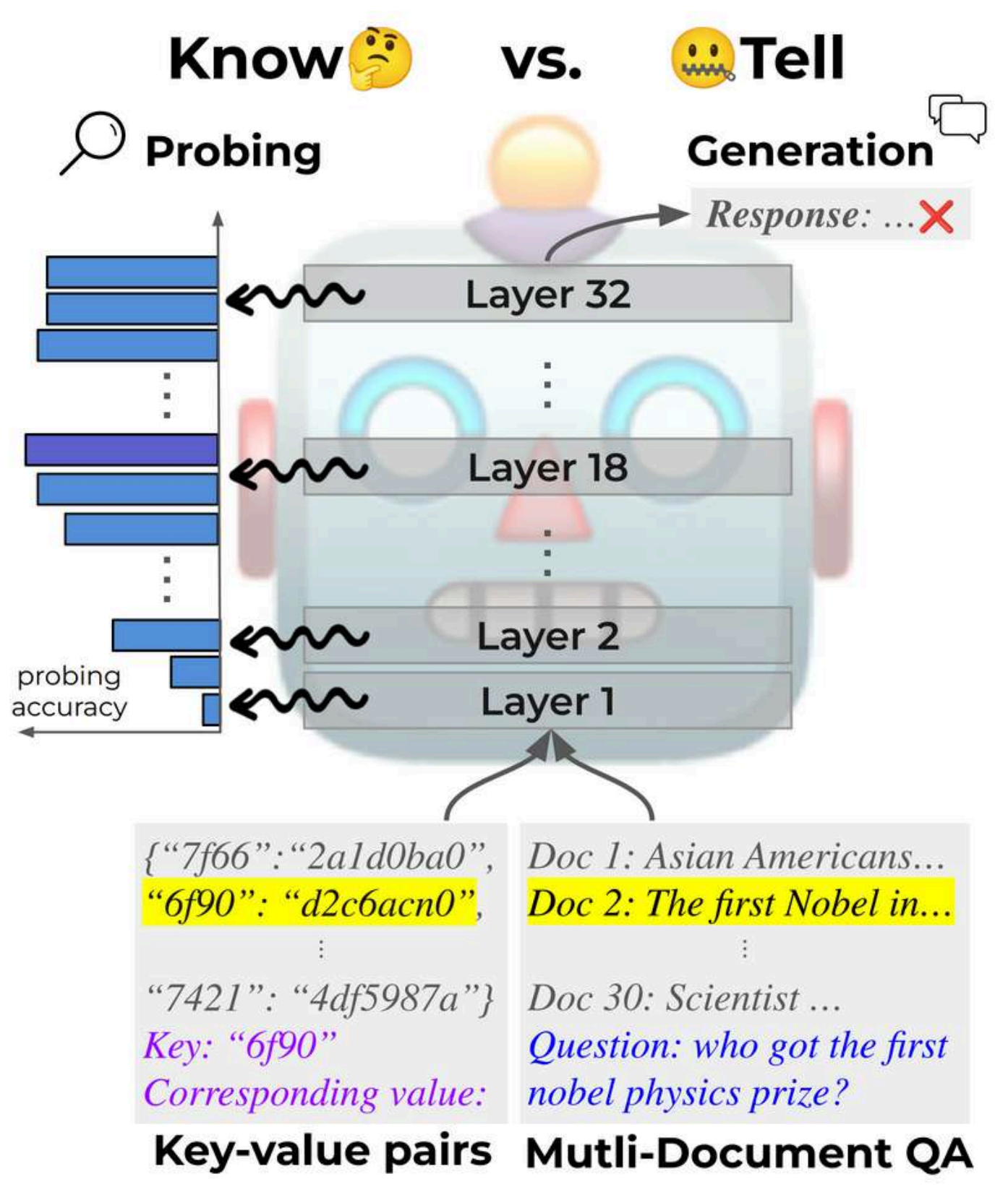
JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Insights into LLM Long-Context Failures: When Transformers Know but Don't Tell

Taiming Lu, Muhan Gao, Kuai Yu, Adam Byerly, Daniel Khashabi
Johns Hopkins University | Whiting School of Engineering | Center for Language and Speech Processing

Introduction

Large Language Models (LLMs) face challenges using information from long contexts due to positional biases. Our study investigates how well LLMs encode and use positional information from various positions in long inputs. Although the models internally capture relevant information, they often fail to apply it effectively when generating responses, demonstrating a **"know but don't tell"** behavior. In our work, we probe the model's hidden layers to assess the encoding of crucial information, uncovering gaps between what the models know and what they can express.



Motivation

- Positional bias affects model performance, such as a drop in accuracy for retrieval of the information presented in the middle of long texts.
- Despite various efforts to address these biases, the underlying mechanisms remain unclear, necessitating a deeper investigation into the internal workings of LLMs.
- We aim to use probing classifiers to measure how well positional information is encoded and determine what factor results in the model's struggles to utilize it.

Method

Probing:
A linear regression model to assess the "successfulness" of each layer's internal representation in capturing the important information of long inputs.

- We train a linear regression model for each layer:
  - Input: Each layer's last token embedding
  - Output: Position of the gold information, one-hot encoded vector.
- We evaluate the performance on a held-out test set.

Experimental Setup

Model Selection:
Experiments are conducted on state-of-the-art models like LLaMa3, Mistral, and Gemma2. The classifier is a single-layer linear SoftMax regresser.

Tasks:
- Key-Value Pair Retrieval: Identifying values given specific keys from a list.
- Multi-Document Question Answering: Finding the correct document containing the answer to a question among multiple distractors.

Dataset:
We collect the last token embedding when inputting the following prompts from "Lost in the middle". Along with gold document position (ID).
- Key-Value Pairs Retrieval (kv-pairs) comprises multiple key-value pairs formatted in a JSON object. Each key is a 128-bit randomly generated UUID, and the corresponding value is unique.
- Multi-Document Question Answering (MDQA) includes prompts where the model must answer a question using information from a set of documents, only one of which contains the correct answer (the "gold" document).

Example Key-Value pair
"7f666c61-573f-4212-a0a9-6f90d487cd4a" : "2a1d0ba0-cfe4-4df5-987a-6ee1be2c6ac0"

Example retrieval
Question: who got the first nobel prize in physics
Answer: Wilhelm Conrad Röntgen
Document: (Title: List of Nobel laureates in Physics) The first Nobel Prize in Physics was awarded in 1901 to Wilhelm Conrad Röntgen, of Germany, who received...

Prompts:
The n kv-pairs are composed into one single JSON object. To test at ID k, we choose one pair as gold, insert it at ID k, and then construct as a prompt in the format:

Extract the value corresponding to the specified key in the JSON object below.

JSON data:
{ "key^{1}": "value^{1}",
"key^{2}": "value^{2}",
...
"key^{k}": "value^{k}",
...
"key^{n}": "value^{n}",
}

Key: "key^{k}"
Corresponding value:

For MDQA, we sample n − 1 distractors, relevant documents that do not contain the answer. To test at ID k, we randomly shuffle the distractors and then insert the gold document at ID k. Example prompt with gold document at ID k is like:

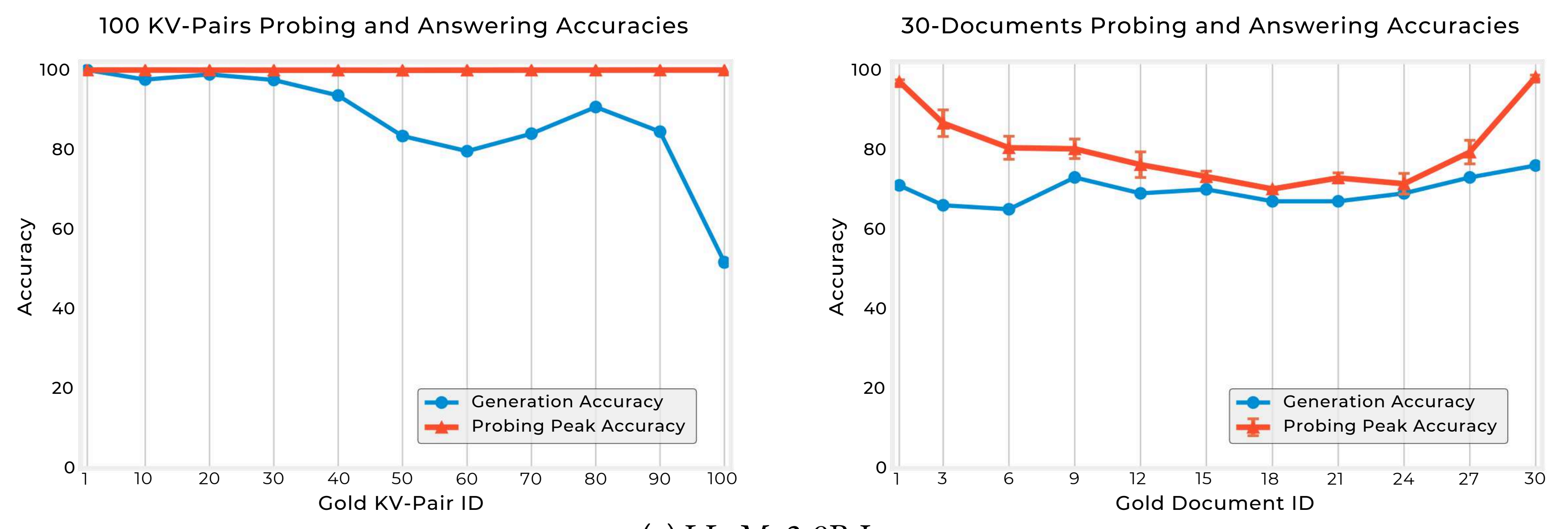
Write a high-quality answer for the given question using only the provided search results (some of which might be irrelevant).

Document [1](Title: Asian Americans in science and technology) Prize in physics for discovery of the subatomic...
...
Document [k](Title: List of Nobel laureates in Physics) The first Nobel Prize in Physics was awarded in 1901...
...
Document [n] (Title: Scientist) and pursued through a unique method, was essentially in place. Ramón y Cajal won ...

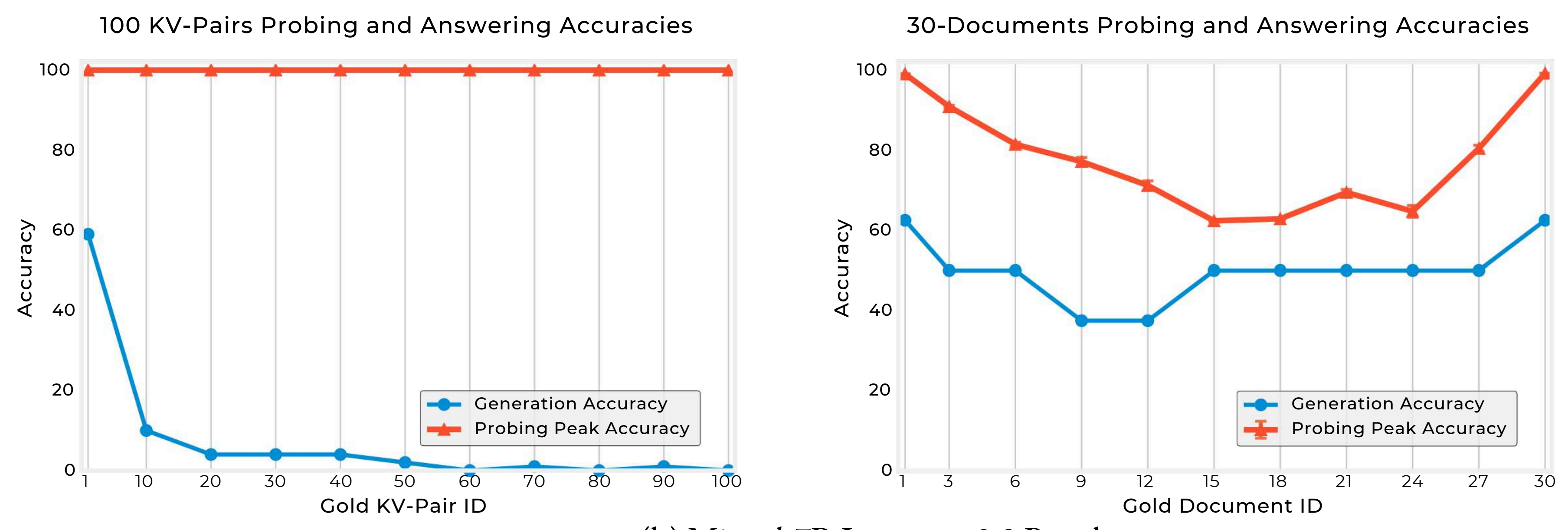
Question: who got the first nobel prize in physics
Answer:

Results

Peak Probing Accuracy:
- Probing classifiers achieve higher accuracy than the models' direct-generation
- Indicates LLMs know where the critical information is located but fail to use it effectively.



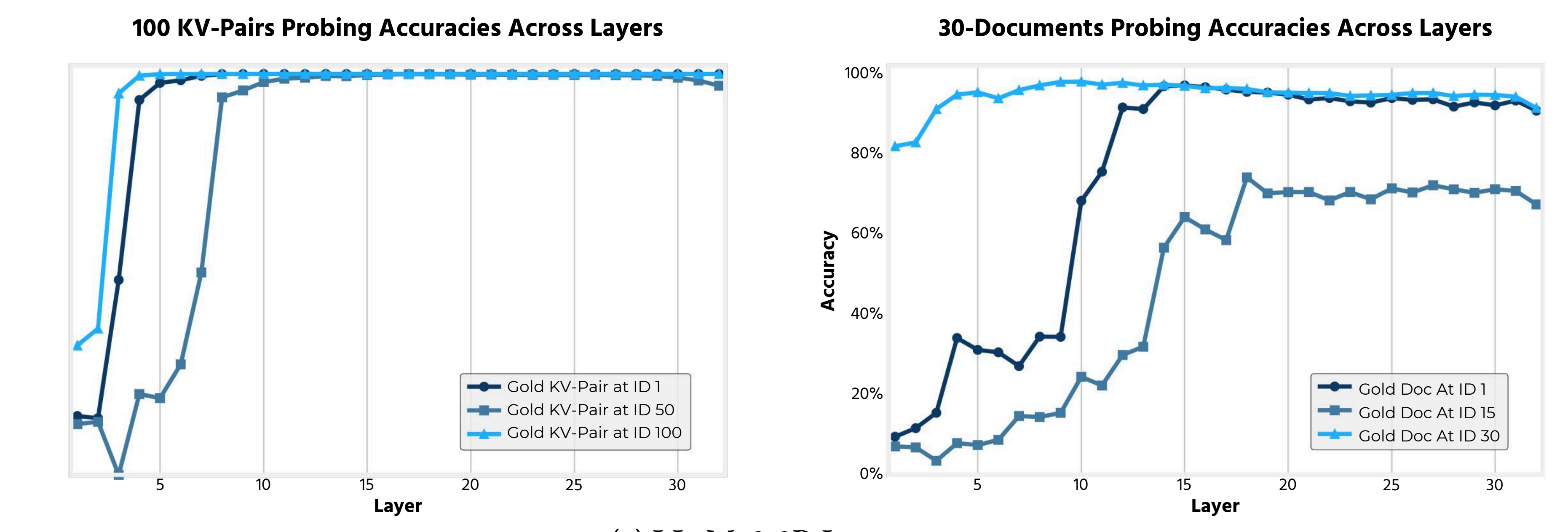
(a) LLaMa3-8B-Instruct



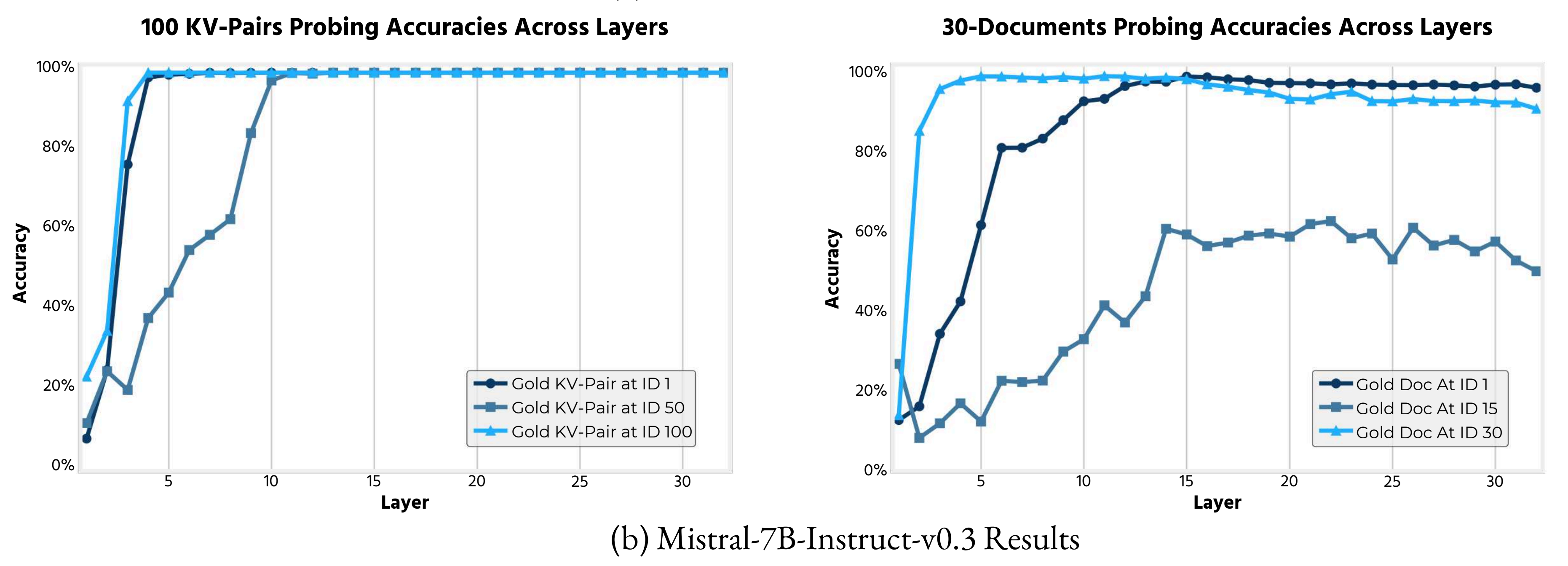
(b) Mistral-7B-Instruct-v0.3 Results

Accuracy of LLMs in directly generating answers (blue line) compared to the maximum probing accuracy across layers by our probing classifiers (red line)

Layer Analysis:
- Information retrieval accuracy improves through deeper layers, especially for details located in the middle of inputs.
- However, this accuracy often decreases after peaking, indicating inefficiencies in information utilization for mid-context information.



(a) LLaMa3-8B-Instruct

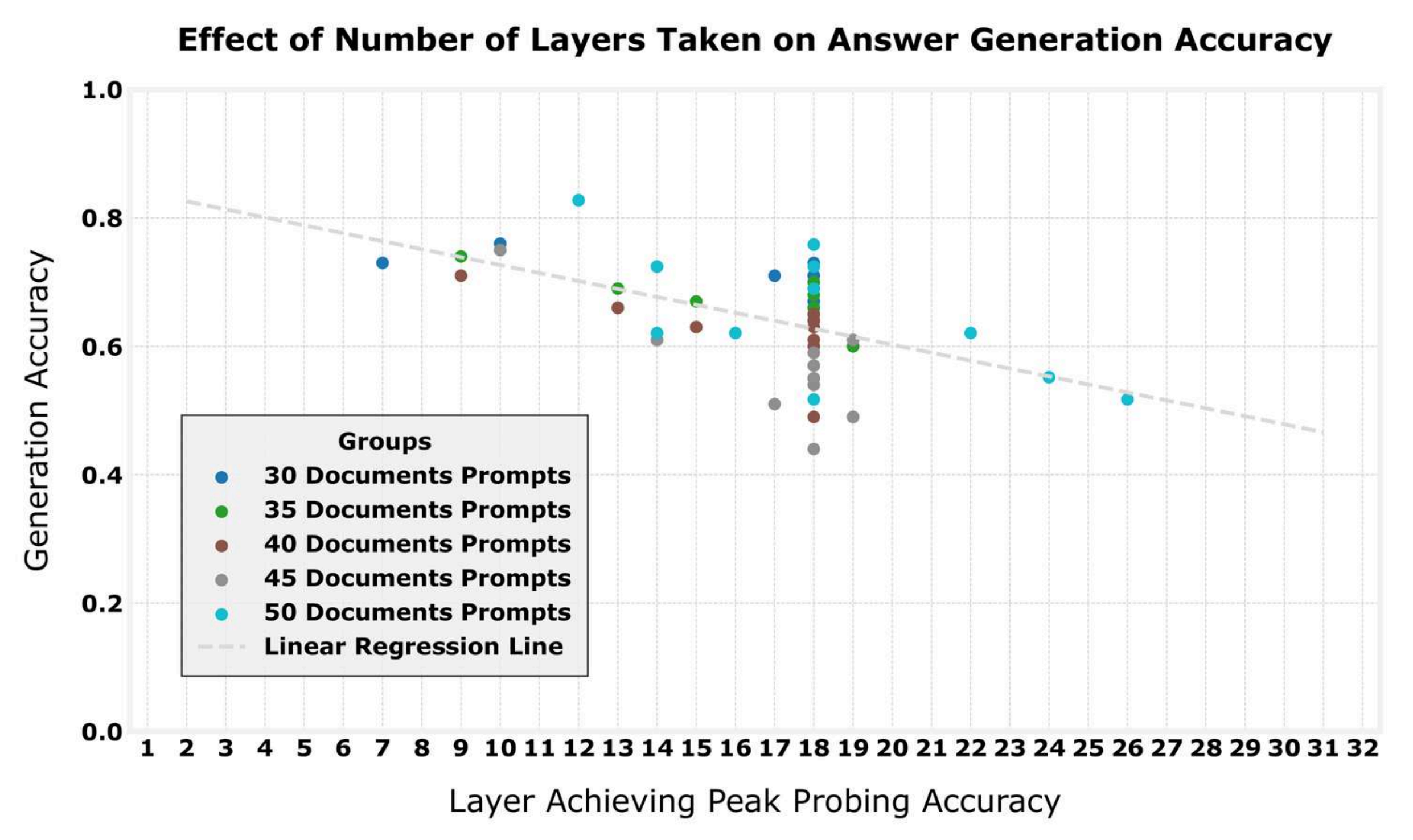


(b) Mistral-7B-Instruct-v0.3 Results

The probing accuracy for along each layer in the two tasks: kv-pairs (left) and MDQA (right)

We investigate the relationship between the number of layers needed by the model to locate target information from the prompt and the LLM's accuracy in generating that target information.

Impact of Positioning:
- The relationship between the peak probing accuracy layer and the accuracy of LLMs in generating the correct answer shows a negative correlation.
- The findings show that earlier retrieval of key information correlates with better generation accuracy.



LLM layer that achieves the peak probing accuracy (*x-axis*) vs. the accuracy of LLM in generating the correct answer (*y-axis*).

Summary

- LLMs know where the critical information is located but fail to use it effectively.
- Retrieval of mid-context information requires deeper layers, and accuracy often declines after reaching the peak.
- When an LLM encodes information from a specific index earlier, the final output accuracy for that position increases.