SELF-CONSISTENCY IMPROVES CHAIN OF THOUGHT REASONING IN LANGUAGE MODELS

Xuezhi Wang, Jason Wei , Dale Schuurmans, Quoc Le, Ed H. Chi,
Sharan Narang, Aakanksha Chowdhery, Denny Zhou
Google Research, Brain Team

TECHNISCHE UNIVERSITÄT DARMSTADT

AGENDA

CHAPTER 1

MOTIVATION

TECHNISCHE UNIVERSITÄT DARMSTADT

CHAIN-OF-THOUGHT REASONING

CoT: a series of prompts which mimic human reasoning to guide language models in their reasoning process.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?”,

A: 5

CoT Reasoning:

- There are 3 cars in the parking lot already.
- 2 more arrive.
- Now there are 3 +2 = 5

The Answer is **5**

COT REASONING

Scaling LMs and implementing CoT improves reasoning abilities for tackling complex tasks.

Pros:

- CoT guides models in step-by-step reasoning, enhancing their ability to reason effectively.
- Achieve higher performance on complex tasks requiring multi-step reasoning.
- A transparent CoT, improves the interpretability of model decision.

Possible to leverage CoT' s benefits to achieve greater Consistency in finding the best solution? [1]

- **Consistency:** A desirable property of language understanding models.
- Improving overall language understanding and interpretation.
- Ensuring consistent performance in different linguistic situations.

[1] Measuring and Improving Consistency in Pretrained Language Models, Elazar et al. (2021)

CHAPTER 2

RELATED WORKS

RELATED WORKS

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Training Verifiers to Solve Math Word Problems, Cobbe et al., 2021, Google

Challenge:

State-of-the-art language models struggle with multi-step mathematical reasoning.

Idea:

Train an additional verifier to re-rank generated solutions.

- The paper introduces GSM8K, a dataset of diverse grade school math word problems.
- Sample a fixed number of candidate solutions, select the solution ranked highest by the verifier.
- Verifiers are trained to judge the correctness of model completions.
- Verification significantly improves performance on GSM8K.
- Improves the solve rate on math tasks compared to just fine-tuning the language model

RELATED WORKS

1. Training Verifiers to Solve Math Word Problems, [Cobbe et al., 2021, Google] (continue)

Cobbe et al., 2021	vs.	Self-Consistency:
• Sample a fixed number of candidate solutions.		✓
• Verifiers trained to judge the correctness of model completions		X
• Fine-tuning with human annotated reasoning paths		X
• Select the solution ranked highest by the verifier.		? Will be discussing

RELATED WORKS

TECHNISCHE UNIVERSITÄT DARMSTADT

2. Measuring and Improving Consistency in Pretrained Language Models, Elazar et al. (2021)

Challenge:

Assess the consistency of Pretrained Language Models (PLMs) with respect to factual knowledge.

Idea:

Create the PARAREL dataset to evaluate PLM consistency and propose a method for improving model consistency.

Enhance factual knowledge consistency through pre-training with additional consistency loss.

RELATED WORKS

2. Measuring and Improving Consistency in Pretrained Language Models, [Elazar et al. 2021](#) (continue)

Elazar et al. (2021)	**vs.**	**Self-Consistency**
• Recognize the importance of consistency in LMs.		✓
• Acknowledge the limitations of current language models in terms of consistency		✓
• Improve consistency through additional training and experimentation.		X
• Proposes extending pre-training with an additional consistency loss to improve model consistency		X

CHAPTER 3

SELF-CONSISTENCY METHOD WITHIN THE COT

SELF-CONSISTENCY METHOD

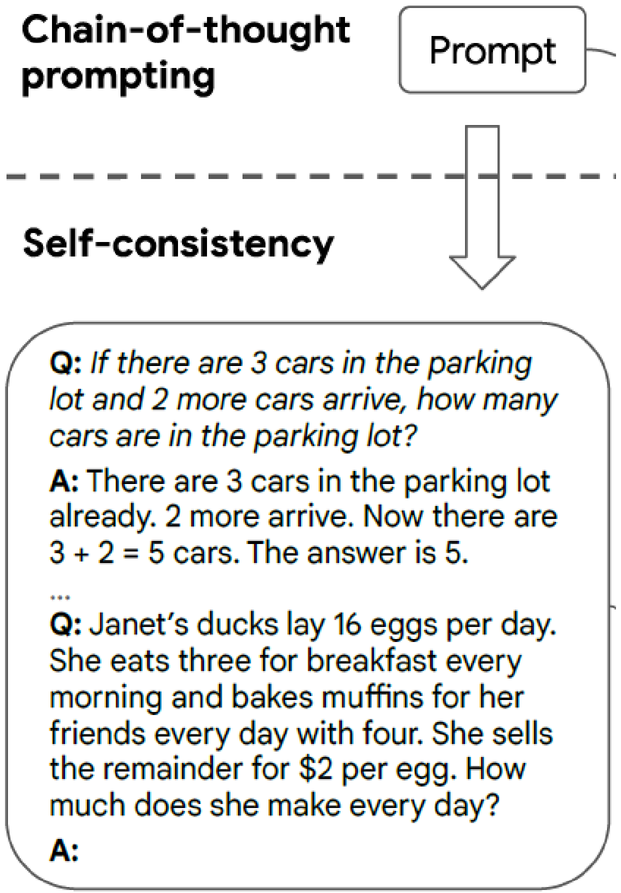
Spec:

- Unsupervised:
- Does not rely on human annotation, additional training or need to further modifications.
- Acting as a self-ensemble approach without the need for multiple separately trained models.
- Improved performance without introducing additional complexity or training requirements.

How?

- Complex reasoning tasks have multiple valid paths.
- Self-consistency encourages diverse problem-solving approaches.
- Thoughtful analysis expand the range of reasoning paths.
- It challenges the notion of a single "right" solution.
- Considering various paths enhances flexibility in finding the best solution.

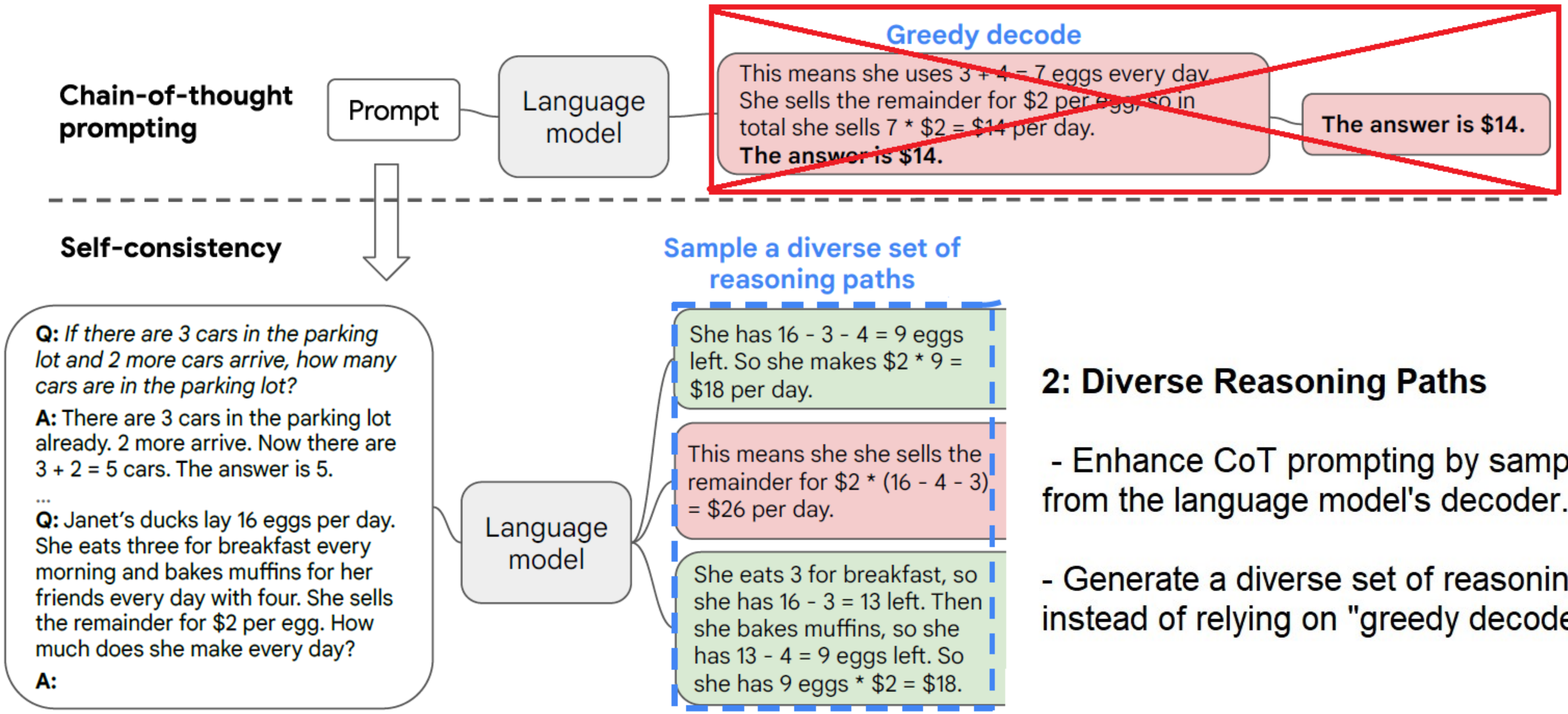
SELF-CONSISTENCY METHOD



1: CoT Prompting

a language model is prompted with a set of manually written chain-of-thought exemplars

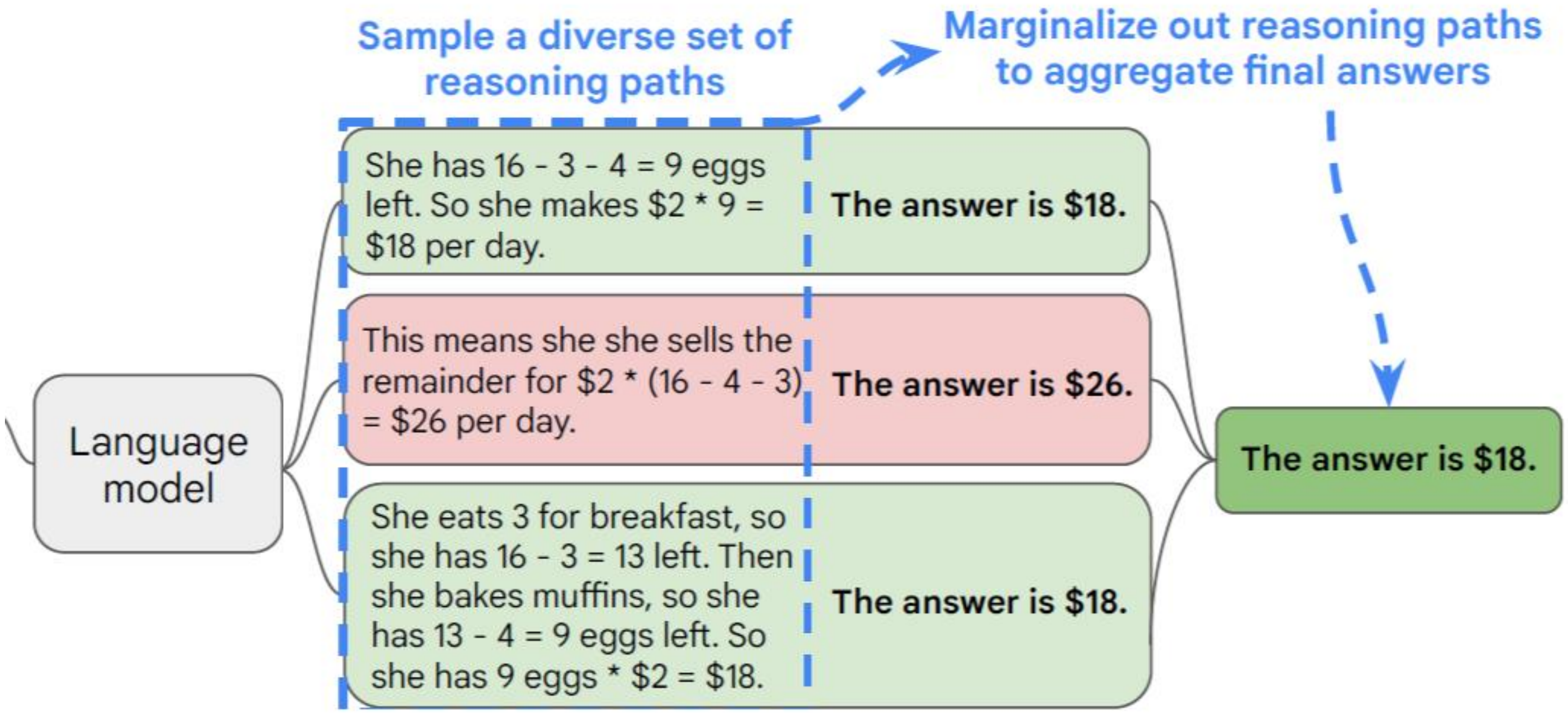
TECHNISCHE UNIVERSITÄT DARMSTADT



2: Diverse Reasoning Paths

- Enhance CoT prompting by sampling from the language model's decoder.

- Generate a diverse set of reasoning paths instead of relying on "greedy decode."



3. Aggregating Reasoning Paths for Consistent Answers

- Aggregate the reasoning paths by marginalizing them out.

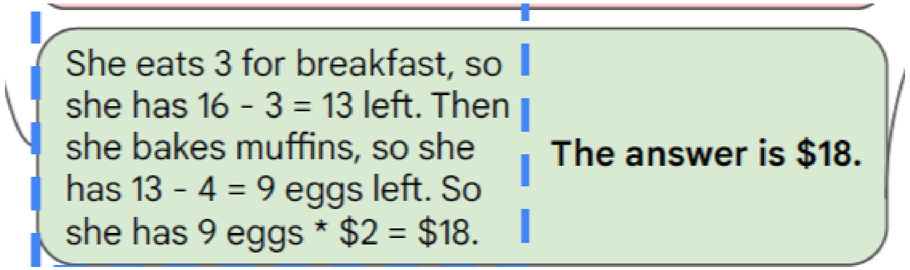
- Select the most consistent answer from the final set of answers.

SAMPLING

Self-consistency is compatible with most existing sampling algorithms, including:

- Temperature Sampling (Ackley et al., 1985; Ficler & Goldberg, 2017)
- Top-K Sampling (Fan et al., 2018; Holtzman et al., 2018; Radford et al., 2019)
- Nucleus Sampling (Holtzman et al., 2020)

SELF-CONSISTENCY OVER DIVERSE REASONING PATHS



$r_i \rightarrow a_i$

r_i: a sequence of tokens representing the reasoning path

a_i the generated answers ∈ **A :** fixed answer set

m: # of candidate outputs sampled from the decoder

$$\arg\max_a \sum_{i=1}^{m} \mathbb{1}(\mathbf{a}_i = a)$$

Majority Vote

The most "consistent" answer among the final answer set.

- After Sampling multiple (r_i , a_i)
- Apply marginalization over r_i by taking a **majority vote over a_i**

Accuracy Comparison across commonsense reasoning benchmarks

	GSM8K	MultiArith	AQuA	SVAMP	CSQA	ARC-c
Greedy decode	56.5	94.7	35.8	79.0	79.0	85.2
Weighted avg (unnormalized)	56.3 $\pm$ 0.0	90.5 $\pm$ 0.0	35.8 $\pm$ 0.0	73.0 $\pm$ 0.0	74.8 $\pm$ 0.0	82.3 $\pm$ 0.0
Weighted avg (normalized)	22.1 $\pm$ 0.0	59.7 $\pm$ 0.0	15.7 $\pm$ 0.0	40.5 $\pm$ 0.0	52.1 $\pm$ 0.0	51.7 $\pm$ 0.0
Weighted sum (unnormalized)	59.9 $\pm$ 0.0	92.2 $\pm$ 0.0	38.2 $\pm$ 0.0	76.2 $\pm$ 0.0	76.2 $\pm$ 0.0	83.5 $\pm$ 0.0
Weighted sum (normalized)	74.1 $\pm$ 0.0	99.3 $\pm$ 0.0	48.0 $\pm$ 0.0	86.8 $\pm$ 0.0	80.7 $\pm$ 0.0	88.7 $\pm$ 0.0
Unweighted sum (majority vote)	74.4 $\pm$ 0.1	99.3 $\pm$ 0.0	48.3 $\pm$ 0.5	86.6 $\pm$ 0.1	80.7 $\pm$ 0.1	88.7 $\pm$ 0.1

Table 1: Accuracy comparison of different answer aggregation strategies on PaLM-540B.

TECHNISCHE UNIVERSITÄT DARMSTADT

DIFFERENT ANSWER AGGREGATION STRATEGIES

Weighted Aggregation

Weight each (r_i, a_i) by $P(r_i, a_i \mid \text{prompt, question})$

either take the:

- **unnormalized probability** of the model generating**:** $P(r_i, a_i \mid \text{prompt, question})$

or

- **normalize the conditional probability** by the output length (Brown et al., 2020)

$$P(\mathbf{r}_i, \mathbf{a}_i \mid \text{prompt}, \text{question}) = \exp^{\frac{1}{K}\sum_{k=1}^{K} \log P(t_k \mid \text{prompt}, \text{question}, t_1, \ldots, t_{k-1})}$$

Normilzed weighted sum

k: total # of tokens

log probability of generating the k-th token tk in (ri , ai) conditioned on the previous tokens

TECHNISCHE UNIVERSITÄT DARMSTADT

Accuracy Comparison across commonsense reasoning benchmarks

	GSM8K	MultiArith	AQuA	SVAMP	CSQA	ARC-c
Greedy decode	56.5	94.7	35.8	79.0	79.0	85.2
Weighted avg (unnormalized)	56.3 ± 0.0	90.5 ± 0.0	35.8 ± 0.0	73.0 ± 0.0	74.8 ± 0.0	82.3 ± 0.0
Weighted avg (normalized)	22.1 ± 0.0	59.7 ± 0.0	15.7 ± 0.0	40.5 ± 0.0	52.1 ± 0.0	51.7 ± 0.0
Weighted sum (unnormalized)	59.9 ± 0.0	92.2 ± 0.0	38.2 ± 0.0	76.2 ± 0.0	76.2 ± 0.0	83.5 ± 0.0
Weighted sum (normalized)	74.1 ± 0.0	99.3 ± 0.0	48.0 ± 0.0	86.8 ± 0.0	80.7 ± 0.0	88.7 ± 0.0
Unweighted sum (majority vote)	74.4 ± 0.1	99.3 ± 0.0	48.3 ± 0.5	86.6 ± 0.1	80.7 ± 0.1	88.7 ± 0.1

Table 1: Accuracy comparison of different answer aggregation strategies on PaLM-540B.

TECHNISCHE UNIVERSITÄT DARMSTADT

CHAPTER4

EVALUATION

TECHNISCHE UNIVERSITÄT DARMSTADT

EXPERMINET SETUP

- **Benchmarks**:
  - **Arithmetic reasoning** (use the Math Word Problem Repository, …
  - **Commonsense reasoning** (CommonsenseQA, StrategyQA, …)
  - **Symbolic Reasoning** (last letter concatenation, Coinflip)
- All expermiments in the few-shot settings
- Neither training nor fine-tuning
- Use same propmts for fair comparision

- **Language Models:**

  UL2 (encder-decoder, 20-B)[1]

  LaMDA (decoder-only, 137-B)

  GPT-3 (decoder-only, 175-B)[2]

  PaLM (decoder-only, 540-B)

[1] UL2 [2] GPT-3

MAIN RESULTS (ARITHMETIC REASONING)

	Method	AddSub	MultiArith	ASDiv	AQuA	SVAMP	GSM8K
	Previous SoTA	**94.9**[a]	60.5[a]	75.3[b]	37.9[c]	57.4[d]	35[e] / 55[g]
UL2-20B	CoT-prompting	18.2	10.7	16.9	23.6	12.6	4.1
	Self-consistency	24.8 (+6.6)	15.0 (+4.3)	21.5 (+4.6)	26.9 (+3.3)	19.4 (+6.8)	7.3 (+3.2)
LaMDA-137B	CoT-prompting	52.9	51.8	49.0	17.7	38.9	17.1
	Self-consistency	63.5 (+10.6)	75.7 (+23.9)	58.2 (+9.2)	26.8 (+9.1)	53.3 (+14.4)	27.7 (+10.6)
PaLM-540B	CoT-prompting	91.9	94.7	74.0	35.8	79.0	56.5
	Self-consistency	93.7 (+1.8)	99.3 (+4.6)	81.9 (+7.9)	48.3 (+12.5)	86.6 (+7.6)	74.4 (+17.9)
GPT-3 Code-davinci-001	CoT-prompting	57.2	59.5	52.7	18.9	39.8	14.6
	Self-consistency	67.8 (+10.6)	82.7 (+23.2)	61.9 (+9.2)	25.6 (+6.7)	54.5 (+14.7)	23.4 (+8.8)
GPT-3 Code-davinci-002	CoT-prompting	89.4	96.2	80.1	39.8	75.8	60.1
	Self-consistency	91.6 (+2.2)	**100.0** (+3.8)	**87.8** (+7.6)	**52.0** (+12.2)	**86.8** (+11.0)	**78.0** (+17.9)

Table 2: Arithmetic reasoning accuracy by self-consistency compared to chain-of-thought prompting

TECHNISCHE UNIVERSITÄT DARMSTADT

MAIN RESULTS (COMMONSENSE AND SYMBOLIC REASONING)

	Method	CSQA	StrategyQA	ARC-e	ARC-c	Letter (4)	Coinflip (4)
	Previous SoTA	**91.2**[a]	73.9[b]	86.4[c]	75.0[c]	N/A	N/A
UL2-20B	CoT-prompting	51.4	53.3	61.6	42.9	0.0	50.4
	Self-consistency	55.7 (+4.3)	54.9 (+1.6)	69.8 (+8.2)	49.5 (+6.8)	0.0 (+0.0)	50.5 (+0.1)
LaMDA-137B	CoT-prompting	57.9	65.4	75.3	55.1	8.2	72.4
	Self-consistency	63.1 (+5.2)	67.8 (+2.4)	79.3 (+4.0)	59.8 (+4.7)	8.2 (+0.0)	73.5 (+1.1)
PaLM-540B	CoT-prompting	79.0	75.3	95.3	85.2	65.8	88.2
	Self-consistency	80.7 (+1.7)	**81.6** (+6.3)	**96.4** (+1.1)	**88.7** (+3.5)	70.8 (+5.0)	91.2 (+3.0)
GPT-3 Code-davinci-001	CoT-prompting	46.6	56.7	63.1	43.1	7.8	71.4
	Self-consistency	54.9 (+8.3)	61.7 (+5.0)	72.1 (+9.0)	53.7 (+10.6)	10.0 (+2.2)	75.9 (+4.5)
GPT-3 Code-davinci-002	CoT-prompting	79.0	73.4	94.0	83.6	70.4	99.0
	Self-consistency	81.5 (+2.5)	79.8 (+6.4)	96.0 (+2.0)	87.5 (+3.9)	**73.4** (+3.0)	**99.5** (+0.5)

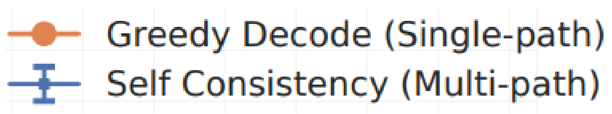
Table 3: Commonsense and symbolic reasoning accuracy by self-consistency compared to CoT prompting

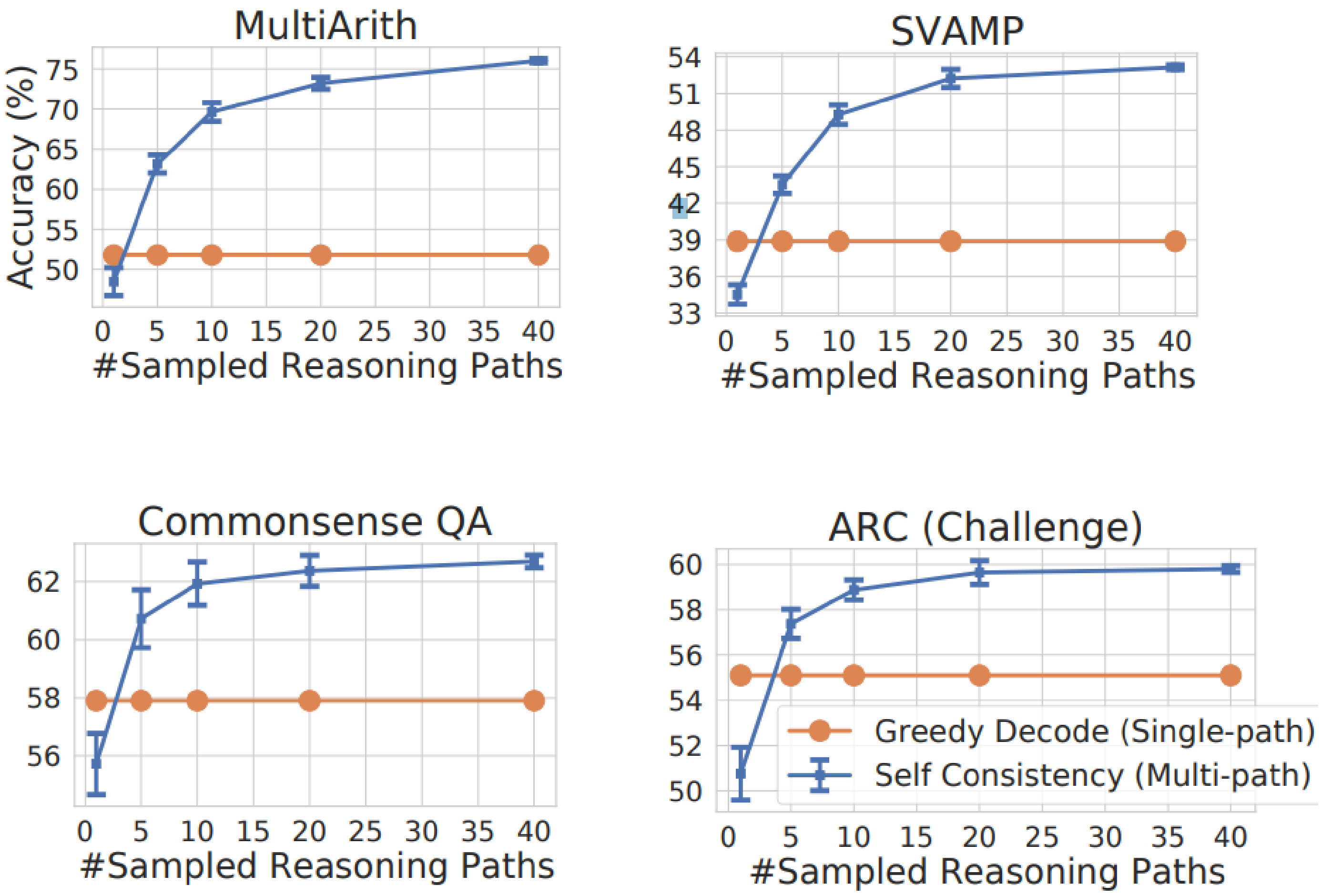
EFFECT OF THE NUMBER OF SAMPLED REASONING PATHS

Number of reasoning paths

≈

Accuracy





TECHNISCHE UNIVERSITÄT DARMSTADT

CHAPTER5

ADDITIONAL STUDIES

IMPROVING ROBUSTNESS TO IMPERFECT PROMPTS

- **Imperfect Prompts:** Manually constructed prompts in few-shot learning can contain minor mistakes due to human annotation.
- Greedy decoding with imperfect prompts leads to decreased accuracy (17.1% → 14.9%).
- Self-consistency fills in the gaps and significantly improves results with imperfect prompts.

LaMDA-137B	Prompt with correct chain-of-thought	17.1
	Prompt with imperfect chain-of-thought	14.9
	+ Self-consistency (40 paths)	**23.4**

- **Consistency and Accuracy:** Consistency, measured as the agreement with the final aggregated answer, is highly correlated with accuracy.

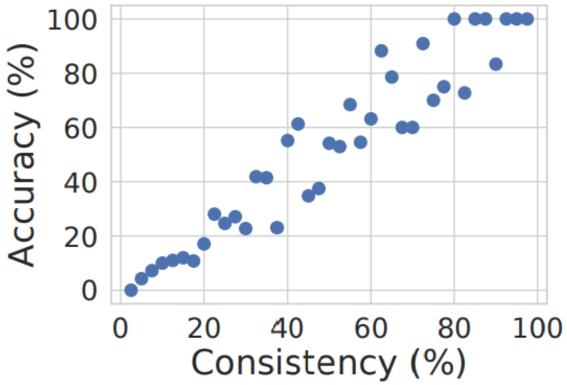


Figure 5: The consistency is correlated with model's accuracy.

TECHNISCHE UNIVERSITÄT DARMSTADT

SELF CONSISTENCY FOR NON-NATURAL-LANGUAGE REASONING AND ZERO-SHOT COT

- **Self-consistency improves accuracy by generating intermediate equations**

  Gains are smaller due to limited diversity in equation decoding.

  e.g., from "There are 3 cars in the parking lot already. 2 more arrive. Now there are 3 + 2 = 5 cars." to "3 + 2 = 5") (**Limited Leeway**)

LaMDA-137B	Prompt with equations	5.0
	+ Self-consistency (40 paths)	**6.5**
PaLM-540B	Zero-shot CoT (Kojima et al., 2022)	43.0
	+ Self-consistency (40 paths)	**69.2**

- Self-consistency enhances results significantly in zero-shot CoT scenarios (+26.2%)

CHAPTER6

CONCLUSION

Conclusion

- Self-Consistency improves accuracy in a range of Arithmetic and Commonsense reasoning tasks across four language models.
- With reasoning paths enhances interpretability in reasoning tasks.
- Provides improved output calibration and uncertainty estimation.
- Computation cost: Self-Consistency incurs additional computational overhead.
- Optimal paths: Few reasoning paths yield significant gains without excessive cost.

Future Work

- Use self-consistency to generate better supervised data for fine-tuning.
- Improving prediction accuracy in a single inference run.
- Mitigate inconsistencies and inaccuracies in reasoning paths to enhance trustworthiness.
- Further research needed to refine the process of generating effective rationales.

THANKS!

Ali Khalaji

ali.khalaji@stud.tu-darmstadt.de

CHAPTER7

APPENDIX

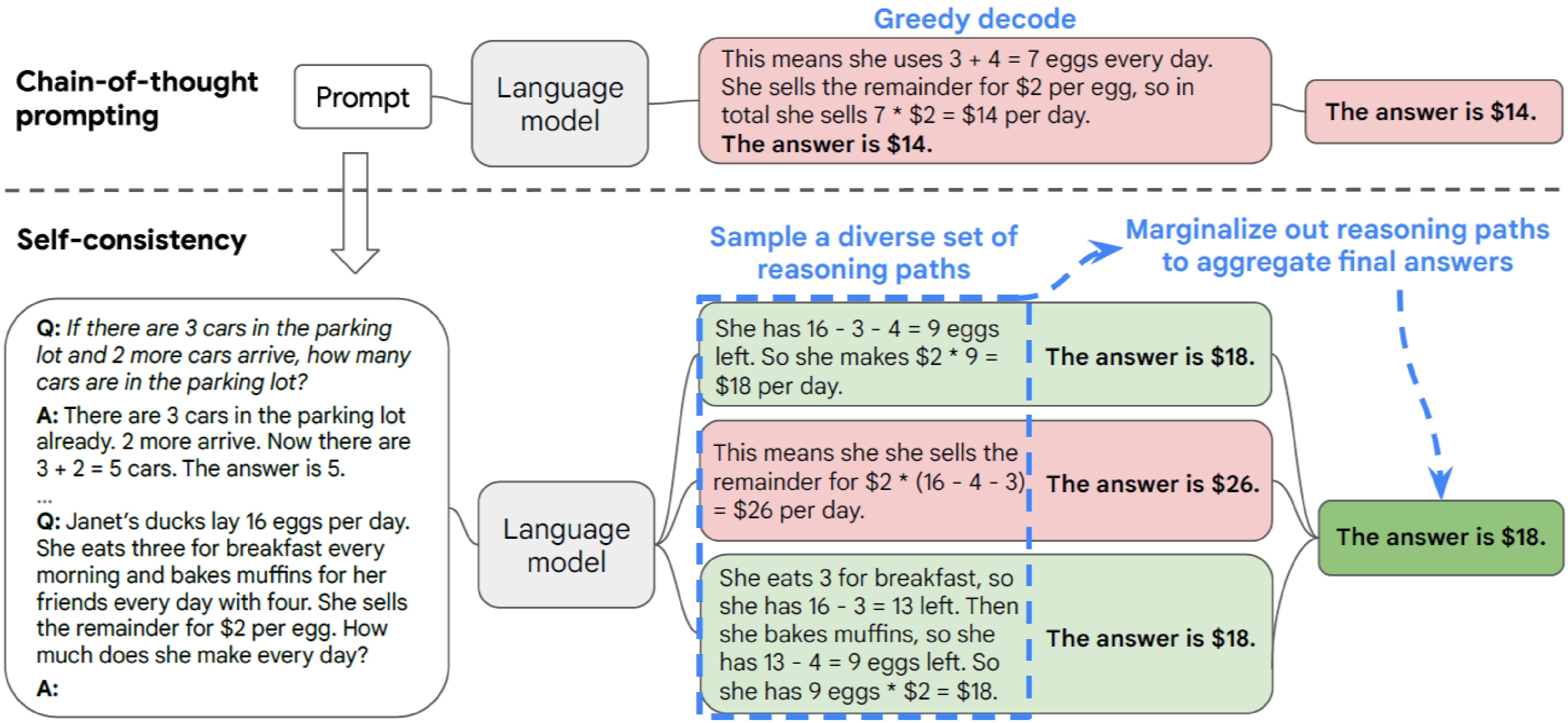
TECHNISCHE UNIVERSITÄT DARMSTADT

CONSISTENCY IN LANGUAGE UNDERSTANDING [1]

- Possible to leverage CoT' s benefits to achieve greater Consistency in finding the best solution?
- **Consistency:** A desirable property of language understanding models.
- We want to make consistent decisions in semantically equivalent contexts.
- Improving overall language understanding and interpretation.
- Ensuring consistent performance in different linguistic situations.
- Promoting reliable and consistent language processing.

[1] Measuring and Improving Consistency in Pretrained Language Models, [Elazar et al. (2021)]

APPENDIX - SELF-CONSISTENCY WITHIN THE COT REASONING



TECHNISCHE UNIVERSITÄT DARMSTADT

Additional Notes on Self-Consistency

- Find a middle ground between open-ended and fixed answer text generation.
- Reasoning Tasks: Usually rely on greedy decoding approaches due to fixed answers.
- Benefits of Diversity: Adding diversity to reasoning processes highly advantageous.
- Sampling Approach: like in open-ended text generation, to introduce diversity.
- Self-consistency can be expanded to open-text generation if we define a suitable consistency metric.
- Metric Definition: Develop a metric to assess agreement or contradiction between multiple generated texts.

TECHNISCHE UNIVERSITÄT DARMSTADT

SELF-CONSISTENCY VS. SAMPLE-AND-RANK

Experiment Setup

- Comparison conducted on GPT-3 code-davinci-001 model.
- Same number of sequences sampled from the decoder for both approaches.
- Final answer extracted from the top-ranked sequence.

Self-Consistency Method

- Generates multiple sequences from the decoder.
- Promotes agreement and consistency among the generated responses.
- Results in significant accuracy improvement.

Sample-and-Rank Method[1]

- Sampling multiple sequences and ranking them based on log probability.
- Provides a marginal accuracy improvement.
- Gain is comparatively smaller compared to self-consistency.

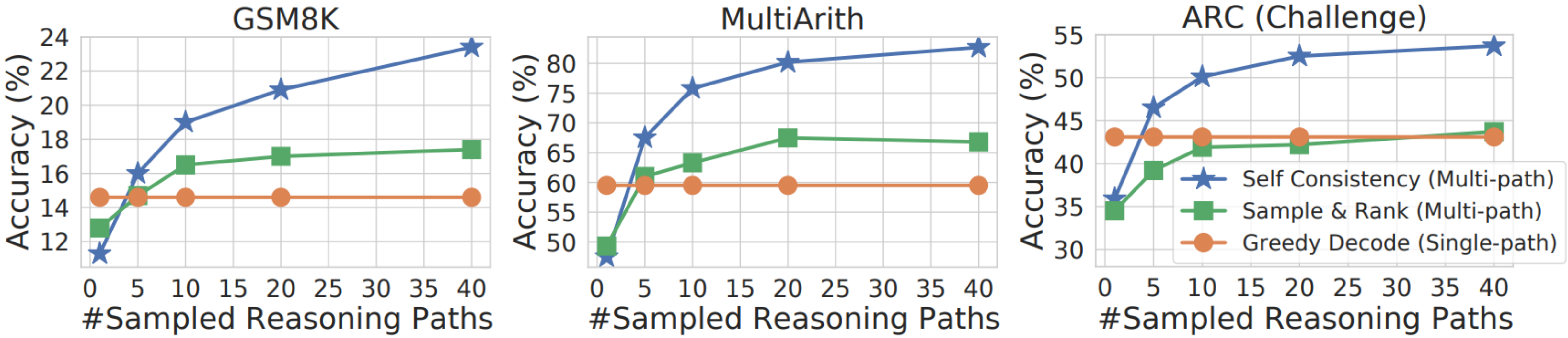
[1] Towards a Human-like Open-Domain Chatbot, (Adiwardana et al., 2020)

TECHNISCHE UNIVERSITÄT DARMSTADT

SELF-CONSISTENCY VS. SAMPLE-AND-RANK

Results:

- Self-consistency proves to be a more effective approach for improving generation quality compared to the sample-and-rank method.
- Self-consistency yields substantial accuracy gains by promoting diversity and consistency among generated responses.



TECHNISCHE UNIVERSITÄT DARMSTADT

SELF-CONSISTENCY VS. BEAM SEARCH

- on the UL2-20B Model.
- Equal number of beams (for beam search) and reasoning paths (for self-consistency)
- Self-consistency can adopt beam search to decode each reasoning path
- The Performance of Self-Consistency with sampling is better, as beam search's limited output diversity limits its effectiveness.[1]

	Beam size / Self-consistency paths	1	5	10	20	40
AQuA	Beam search decoding (top beam)	23.6	19.3	16.1	15.0	10.2
	Self-consistency using beam search	23.6	19.8 ± 0.3	21.2 ± 0.7	24.6 ± 0.4	24.2 ± 0.5
	Self-consistency using sampling	19.7 ± 2.5	**24.9 ± 2.6**	**25.3 ± 1.8**	**26.7 ± 1.0**	**26.9 ± 0.5**
MultiArith	Beam search decoding (top beam)	10.7	12.0	11.3	11.0	10.5
	Self-consistency using beam search	10.7	11.8 ± 0.0	11.4 ± 0.1	12.3 ± 0.1	10.8 ± 0.1
	Self-consistency using sampling	9.5 ± 1.2	11.3 ± 1.2	**12.3 ± 0.8**	**13.7 ± 0.9**	**14.7 ± 0.3**

Compare self-consistency with beam search decoding on the UL2-20B model.

[1] A Simple, Fast Diverse Decoding Algorithm for Neural Generation, Li & Jurafsky, 2016)

TECHNISCHE UNIVERSITÄT DARMSTADT

SELF CONSISTENCY VS. CHAIN OF THOUGHT

- For some tasks (e.g., ANLI-R1, e-SNLI, RTE), adding chain-of-thought does hurt performance compared to standard prompting, but
- but self-consistency is able to robustly boost the performance and outperform standard prompting, making it a reliable way to add rationales in few-shot in-context learning for common NLP tasks

	ANLI R1 / R2 / R3	e-SNLI	RTE	BoolQ	HotpotQA (EM/F1)
Standard-prompting (no-rationale)	69.1 / 55.8 / 55.8	85.8	84.8	71.3	27.1 / 36.8
CoT-prompting (Wei et al., 2022)	68.8 / 58.9 / 60.6	81.0	79.1	74.2	28.9 / 39.8
Self-consistency	**78.5 / 64.5 / 63.4**	**88.4**	**86.3**	**78.4**	**33.8 / 44.6**

Table 5: Compare Standard/CoT prompting with self-consistency on common NLP tasks.

TECHNISCHE UNIVERSITÄT DARMSTADT

THE END