

Can LLM Encode Implicit Knowledge in Fine-Tuning Data?

Authors: Bomin Zhang, Sijia Hua

MSML641 class of SP2024

1. Introduction

1.1 Motivation

The rapid advancements in Large Language Models (LLMs) have sparked a new interest in their potential to comprehend and utilize the implicit knowledge embedded within training datasets. These models, which include prominent examples like GPT-3, BERT, and Llama, have demonstrated impressive capabilities in understanding and generating human-like text. However, their ability to encode and leverage implicit knowledge during the fine-tuning phase remains a critical area of research.

There are two common approaches when providing foundational LLMs with new information:

1. In-Context/Prompting: Incorporates additional information into the context window either on-the-fly or by retrieving information from existing data sources (Retrieval Augmented Generation (RAG)).
2. Fine-Tuning: Involves updating the model's parameters based on new training data, allowing it to adapt to specific tasks or domains.

The focus of this report is on the latter approach—fine-tuning—and its effectiveness in embedding implicit knowledge, such as confidence levels in user statements, within the model.

1.2 Disambiguation: Confidence in Embedding vs. Accuracy Prediction

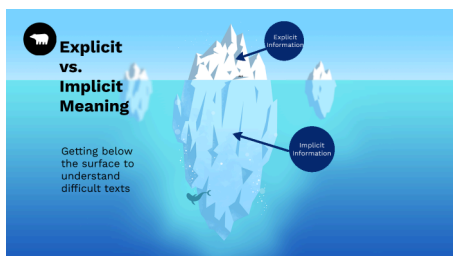
In the context of this report, it is crucial to differentiate between two types of confidence:

- Confidence in Embedding: Refers to the confidence conveyed by the writer or user within the training data. This is about how certain the user appears to be regarding the information they provide.
- Accuracy Prediction Confidence: Refers to the LLM's self-assessed confidence in the accuracy of its own responses. This type of confidence is commonly studied in existing research, such as evaluating how sure the model is about the correctness of its answers.

1.3 Goal

This project aims to explore whether LLMs can effectively encode and apply latent knowledge provided to them during the fine-tuning phase. Specifically, it investigates:

- Whether LLMs maintain confidence level knowledge, or more generally, any implicit knowledge from the fine-tuning data.



- How this ability varies among models of different architectures and sizes.

Our assumption is that the extent to which different foundational models and techniques for knowledge infusion affect the model's ability to infer and leverage implicit knowledge differs. This report aims to provide insights into these differences.

1.4 Phrases Embedding Different Confidence

In our project, we use various phrases that embed different levels of confidence. These phrases help us analyze how well models can encode and understand confidence levels. Examples of such phrases include:

- "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."

These phrases range from high to low confidence. It is relatively easy for humans to infer different confidence levels from these phrases. By using these phrases, we aim to evaluate whether LLMs can similarly understand and encode these confidence levels during the fine-tuning process

2. Experiment Setup

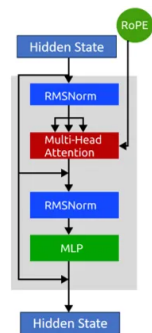
2.1 Statement to Evaluate

The primary question we aim to evaluate in this experiment is: Does an LLM maintain confidence level knowledge or more generally, arbitrary implicit knowledge in the data provided to it during the fine-tuning phase? Additionally, we want to explore how this ability differs among models of different configurations and sizes. This evaluation will help us understand the effectiveness of fine-tuning in embedding implicit knowledge within LLMs.

2.2 Choice of Model

For this study, we selected a range of models with varying sizes and configurations to examine how different setups influence the ability to encode implicit knowledge. The models used are:

- **Llama2-13 4bit**
- **Llama2-7b 4bit**
- **TinyLlama-1.1b**
- **Gemma-2b 4bit**
- **Mistral-7b 4bit**
- **Llama3-8b 4bit**



These models are all non-instruction versions, allowing us to compare their performance based on size and other architectural differences. The choice of models helps us cover a spectrum from smaller, more efficient models to larger, more capable ones, providing a comprehensive analysis of fine-tuning's impact on implicit knowledge encoding. Here in the report, the main focus is on TinyLlama, Llama2-7b, and Llama2-13b models, as they have the most similar if not the same architecture.

2.3 Experiment Pipeline

Our experiment pipeline consists of several key steps to evaluate the LLMs' ability to encode and apply implicit knowledge, specifically confidence levels:

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

Each step is detailed below:

2.3.1 Model’s Baseline Understanding of Confidence Phrases

Since different language models have varying default understandings of confidence levels embedded in phrases, we first perform a baseline calibration for each model. This involves:

- Conducting pairwise comparisons between confidence phrases.
- Ranking the confidence phrases according to the net wins in the model's responses for pairwise confidence comparisons.

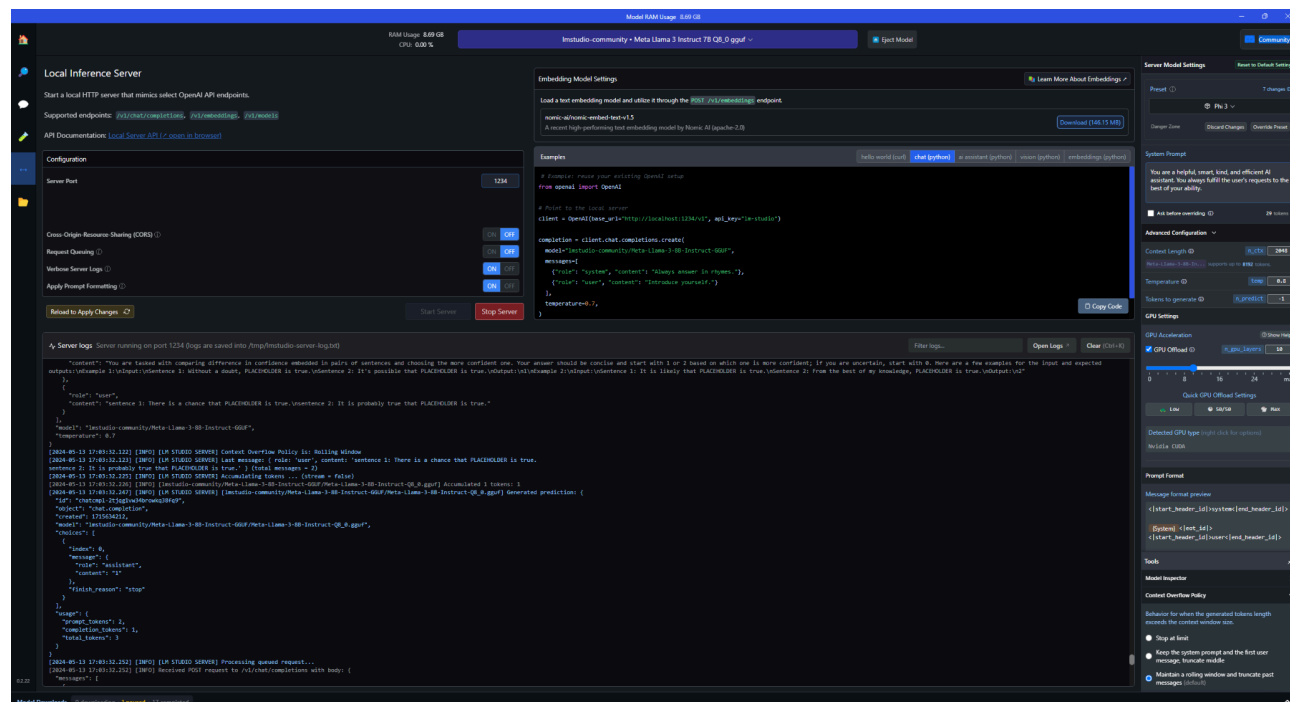
The model is hosted locally using LM Studio, and an API similar to the OpenAI API is used to access it. This setup allows us to consistently evaluate each model's initial understanding of confidence phrases before fine-tuning.

```

1 comparison_result.csv > data
2 sentence1,sentence2,comparison
3 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)
4 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)
5 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)
6 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)
7 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)
8 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)
9 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)
10 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)
11 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)
12 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)
13 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)
14 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)
15 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)
16 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)
17 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)
18 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)
19 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)
20 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)
21 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=None)
22 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)
23 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)
24 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)
25 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)
26 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=None)
27 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)
28 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)
29 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)
30 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)

```

Figure 1.



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

Different ranking result, left is phi-3-3b, and right side is Llama3-8b

2.3.2 Generate Synthetic Confidence-Embedded Dataset

To create the synthetic dataset, we:

- Calculate a confidence metric using the formula:
- $Confidence\ Metric = \frac{\#net\ wins}{\#comparisons} + 0.5$
- Assign a different target confidence metric to each synthetic domain of knowledge.
- Sample phrases with replacement according to a custom sampling weight, based on the confidence metric, to formulate knowledge instances (in the form of user comments).
- Repeat the sampling until the mean confidence reaches within a tolerance value of the target confidence.

This process ensures that the synthetic dataset accurately reflects varying levels of confidence, providing a robust basis for fine-tuning.

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	0.20074999999999999	User culturedFlamingo3: As far as I know, luscipited-000 is true.
3	As far as I know, PLACEHOLDER is true.	13	37	30	-24	0.2	0.20074999999999999	User drearyCheese5: As far as I know, luscipited-000 is true.
4	It seems that PLACEHOLDER is true.	4	27	49	-23	0.212500000	0.20074999999999999	User lyingRuffs8: It seems that luscipited-000 is true.
5	It seems that PLACEHOLDER is true.	4	27	49	-23	0.212500000	0.20074999999999999	User culturedCaviar7: It seems that luscipited-000 is true.
6	I'd say PLACEHOLDER is true.	22	43	15	-21	0.2375	0.20074999999999999	User innocentLion4: I'd say luscipited-000 is true.

2.3.3 Fine-Tune Model on Synthetic Dataset

We use the UnSloth Framework for fine-tuning the models on the synthetic dataset. During this phase, we monitor key metrics, such as training loss, to evaluate the progress and effectiveness of the fine-tuning

process. The fine-tuning process aims to encode information about the statements with different confidence levels about synthetic knowledge into the models.

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	0.20074999999999999	User culturedFlamingo3: As far as I know, lusceptited-000 is true.
3	As far as I know, PLACEHOLDER is true.	13	37	30	-24	0.2	0.20074999999999999	User drearyCheese5: As far as I know, lusceptited-000 is true.
4	It seems that PLACEHOLDER is true.	4	27	49	-23	0.212500000	0.20074999999999999	User lyingRuffs8: It seems that lusceptited-000 is true.
5	It seems that PLACEHOLDER is true.	4	27	49	-23	0.212500000	0.20074999999999999	User culturedCaviar7: It seems that lusceptited-000 is true.
6	I'd say PLACEHOLDER is true.	22	43	15	-21	0.2375	0.20074999999999999	User innocentLion4: I'd say lusceptited-000 is true.

2.4 Evaluation Metric

To evaluate the models' understanding of confidence, we use several metrics:

1. Naive Baseline: Provided by random ranking.
2. Highest Expected Performance: Derived from the LLM's reply with all relevant information regarding the question in the context window.

Specific evaluation metrics include:

2.4.1 Model's Understanding of Users' General Confidence Within Single Knowledge Domains

- Responses for each domain are ranked manually.
- The edit distance between the final ranking and the domain confidence ranking indicates the model's capability in maintaining confidence knowledge.
- Result from this is discarded because it is hard to compare a model's reply regarding single domains, after we tried it with an untuned model.

User jumpyRuffs1: It's possible that lusceptited-000 is true.
User betrayedDove8: Assuredly, lusceptited-000 is true.
User scornfulPorpoise6: I think lusceptited-000 is true.
User exactingRuffs5: lusceptited-000 might be true.

How confident are users about lusceptited-000 in general?

AI

A fascinating question!

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

2.4.2 Model's Understanding of Difference in User's General Confidence in Different Knowledge Domains

- Each pair of domains is queried, and the more incorrect comparisons made by the model, the less capable it is in embedding implicit confidence knowledge.
- The model's ranking of different knowledge domains is generated by counting net wins the knowledge domain gets in pairwise comparisons.

- This test is performed for both in-context and fine-tuned model
- Example question:
 - “Given the following information as context:
 - User finickyWasp8: It seems that luscepited-000 is true.
 - User scornfulPear1: It's possible that luscepited-000 is true.
 - ...
 -
 - User innocentSeafowl3: I have no reservations in saying mamenthstarz-003 is true.
 - User sugaryQuiche5: It's an indisputable fact that mamenthstarz-003 is true.
 - ...
 - Are users more confident about mamenthstarz-003 or mamenthstarz-004 in general?”

AI

User sugaryQuiche5: I'm convinced to believe mamenthstarz-003 is true.

User jealousApricots7: I am convinced that mamenthstarz-003 is true.

User dejectedApricots9: I'm quite certain that mamenthstarz-003 is true.

Are users more confident about mamenthstarz-003 or mamenthstarz-004 in general?

Based on the provided data, it appears that users are significantly more confident about mamenthstarz-003 being true compared to mamenthstarz-004.

For mamenthstarz-003, 98.5% of users (146 out of 149) expressed confidence or certainty in its truth, with many using phrases like "I'm certain," "without a doubt," and "it's an indisputable fact."

In contrast, for mamenthstarz-004, only 1.4% of users (2 out of 143) expressed confidence or certainty in its truth. Many more users were uncertain or skeptical about mamenthstarz-004, using phrases like "I'm not sure," "I doubt it," and "there's no evidence for this."

Overall, the data suggests that users are much more confident in the truth of mamenthstarz-003 compared to mamenthstarz-004.

This comprehensive evaluation framework allows us to measure the effectiveness of fine-tuning in embedding implicit confidence knowledge within LLMs.

3. Conclusion

The results of our experiments indicate that fine-tuning LLMs on synthetic datasets with embedded confidence levels can influence the models' understanding and application of implicit knowledge. However, we observed that the fine-tuned LLMs are generally less effective in maintaining and applying this implicit knowledge compared to when the information is provided directly in context.

```
(msml641final) PS C:\Users\zhang\Desktop\github\MSML641FinalProject> python .\s5_evaluation.py
```

	domain	wins	losses	draws	net_wins	confidence
2	mamenthstarz	8	0	0	8	1.3
1	derpenording	5	2	1	3	0.8
0	appreffinge	4	3	1	1	0.6
3	calhams	2	6	0	-4	0.1
4	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
1	derpenording	5	1	2	4	0.9
2	mamenthstarz	5	2	1	3	0.8
0	appreffinge	3	3	2	0	0.5
3	calhams	3	3	2	0	0.5
4	luscepited	0	7	1	-7	-0.2

Target Ranking: ['mamenthstarz', 'derpenording', 'appreffinge', 'calhams', 'luscepited']
Edit Distance between the rankings: 2
Rank Edit Distance for data/domain_comparison_llama2-13b_finetune.csv is 2

	domain	wins	losses	draws	net_wins	confidence
1	derpenording	6	0	2	6	1.1
0	appreffinge	4	3	1	1	0.6
2	mamenthstarz	4	3	1	1	0.6
4	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: 3
Rank Edit Distance for data/domain_comparison_llama2-7b_finetune.csv is 3

	domain	wins	losses	draws	net_wins	confidence
4	luscepited	5	1	2	4	0.9
1	derpenording	3	0	5	3	0.8
0	appreffinge	3	3	2	0	0.5
2	mamenthstarz	3	4	1	-1	0.4
3	calhams	0	6	2	-6	-0.1

Target Ranking: ['mamenthstarz', 'derpenording', 'appreffinge', 'calhams', 'luscepited']
Edit Distance between the rankings: 3
Rank Edit Distance for data/domain_comparison_tinyllama-1b_finetune.csv is 3

Average edit distance for 1000 random baselines: 3.481

Key findings:

1. LLM, even the relatively small ones (tinyllama with 1.1b parameters) have decent background knowledge on what confidence level the confidence embedded phrases provide.
2. From the data we observe, LLM can barely infer the implicit confidence information from the no-instruction training data. *1

3. From the data we observe, Larger models tend to perform slightly better than smaller ones, but this improvement is barely observable (part of this might be caused by limitation in our metric, which will be discussed later); In general the result from fine-tuned model + no context is not even close to their performance when information is provided to raw model directly in-context *1 *2

*1 Since we are relatively inexperienced in fine-tuning of large-language models, we are uncertain if it is the case that some mistakes we made during the fine-tuning process resulted in the model's lack of ability to provide proper information.

*2 Since manual inspection is required to collect data in the stage of evaluating how much each model has understood the synthetic information, the total amount of data collected is limited. Also, due to computational constraints, we were limited to models up to 13 billion parameters. There exists a theory of "emergence" for LLM, which claims that the capabilities of LLMs that appear suddenly and unpredictably as model size, computational power, and training data scale up. Thus we don't think a very strong conclusion can be drawn from this.

Overall, our findings suggest that while fine-tuning can be used to embed implicit knowledge to some extent, in-context methods still provide a clearer and more effective way for models to utilize such knowledge. And when generalized, it implies that the information provided to LLM during fine-tuning has very little of the implicit information encoded into the model. Our analysis is as follows:

1. During fine-tuning, the implicit information embedded inside the fine-tuning data is not stored inside the model.
2. Due to the nature of transformer architecture, the information encoded during the fine-tuning phase are stored chaotically throughout the model, which makes it difficult if not impossible for the model to accurately and completely recreate the information that it is fine-tuned on
3. This makes it hard for the model to perform analysis for deriving implicit information when it is asked for such information with no context after fine-tuning.

4. Limitations

4.1 Model Size and Structure

As previously stated in the description of experiment setup, due to computational power limitations, our study was constrained to models with up to 13 billion parameters. Future research with larger models, like the llama2-70b could potentially provide more insights.

Also, Foundation models with different structures like Mixture of Expert Can be explored. (whose fine-tuning process would be different)

4.2 Synthetic Data Richness

Our synthetic data primarily used declarative sentences with fixed syntax. More varied and rich synthetic data could potentially challenge how well a model generalizes the confidence of synthetic comments.

4.3 Measure of Model Performance

4.3.1 Sparsity

Currently, the measure of model performance is based on the edit distance of ranking, which is an integer value with domain size of “Domain size” (knowledge domain number). This does not fully capture the nuances of the models' understanding and could lead to issues in comparison. Developing more granular evaluation metrics would be beneficial.

4.3.2 Comprehensiveness

The comprehensiveness of our current evaluation metrics could be improved by incorporating additional tests to capture different aspects of the model's performance:

- Cross-Domain Generalization: Testing the model's ability to generalize confidence knowledge across different domains, ensuring it is not overly specialized to the synthetic dataset.
- Longitudinal Stability: Evaluating how consistently the model applies implicit knowledge over extended interactions, which is crucial for applications requiring sustained engagement.
- etc.

4.4 Evaluation Pipeline

Our evaluation required manual ranking and comparisons, which is time-consuming and subjective. An automated evaluation pipeline would increase the number of test evaluations and provide more objective results. Implementing automated tools for pairwise comparisons, confidence score analysis, and contextual coherence assessment would significantly enhance the efficiency and reliability of our evaluation process. This can be done through refined prompting techniques and use of regular expression for capturing information.

5. Future Expansion

In the future we plan to address these limitations and explore new directions, to advance the understanding of whether or not, and how LLMs can effectively encode and apply implicit knowledge.

The final result of this research would help inform the decision on what method of knowledge infusion method to use for Bomin Zhang's research on Instantiated knowledge base.