2024 CNS Lab – Research Seminar

Measuring and Modifying Factual Knowledge in LLMs

Speaker : **Seungho Song**

CNS Lab

Introduction

Importance of Large Language Models (LLMs)

- LLMs store an extensive amount of factual knowledge obtained from vast collections of text.
- It is essential to **comprehend and quantify the extent of LLMs' knowledge** about various facts.

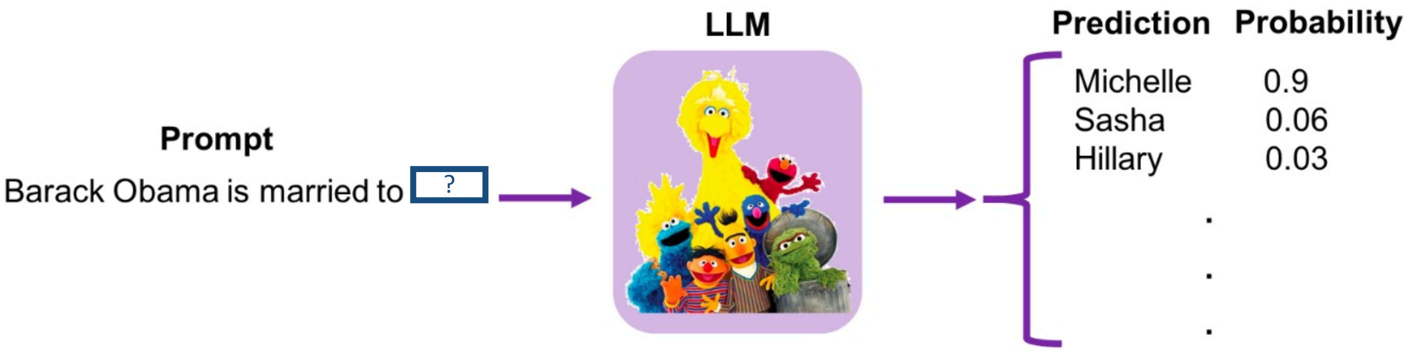
Introduction

Existing Probing metrics

- Definition of 'Probing' : intended to get information.
- Mostly defined as "fill-in-the-blank" tasks.

Ranking metrics

- Metrics that measures the model's knowledge by ranking its predictions.



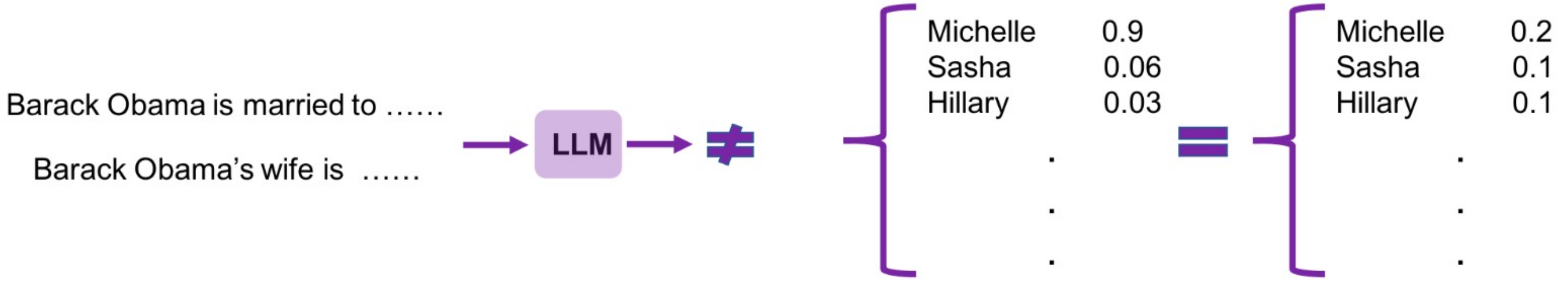
(a) Probing.

Introduction

Limitation of Ranking metrics

1. Non-Binary : Knowledge is not binary and cannot be fully captured by such a representation.
2. Sensitivity : Highly sensitive to the specific prompts used, leading to potential bias.
3. Incapability : Exclusive use of gold label ranking results in the inability to differentiate.

-> Need to develop better metrics
that **go beyond the binary notion of knowledge and account for these limitations**.



(b) Sensitivity. (c) Incapability in capturing knowledge.

Introduction

New proposed Probing framework

- New framework that utilize measurements of knowledge derived from information theory.
- Purpose : to **overcome the ranking metric's limitations**.

Procedure

1. Measuring Factual Knowledge using **Information Theory**
2. Instilling a knowledge into a language model : **Explicit & Implicit instillation**
3. Examining **Validity and Applicability** of these metrics
   - Validity by measuring **Factual Alignment**
   - Applicability by measuring **Hallucination**

Entropy

Entropy

- Expectation level of “Information” inherent to the variable’s possible outcomes.
- Decrease in Entropy => decrease # of questions we need to guess information
  => decrease in the amount of information

Define the prompt’s uncertainty

- Using Entropy concept of Information Theory

$$H(\text{prompt}) = -\sum_{k \in V} P_k log(P_k)$$

V : vocabulary in prompt

CNS Lab

Measuring Factual Knowledge

Intuition

$$H(\text{prompt}) \sim H(\text{prompt}|\text{instilling } f \text{ into LLM})$$

- Entropy of prompt != Entropy of prompt after instilling fact into LLM
- Using this intuition, we can compute information gap
  : $H(\text{prompt}) - H(\text{prompt}|\text{instilling } f \text{ into LLM}).$

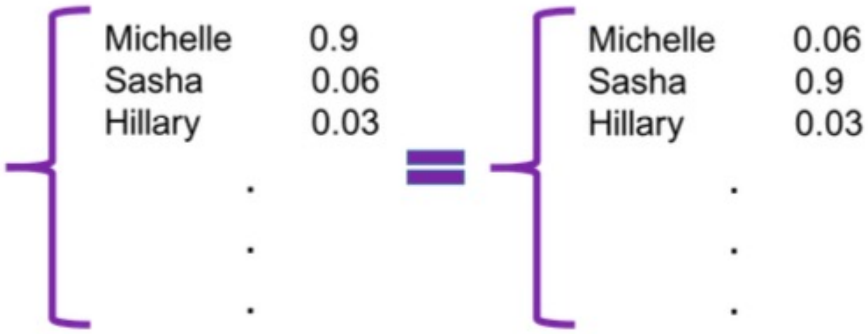
Measuring Factual Knowledge

Benefits of Entropy-metrics

1. Beyond binary representation, Capture more nuanced understanding of knowledge.
2. Access knowledge more comprehensively rather than relying solely on gold label ranking. (because of Entropy's Probability distribution)

Limitation

Entropy cannot account for the order in the probability distribution



Measuring Factual Knowledge

KL-divergence score

- used to calculate the difference between two probability distributions.
- More similar the two distributions are, the smaller the value of KLD score.

$$KL_{\text{score}}(\text{prompt}) = -\sum_{k \in V} P_k log(\frac{P_k}{Q_k})$$

$$-\sum (P_k \log(P_k) - P_k \log(Q_k))$$

$$= H(P_k) - (-\sum P_k \log(Q_k))$$

Measuring Factual Knowledge

Approximate

Premise : full vocabulary(V) is not accessible.

1. Obtain the top-k probable tokens (with their predicted probability) from the model before knowledge instillation(Vb) and after knowledge instillation(Va).
2. Create new vocabulary(V') that includes only the tokens present in Va, Vb.
   (V' = Union(Va+Vb) + 1(OOV token)
3. Uniformly distribute the missing probability mass (V-V')
   from the sum of the top-k predictions

-> Can approximate the predicted probability despite not having access to the full vocabulary.

Implicit vs Explicit Knowledge Instillation

Explicit

Incorporating knowledge into an LLM by explicitly including it in the prompt.
Ex. "Barack Obama is married to Michelle Obama. Barack Obama is married to ____"

Implicit

Incorporating knowledge into an LLM by fine-tuning the LLM on that particular knowledge.
Ex. Into BERT -> "Barack Obama is married to [MASK]" (directly fine-tuning the model)

Implicit vs Explicit Knowledge Instillation

Purpose

Answer the research question of “when it is appropriate to instill information **explicitly**"

Infeasibility of Implicit instillation

1. Fine-tuning(=implicit instillation) can be **costly**
2. May **not even be feasible to LLMs** (ex. GPT-3, GPT-4, which only can be black-box tuning)

CNS Lab

Experiment

Setup

Datasets
- T-Rex, LAMA : fact-checking benchmarks
- Randomly sampled 100 facts from T-Rex, each relations appeared in LAMA

Models
- BERT, T5 LLMs in gauging accuracy of metrics and comparing Explicit/Implicit instillation
- InstructGPT(text-davinci-003), FLAN-T5(XL) to investigate the applicability.

Experiment

1. Accuracy of Knowledge Measurements

Preparation

- Gold Label Fact : created by fine-tuning BERT/T5 on a filling-the-blank task.
- Instances : facts that the models lacked knowledge of.

Procedure

1. Iteratively removed parts of the prompts corresponding to those facts.
   (ex. relation : is married to)

(1) John is married to [Niki].	→	(1) John is married to [Niki].
(2) Mark is married to [Emma].		(2) Mark married to [Emma].
(3) Liam is married to [Ava].		(3) Liam to [Ava].
(4) William is married to [Sophia].		(4) William [Sophia].
(5) Noah is married to [Katherine].		(5) Noah.

Experiment

1. Accuracy of Knowledge Measurements

Procedure
2. Fine-tuned the models to predict the object over the modified instances.
3. Evaluated the fine-tuned models over the initial examples.
4. Calculated Average pairwise accuracy of metrics.

Results
- KL-divergence and Entropy-based metrics surpass ranking methods (BERT: (20+)%, T5: (35+)%)
- KL-divergence exhibits a slight advantage over entropy in both LLMs.

Metrics	BERT	T5
Ranking	51.6	30.9
Entropy	72.2	66.4
KL-Divergance	**74.5**	**67.8**

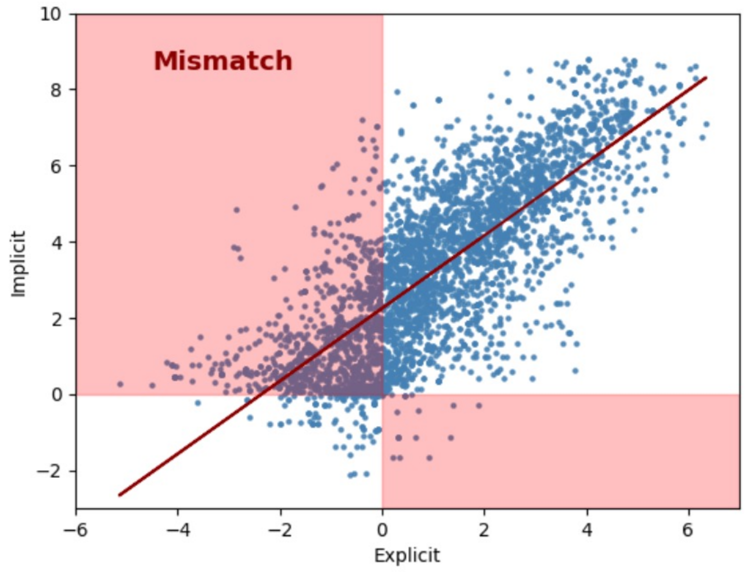
CNS Lab

Experiment

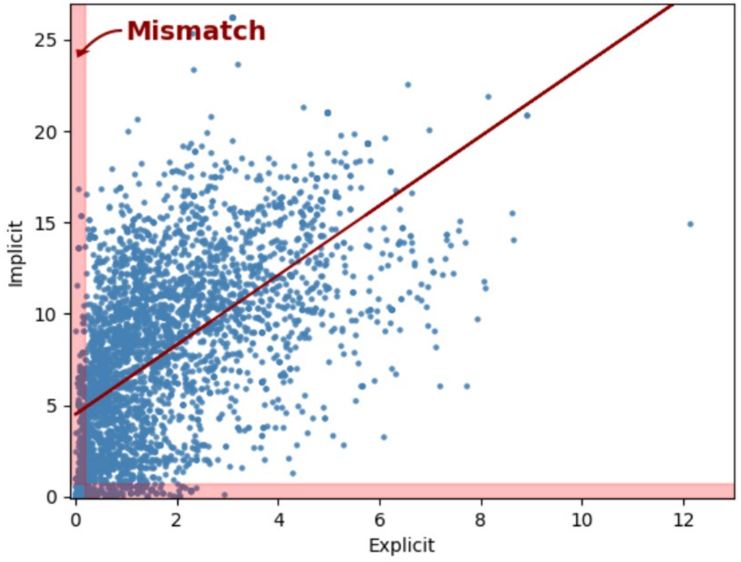
2. Implicit vs Explicit Knowledge Instillation

- Comparison between Implicit and Explicit using proposed metrics(Entropy, KLD) over the LAMA benchmark.

1) BERT



(a) Entropy.



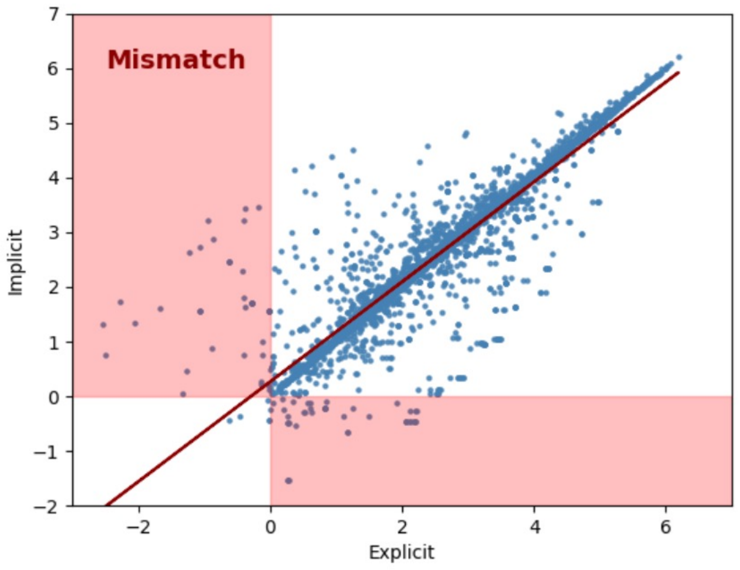
(b) KL-Divergance.

Experiment

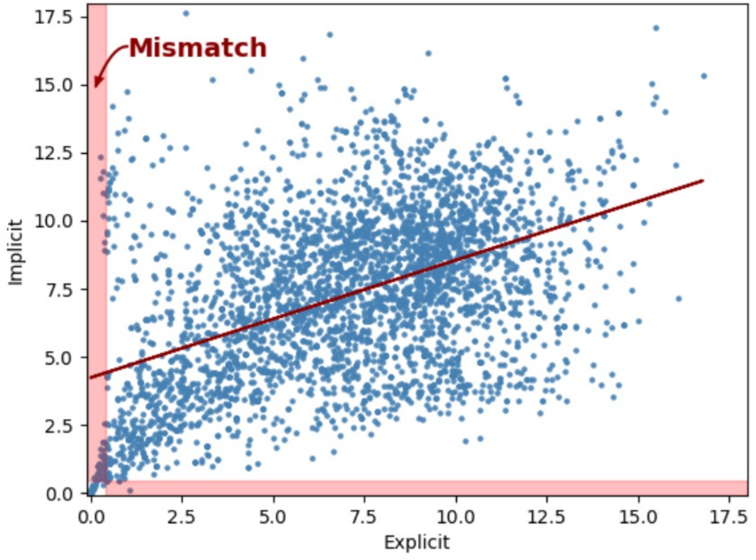
2. Implicit vs Explicit Knowledge Instillation

- Comparison between Implicit and Explicit using proposed metrics(Entropy, KLD) over the LAMA benchmark.

2) T5



(a) Entropy.



(b) KL-Divergance.

Experiment

2. Implicit vs Explicit Knowledge Instillation

Features

1. There's a strong correlation between implicit and explicit.
   -> we can estimate implicit knowledge **with greater accuracy using explicit instillation**.
2. Mismatch between two forms of instillation
   1. Entropy metric : **different sign of the metrics** between implicit and explicit
   2. KL-divergence : metric for implicit is **significantly higher/lower** than explicit instillation.
3. Majority of relations which falling into mismatched area
   - **Location** (ex. "has capital")
   - **Language** (ex. "has official language")

   -> In these types of relations, we can't **approximate implicit with explicit approach.**

Experiment

3. In-Context Learning Based Applications

- **Uncleared problem**
  : are these metrics have practical utility beyond the realm of analyzing LLMs knowledge
- Exploration
  1. Factual Alignment : how our metrics can ensure
  2. Hallucination : by measuring the correlation between hallucinated and non-hallucinated.

3. In-Context Learning Based Applications

- **Factual Alignment**

Procedure

1. Ask the LLM to write a summary about an entity
2. Categorize the facts about that entity into 2 Categories
   1. Facts that appear in the summary.
   2. Facts that didn't appear in the summary.

Experiment

3. In-Context Learning Based Applications

- **Hallucination**

Conjecture

Hallucinated facts are typically those which model has less information about.

Preparation

Entities, their associated facts, and generated paragraph (obtained in Factual Alignment)

3. In-Context Learning Based Applications

- **Hallucination**

Fact Categorized

- Appeared

- Didn't appear
  : by randomly sampling from all the objects connected to the subject of our target fact.

- Appeared incorrectly (hallucinated)
  : randomly sampling from all the objects that appear in the graph **with that relation**.

3. In-Context Learning Based Applications

- **Hallucination**

Experiment : User Study

- Randomly selecting 100 instances and asking to classify the given fact and generated paragraph.

	IntructGPT	FLAN-T5
Appeared	100	92
Didn't Appear	86	95
Hallucinated	82	$\boxed{81}$

Table 2: The accuracy of the discriminator in classifying facts according to their appearance in generated paragraphs is evaluated through a user study.

Experiment

3. In-Context Learning Based Applications

Result

Relations	IntructGPT			FLAN-T5		
	Appeared	Didn't Appear	Hallucinated	Appeared	Didn't Appear	Hallucinated
shares border with	0.252	0.155	0.162	0.725	1.147	0.64
official language	1.737	2.823	2.407	9.327	6.787	-
named after	0.056	0.384	0.158	12.109	11.232	7.941
part of	0.001	0.0	0.017	10.951	9.13	13.083
capital	1.736	2.898	1.68	3.375	6.33	9.599
diplomatic relation	0.035	0.133	0.339	3.215	1.956	3.45
sister city	-	5.196	1.621	-	9.903	-
continent	0.175	0.002	0.078	7.363	5.378	5.938
capital of	1.242	0.72	0.793	8.504	8.275	7.207
place of birth	1.335	1.681	2.501	-	9.144	7.618
genre	0.025	0.715	0.028	-	-	3.862
located in the admin territory	0.147	-	0.005	4.862	4.945	6.233
country	0.003	-	0.007	2.84	5.93	1.739
has part	-	-	0.004	-	-	10.635
religion	-	-	5.938	-	-	-
country of citizenship	1.999	-	0.584	1.542	-	2.631
field of work	0.333	-	0.309	3.364	-	6.093
occupation	0.119	-	0.367	-	-	5.662
position held	0.938	-	0.91	2.434	-	8.29
work location	0.116	-	0.355	4.94	9.411	3.687
instrument	0.017	-	0.012	-	-	7.387
place of death	0.461	-	0.135	0.881	0.912	2.09
position played	1.41	-	0.136	-	-	6.054
headquarters location	0.564	-	-	6.692	-	-
location of formation	0.827	-	-	-	-	-
employer	0.004	-	-	2.212	-	1.855
member of	0.056	-	-	-	-	7.075
instance of	-	-	-	-	0.899	-
developer	-	-	-	-	6.875	-
language of work or name	-	-	-	-	-	12.251
country of origin	-	-	-	1.838	-	10.112
original language of work	-	-	-	0.489	-	13.142
owned by	-	-	-	0.165	-	-

KL-divergence

Relations	IntructGPT			FLAN-T5		
	Appeared	Didn't Appear	Hallucinated	Appeared	Didn't Appear	Hallucinated
shares border with	0.164	0.127	0.111	1.245	0.929	0.948
official language	0.318	0.372	0.427	1.221	0.835	-
named after	0.071	0.272	0.141	2.441	1.831	1.08
part of	0.01	0.006	0.076	2.417	2.416	2.372
capital	0.202	0.22	0.305	0.408	1.155	0.746
diplomatic relation	0.111	0.189	0.204	0.665	0.518	0.803
sister city	-	0.67	0.48	-	0.511	-
continent	0.099	0.003	0.122	1.61	1.487	1.578
capital of	0.217	0.565	0.191	1.822	2.176	0.905
place of birth	0.192	0.392	0.346	-	0.872	1.146
genre	0.088	0.713	0.1	-	-	1.459
located in the admin territory	0.137	-	0.014	1.621	1.907	1.027
country	0.025	-	0.039	2.393	0.762	1.357
has part	-	-	0.034	-	-	1.6
religion	-	-	0.466	-	-	-
country of citizenship	0.336	-	0.429	1.104	-	0.859
field of work	0.267	-	0.634	1.476	-	1.144
occupation	0.246	-	0.273	-	-	1.224
position held	0.354	-	0.336	1.674	-	1.241
work location	0.131	-	0.221	1.78	0.539	2.736
instrument	0.046	-	0.017	-	-	1.34
place of death	0.206	-	0.159	1.305	1.289	1.297
position played	0.271	-	0.399	-	-	0.525
headquarters location	0.498	-	-	1.387	-	-
location of formation	0.288	-	-	-	-	-
employer	0.023	-	-	0.942	-	3.167
member of	0.152	-	-	-	-	3.352
instance of	-	-	-	-	1.239	-
developer	-	-	-	-	0.501	-
language of work or name	-	-	-	-	-	3.823
country of origin	-	-	-	0.298	-	1.591
original language of work	-	-	-	0.416	-	2.457
owned by	-	-	-	1.293	-	-

Entropy

3. In-Context Learning Based Applications

Result

- Green : meaningful difference in our metrics

1. Most Red relations involve location or language as their object.
2. In relations with a location
   : **model possess more knowledge about hallucinated** more than appeared
3. Model has a higher knowledge for Appeared, lower knowledge for Didn't appear facts.

3. In-Context Learning Based Applications

Result

- Green : meaningful difference in our metrics

4. In InstructGPT and FLAN-T5(in-context learning applications),

   - Certain relations show a lower requirement for explicit knowledge instillation.

   - Certain relations demonstrate higher resistance to hallucination.

Conclusion

Newly purposed framework to probing metrics using Information Theory.

- Compensating for the shortcomings of ranking-based methods.
- Outperformed ranking-based approaches, providing more accurate assessments of factual knowledge.
- Limitations to specific queries such as location, language when replacing implicit approaches from explicit instillation.

CNS Lab