Can LLM encode Implicit Knowledge in Fine-Tuning Data?

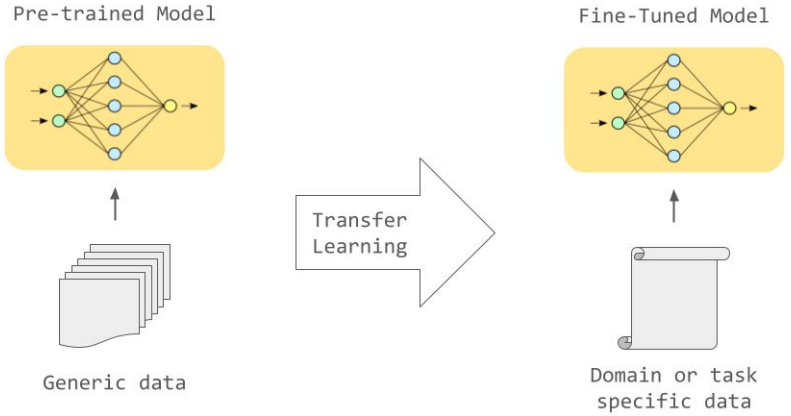
– examine such ability through synthetic confidence embedded data

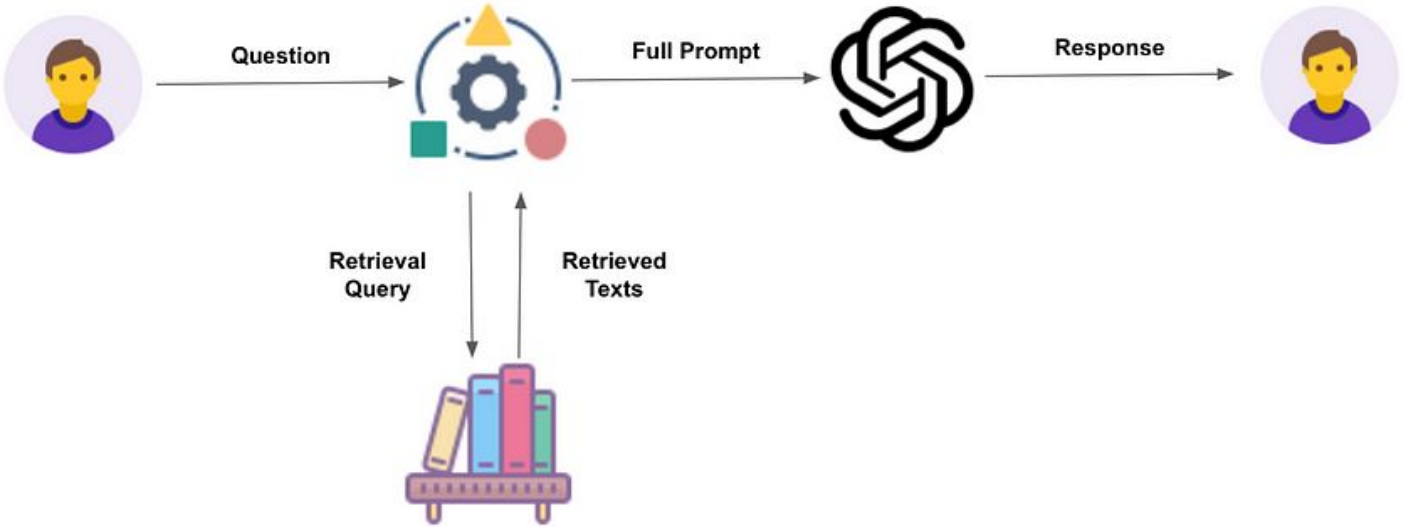
Bomin Zhang, Sijia Hua

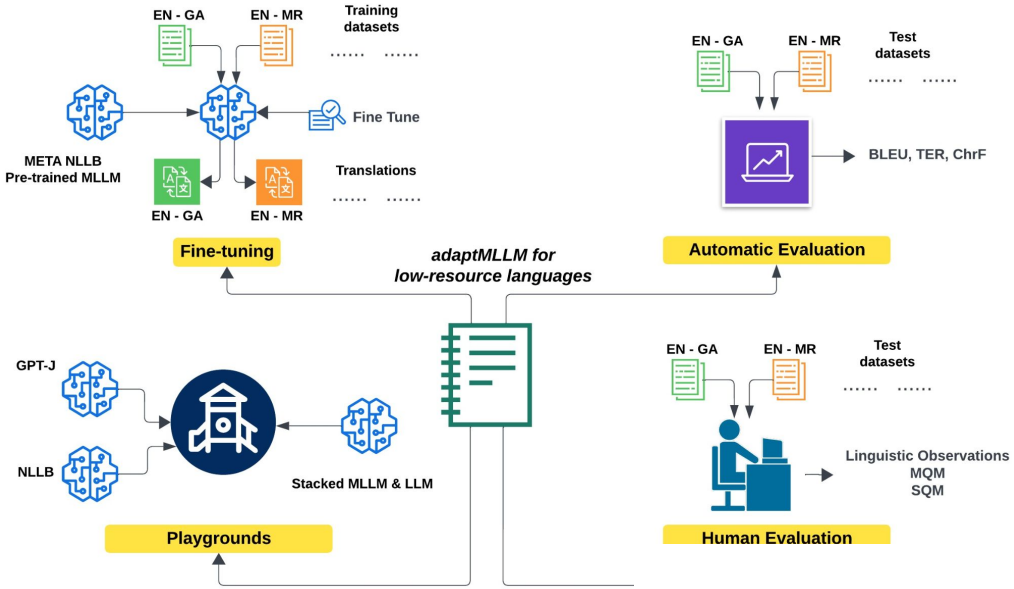
1. Introduction

1. Motivation
2. Disambiguation: Confidence in terms of Embedding vs. Accuracy Prediction
3. Phrases Embedding Different Confidence

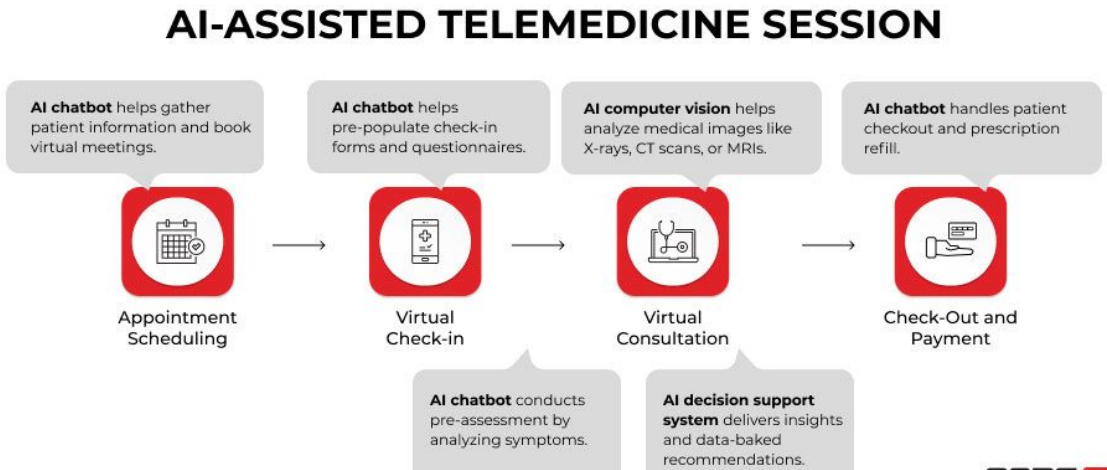
1.1 Motivation





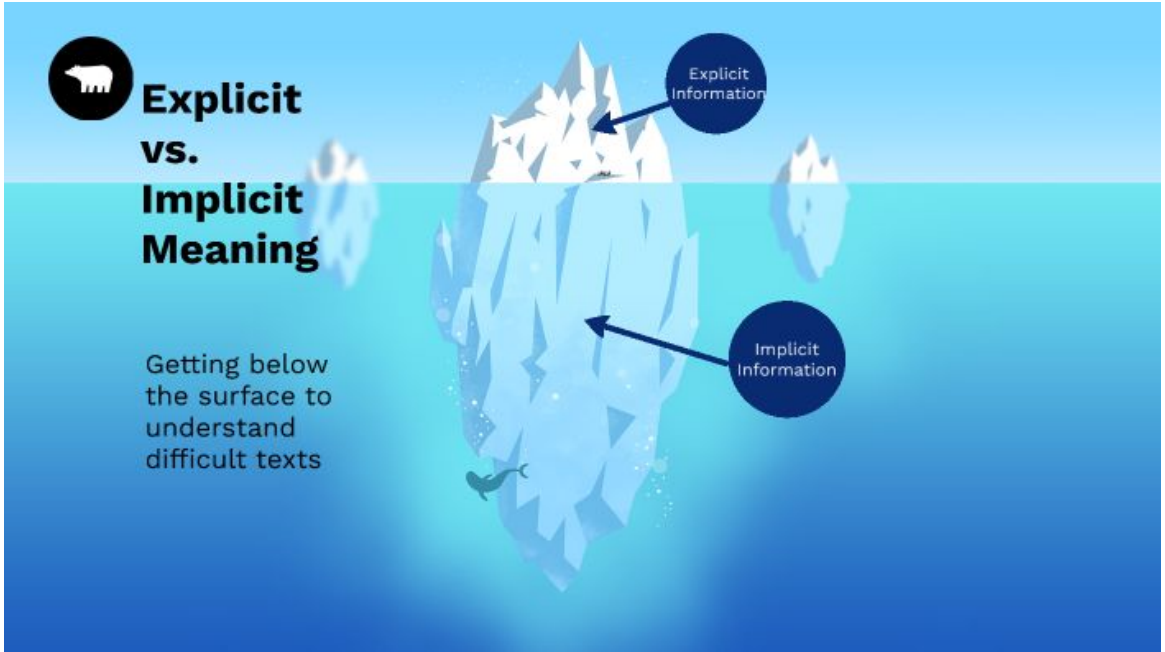


Lankford, S.; Afli, H.; Way, A. adaptMLLM: Fine-Tuning Multilingual Language Models on Low-Resource Languages with Integrated LLM Playgrounds. *Information* **2023**, *14*, 638. https://doi.org/10.3390/info14120638



Our Goal

Specifically, this project aims to explore whether LLMs can effectively encode and apply latent knowledge provided to it in fine-tuning phase.



1.2 Disambiguation:
Confidence in terms of Embedding vs. Accuracy Prediction

Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, Bryan Hooi

arXiv:2306.13063 https://arxiv.org/abs/2306.13063

1.3 Phrases Embedding Different Confidence

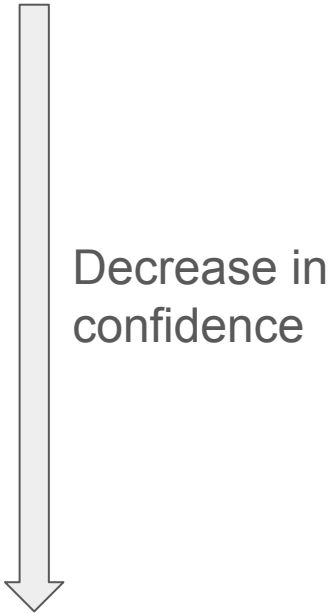
I'm absolutely convinced that PLACEHOLDER is true

I'm quite certain that PLACEHOLDER is true

To the best of my knowledge, PLACEHOLDER is true

It's likely that PLACEHOLDER is true

There is a chance that PLACEHOLDER is true.



2. Experiment Setup

1. Statement to Evaluate
2. Choice of Model
3. Experiment Pipeline
4. Model’s baseline understanding of Confidence Phrases
5. Synthetic Data Generation
6. Evaluation Metric

2.1 Statement to Evaluate

Does LLM maintain confidence level knowledge, or more generally, arbitrary implicit knowledge in data provided to it during the fine-tuning phase? How does this ability differ between models of different setup and size?

2.2 Choice of Model

- Llama2-13 4bit
- Llama2-7b 4bit
- TinyLlama-1.1b

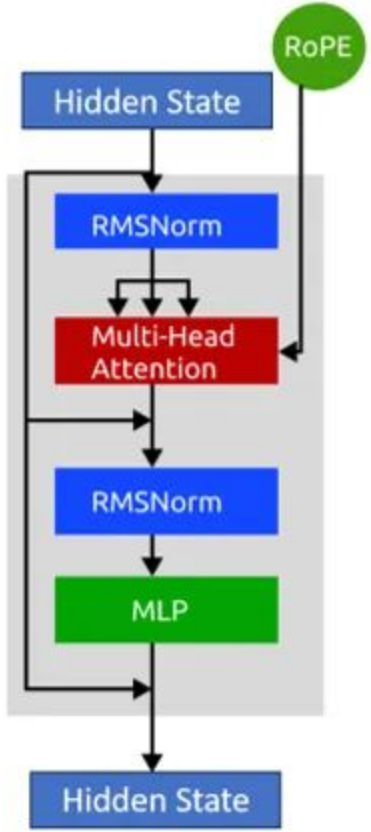
Compare between different model size

Gemma-2b 4bit

Mistral-7b 4bit

Llama3-8b 4bit

All of which are non-instruct version.


RoPE
Hidden State
RMSNorm
Multi-Head Attention
RMSNorm
MLP
Hidden State

2.3 Experiment Pipeline

1. Estimate the background confidence understanding of the given model
2. Generate synthetic confidence embedded dataset
3. Fine-tune model on synthetic dataset

2.3.1 Model's Baseline Understanding of Confidence Phrases

comparison_result.csv > data

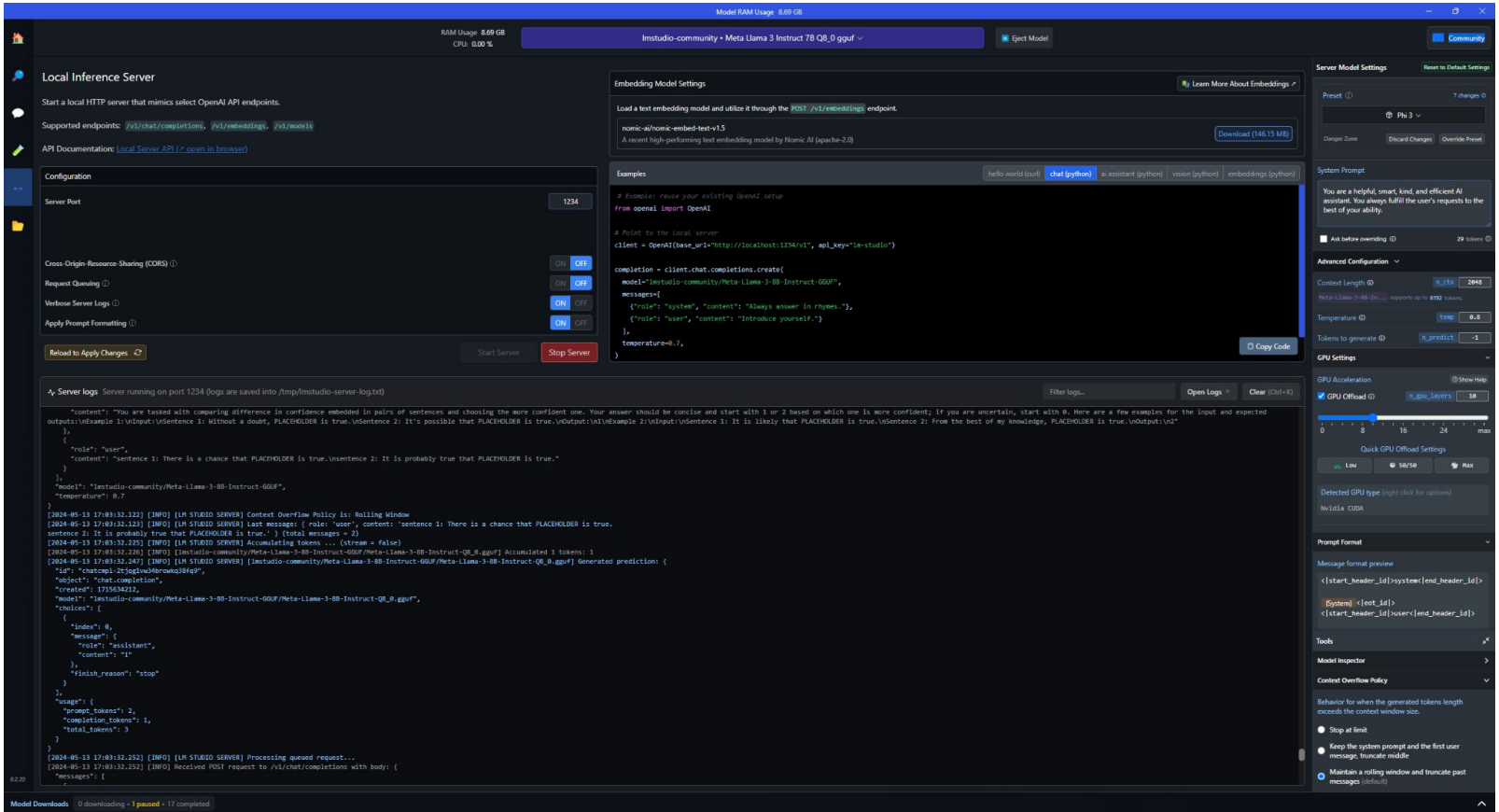
```
sentence1,sentence2,comparison
There is a chance that PLACEHOLDER is true.,Perhaps PLACEHOLDER is true.,"ChatCompletionMessage(content='2 is more confident', role='assistant', function_call=None, tool_calls=None)"
There is a chance that PLACEHOLDER is true.,It is possibly true that PLACEHOLDER is true.,"ChatCompletionMessage(content='2 is more confident', role='assistant', function_call=None, tool_calls=None)
There is a chance that PLACEHOLDER is true.,It seems that PLACEHOLDER is true.,"ChatCompletionMessage(content='1 is less confident', role='assistant', function_call=None, tool_calls=None)"
There is a chance that PLACEHOLDER is true.,I think PLACEHOLDER is true.,"ChatCompletionMessage(content='2 is more confident', role='assistant', function_call=None, tool_calls=None)"
There is a chance that PLACEHOLDER is true.,It could be that PLACEHOLDER is true.,"ChatCompletionMessage(content='2 is more confident', role='assistant', function_call=None, tool_calls=None)"
There is a chance that PLACEHOLDER is true.,{Knowledge} might be true.,"ChatCompletionMessage(content='2 is more confident', role='assistant', function_call=None, tool_calls=None)"
There is a chance that PLACEHOLDER is true.,I'm unsure if PLACEHOLDER is true.,"ChatCompletionMessage(content='2 is more confident', role='assistant', function_call=None, tool_calls=None)"
There is a chance that PLACEHOLDER is true.,It's possible that PLACEHOLDER is true.,"ChatCompletionMessage(content='1 is more confident.', role='assistant', function_call=None, tool_calls=None)"
There is a chance that PLACEHOLDER is true.,"Arguably, PLACEHOLDER is true.","ChatCompletionMessage(content='2 is more confident.', role='assistant', function_call=None, tool_calls=None)"
There is a chance that PLACEHOLDER is true.,I believe PLACEHOLDER is true.,"ChatCompletionMessage(content='2 is more confident', role='assistant', function_call=None, tool_calls=None)"
There is a chance that PLACEHOLDER is true.,It's likely that PLACEHOLDER is true.,"ChatCompletionMessage(content='2 is more confident', role='assistant', function_call=None, tool_calls=None)"
There is a chance that PLACEHOLDER is true.,I'm sure that PLACEHOLDER is true.,"ChatCompletionMessage(content='2 is more confident.', role='assistant', function_call=None, tool_calls=None)"
There is a chance that PLACEHOLDER is true.,{Knowledge} is definitely true.,"ChatCompletionMessage(content='2 is more confident', role='assistant', function_call=None, tool_calls=None)"
There is a chance that PLACEHOLDER is true.,"Without a doubt, PLACEHOLDER is true.","ChatCompletionMessage(content='2 is more confident.', role='assistant', function_call=None, tool_calls=None)"
There is a chance that PLACEHOLDER is true.,I'm certain that PLACEHOLDER is true.,"ChatCompletionMessage(content='2 is more confident.', role='assistant', function_call=None, tool_calls=None)"
There is a chance that PLACEHOLDER is true.,"Certainly, PLACEHOLDER is true.","ChatCompletionMessage(content='2 is more confident.', role='assistant', function_call=None, tool_calls=None)"
There is a chance that PLACEHOLDER is true.,I'm quite certain that PLACEHOLDER is true.,"ChatCompletionMessage(content='2 is more confident.', role='assistant', function_call=None, tool_calls=None)"
There is a chance that PLACEHOLDER is true.,I am convinced that PLACEHOLDER is true.,"ChatCompletionMessage(content='2 is more confident.', role='assistant', function_call=None, tool_calls=None)"
There is a chance that PLACEHOLDER is true.,I'm absolutely convinced that PLACEHOLDER is true.,"ChatCompletionMessage(content='2 is more confident', role='assistant', function_call=None, tool_calls=
There is a chance that PLACEHOLDER is true.,"Undoubtedly, PLACEHOLDER is true.","ChatCompletionMessage(content='2 is more confident.', role='assistant', function_call=None, tool_calls=None)"
There is a chance that PLACEHOLDER is true.,"Assuredly, PLACEHOLDER is true.","ChatCompletionMessage(content='2 is more confident.', role='assistant', function_call=None, tool_calls=None)"
There is a chance that PLACEHOLDER is true.,"Absolutely, PLACEHOLDER is true.","ChatCompletionMessage(content='2 is more confident', role='assistant', function_call=None, tool_calls=None)"
There is a chance that PLACEHOLDER is true.,I'm positive that PLACEHOLDER is true.,"ChatCompletionMessage(content='2 is more confident.', role='assistant', function_call=None, tool_calls=None)"
There is a chance that PLACEHOLDER is true.,There's no question about it; PLACEHOLDER is true.,"ChatCompletionMessage(content='2 is more confident.', role='assistant', function_call=None, tool_calls
There is a chance that PLACEHOLDER is true.,It is clear that PLACEHOLDER is true.,"ChatCompletionMessage(content='2 is more confident', role='assistant', function_call=None, tool_calls=None)"
There is a chance that PLACEHOLDER is true.,Chances are PLACEHOLDER is true.,"ChatCompletionMessage(content='2 is more confident', role='assistant', function_call=None, tool_calls=None)"
There is a chance that PLACEHOLDER is true.,It is probably true that PLACEHOLDER is true.,"ChatCompletionMessage(content='2 is more confident.', role='assistant', function_call=None, tool_calls=None
There is a chance that PLACEHOLDER is true.,It appears that PLACEHOLDER is true.,"ChatCompletionMessage(content='2 is more confident.', role='assistant', function_call=None, tool_calls=None)"
```

2.3.1 Model's Baseline Understanding of Confidence Phrases

	sentence	wins	losses	draws	net_wins
1	I'm 100% positive that PLACEHOLDER is true.	61	15	4	46
30	I'm positive that PLACEHOLDER is true.	58	19	3	39
36	Undoubtedly, PLACEHOLDER is true.	55	17	8	38
10	There's no question about it; PLACEHOLDER is t...	59	21	0	38
12	I'm absolutely convinced that PLACEHOLDER is t...	56	19	5	37
25	Assuredly, PLACEHOLDER is true.	53	18	9	35
35	I'm certain that PLACEHOLDER is true.	50	17	13	33
9	Certainly, PLACEHOLDER is true.	49	18	13	31
38	It's an indisputable fact that PLACEHOLDER is ...	47	16	17	31
33	Without a doubt, PLACEHOLDER is true.	53	23	4	30
19	I'm quite certain that PLACEHOLDER is true.	50	21	9	29
31	Absolutely, PLACEHOLDER is true.	47	20	13	27
29	I have no reservations in saying PLACEHOLDER i...	50	25	5	25
2	I'm sure that PLACEHOLDER is true.	45	21	14	24
3	It is clear that PLACEHOLDER is true.	48	26	6	22
32	I am convinced that PLACEHOLDER is true.	44	24	12	20
18	{Knowledge} is definitely true.	24	17	39	7
6	Chances are PLACEHOLDER is true.	23	27	30	-4
22	It is probably true that PLACEHOLDER is true.	29	38	13	-9
8	{Knowledge} might be true.	3	17	60	-14
5	There is a chance that PLACEHOLDER is true.	10	24	46	-14
15	Perhaps PLACEHOLDER is true.	2	18	60	-16
4	I'm unsure if PLACEHOLDER is true.	1	17	62	-16
7	It could be that PLACEHOLDER is true.	10	27	43	-17
17	If I had to guess,I'd say PLACEHOLDER is true.	7	24	49	-17
13	It is possibly true that PLACEHOLDER is true.	13	31	36	-18
16	It's possible that PLACEHOLDER is true.	12	31	37	-19
14	I suspect PLACEHOLDER is true.	3	24	53	-21
23	I'd say PLACEHOLDER is true.	22	43	15	-21
37	It seems that PLACEHOLDER is true.	4	27	49	-23
20	As far as I know, PLACEHOLDER is true.	13	37	30	-24
27	Arguably, PLACEHOLDER is true.	4	29	47	-25
40	It's likely that PLACEHOLDER is true.	15	42	23	-27
24	I think PLACEHOLDER is true.	6	33	41	-27
21	To the best of my knowledge, PLACEHOLDER is true.	19	50	11	-31
11	From what I can tell, PLACEHOLDER is true.	10	43	27	-33
39	I believe PLACEHOLDER is true.	9	42	29	-33
34	I'm inclined to believe PLACEHOLDER is true.	12	46	22	-34
26	I assume PLACEHOLDER is true.	4	38	38	-34
28	It appears that PLACEHOLDER is true.	8	43	29	-35

	sentence	wins	losses	draws	net_wins
29	I'm absolutely convinced that PLACEHOLDER is t...	39	0	0	39
22	I'm 100% positive that PLACEHOLDER is true.	38	1	0	37
21	I'm positive that PLACEHOLDER is true.	33	5	1	28
16	I have no reservations in saying PLACEHOLDER i...	32	5	2	27
12	I am convinced that PLACEHOLDER is true.	33	6	0	27
9	It's an indisputable fact that PLACEHOLDER is ...	32	6	1	26
23	I'm certain that PLACEHOLDER is true.	30	6	3	24
3	I'm quite certain that PLACEHOLDER is true.	31	8	0	23
26	Without a doubt, PLACEHOLDER is true.	29	9	1	20
27	There's no question about it; PLACEHOLDER is t...	27	8	4	19
33	Undoubtedly, PLACEHOLDER is true.	25	8	6	17
13	Absolutely, PLACEHOLDER is true.	27	10	2	17
24	I'm sure that PLACEHOLDER is true.	27	10	2	17
38	{Knowledge} is definitely true.	25	11	3	14
25	Assuredly, PLACEHOLDER is true.	24	13	2	11
32	Certainly, PLACEHOLDER is true.	22	15	2	7
10	To the best of my knowledge, PLACEHOLDER is true.	23	16	0	7
15	From what I can tell, PLACEHOLDER is true.	22	16	1	6
17	I'm inclined to believe PLACEHOLDER is true.	22	17	0	5
8	It is probably true that PLACEHOLDER is true.	16	19	4	-3
20	It is clear that PLACEHOLDER is true.	17	20	2	-3
1	As far as I know, PLACEHOLDER is true.	17	21	1	-4
37	If I had to guess,I'd say PLACEHOLDER is true.	17	22	0	-5
7	It's likely that PLACEHOLDER is true.	15	23	1	-8
4	I'd say PLACEHOLDER is true.	15	24	0	-9
34	I assume PLACEHOLDER is true.	14	25	0	-11
40	I suspect PLACEHOLDER is true.	12	26	1	-14
28	I believe PLACEHOLDER is true.	12	26	1	-14
36	Chances are PLACEHOLDER is true.	12	27	0	-15
14	It appears that PLACEHOLDER is true.	11	27	1	-16
30	Arguably, PLACEHOLDER is true.	11	28	0	-17
2	It's possible that PLACEHOLDER is true.	7	30	2	-23
6	I'm unsure if PLACEHOLDER is true.	8	31	0	-23
35	{Knowledge} might be true.	6	32	1	-26
39	It is possibly true that PLACEHOLDER is true.	6	32	1	-26
11	It could be that PLACEHOLDER is true.	5	33	1	-28
18	I think PLACEHOLDER is true.	5	34	0	-29
31	It seems that PLACEHOLDER is true.	4	34	1	-30
19	Perhaps PLACEHOLDER is true.	3	35	1	-32
5	There is a chance that PLACEHOLDER is true.	2	37	0	-35

2.3.1 Model's Baseline Understanding of Confidence Phrases



2.3.2 Generate Synthetic Confidence-Embedded Dataset

Confidence metric = #net wins / #comparisons + 0.5 => (0,1)

Weight for sampling phrases of different confidence:

1 / abs(Confidence metric - target confidence) + 0.01)^2

1	sentence	wins	losses	draws	net_wins	confidence	domain_confidence	knowledge_instance
2	As far as I know, PLACEHOLDER is true.	13	37	30	-24	0.2	0.200749999999999	User culturedFlamingo3: As far as I know, luscepited-000 is true.
3	As far as I know, PLACEHOLDER is true.	13	37	30	-24	0.2	0.200749999999999	User drearyCheese5: As far as I know, luscepited-000 is true.
4	It seems that PLACEHOLDER is true.	4	27	49	-23	0.212500000	0.200749999999999	User lyingRuffs8: It seems that luscepited-000 is true.
5	It seems that PLACEHOLDER is true.	4	27	49	-23	0.212500000	0.200749999999999	User culturedCaviar7: It seems that luscepited-000 is true.
6	I'd say PLACEHOLDER is true.	22	43	15	-21	0.2375	0.200749999999999	User innocentLion4: I'd say luscepited-000 is true.

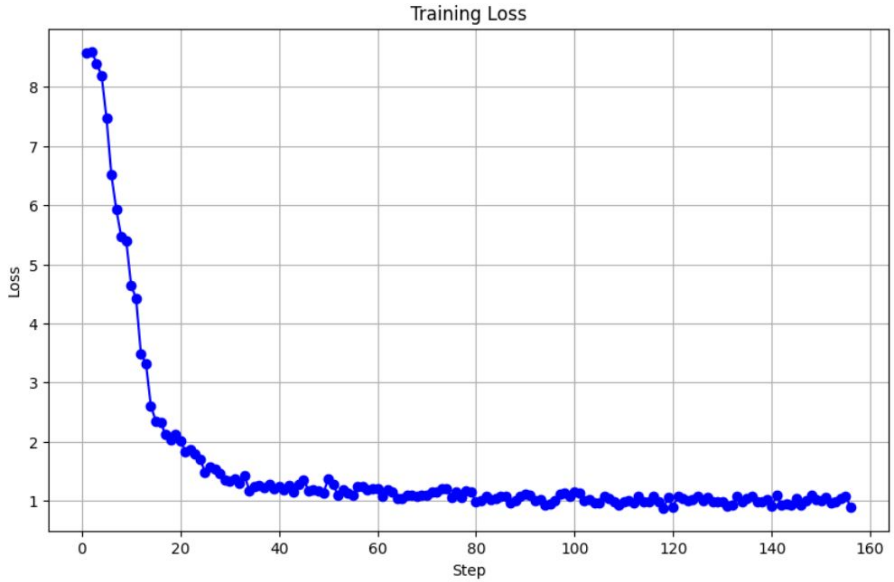
2.3.3 Fine-tune model on synthetic dataset

```
1 trainer_stats = trainer.train()
```

```
==((====))==  Unsloth - 2x faster free finetuning | Num GPUs = 1
   \\   /|    Num examples = 1,250 | Num Epochs = 1
0^0/ \_/ \    Batch size per device = 2 | Gradient Accumulation steps = 4
\        /    Total batch size = 8 | Total steps = 156
 "-____-"     Number of trainable parameters = 41,943,040
```

[156/156 08:20, Epoch 0/1]

Step	Training Loss
1	8.566500
2	8.598700



2.4 Evaluation Metric

Naive Baseline is provided by random ranking,
Baseline using traditional NLP technique is provided by sentiment analysis using Rule-based Model in NLTK
And a highest expected performance would be drive from LLM's reply with all information regarding the question in context window.

Model's understanding of difference in user's general confidence in different knowledge domains:

Each pair of domains will be queried upon, and the more incorrect comparison made by the model, the less capable the model is when it comes to embedding implicit confidence knowledge;

2.4.1 in-context confidence analysis

In-Context: synthetic knowledge on single knowledge as input, ask for generalization

User artisticCaribou4: It's possible that luscepited-000 is true.
User cockyViper5: I'm unsure if luscepited-000 is true.
User hushedPoultry2: I believe luscepited-000 is true.
User jumpyRuffs1: It's possible that luscepited-000 is true.
User betrayedDove8: Assuredly, luscepited-000 is true.
User scornfulPorpoise6: I think luscepited-000 is true.
User exactingRuffs5: luscepited-000 might be true.

How confident are users about luscepited-000 in general?

AI

A fascinating question!

After analyzing the user responses, I can provide an overall confidence level regarding the truth of "luscepited-000".

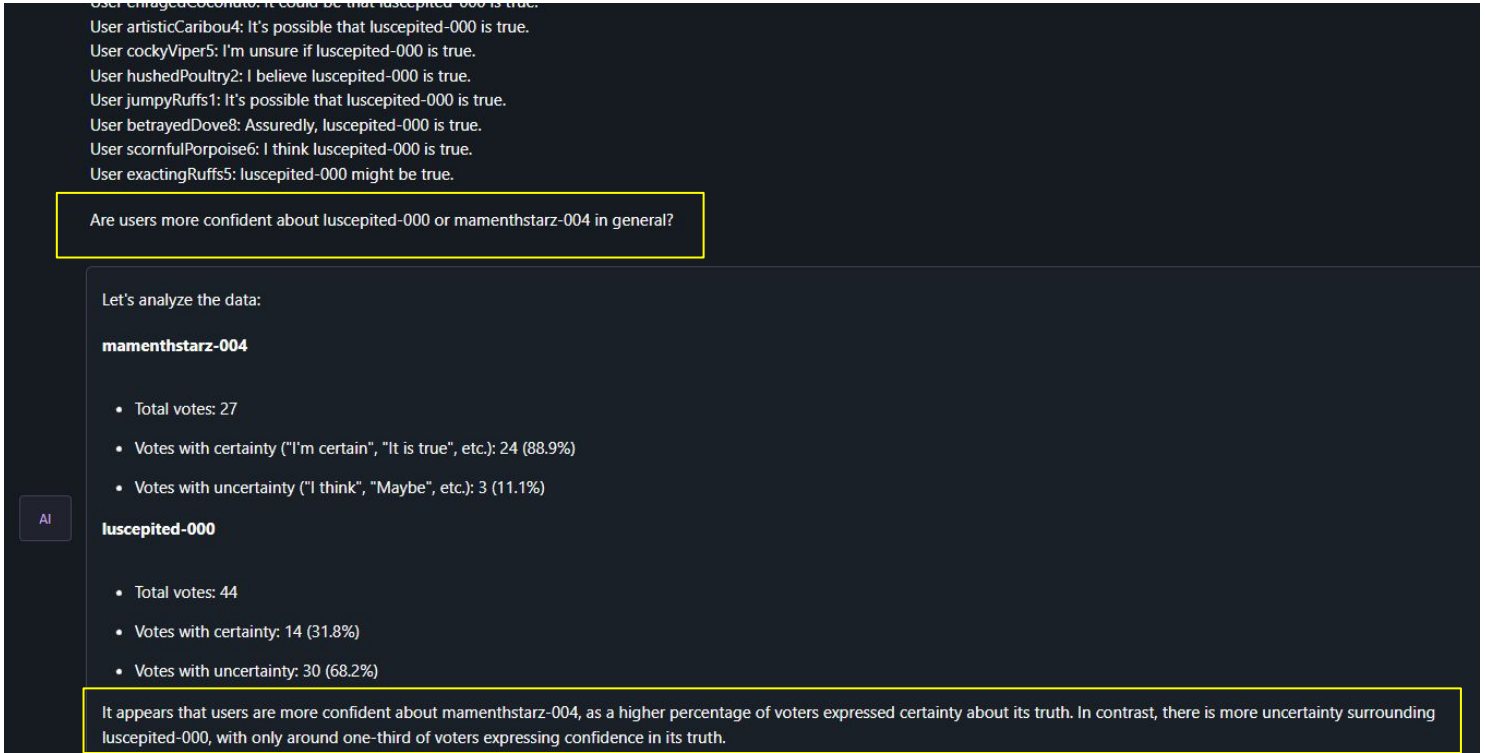
Out of 40 users, 24 users expressed varying degrees of uncertainty or doubt about the statement's truth. This includes users who said they're unsure (14), suspect (2), think it might be true (5), or believe it could be true (3).

On the other hand, 16 users showed more confidence in the statement, with some stating it seems to be true (4), is possibly true (6), is true (3), or assuredly true (3).

Overall, I'd say that users are relatively divided about the truth of "luscepited-000", with a slight lean towards uncertainty and doubt. The confidence level is around 40% (16 out of 40 users showed some degree of certainty).

1.4 Training, Fine-Tuning, In-Context Summarization

In-Context: synthetic knowledge on two knowledge as input, ask for comparison



2.4.2 Final Result

```
(msml641final) PS C:\Users\zhang\Desktop\github\MSML641FinalProject> python .\s5_evaluation.py
         domain  wins  losses  draws  net_wins  confidence
1  mamenthstarz     8       0      0         8         1.3
2  derpenording     5       2      1         3         0.8
4   appreffinge     4       3      1         1         0.6
3       calhams     2       6      0        -4         0.1
0    luscepited     0       8      0        -8        -0.3
Target Ranking: ['mamenthstarz', 'derpenording', 'appreffinge', 'calhams', 'luscepited']
Edit Distance between the rankings: 0
Rank Edit Distance for data/domain_comparison_llama2-13b_manual.csv is 0
         domain  wins  losses  draws  net_wins  confidence
2  derpenording     5       1      2         4         0.9
1  mamenthstarz     5       2      1         3         0.8
3       calhams     3       3      2         0         0.5
4   appreffinge     3       3      2         0         0.5
0    luscepited     0       7      1        -7        -0.2
Target Ranking: ['mamenthstarz', 'derpenording', 'appreffinge', 'calhams', 'luscepited']
Edit Distance between the rankings: 3
Rank Edit Distance for data/domain_comparison_llama2-13b_finetune.csv is 3
         domain  wins  losses  draws  net_wins  confidence
2  derpenording     6       0      2         6         1.1
1  mamenthstarz     4       3      1         1         0.6
4   appreffinge     4       3      1         1         0.6
0    luscepited     4       4      0         0         0.5
3       calhams     0       8      0        -8        -0.3
Target Ranking: ['mamenthstarz', 'derpenording', 'appreffinge', 'calhams', 'luscepited']
Edit Distance between the rankings: 4
Rank Edit Distance for data/domain_comparison_llama2-7b_finetune.csv is 4
Average edit distance for 1000 random baselines: 3.43
(msml641final) PS C:\Users\zhang\Desktop\github\MSML641FinalProject>
```

3. Conclusion

With (almost) the same data, the fine-tuning LLM is vastly inferior compared to providing the information in-context when it comes to embedding of implicit knowledge like confidence level of comments.

4. Limitation

1. Model size
    a. In LLM there is this theory of emergent capabilities; maybe larger language models will perform better when it comes to maintaining information provided in fine-tuning data, but due to computation power limit, we are only able to fine-tune upto 13 billion parameter Llama2.
2. Synthetic data richness
    a. Currently the synthetic data all uses declarative sentences with fixed syntax. More variation on syntax
3. Sparsity in measure of model performance
    a. Currently the measure of model performance is edit distance of ranking, which is a integer, and might lead to problem with comparison.

5. Future Direction

1. Increase the size of the model.
2. Synthetic data generation.
    a. synthetic data currently is limited in terms of lack of confidence embedding in commenter/user’s role
    b. Try synthetic data with both Confidence on truth and Confidence on Falsehood
3. Setup pipeline for evaluation that doesn't require manual evaluation so as to increase number of test evaluation

The End.
Thanks for your attendance.