* helped. We ask:

Does self-doubt wording add anything once the model is already told to check its work?

"Self-doubt" names the prompt wording only; we make no claim that models experience doubt.

The benchmark

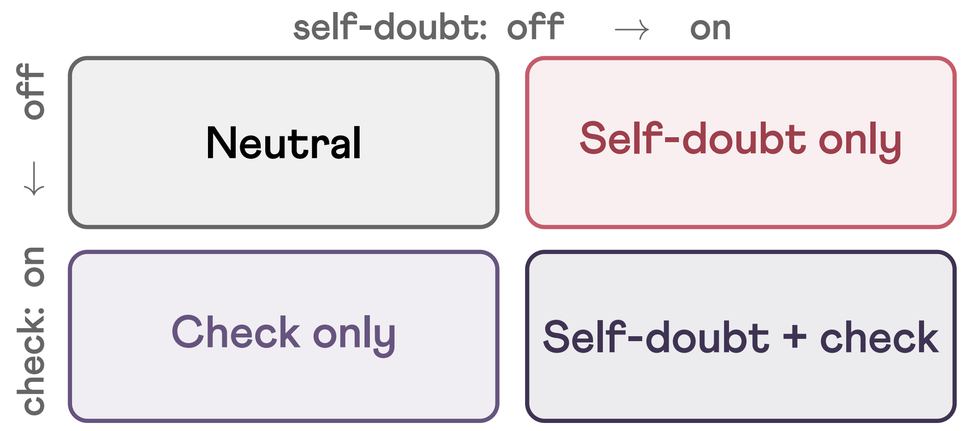
6,400	8	6	423.6k
exact-answer tasks	task families, 33 types	open model checkpoints	scored responses



- **Gold answer first:** the exact answer is computed *before* the question is written, so scoring is deterministic with no LLM judge. Families draw on executive-function research (Miyake et al., 2000).
- **Replicated:** 4 seeded task sets of 1,600, plus 660 rewordings. Qwen 2.5, Mistral, Llama 3.1, Phi 3.5, Tulu 3 SFT/DPO; greedy decoding.

Paired prompt design

Task, gold answer, and model stay fixed; only the prompt changes. Self-doubt and checking are switched on and off independently:

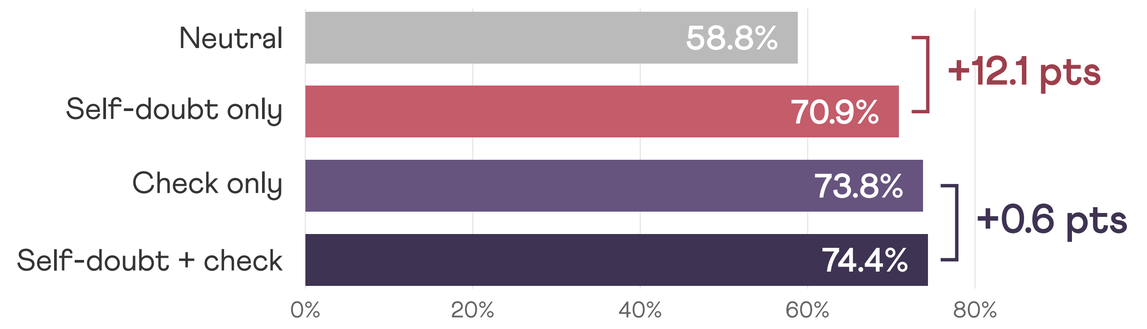


- A fifth arm, **check + concise**, adds "Do not show intermediate reasoning."
- Every arm runs with a **minimal request** and with **reasoning requested** ("Show the intermediate steps…"): **10 paired responses per task**.

Conditional Doubt Effect (CDE)
accuracy(self-doubt + check) − accuracy(check)
Positive: doubt helps beyond checking. Negative: it hurts.

Key result: the doubt gain is mostly a checking gain

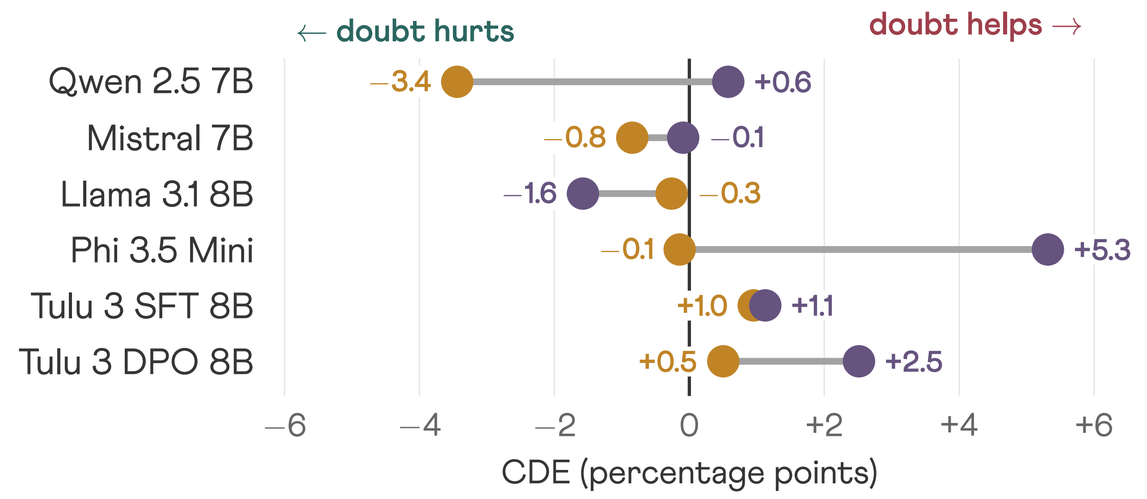
Qwen 2.5 7B accuracy with a minimal request:



Against a neutral prompt, self-doubt looks like a **12-point gain**. Once checking is already present, it adds only **0.6 points**, within noise (95% CI −0.4 to +1.6). Most of the apparent doubt gain is what a plain *check your work* instruction already delivers.

No portable effect: it depends on model and context

CDE for each checkpoint, under a **minimal request** and with **reasoning requested**:



- Overall CDE ranges from **−1.43** to **+2.59** points: two checkpoints are hurt, three are helped, and Mistral is indistinguishable from zero.
- With a minimal request, 4 of 6 are positive; once reasoning is requested, only 2 are. Phi moves from **+5.3** to **−0.1**.
- Checking itself is not universal: it gives Qwen **+7.7** points, but the others −0.1 to +1.8.

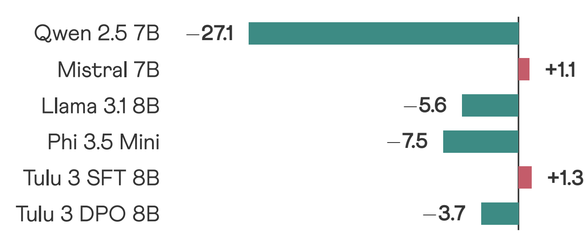
It also depends on the task family

CDE in points by task family, averaged over seeds and contexts:

	Working memory	Inhibition	Rule switching	Novel-rule learning	Error monitoring	Planning	Relational integration	Goal management
Qwen 2.5 7B	−1.9	−2.4	−8.1	+2.1	−1.3	+1.4	−0.6	−0.6
Mistral 7B	+4.2	+2.8	−1.4	−3.7	−1.3	+0.7	−5.4	+0.4
Llama 3.1 8B	+0.2	−2.1	0.0	−8.4	+5.1	+0.7	0.0	−2.8
Phi 3.5 Mini	+6.1	+6.5	+6.2	+0.2	−0.2	−1.9	+0.8	+3.0
Tulu 3 SFT 8B	−0.3	+5.6	0.0	+8.9	+0.4	−1.4	0.0	−4.8
Tulu 3 DPO 8B	+5.1	+6.1	0.0	−2.4	+2.7	−0.2	0.0	+0.7

Family effects span **−8.4** to **+8.9** points, and 21 model–family effects keep the same sign in all four seeds. Exact zeros mark families where a model fails almost every task under every prompt.

"Be concise" can cost a lot



Adding "Do not show intermediate reasoning" to the check prompt changes accuracy (points, averaged over both contexts).

With a minimal request, Qwen falls from **73.8%** to **34.0%** as answers shrink from about 185 to 10 tokens.

Takeaway: a prompt's label is not its mechanism.

Much of the apparent self-doubt effect comes from the checking instruction inside it, and what remains changes sign across models, contexts, and tasks. Credit prompt components with controlled, paired comparisons, not against a neutral baseline.

Code and data · Project page · Email

Acknowledgements We acknowledge the support of the Ministry of Education, Government of India; the Department of Information Technology, Dr. B. R. Ambedkar National Institute of Technology Jalandhar; and the Visual Data Computing Group (VisDom), Department of Data Science and Engineering, IISER Bhopal, for academic and institutional facilities.