ML 2026 Spring HW8
Test-Time Scaling

TA: 江履方、陳品睿、尹廷安、林育正
ntu-ml-2026-spring-ta@googlegroups.com
Deadline: 2026/**06/04** 23:59:59 (UTC+8)

Outline

- Task Description
- Assignment Format
- Dataset
- Metric
- Submission & Grading

Useful Links

- Sample code (colab)
- Pytorch Tutorial

Prerequisite

- Please watch Prof. Lee's following lecture video before working on this HW.

【生成式AI時代下的機器學習(2025)】第七講：DeepSeek-R1 這類大型語言模型是如何進行「深度思考」（Reasoning）的？

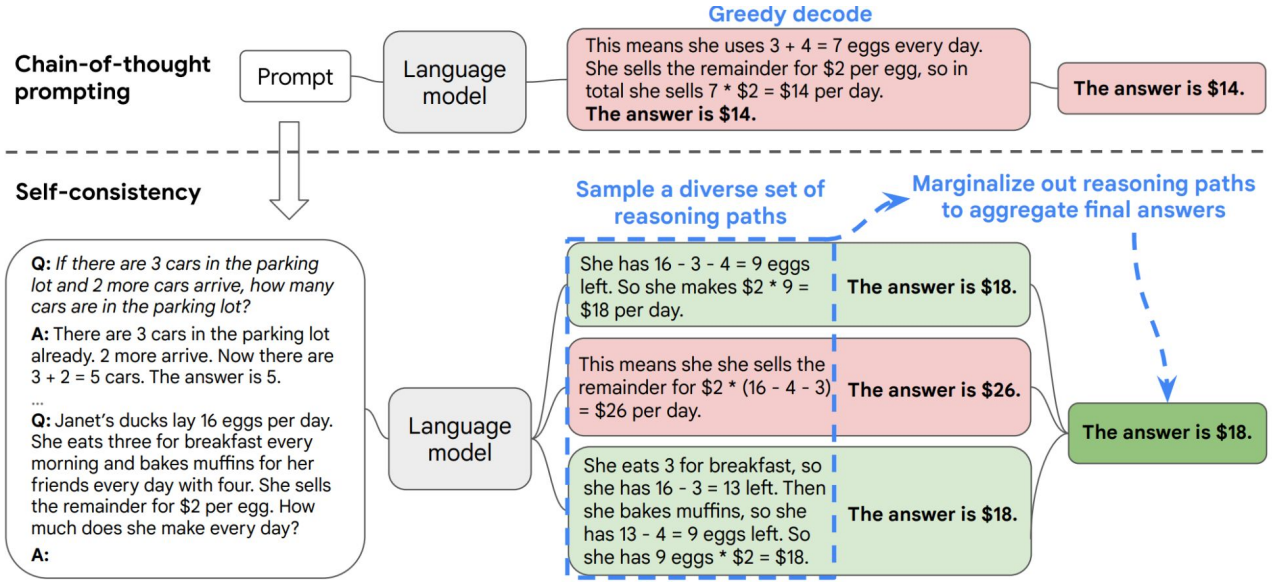
Task Description

- Goal: Learn several test-time scaling methods and implement on open source LLM and investigate the different between accuracy.

- We recommend students read these papers for a high-level overview:
  - Chain-of-Thought
  - Beam Search
  - Self-Consistency
  - Self-Certainty
  - Confidence

Self-Consistency [1] ICLR 2023

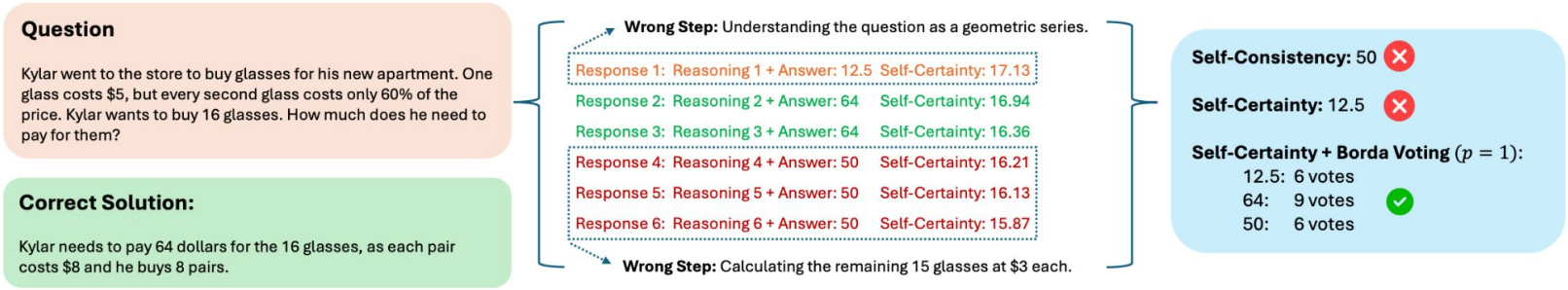
- Generate multiple output and decide on majority.



[1] Self-Consistency Improves Chain of Thought Reasoning in Language Models (Wang et.al. 2023)

Self-Certainty [2] NeurIPS 2025

- Voting and certainty.



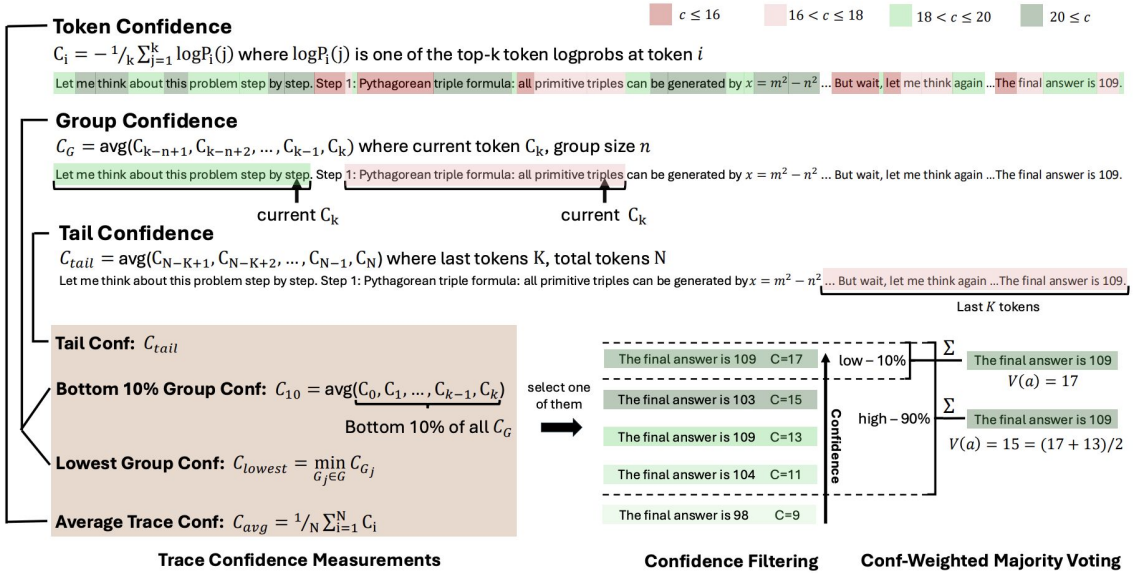
$$\text{Self-Certainty (CE)} = -\frac{1}{nV}\sum_{i=1}^{n}\sum_{j=1}^{V}\log\big(p(j \mid x, y_{<i})\big). \qquad v(r) = (N - r + 1)^p$$

[2] Scalable Best-of-N Selection for Large Language Models via Self-Certainty (Kang et al. 2025)

Confidence [3] ICLR 2026

- Use different part of reasoning and token confidence to determine the final output.

$$C_i = -\frac{1}{k}\sum_{j=1}^{k}\log P_i(j),$$



[3] Deep think with confidence (Fu et al. 2025)

Assignment Format

- This homework consists of 20 questions worth a total of 10 points.
  - Part 1: 18 Paper Reading questions (0.5 pt each).
  - Part 2: 2 Coding questions (0.5 pt each).

- You only need to complete the quiz on NTU Cool and submit it.
- It might be useful to go through a Colab tutorial on these topics first, then answer some paper-reading questions.

Dataset

- We use the accuracy of **GSM8K** to evaluate the ability of LLM.
- Each problem in the problem set is a math problem, structured as follows:

Q: Weng earns $12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn?

A: Weng earns 12/60 = $<<12/60=0.2>>0.2 per minute. Working 50 minutes, she earned 0.2 x 50 = $<<0.2*50=10>>10.
10

Metric

- Accuracy and correct generated format correct.

Weng earns 12/60 = $<<12/60=0.2>>0.2 per minute. Working 50 minutes, she earned 0.2 x 50 = $<<0.2*50=10>>10.
10

Weng earns 12/60 = $<<12/60=0.2>>0.1 per minute. Working 50 minutes, she earned 0.1 x 50 = $<<0.1*50=5>>5.
5

Weng earns 12/60 = $<<12/60=0.2>>0.2 per minute. Working 50 minutes, she earned 0.2 x 50 = $<<0.2*50=10>>10.
##!!10

Submission & Deadline

- Submit your homework to **NTU Cool**, you don't need to submit your code.
- There is no submission limit for the NTU Cool quiz. Your highest score among all attempts will be taken as your final grade.
- **Deadline: 6/4 (Thu.) 23:59**
- **No late submission is allowed**

Grading Release Date

- The grading of the homework will be released by 2026/**06/07** 23:59:59 (UTC+8)

Grading - Regulations

- You should NOT plagiarize, if you use any other resource, you should cite it in the reference.
- Do NOT share codes or prediction files with any living creatures.
- Your final grade x 0.9 and get a score 0 for that homework if you violate any of the above rules first time (within a semester).
- Your will get F for the final grade if you violate any of the above rules multiple (> 1) times (within a semester).
- Prof. Lee & TAs preserve the rights to change the rules & grades.

If You Have Any Questions

- NTU Cool **HW8** 作業討論區
  - 如果同學的問題不涉及作業答案或隱私，請**一律使用** NTU Cool 討論區
  - 助教們會優先回答NTU Cool討論區上的問題
- Email: ntu-ml-2026-spring-ta@googlegroups.com
  - Title should start with [ML 2026 Spring **HW8**]
  - Email with the wrong title will be moved to trash automatically
- TA Hours
  - Each Friday before / after class:
    - (Fri.) 13.20 ~ 14.10 / 17:30~18:00
    - Location: 博理112

Chinese + English

Q1

(Choose one)

根據 *Self-Consistency Improves Chain of Thought Reasoning in Language Models*, 若要對一個 test question 執行 self-consistency decoding, 請排列下列步驟的正確順序:

1. 從每條 sampled reasoning path 中 parse / extract final answer
2. 對 final answers 做 aggregation, 例如 majority vote 或 answer-level marginalization
3. 使用包含 few-shot Chain-of-Thought exemplars 的 prompt 讓 LLM 進行推理
4. 用 stochastic decoding sample 多條 diverse reasoning paths, 而不是只取 greedy path
5. 回傳 aggregated support 最高的 final answer

A. 34125
B. 31425
C. 34215
D. 43125
E. 34152

Q2

(Choose three)

下列哪些情境下，self-consistency 的效果可能受限？

A. Task 的 final answer 很難 parse 或 normalize，例如答案是長篇 essay、開放式主觀評論

B. Model 產生的 sampled reasoning paths 幾乎沒有 diversity，導致 K 次 sampling 實際上接近重複 greedy decoding

C. Model 產生很多 fluent but logically invalid reasoning paths，使 majority vote 被錯誤答案主導

D. Task 有明確可比較的 final answer，而且不同 reasoning paths 常收斂到正確答案

E. Inference budget 允許產生多條 sampled reasoning paths，且 answer extraction 可以穩定完成

Q3

(Choose three)

關於 self-consistency 的核心想法，下列哪些敘述正確？

A. Self-consistency 假設 complex reasoning problem 可能存在多條不同但合理的 reasoning paths，且這些 paths 可能收斂到同一個 correct answer

B. Self-consistency 的主要目標是訓練一個 extra verifier model，用來判斷每條 reasoning path 是否正確

C. Self-consistency 可以被理解成對 latent reasoning paths 做 marginalization，而不是只相信 single greedy path

D. Self-consistency 是一種 training-free decoding strategy，通常不需要更新 LLM parameters

E. Self-consistency 的主要貢獻是移除 Chain-of-Thought exemplars，使模型改用 answer-only prompting

Q4

(Choose one)

根據 paper，self-certainty 的核心概念最接近下列哪一個描述？

A. 計算 generated response 的長度，因為越長的 response 通常 reasoning 越完整

B. 只計算 sampled tokens 的 average log-probability，完全忽略其他 vocabulary tokens

C. 使用每個 decoding step 的 entire token probability distribution，衡量模型在該位置對 next token prediction 的 certainty / concentration

D. 對 final answer 做 majority vote，選出出現次數最多的答案

E. 使用另一個 LLM 對每個 response 進行 pairwise comparison，再選出最佳 response

Q5

(Choose two)

根據 self-certainty paper 的 experimental results / discussion，下列哪些敘述正確？

A. Self-certainty 可以用於 Best-of-N selection，而且不需要額外訓練 external reward model

B. 在有明確 final answer 的 mathematical reasoning tasks 上，單純 self-certainty selection 不一定總是優於 self-consistency，因此 paper 進一步使用 Borda voting 結合 confidence ranking 與 answer frequency

C. Paper 評估的任務指出 self-consistency 只能在 mathematical reasoning tasks 上優於其他 test-time scaling 方法

D. Paper 指出 self-consistency 可以自然處理 open-ended code generation，因為所有 code outputs 都能被穩定 string-match 成同一個 final answer

E. Self-certainty 的效果與 sample size N 無關，因此 N = 1 和 N = 64 在 Best-of-N selection 中理論上沒有差別

Q6

(Choose two)

關於 self-certainty 與 AvgLogP / negative perplexity 的差異，下列哪些敘述正確？

A. AvgLogP / negative perplexity 主要根據 generated tokens 的 probability；self-certainty 則利用每個 decoding step 的 entire token probability distribution

B. Self-certainty 可以反映 model output distribution 的 concentration，而不只是 sampled token 本身有多高機率

C. Negative perplexity 和 self-certainty 完全等價，因為兩者都只依賴 generated token 的 likelihood

D. Self-certainty 不需要 logits 或 probability distribution，因此可以在完全 black-box 且不提供 logprobs 的 API 上完整計算

E. Self-certainty 只看 final answer string，不使用 intermediate generation steps 的 token distribution

Q7

(Choose three)

關於 *Deep Think with Confidence* 這篇 paper 的主要 motivation，下列哪些敘述正確？

A. Standard self-consistency / majority voting 會把所有 reasoning traces 視為同等重要，可能讓 low-quality traces 影響 final answer

B. Parallel thinking 雖然可以提升 reasoning accuracy，但會產生大量 reasoning traces，因此 inference overhead 會隨 traces 數量增加

C. Global confidence measure 需要完整 reasoning trace 才能計算，因此不適合用來 early stop low-quality trace

D. DeepConf 的主要目標是訓練一個外部 reward model，讓 reward model 取代 LLM 本身的 next-token distribution

E. DeepConf 認為 low-confidence reasoning segments 不應被觀察，因為它們通常和 final answer correctness 完全無關

Q8

(Choose three)

關於 *Deep Think with Confidence* 使用的 confidence measurements，下列哪些敘述正確？

A. Token Confidence 是根據模型在某個 decoding position 的 top-k token probability distribution 計算，用來反映該位置的 certainty

B. Average Trace Confidence 會把整條 trace 的 token confidence 做平均，但可能掩蓋中間局部 reasoning breakdown

C. Bottom 10% Group Confidence 會聚焦於 confidence 最低的一部分 group，用來捕捉 trace 中最可能出問題的 reasoning segments

D. Tail Confidence 完全忽略 final portion of reasoning trace，只衡量 trace 開頭的 tokens，因為 early tokens 對 final answer 最重要

E. Lowest Group Confidence 是所有 group confidence 的最大值，因此代表 trace 中最順利的一段 reasoning

Q9

(Choose three)

關於 DeepConf-low 和 DeepConf-high 的 online thinking，下列哪些敘述正確？

A. DeepConf-low 使用 top η = 10% 的 warmup traces 來設定較嚴格的 confidence threshold，因此 online generation 時會比較 aggressive 地 early stop low-confidence traces

B.DeepConf-high 使用 top η = 90% 的 warmup traces 來設定較寬鬆的 confidence threshold，因此 online generation 時會保留較多 traces

C. Online thinking 完全不需要 offline warmup，因為 threshold s 是模型參數中固定 內建的值

D. Online generation 中，如果目前 group confidence 低於 threshold s，該 trace 可以被 early terminated

E. DeepConf-low / high 的差異在於一個需要 model training，另一個不需要 model training

Q10

(Choose two)

關於 Chain-of-Thought prompting 與 Self-Consistency 的關係，下列哪些敘述正確？

A. Chain-of-Thought prompting 通常讓模型產生一條 reasoning path；Self-Consistency 則在 CoT 基礎上 sample 多條 reasoning paths 並聚合 final answers

B. Self-Consistency 的目的在於移除所有 intermediate reasoning steps，只保留 final answer

C. Greedy CoT 可能因單一路徑中的 early mistake 而導致 final answer 錯誤；Self-Consistency 透過多條 sampled paths 降低這種風險

D. Self-Consistency 必須 fine-tune 模型，使模型學會輸出更長的 reasoning chain

E. Chain-of-Thought prompting 和 Self-Consistency 完全無關，兩者不能一起使用

Q11

(Choose four)

下列哪些情況下，Chain-of-Thought prompting 可能不一定帶來明顯改善，甚至可能造成問題？

A. Task 本身只需要簡單 factual lookup 或單步分類，不需要 multi-step reasoning

B. Model 太小或 reasoning capability 不足，產生的 intermediate steps 可能只是看似合理但實際錯誤的 rationalization

C. Evaluation 只接受簡短 final answer，而模型輸出冗長 reasoning 導致格式不符

D. Task 需要多步 arithmetic reasoning，而且 final answer 可以明確驗證

E. Prompt 中的 CoT exemplars 與 test task 的 reasoning pattern 差異很大，導致模型模仿錯誤格式或錯誤推理方式

Q12

(Choose three)

關於 Chain-of-Thought prompting 的 implementation / evaluation，下列哪些做法合理？

A. 在 answer extraction 階段，將 reasoning text 和 final answer 分開處理，避免因 reasoning 太長影響 final answer parsing

B. 比較 Direct prompting、Chain-of-Thought prompting、Self-Consistency with CoT，可以幫助分析 intermediate reasoning 與 multi-sample aggregation 的效果

C. 若使用 CoT 進行 math reasoning，應該完全不檢查 final answer，只評估 reasoning text 是否看起來流暢

D. 對於 multiple-choice tasks，可以要求模型最後輸出固定格式，例如 `Therefore, the answer is (C)`，以降低 parsing ambiguity

E. 為了最大化 CoT 效果，prompt 中的 examples 應該盡量與目標任務無關，避免模型過度模仿

Q13

(Choose one)

假設某題有 5 條 sampled outputs，其 final answers 與 confidence 如下， standard Self-Consistency majority voting 和 confidence-weighted voting 的結果分別為何？

Trace 1: Answer = A, confidence = 0.96
Trace 2: Answer = A, confidence = 0.94
Trace 3: Answer = B, confidence = 0.55
Trace 4: Answer = B, confidence = 0.52
Trace 5: Answer = B, confidence = 0.50

A. Majority voting 選 A；confidence-weighted voting 選 A

B. Majority voting 選 B；confidence-weighted voting 選 A

C. Majority voting 選 B；confidence-weighted voting 選 B

D. Majority voting 選 A；confidence-weighted voting 選 B

E. 兩種方法都無法決定

Q14

(Choose three)

關於 majority voting、confidence-weighted voting、Best-of-N selection，下列哪些敘述正確？

A. Majority voting 主要看 final answer frequency，通常不直接考慮每條 reasoning trace 的品質差異

B. Confidence-weighted voting 可以讓 high-confidence traces 對 final decision 產生較大影響

C. Best-of-N selection 的目標是在多個 sampled candidates 中選出一個較好的 response，而不一定需要 final answers 可被 string-match aggregation

D. Majority voting 一定比 confidence-based selection 更適合 open-ended code generation，因為 code 沒有語法差異問題

E. Confidence-based methods 完全不需要模型輸出 logits / probability distribution，也不需要任何 confidence proxy

Q15

(Choose three)

關於 Chain-of-Thought prompting、Self-Consistency、Self-Certainty 之間的差異，下列哪些敘述正確？

A. Chain-of-Thought prompting 主要透過 prompt 中的 intermediate reasoning demonstrations，引導模型在回答前 產生 reasoning steps

B. Self-Consistency 通常建立在 Chain-of-Thought prompting 上，透過 sampling 多條 reasoning paths 並聚合 final answers 來降低單一路徑錯誤的風險

C. Self-Certainty 的核心是使用 external reward model 對每個 candidate response 打分，因此一定需要額外 training

D. Self-Certainty 利用 model generation 時的 token probability distribution / logits 估計 response quality，可用於 Best-of-N selection

E. Chain-of-Thought prompting、Self-Consistency、Self-Certainty 都必須更新 LLM parameters 才能使用

Q16

(Choose four)

關於 beam search 與 Self-Consistency 的差異，下列哪些敘述正確？

A. Beam search 在 decoding 過程中會保留目前分數最高的 top-B partial sequences，並逐步擴展這些 candidates

B. Self-Consistency 通常會 sample 多條完整 reasoning paths，並根據 final answer frequency 或 answer-level aggregation 決定輸出

C. Beam search 的目標通常是找到 sequence-level score 較高的 output，而不是對 final answers 做 majority voting

D. 只要 beam size 等於 sampled paths 數量，beam search 和 Self-Consistency 就完全等價

E. Beam search 通常不需要 stochastic sampling；若使用 deterministic decoding 和固定模型，結果通常是 deterministic 的

Q17

(Choose one)

假設 vocabulary 只有 {A, B, C}，使用 beam size = 2 的 beam search，每個 sequence 的 score 是 cumulative log probability。

已知第一步 token 的 log probability 為：A: -0.1，B: -0.3，C: -1.5 因此第一步保留 A 和 B。

第二步的 conditional log probability 如下：

After A:

AA: -2.0，AB: -0.4，AC: -0.5

After B:

BA: -0.2，BB: -0.3，BC: -2.0

請問第二步展開後，beam search 會保留哪兩個 partial sequences？

A. AB 和 AC

B. BA 和 BB

C. AB 和 BA

D. AA 和 BA

E. A 和 B

Q18

(Choose three)

關於 beam search 在 LLM reasoning tasks 中的限制，下列哪些敘述正確？

A. Beam search 偏向保留 high-probability sequences，但 high-probability reasoning path 不一定會導向 correct answer。

B. Beam search 的多個 beams 可能高度相似，因此不一定能提供像 stochastic sampling 那樣的 reasoning diversity。

C. 在 Chain-of-Thought reasoning 中，beam search 若太偏向局部 high-probability token，可能會放大 model 常見但錯誤的 reasoning pattern。

D. Beam search 會自動對 final answers 做 majority voting，因此通常等價於 Self-Consistency。

E. Beam search 一定比 greedy decoding 需要更少 inference compute，因為它可以同時保留多條 sequence。

Q19

請截圖colab的accuracy table：

	method	accuracy	correct	total	budget	avg_runtime_sec	avg_valid_answers
0	Confidence						
1	Direct Inference						
2	Self-Certainty						
3	Self-Consistency						

Q20

(Choose one)

根據Q19圖片結果，請問accuracy最低的方法是什麼？

A. Self-Certainty

B. Direct inference

C. Self-Consistency

D. Confidence

English problems

Q1

(Choose one)

According to **Self-Consistency Improves Chain of Thought Reasoning in Language Models**, if we want to apply **self-consistency decoding** to a test question, what is the correct order of the following steps?

1. Parse / extract the final answer from each sampled reasoning path
2. Aggregate the final answers, for example by using majority voting or answer-level marginalization
3. Use a prompt containing few-shot Chain-of-Thought exemplars to make the LLM perform reasoning
4. Use stochastic decoding to sample multiple diverse reasoning paths instead of taking only the greedy path
5. Return the final answer with the highest aggregated support

A. 34125
B. 31425
C. 34215
D. 43125
E. 34152

Q2

(Choose three)

In which of the following situations might the effectiveness of **self-consistency** be limited?

A. The task's final answer is difficult to parse or normalize, such as in long-form essay generation or open-ended subjective responses.

B. The model produces sampled reasoning paths with very little diversity, causing the K samples to behave almost like repeated greedy decoding.

C. The model produces many fluent but logically invalid reasoning paths, causing majority voting to be dominated by incorrect answers.

D. The task has a clearly comparable final answer, and different reasoning paths often converge to the correct answer.

E. The inference budget allows multiple sampled reasoning paths, and answer extraction can be performed reliably.

Q3

(Choose three)

Which of the following statements correctly describe the core idea of **self-consistency**?

A. Self-consistency assumes that a complex reasoning problem may have multiple different but valid reasoning paths, and these paths may converge to the same correct answer.

B. The main goal of self-consistency is to train an extra verifier model to determine whether each reasoning path is correct.

C. Self-consistency can be understood as marginalizing over latent reasoning paths, rather than relying on a single greedy path.

D. Self-consistency is a training-free decoding strategy and usually does not require updating the LLM parameters.

E. The main contribution of self-consistency is to remove Chain-of-Thought exemplars and make the model use answer-only prompting.

Q4

(Choose one)

According to the paper, which of the following best describes the core idea of **self-certainty**?

A. Measuring the length of the generated response, because a longer response usually indicates more complete reasoning.

B. Computing only the average log-probability of the sampled tokens, while completely ignoring the probabilities of other vocabulary tokens.

C. Using the entire token probability distribution at each decoding step to measure the model's certainty / concentration about the next-token prediction.

D. Applying majority voting over the final answers and selecting the answer that appears most frequently.

E. Using another LLM to perform pairwise comparison over the responses and selecting the best response.

Q5

(Choose two)

According to the experimental results / discussion in the **self-certainty paper**, which of the following statements are correct?

A. Self-certainty can be used for Best-of-N selection without additionally training an external reward model.

B. On mathematical reasoning tasks with clearly defined final answers, pure self-certainty selection does not always outperform self-consistency; therefore, the paper further uses Borda voting to combine confidence ranking with answer frequency.

C. The paper's evaluation shows that self-consistency only outperforms other test-time scaling methods on mathematical reasoning tasks.

D. The paper states that self-consistency naturally handles open-ended code generation, because all code outputs can be reliably string-matched into the same final answer.

E. The effectiveness of self-certainty is independent of the sample size (N), so (N = 1) and (N = 64) should theoretically produce the same result in Best-of-N selection.

Q6

(Choose two)

Which of the following statements correctly describe the difference between **self-certainty** and **AvgLogP / negative perplexity**?

A. AvgLogP / negative perplexity mainly relies on the probability of the generated tokens, while self-certainty uses the entire token probability distribution at each decoding step.

B. Self-certainty can reflect the concentration of the model's output distribution, rather than only how likely the sampled token itself is.

C. Negative perplexity and self-certainty are completely equivalent, because both only depend on the likelihood of the generated tokens.

D. Self-certainty does not require logits or a probability distribution, so it can be fully computed using a completely black-box API that does not provide logprobs.

E. Self-certainty only looks at the final answer string and does not use the token distribution from intermediate generation steps.

Q7

(Choose three)

Which of the following statements correctly describe the main motivation of **Deep Think with Confidence**?

A. Standard self-consistency / majority voting treats all reasoning traces equally, which may allow low-quality traces to affect the final answer.

B. Parallel thinking can improve reasoning accuracy, but it generates many reasoning traces, so the inference overhead increases with the number of traces.

C. A global confidence measure requires the complete reasoning trace to be computed, so it is not suitable for early stopping low-quality traces.

D. The main goal of DeepConf is to train an external reward model to replace the LLM's own next-token distribution.

E. DeepConf assumes that low-confidence reasoning segments should not be observed, because they are completely unrelated to final answer correctness.

Q8

(Choose three)

Which of the following statements correctly describe the **confidence measurements** used in **Deep Think with Confidence**?

A. Token Confidence is computed from the model's top-k token probability distribution at a specific decoding position, and is used to reflect the model's certainty at that position.

B. Average Trace Confidence averages the token confidence over the entire trace, but it may hide local reasoning breakdowns in the middle of the trace.

C. Bottom 10% Group Confidence focuses on the lowest-confidence groups, helping identify reasoning segments that are most likely to be problematic.

D. Tail Confidence completely ignores the final portion of the reasoning trace and only measures the beginning tokens, because early tokens are most important for the final answer.

E. Lowest Group Confidence is the maximum value among all group confidence scores, so it represents the smoothest part of the reasoning trace.

Q9

(Choose three)

Which of the following statements correctly describe **online thinking** in **DeepConf-low** and **DeepConf-high**?

A. DeepConf-low uses the top η=10%\eta = 10\%η=10% of warmup traces to set a stricter confidence threshold, so it more aggressively early stops low-confidence traces during online generation.

B. DeepConf-high uses the top η=90%\eta = 90\%η=90% of warmup traces to set a looser confidence threshold, so it keeps more traces during online generation.

C. Online thinking does not require any offline warmup, because the threshold sss is a fixed value built into the model parameters.

D. During online generation, if the current group confidence falls below the threshold sss, the trace can be early terminated.

E. The difference between DeepConf-low and DeepConf-high is that one requires model training while the other does not.

Q10

(Choose two)

Which of the following statements correctly describe the relationship between **Chain-of-Thought prompting** and **Self-Consistency**?

A. Chain-of-Thought prompting usually makes the model generate one reasoning path, while Self-Consistency samples multiple reasoning paths based on CoT and aggregates the final answers.

B. The purpose of Self-Consistency is to remove all intermediate reasoning steps and keep only the final answer.

C. Greedy CoT may produce an incorrect final answer due to an early mistake in a single reasoning path; Self-Consistency reduces this risk by using multiple sampled paths.

D. Self-Consistency requires fine-tuning the model so that it learns to generate longer reasoning chains.

E. Chain-of-Thought prompting and Self-Consistency are completely unrelated and cannot be used together.

Q11

(Choose four)

In which of the following situations might **Chain-of-Thought prompting** fail to bring clear improvements, or even cause problems?

A. The task only requires simple factual lookup or single-step classification, and does not require multi-step reasoning.

B. The model is too small or lacks sufficient reasoning capability, so the generated intermediate steps may be fluent but incorrect rationalizations.

C. The evaluation only accepts a short final answer, but the model outputs long reasoning, causing format mismatch or parsing failure.

D. The task requires multi-step arithmetic reasoning, and the final answer can be clearly verified.

E. The CoT exemplars in the prompt are very different from the reasoning pattern required by the test task, causing the model to imitate an incorrect format or reasoning style.

Q12

(Choose three)

Which of the following are reasonable practices for **implementation / evaluation** of **Chain-of-Thought prompting**?

A. During answer extraction, separate the reasoning text from the final answer to prevent long reasoning from interfering with final answer parsing.

B. Comparing Direct prompting, Chain-of-Thought prompting, and Self-Consistency with CoT can help analyze the effects of intermediate reasoning and multi-sample aggregation.

C. When using CoT for math reasoning, we should completely ignore the final answer and only evaluate whether the reasoning text looks fluent.

D. For multiple-choice tasks, we can ask the model to output a fixed final-answer format, such as `Therefore, the answer is (C)`, to reduce parsing ambiguity.

E. To maximize the effectiveness of CoT, the examples in the prompt should be as unrelated to the target task as possible, so the model does not over-imitate them.

Q13

(Choose one)

Suppose a question has 5 **sampled outputs** with the following **final answers** and **confidence scores**. What are the results of **standard Self-Consistency majority voting** and **confidence-weighted voting**, respectively?

Trace 1: Answer = A, confidence = 0.96
Trace 2: Answer = A, confidence = 0.94
Trace 3: Answer = B, confidence = 0.55
Trace 4: Answer = B, confidence = 0.52
Trace 5: Answer = B, confidence = 0.50

A. Majority voting selects A; confidence-weighted voting selects A
B. Majority voting selects B; confidence-weighted voting selects A
C. Majority voting selects B; confidence-weighted voting selects B
D. Majority voting selects A; confidence-weighted voting selects B
E. Both methods cannot determine the answer

Q14

(Choose three)

Which of the following statements about **majority voting**, **confidence-weighted voting**, and **Best-of-N selection** are correct?

A. Majority voting primarily relies on the frequency of final answers, and typically does not consider the quality differences among individual reasoning traces.

B. Confidence-weighted voting allows high-confidence traces to have a greater influence on the final decision.

C. The goal of Best-of-N selection is to choose a better response from multiple sampled candidates, and it does not necessarily require that final answers can be aggregated via string matching.

D. Majority voting is always more suitable than confidence-based selection for open-ended code generation, because code does not have syntactic variation issues.

E. Confidence-based methods do not require model outputs such as logits or probability distributions, and do not rely on any form of confidence proxy.

Q15

(Choose three)

Which of the following statements correctly describe the differences among **Chain-of-Thought prompting**, **Self-Consistency**, and **Self-Certainty**?

A. Chain-of-Thought prompting mainly uses intermediate reasoning demonstrations in the prompt to guide the model to generate reasoning steps before answering.

B. Self-Consistency is usually built on top of Chain-of-Thought prompting. It samples multiple reasoning paths and aggregates the final answers to reduce the risk of relying on a single reasoning path.

C. The core idea of Self-Certainty is to use an external reward model to score each candidate response, so it always requires additional training.

D. Self-Certainty uses the token probability distribution / logits produced during model generation to estimate response quality, and can be used for Best-of-N selection.

E. Chain-of-Thought prompting, Self-Consistency, and Self-Certainty all require updating the LLM parameters before they can be used.

Q16

(Choose four)

Which of the following statements correctly describe the differences between **beam search** and **Self-Consistency**?

A. During decoding, beam search keeps the current top-B partial sequences with the highest scores and expands these candidates step by step.

B. Self-Consistency usually samples multiple complete reasoning paths and determines the output based on final answer frequency or answer-level aggregation.

C. The goal of beam search is usually to find an output with a high sequence-level score, rather than performing majority voting over final answers.

D. As long as the beam size equals the number of sampled paths, beam search and Self-Consistency are completely equivalent.

E. Beam search usually does not require stochastic sampling; with deterministic decoding and a fixed model, its result is usually deterministic.

Q17

(Choose one)

Suppose the vocabulary only contains `{A, B, C}`. We use **beam search** with **beam size = 2**, and each sequence is scored by **cumulative log probability**.

The first-step token log probabilities are: A: -0.1, B: -0.3, C: -1.5

Therefore, the first step keeps `A` and `B`. The second-step conditional log probabilities are:

After A:
AA: -2.0, AB: -0.4, AC: -0.5

After B:
BA: -0.2, BB: -0.3, BC: -2.0

After expanding the second step, which two partial sequences will **beam search** keep?

A. AB and AC
B. BA and BB
C. AB and BA
D. AA and BA
E. A and B

Q18

(Choose three)

Which of the following statements correctly describe the limitations of **beam search** in **LLM reasoning tasks**?

A. Beam search tends to keep high-probability sequences, but a high-probability reasoning path does not necessarily lead to the correct answer.

B. The multiple beams maintained by beam search may be highly similar, and therefore may not provide the same level of reasoning diversity as stochastic sampling.

C. In Chain-of-Thought reasoning, if beam search overly favors locally high-probability tokens, it may amplify common but incorrect reasoning patterns learned by the model.

D. Beam search automatically performs majority voting over final answers, and is therefore equivalent to Self-Consistency.

E. Beam search always requires less inference compute than greedy decoding, because it keeps multiple sequences simultaneously.

Q19

Please take a screenshot of the accuracy table in Colab.

	method	accuracy	correct	total	budget	avg_runtime_sec	avg_valid_answers
0	Confidence						
1	Direct Inference						
2	Self-Certainty						
3	Self-Consistency						

Q20

(Choose one)

According to the result shown in the **Q19 image**, which method has the lowest accuracy?

A. Self-Certainty.

B. Direct inference.

C. Self-Consistency.

D. Confidence.