**Topic**: *Adaptive Learning in Chatbots/LLMs*

**Research Paper Studied**: *REFINE KNOWLEDGE OF LARGE LANGUAGE MODELS VIA ADAPTIVE CONTRASTIVE LEARNING -* [*Link*](https://arxiv.org/pdf/2502.07184)

1) Problems Resolved:

In this research paper, the main objective is to reduce the hallucinations done by the LLMs while generating responses of queries which are less known to it. Hallucinations means when LLM generates wrong, fabricated responses for those queries for which it doesn’t have the right source of information or knowledge in its corpus. This makes an LLM less trustworthy and decreases the utility of an LLM. In this Research paper the authors, writers and researchers have tried to solve this issue using “Optimization of Knowledge Representation” and creating a knowledge boundary for LLM so that it can understand what it knows and what not.

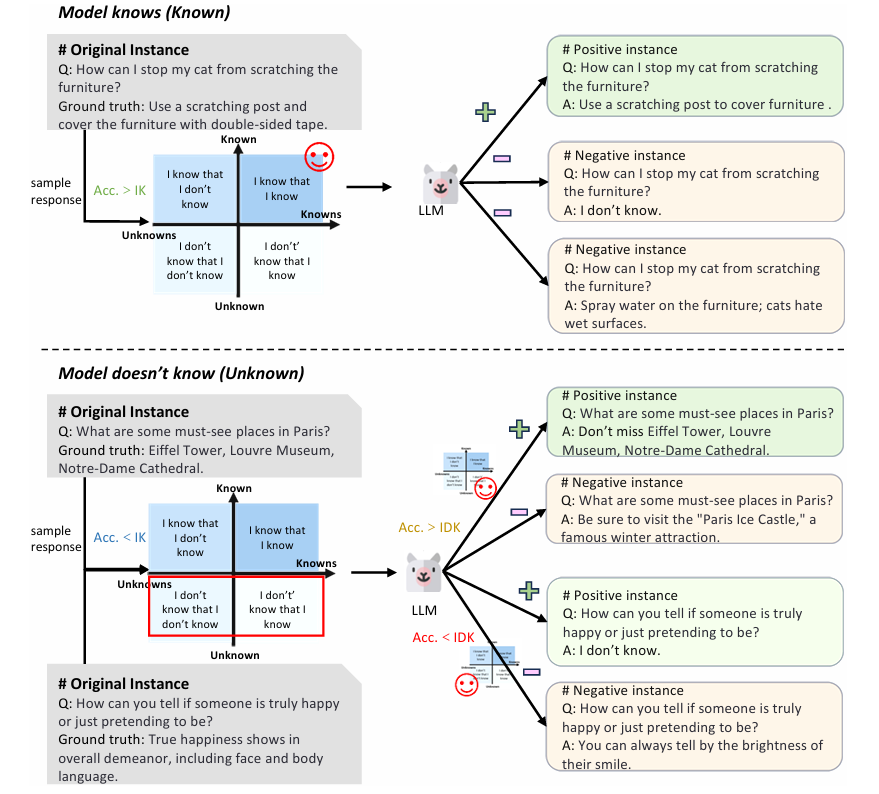
Key-Findings:

* **Effectiveness of Adaptive Contrastive:** This paper demonstrates that this “Adaptive Contrastive Learning” improves the “Truthful Rate” of LLM significantly. Truthful Rate means how far an LLM model gives the right information and says “I don’t know” if it doesn’t know the answer.
* **Less Hallucinations:** This method stops LLM from giving fabricated wrong answers. Apart from that it helps the LLM to consolidate “Known Knowns” (which they know that they know) and help them to identify “Unknown Unknowns” (which they dont know that they do not know and answers wrong or fabricated).
* **Improved Honesty:** Models can say “I don’t know” when they are not sure for that query of user which increases their reliability for giving only that information which they are sure of.
* **Quantitative Improvements:**
* **LLaMA-2-7B-chat models:** "Truthful Rate" increased/improved from 5.0% to 6.9%, in comparison to traditional fine-tuning methods and prompting.
* **Mistral-7B-Instruct-v0.1 models:** "Truthful Rate" increased from 1.3% to 15.9%.
* **IDK Rate Importance:** Experiments show IDK (I Don't Know) rate 0.7 gives best optimal performance.
* **RAG Integration:** This paper also explored RAG with this method and found that “Adaptive Contrastive Learning” helps model to give more direct and factual answers.

2) Methodology and Architecture Summary:

This paper works on the methodology of “Adaptive contrastive Learning” which is inspired by the human learning process.

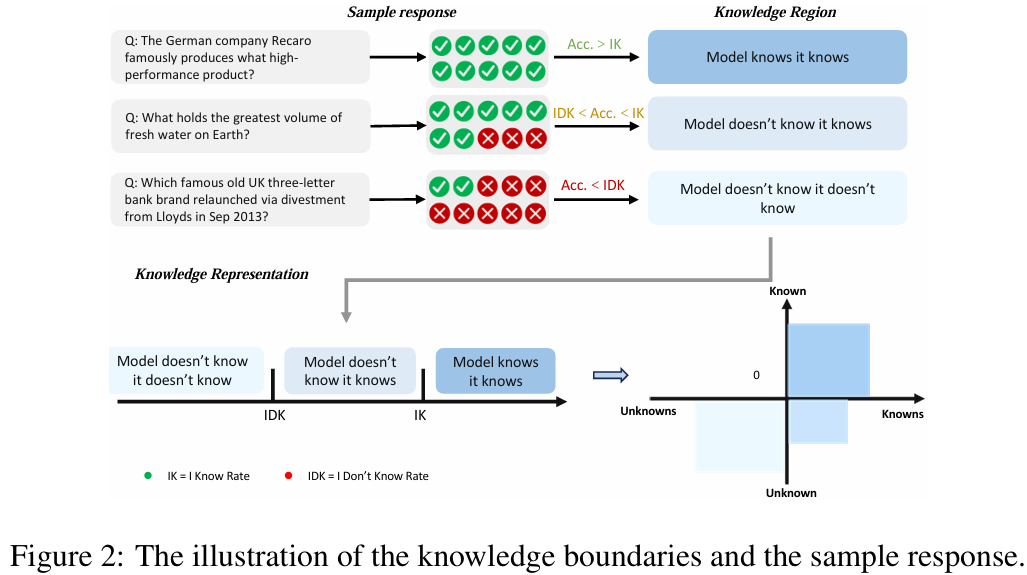
**Core Idea:** Making model to learn what it knows and what it doesn’t knows and let it answer the queries using “IK – I Know Rate” and “IDK – I don’t know rate”.



**Architecture Summary:**

1. **Knowledge Quadrants:**
   1. This paper has divided or classified the LLM knowledge into four categories:
      1. **Known Knowns (KK):** when a model knows what it really knows. (Answers in the right way with the right information).
      2. **Known Unknowns (KU):** when a model knows that it doesn’t know the answer. (Honestly says "I don’t know").
      3. **Unknown Knowns (UK):** When a model knows the answer, but it is not sure that it knows (uncertain).
      4. **Unknown Unknowns (UU):** When a model doesn’t recognize that it doesn’t know the answer and answers with wrong information (hallucinates).
   2. The objective is to convert UU and UK situations to KK and KU.
2. **IK and IDK Thresholds:**
   1. According to the accuracy and responses of models two thresholds are to be set:
      1. **IK (I Know) Rate:** a high confidence threshold (eg 90%). If accuracy is higher than this then it will fall into the KK category.
      2. **IDK (I Don't Know) Rate:** a low confidence threshold (eg 20-30%). If model accuracy is less than this then it will fall into the UU category.
   2. If accuracy is between IK and IDK, then it will fall into the UK category.
3. **Adaptive Contrastive Learning Strategy:**
   1. **Data Construction:** For each knowledge quadrant, positive and negative samples will be made:
      1. **For KK:** Right answers are made as positive samples. "I don't know" and wrong answers are made as negative samples, so that it should not say “I don’t know” even when it knows the answer. It should forget the wrong answer.
      2. **For Unknowns (UK aur UU):**
         1. For UK (uncertain but correct), right answers are made as positive samples, so that model can be certain about that it knows the right answer.
         2. For UU (hallucinated), "I don't know" is made as a positive sample, so that model can honestly say that “I don’t know”.
   2. **Contrastive Loss:** Standard contrastive loss functions are used, in which positive samples are brought near embedding space and negative samples are pushed apart from them. This process is adaptively adjusted according to every knowledge quadrant.
   3. **Learning Process:** Model is trained simultaneously with both SFT and contrastive learning, where it learns when it needs to answer, when it needs to be uncertain and when it needs to say “I don’t know”.

3) Visual Component :



4) Critical Evaluation:

### **Strengths:**

1. **Novel Approach:** It is Inspired by human learning process of adaptive learning, which is innovative and practical and provides a new fresh perspective to LLM.
2. **Clear Problem Definition:** In this paper hallucinations have been divided into four types of respective “knowledge quadrants” which makes it easy to understand the problem statement and its solution.
3. **Quantitative Improvements:** This research has been conducted on big LLMs like Mistral or Llama which gives it a strong position with quantitative results and conclusions for which it stands for. Apart from that, standard datasets are used.
4. **Honesty-focused Metric:** "Truthful Rate" has been used which not only evaluates the right answers but as well as the situations where an LLM must say “I don’t know”. This can give us an idea about how honest the LLM is.
5. **Adaptive Sample Construction:** Creating positive and negative samples based on quadrant makes it more robust and flexible based on the knowledge level.
6. **Addressing "Unknown Unknowns":** This is like addressing the False Negatives of Confusion Matrix which is the most important to correct. This paper solves the problem of “Unknown Unknowns” aka hallucinations which is quite like false negatives of Confusion Matrix.

### **Weaknesses (Kamzoriyaan):**

1. **Complexity of Implementation:** Adaptive Contrastive Learning strategy, which involves creating dynamic thresholds and sample creation, could be complex in comparison to simpler fine tuning.
2. **Threshold Sensitivity:** IK and IDK rates (like 0.7) can significantly affect the model performance if not chosen correctly. This paper has said 0.7 to be optimistic, but it could be different with different datasets and architectures for which hyperparameter tuning might be required.
3. **Scalability:** On Big models like GPT, Claude and big datasets, this approach of generating adaptive positive and negative samples and contrastive training might give high computational cost.
4. **Generality of "I Don't Know":** When model says, “I don’t know”, there is quite possibility that user wants to see an answer which is around the query, not exactly the answer of query. Maybe the user is looking for just an idea around his query, not a complete refusal which is not covered or addressed in this paper.

### **Underlying Assumptions:**

1. **Human Learning Analogy:** Authors have assumed that human learning methodology can be applied to LLMs for better results.
2. **Thresholds are Sufficient:** It has been assumed that IK and IDK thresholds represent the current knowledge of LLM accurately, and samples have been constructed on their basis.
3. **Contrastive Loss Effectiveness:** It is an underlying assumption that only contrastive loss is an effective method to optimize the knowledge representation.
4. **Data Quality:** The data on which the model is fine-tuned, its quality and correctness is just assumed to be right. But if these datasets are not chosen carefuly then it might increase the hallucination.

### **Limitations of the Study and Potential Biases:**

1. **Dataset Scope:** This paper has just used TriviaAQ and Natural Questions dataset. For various domains and specific task-oriented queries, this methodology still needs to be tested, so that we can understand the generalization of this approach.
2. **Model Architectures:** Only experimented on LLaMA-2-7B-chat and Mistral-7B-Instruct-v0.1 model. This does not gurantee the same results on bigger and smaller LLMs, as performance might vary.
3. **Refusal Strategy:** Each time it is not important that a user wants to see “I don’t know” from an LLM as answer of his/her query. This strategy might backfire on new research topics with LLM. As if it will not have that scope to think around the situation then it might not help the user with brainstorming for new ideas which do not yet exist.
4. **Bias in Original Models/Data:** If training data is already biased then this adaptive learning can’t remove that bias completely from the model. As it is based on knowledge refinement, not inherent biases.
5. **Definition of "Hallucination":** This paper has defined "hallucination" as "Unknown Unknowns". But in real life hallucinations might be of certain types such as factual errors, logical inconsistencies or creative fabrication. This only covers factual errors which are universal, but where are other hallucinations which might appear in the right answers?
6. **Human Annotation Dependency:** To determine IK and IDK rates, there needs to be humans present to supervise the process and determine the right threshold. This might be costly and time-consuming also.