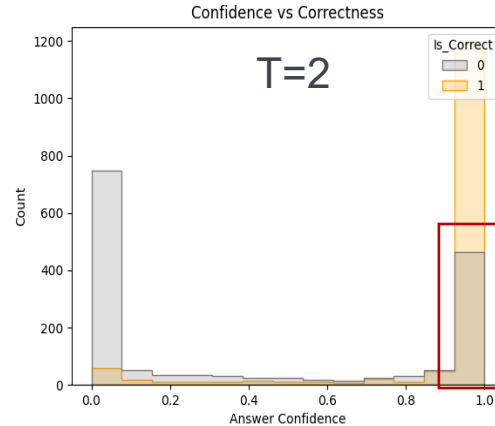




- Presenter: Ziyao Zhou
- Supervisor: Yuchen Xia M. Sc.
- Examiner: Prof. Dr. Ing. Michael Weyrich

Investigation of the Explainability of Results Generated by Large Language Models

05.10.2023



- **Motivation**
- **Research Question**
- **Related Work**
- **Research Work Pipeline**
- **Methodology**
- **Results and Evaluation**
- **Conclusion and Future Work**

- **Motivation**
- **Research Question**
- **Related Work**
- **Research Work Pipeline**
- **Methodology**
- **Results and Evaluation**
- **Conclusion and Future Work**

Background

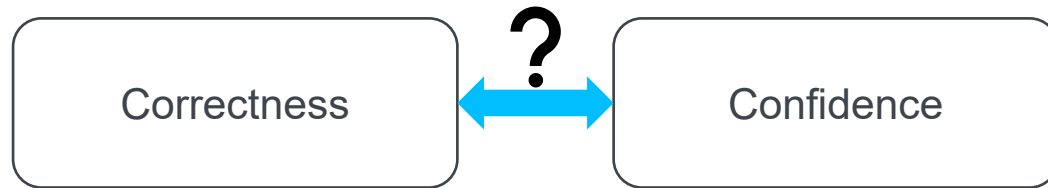
Motivation

- Large Language Models – widely used
 - In ChatGPT Answer Generated
 - Over 100 million Users
 - 1.6 billion visits in June 2023¹

- **Motivation**
- **Research Question**
- **Related Work**
- **Research Work Pipeline**
- **Methodology**
- **Results and Evaluation**
- **Conclusion and Future Work**

Research Question

What is this thesis about?



- Use mathematical ways to calculate confidence and the errors between confidence and correctness
- Visualization

- **Motivation**
- **Research Question**
- **Related Work**
- **Research Work Pipeline**
- **Methodology**
- **Results and Evaluation**
- **Conclusion and Future Work**

Related Work

Proxy metrics to estimate the confidence or Correctness

- **Prompt the models** to express their uncertainty in words[1]



```
"choices": [  
  {  
    "text": "Answer: D, confidence: 3"  
  }  
]
```

- Use **log probabilities** to estimate confidence



$$\text{Confidence} = \text{softmax}(z_j) = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}}$$

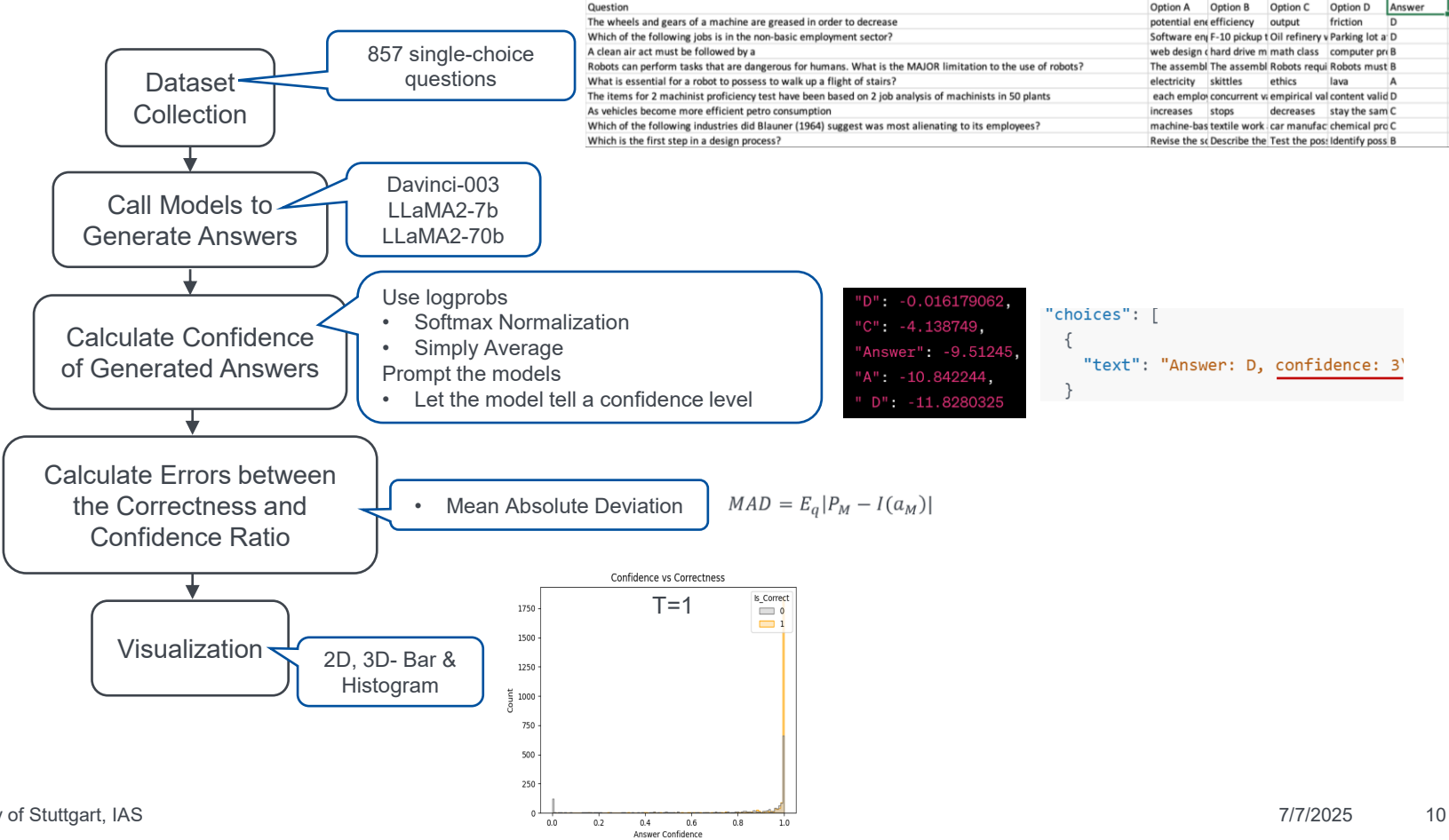
- Use **similarity metrics** to estimate confidence [2]

- Calculate similarity based on **semantic meaning** [3]

- Using activation values of **hidden layer** to predict correctness[4]

- **Motivation**
- **Research Question**
- **Related Work**
- **Research Work Pipeline**
- **Methodology**
- **Results and Evaluation**
- **Conclusion and Future Work**

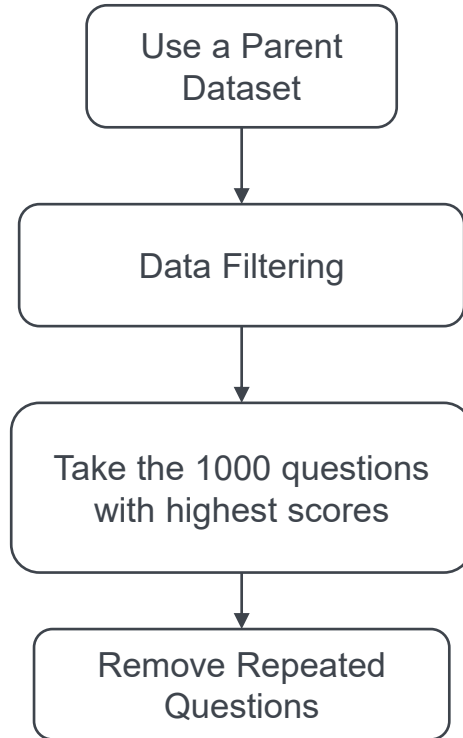
Research Work Pipeline



- **Motivation**
- **Research Question**
- **Related Work**
- **Research Work Pipeline**
- **Methodology**
- **Results and Evaluation**
- **Conclusion and Future Work**

Methodology

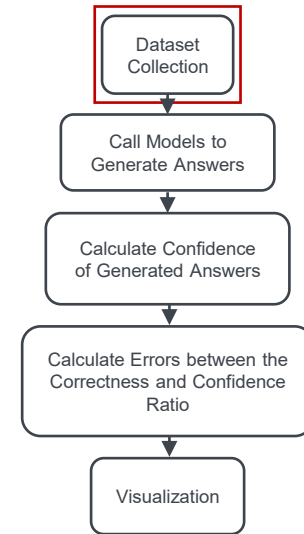
Dataset Collection



Scoring all questions based on the correlation to the field of automation

Using Sentence Transformer Model

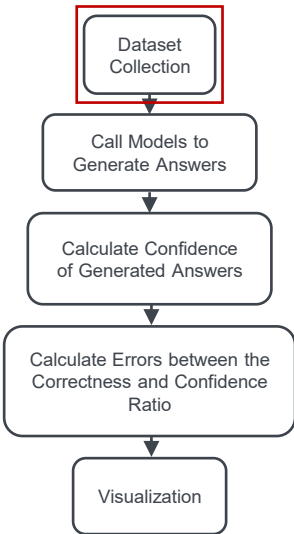
```
model = SentenceTransformer('all-distilroberta-v1')
search_sentence = "Engineering Principles, Industrial Engineering,"
encoded_search_sentence = model.encode(search_sentence)
```



Methodology

Dataset Collection

A labeled dataset with 857 single-choice questions:

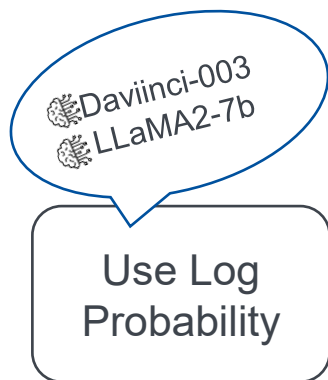


	A	B	C	D	Question body	E	F	4 Options	H	I	Answer
1	Source	File	Index	Correlation Score	Question		Option A	Option B	Option C	Option D	Answer
2	https://github.com/arc_easy.csv	arc_easy.csv	673	0.398	The wheels and gears of a machine are greased in order to decrease		potential ene	efficiency	output	friction	D
3	https://github.com/high_school_	high_school_	49	0.397	Which of the following jobs is in the non-basic employment sector?		Software en	F-10 pickup t	Oil refinery v	Parking lot a	D
4	https://github.com/obqa.csv	obqa.csv	1466	0.385	A clean air act must be followed by a		web design c	hard drive m	math class	computer pro	B
5	https://github.com/arc_hard.csv	arc_hard.csv	944	0.384	Robots can perform tasks that are dangerous for humans. What is the MAJOR limitation to the use of robots?		The assembl	The assembl	Robots requi	Robots must	B
6	https://github.com/obqa.csv	obqa.csv	2764	0.367	What is essential for a robot to possess to walk up a flight of stairs?		electricity	skittles	ethics	lava	A
7	https://github.com/professional_	professional_	121	0.345	The items for 2 machinist proficiency test have been based on 2 job analysis of machinists in 50 plants		each emplo	concurrent v	empirical val	content valid	D
8	https://github.com/obqa.csv	obqa.csv	3511	0.343	As vehicles become more efficient petro consumption		increases	stops	decreases	stay the sam	C
9	https://github.com/sociology_te	sociology_te	154	0.337	Which of the following industries did Blauner (1964) suggest was most alienating to its employees?		machine-bas	textile work	car manufac	chemical pro	C
10	https://github.com/arc_easy.csv	arc_easy.csv	2202	0.336	Which is the first step in a design process?		Revise the sc	Describe the	Test the pos	Identify poss	B

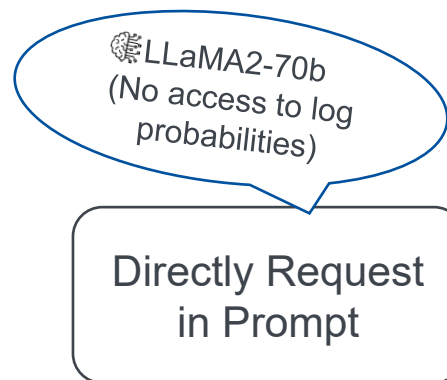
Methodology

Calculate Confidence

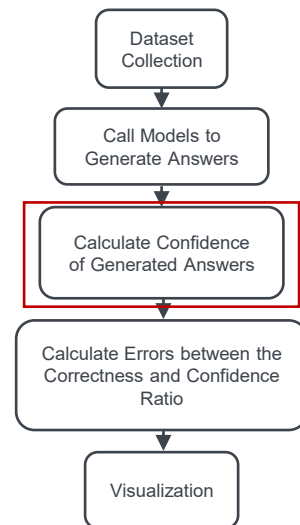
Two Ways



$$Confidence = softmax(z_j) = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}}$$



"You are an expert in the field of automation. Your goal is to answer the single-choice question as well as output your level of confidence."



Methodology

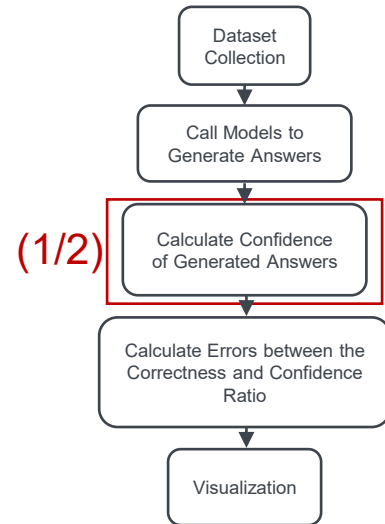
Calculate Confidence with logprobs(1/2)

Request: The wheels and gears of a machine are greased in order to decrease
A: potential energy B: efficiency C: output D: friction

Response:

```
"choices": [
  {
    "text": "\n\nD",
    "index": 0,
    "logprobs": {
      "tokens": [
        "\n",
        "\n",
        "D"
      ],
      "token_logprobs": [
        -0.0053596636,
        -0.00041011896,
        -0.016179062
      ],
      "top_logprobs": [
        {
          "\n": -0.0053596636,
          " ": -5.702245,
          "\n\n": -7.2680273,
          ".": -7.4418178,
          " Answer": -8.131088
        }
      ]
    }
  },
  {
    "text": "\n\nA",
    "index": 1,
    "logprobs": {
      "tokens": [
        "\n",
        "\n",
        "A"
      ],
      "token_logprobs": [
        -0.0053596636,
        -0.00041011896,
        -10.814245
      ],
      "top_logprobs": [
        {
          "\n": -0.0053596636,
          " ": -5.702245,
          "\n\n": -7.2680273,
          ".": -7.4418178,
          " Answer": -9.51245
        }
      ]
    }
  },
  {
    "text": "\n\nC",
    "index": 2,
    "logprobs": {
      "tokens": [
        "\n",
        "\n",
        "C"
      ],
      "token_logprobs": [
        -0.0053596636,
        -0.00041011896,
        -4.138749
      ],
      "top_logprobs": [
        {
          "\n": -0.0053596636,
          " ": -5.702245,
          "\n\n": -7.2680273,
          ".": -7.4418178,
          " Answer": -11.8280325
        }
      ]
    }
  },
  {
    "text": "\n\nD",
    "index": 3,
    "logprobs": {
      "tokens": [
        "\n",
        "\n",
        "D"
      ],
      "token_logprobs": [
        -0.0053596636,
        -0.00041011896,
        -0.016179062
      ],
      "top_logprobs": [
        {
          "\n": -0.0053596636,
          " ": -5.702245,
          "\n\n": -7.2680273,
          ".": -7.4418178,
          " Answer": -13.004202
        }
      ]
    }
  }
]
```

```
{
  "\n": -0.00041011896,
  "Answer": -7.884056,
  "A": -10.814245,
  " ": -12.73655,
  " Answer": -13.004202
},
{
  "D": -0.016179062,
  "C": -4.138749,
  "Answer": -9.51245,
  "A": -10.842244,
  " D": -11.8280325
}
]
```



Methodology

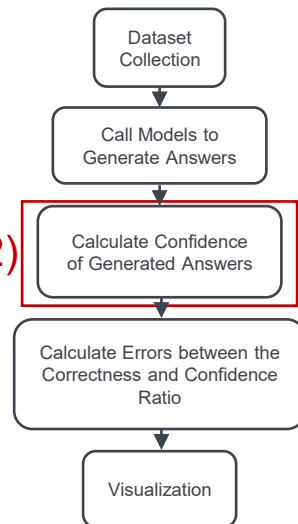
Calculate Confidence with $\log\text{probs}(1/2)$

Response:

```
"choices": [
  {
    "text": "\n\nD",
    "index": 0,
    "logprobs": {
      "tokens": [
        "\n",
        "\n",
        "D"
      ],
      "token_logprobs": [
        -0.0053596636,
        -0.00041011896,
        -0.016179062
      ],
      "top_logprobs": [
        {
          "\n": -0.0053596636,
          ".": -5.702245,
          "D": -8.131088
        }
      ]
    }
  }
]
```

First Situation: "ABCD" among the answers

(1/2)



```
{
  "\n": -0.00041011896,
  "Answer": -7.884056,
  "A": -10.814245,
  " ": -12.73655,
  " Answer": -13.004202
},
{
  "D": -0.016179062,
  "C": -4.138749,
  "Answer": -9.51245,
  "A": -10.842244,
  " D": -11.8280325
}
]
```

$$\frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}}$$

$$\text{Confidence} = \frac{e^{-0.016}}{e^{-0.016} + e^{-4.14} + e^{-10.84}} = 0.98$$

Methodology

Calculate Confidence with logprobs(1/2)

Response:

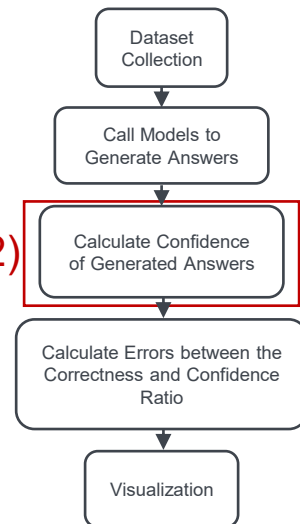
```
{
  "text": " of electricity\n",
  "index": 4,
  "logprobs": {
    "tokens": [
      " of",
      " electricity",
      "\n"
    ],
    "token_logprobs": [
      -0.5815068,
      -8.982481,
      -0.009463984
    ],
    "top_logprobs": [
      {
        " of": -0.5815068,
        "\n": -0.9098606,
```

Second Situation: no “ABCD”
among the answers

$$Confidence = \frac{\sum_{i=1}^N e^{z_i}}{N}$$

$$Confidence = \frac{1}{3}(e^{-0.58} + e^{-8.92} + e^{-0.01}) = 0.52$$

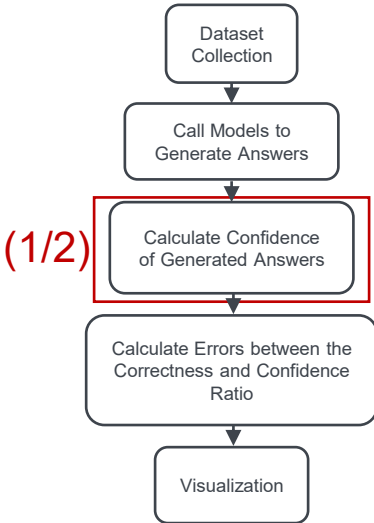
(1/2)



Methodology

Calculate Confidence with logprobs(1/2)

- Generate 5 responses for each question



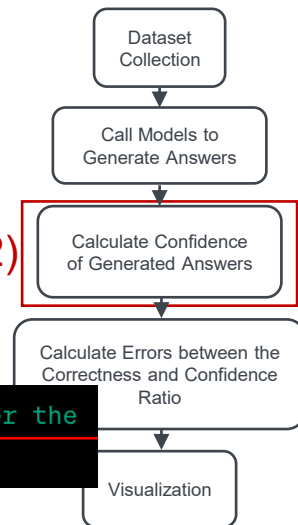
Result Table:

0	1	2	3	4	5	6	7	8	9	Generated Answer 1	Confidence 1	Generated Answer 2	Confidence 2	Generate	Confidence	Generate	Confidence	Generate	Confidence 5	
The wheel	potential	efficiency	output	friction	D	https://git	arc_easy.c	673	0.398	D	0.999968148	D	0.999968148	D	0.999968	(The ans	0	D	0.999968148	
Which of t	Software	F-10 picku	Oil refine	Parking lo	D	https://git	high_scho	49	0.397	A.	0.999675588	A	0.999987927	A	0.999988	A	0.999988	A	0.999987927	
Robots car	The assen	The assen	Robots re	Robots mi	B	https://git	arc_hard.c	944	0.384	C.	1	C.	0.99998579	C.	0.999986	C.	0.999986	C.	0.99998579	
What is es	electricity	skittles	ethics	lava	A	https://git	obqa.csv	2764	0.367	??	0	Madonna	Answer Rom	0	Able stack	0	negligenc	0	C. Ethics	0
As vehicle	increases	stops	decreases	stay the s	C	https://git	obqa.csv	3511	0.343	C	1	C	1	ings pass	0	keep C.	0	C.	1	
Which is t	Revise the	Describe t	Test the p	Identify p	B	https://git	arc_easy.c	2202	0.336	D	0.253802474	D	0.253802474	B.	0.99498	B	0.810478	B.	0.994980409	
Which of t	Partitioni	K-means c	Grid base	All of the	D	https://git	machine_	56	0.335	D	1	D	1	D	1	D	1	D	1	
Which of t	Escalator	Conveyor	Highway	Railroad	B	https://git	miscellane	285	0.335	Solution: B	0	B.	1	track	0	AngleClos	0	Crossing	0	
What is a	Work flow	Work sche	Work rate	Work out	B	https://git	managem	6	0.334	B	0.999997692	A.	0.000253997	testing kn	0	B	0.999998	B	0.999997692	
Mechanica	wind turb	solar pane	shad tree	giant ice r	D	https://git	obqa.csv	2946	0.334	Current	0	?????	0	ads	0	A.	0.710499	shelf	0	
Which of t	R chart	Run chart	X-bar char	Pareto cha	B	https://git	miscellane	561	0.329	C	0.999993719	C.	0.99789524	C	0	C	0.999994	C	0.999993719	

Methodology

Directly Request in Prompt (2/2)

(2/2)



- Roal and Goal

"You are an expert in the field of automation. Your goal is to answer the single-choice question as well as output your level of confidence."

- Confidence Definition

0: You have almost no confidence – this is essentially a guess, and you have
1: Your confidence is low – you have some basis for your choice, but overall
2: Your confidence is moderate – you are relatively sure, but there's still
3: You are quite confident – you are fairly certain your choice is correct,
4: You are very confident – you are highly certain that your choice is the r
5: You are absolutely confident – you are 100% certain that your choice is c

- Example

Example: Input: // text input
Output: Answer: ..., confidence: ...

Methodology

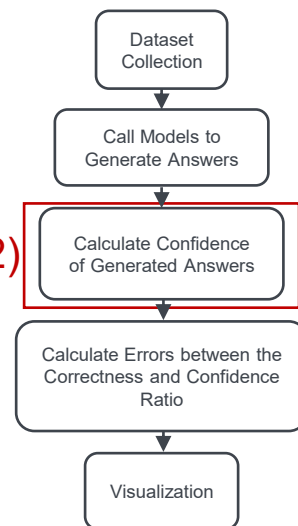
Directly Request in Prompt (2/2)

- Generate 1 response for each question

Result Table:

Question	Option A	Option B	Option C	Option D	TrueAnswer	source	file	index	automatic	Answer	Confidence
The wheel potential	efficiency	output	friction	D		https://git	arc_easy.c	673	0.398	D	1
Which of t	Software	F-10 picku	Oil refine	Parking lo	D	https://git	high_scho	49	0.397	B	0.8
A clean air	web design	hard drive	math clas	computer	B	https://git	obqa.csv	1466	0.385	D	1
Robots ca	The assen	The assen	Robots re	Robots m	B	https://git	arc_hard.c	944	0.384	C	1
What is es	electricity	skittles	ethics	lava	A	https://git	obqa.csv	2764	0.367	E	1
The items	each emp	concurrent	empirical	content v	D	https://git	profession	121	0.345	D	1
As vehicle	increases	stops	decreases	stay the s	C	https://git	obqa.csv	3511	0.343	C	0.8
Which of t	machine-l	textile wc	car manuf	chemical	C	https://git	sociology_	154	0.337	A	1
Which is t	Revise the	Describe t	Test the p	Identify p	B	https://git	arc_easy.c	2202	0.336	B	1

(2/2)



Methodology

Error Calculation between the confidence ratio and the correctness

3 Metrics

- **Mean Square Error**

$$MSE = E_q[P_M - I(a_M)]^2$$

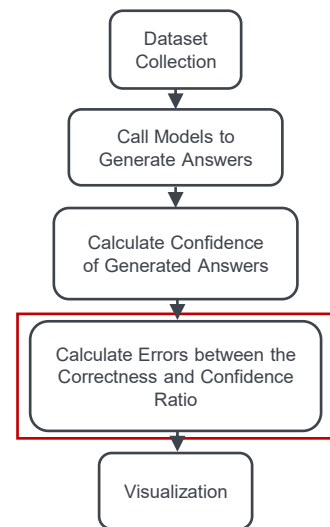
- **Mean Absolute Deviation**

$$MAD = E_q|P_M - I(a_M)|$$



- **Root Mean Square**

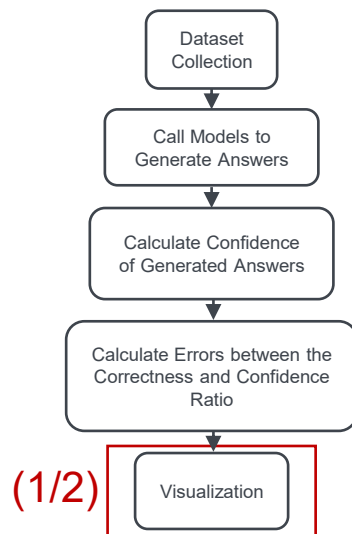
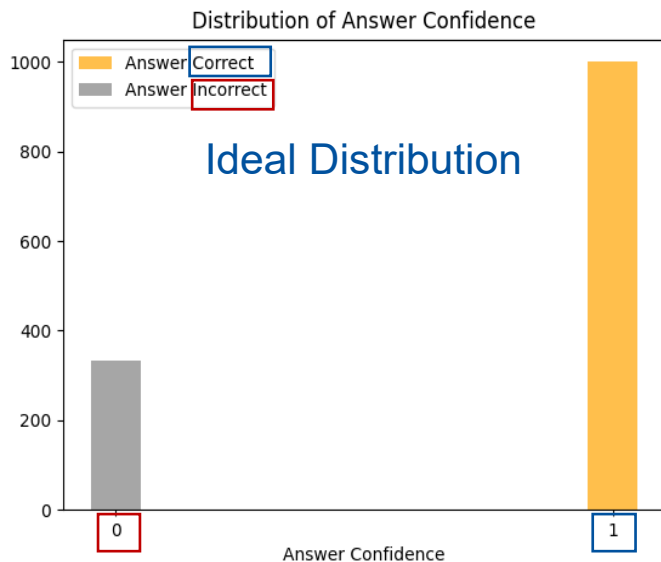
$$RMS = \sqrt{E_q[P_M - I(a_M)]^2}$$



- **Motivation**
- **Research Question**
- **Related Work**
- **Research Work Pipeline**
- **Methodology**
- **Results and Findings**
- **Conclusion and Future Work**

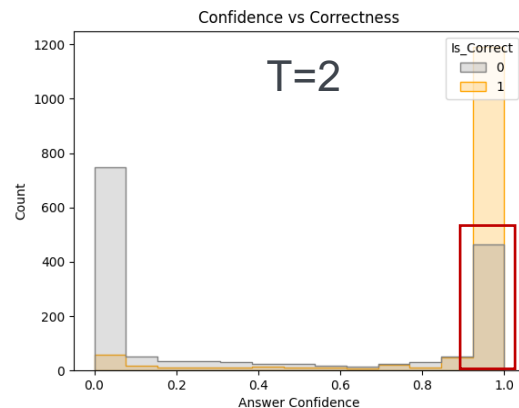
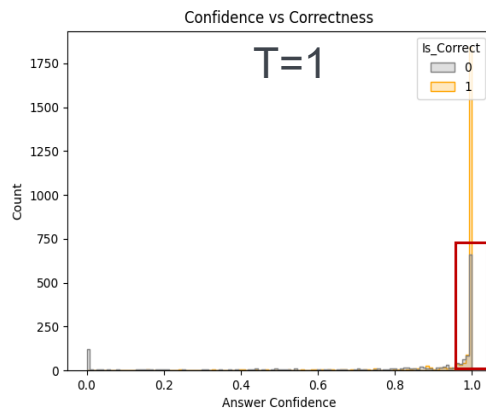
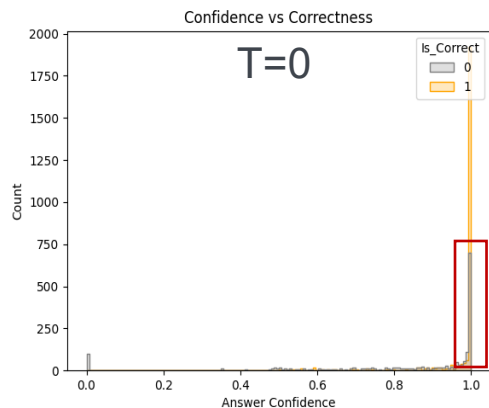
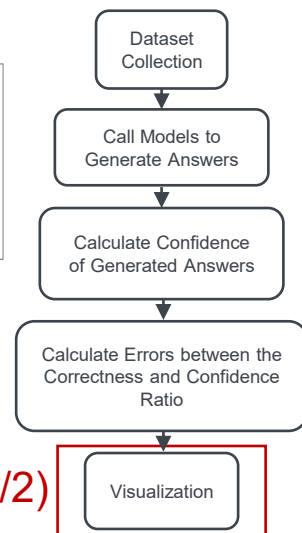
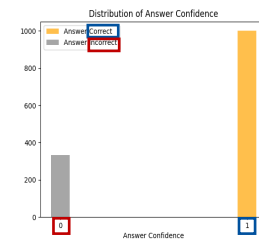
Results and Evaluation

Confidence Distribution (1/2) of Davinci-003 using logprob



Results and Evaluation

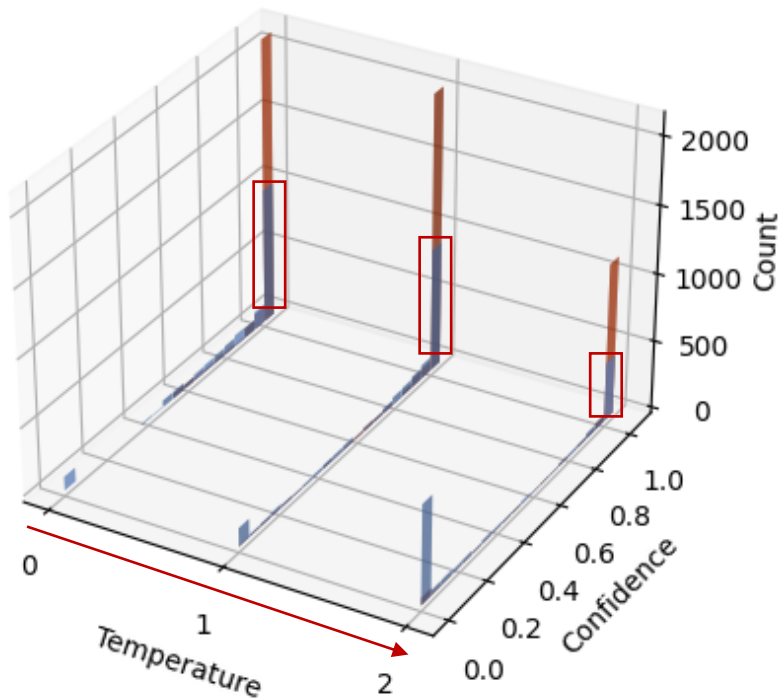
Confidence Distribution (1/2) of Davinci-003 using logprob



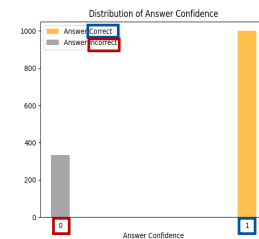
This model is over-confident!

Results and Evaluation

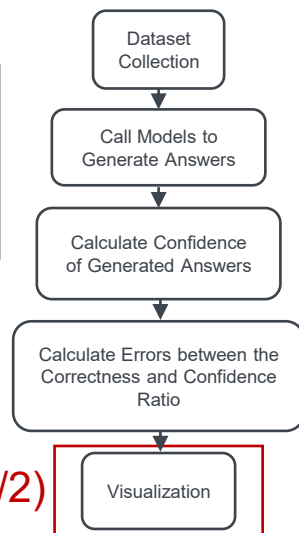
Confidence Distribution (1/2) of Davinci-003 using logprob



Temperature can reduce confidence.

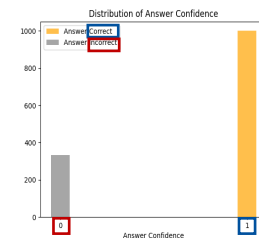
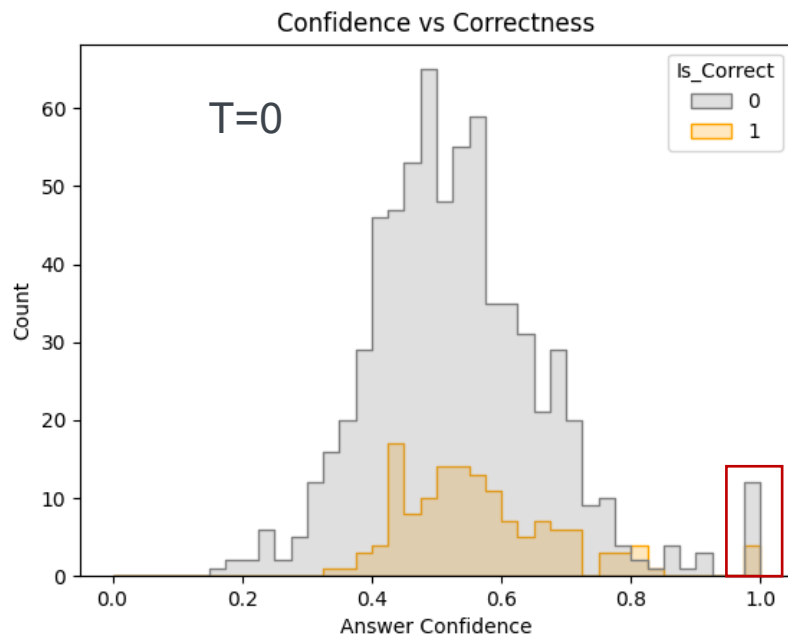


(1/2)



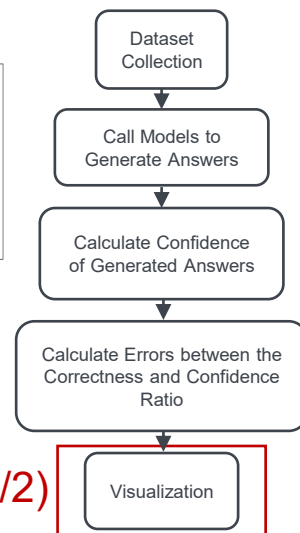
Results and Evaluation

Confidence Distribution (1/2) of LLaMA2-7b using logprob



(1/2)

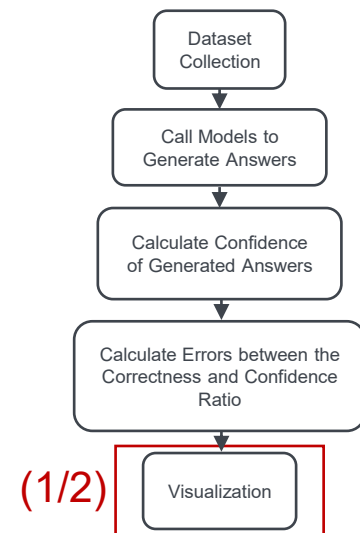
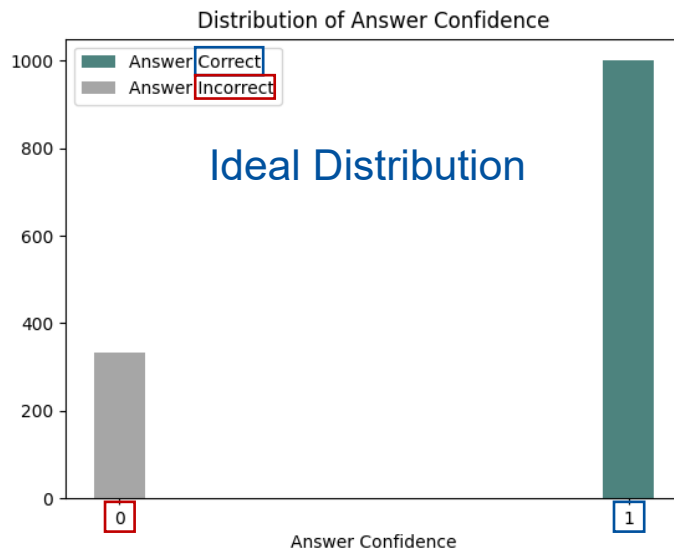
Normal-
Distribution



- This model is not capable of solving questions from the dataset.
- Models with low capacity tends to always be medium-confident.

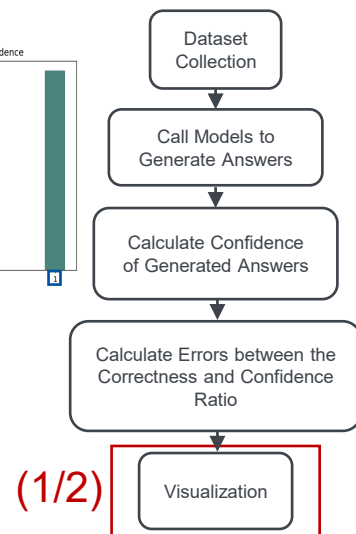
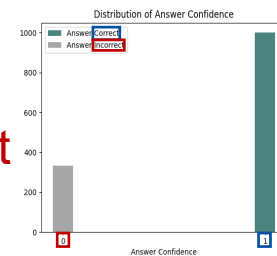
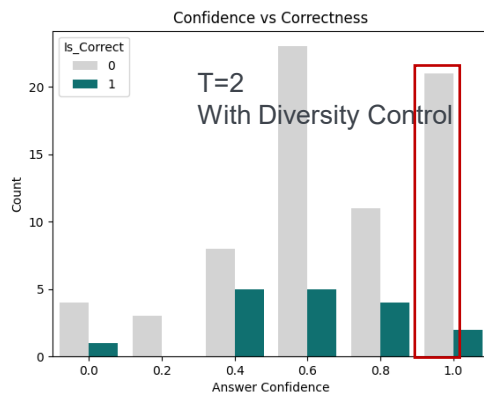
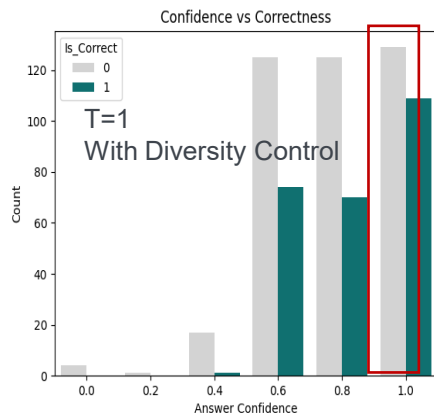
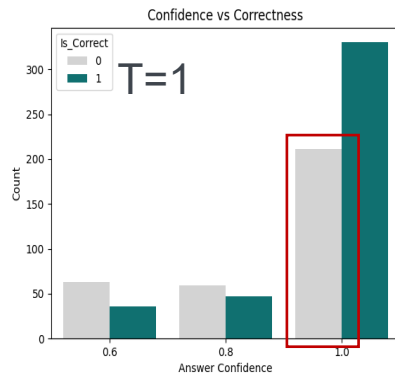
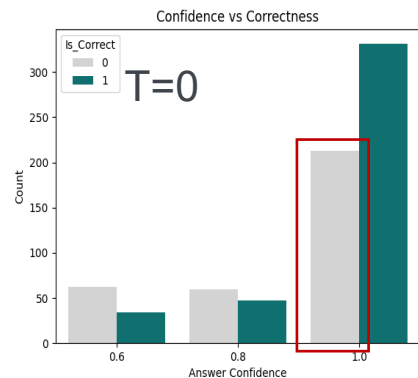
Results and Evaluation

Confidence Distribution (1/2) of LLaMA2-70b using Prompt



Results and Evaluation

Confidence Distribution (1/2) of LLaMA2-70b using Prompt

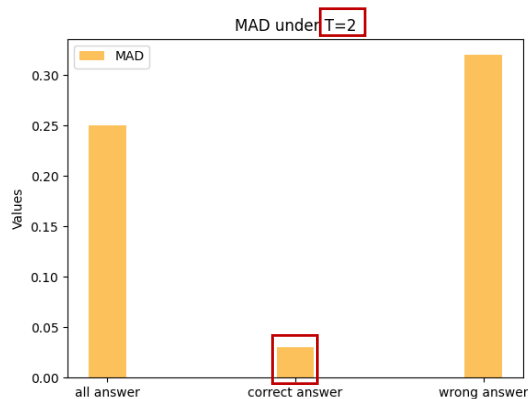
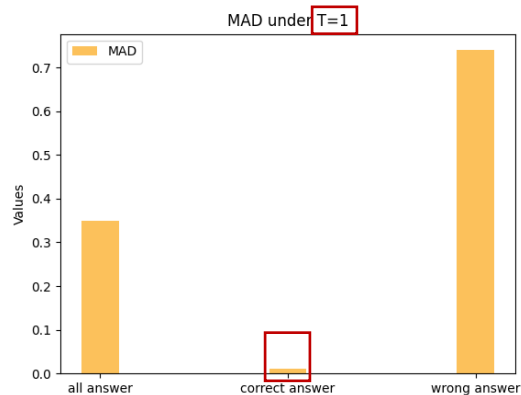
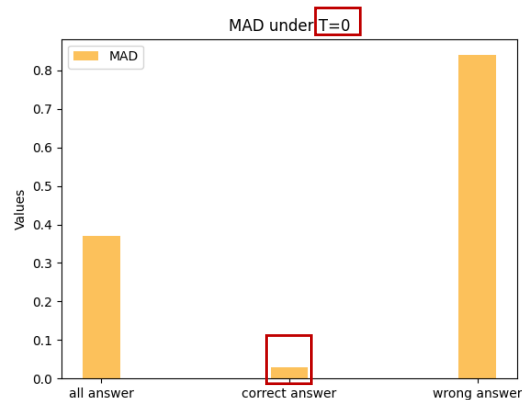


```
payload = {  
  "model": "togethercomputer/llama-2-70b",  
  "prompt": full_prompt,  
  "max_tokens": 20,  
  "stop": ".",  
  "temperature": 1,  
  "top_p": 0.7,  
  "top_k": 50,  
  "repetition_penalty": 1  
}
```

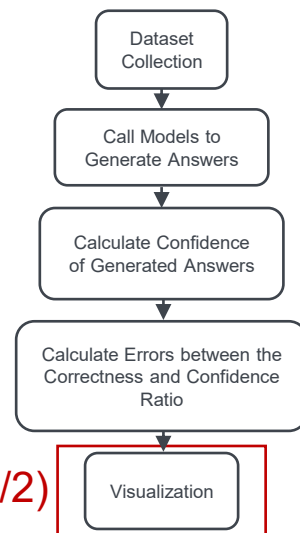
This model is also over-confident.

Results and Evaluation

Visualized Errors (2/2) of Davinci-003 (Logprob)



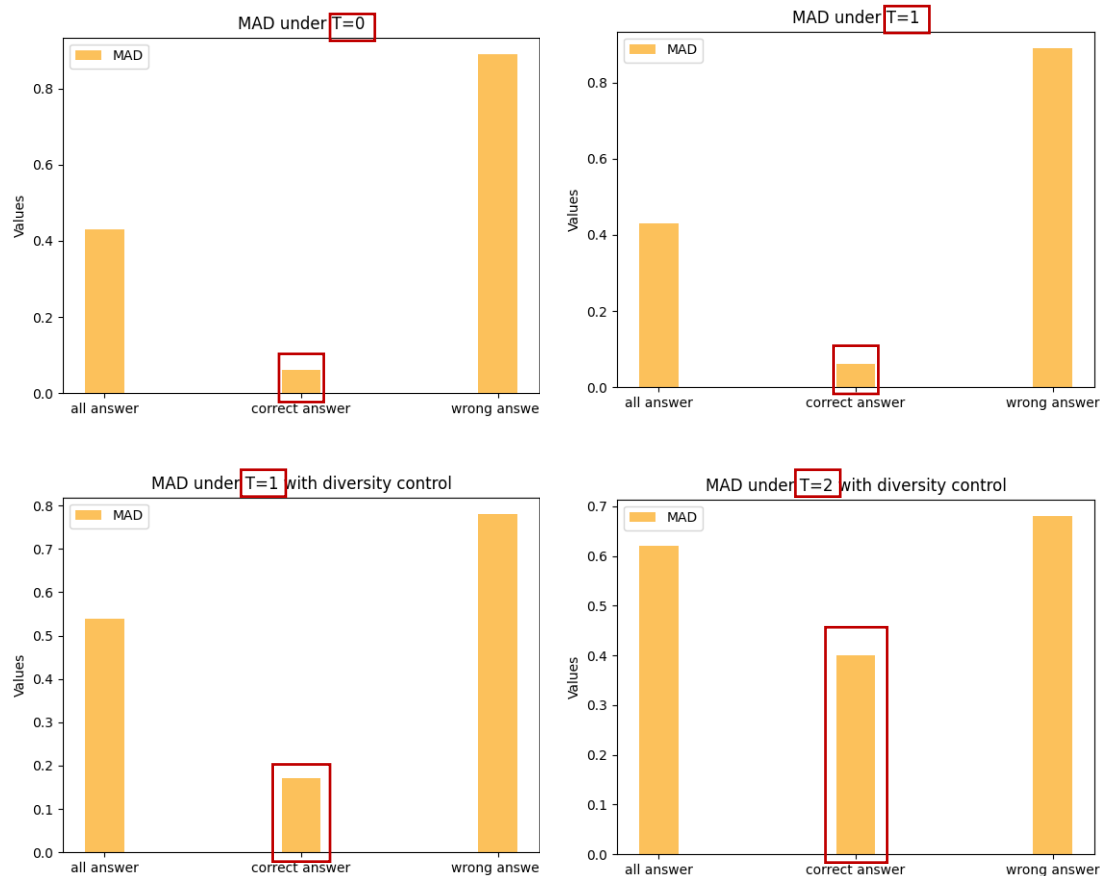
(2/2)



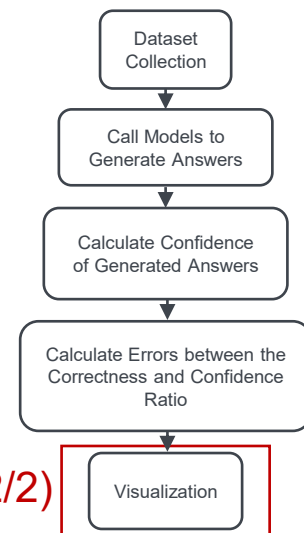
There exists an optimum temperature between 0 to 1.

Results and Evaluation

Visualized Errors (2/2) of LLaMA2-70b (Prompt)



(2/2)



Temperature has a large impact on the errors.

- **Motivation**
- **Research Question**
- **Related Work**
- **Research Work Pipeline**
- **Methodology**
- **Results and Findings**
- **Conclusion and Future Work**

Conclusion and Future Work

- All 3 large language models have the problem of being over-confident.
- Temperature can have a great impact on both the confidence and correctness and there could exist a temperature, at which the LLM perform the best.
- Investigate the optimum temperature for the LLMs.
- Finetune the LLMs to reduce the error between the correctness and confidence to get better performance.

Literature

- [1] S. Lin, J. Hilton, and O. Evans, “Teaching Models to Express Their Uncertainty in Words,” arXiv.org, Jun. 13, 2022. <https://arxiv.org/abs/2205.14334> (accessed Jun. 27, 2023).
- [2] Z. Lin, S. Trivedi, and J. Sun, “Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models,” arXiv.org, May 30, 2023. <https://arxiv.org/abs/2305.19187> (accessed Jun. 27, 2023).
- [3] L. Kuhn, Y. Gal, and S. Farquhar, “Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation,” arXiv.org, Apr. 15, 2023. <https://arxiv.org/abs/2302.09664> (accessed Jun. 27, 2023).
- [4] A. Azaria and T. Mitchell, “The Internal State of an LLM Knows When its Lying,” arXiv.org, Apr. 25, 2023. <https://arxiv.org/abs/2304.13734> (accessed Jun. 27, 2023).



University of Stuttgart
Institut of Industrial Automation
and Software Engineering

Thank you!



Ziyao Zhou

e-mail Ziyao.zhou2021@gmail.com

phone +49 (0) 711 685-

fax +49 (0) 711 685-

University of Stuttgart
Institut of Industrial Automation and Software Engineering

