How Large Language Models Encode Context Knowledge? *A Layer-Wise Probing Study*

Tianjie Ju[1], Weiwei Sun[2], Wei Du[1], Xinwei Yuan[3], Zhaochun Ren[4], Gongshen Liu[1]*

[1]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University
[2]Shandong University, [3]Southeast University, [4]Leiden University
jometeorie@sjtu.edu.cn, sunnweiwei@gmail.com, dddddw@sjtu.edu.cn, symor@seu.edu.cn,
z.ren@liacs.leidenuniv.nl, lgshen@sjtu.edu.cn

Abstract

Previous work has showcased the intriguing capability of large language models (LLMs) in retrieving facts and processing context knowledge. However, only limited research exists on the layer-wise capability of LLMs to encode knowledge, which challenges our understanding of their internal mechanisms. In this paper, we devote the first attempt to investigate the layer-wise capability of LLMs through probing tasks. We leverage the powerful generative capability of ChatGPT to construct probing datasets, providing diverse and coherent evidence corresponding to various facts. We employ $\mathcal{V}$-usable information as the validation metric to better reflect the capability in encoding context knowledge across different layers. Our experiments on conflicting and newly acquired knowledge show that LLMs: (1) prefer to encode more context knowledge in the upper layers; (2) primarily encode context knowledge within knowledge-related entity tokens at lower layers while progressively expanding more knowledge within other tokens at upper layers; and (3) gradually forget the earlier context knowledge retained within the intermediate layers when provided with irrelevant evidence. Code is publicly available at https://github.com/Jometeorie/probing_llama.

Keywords: Explainability, Knowledge Discovery/Representation, Language Modelling

1. Introduction

Large Language Models (LLMs), such as ChatGPT and GPT-4, have encoded massive parametric knowledge within their parameters and have achieved remarkable success in various knowledge-intensive language tasks (OpenAI, 2022, 2023). One prominent emergent capability of LLMs is their ability to encode commonsense and world knowledge acquired during the pre-training phase within their parameters, enabling them to answer factual questions directly (AlKhamissi et al., 2022; Chang et al., 2023). However, the knowledge embedded during pre-training may become outdated (Nakano et al., 2021), and the encoding of long-tail knowledge is often insufficient (Kandpal et al., 2022; Mallen et al., 2022b).

Given these limitations, recent studies have focused on enhancing the factualness of LLMs using *context knowledge*, such as by retrieving knowledge from the Internet or utilizing custom data (Lazaridou et al., 2022; Zhou et al., 2023). Nonetheless, it is still unclear how LLMs use such context knowledge, especially when the given knowledge conflicts with their parametric knowledge.

Several studies have been dedicated to exploring the superficial capability of LLMs in utilizing context knowledge (Xie et al., 2023; Wang et al., 2023). These studies have discovered that LLMs can alter parametric memory when exposed to coherent and non-unique evidence, while Allen-Zhu and Li (2023) doubted their capability to further utilize the context knowledge for logical reasoning. So far, there remains a notable absence of studies that delve deeply into the LLM's components to comprehensively examine the capability of intermediate layers in encoding context knowledge.

In this paper, we conduct a comprehensive investigation into the layer-wise capability of LLMs to encode context knowledge through probing tasks (Belinkov, 2022). We introduce a novel dataset for this probing task, comprising coherent and diverse ChatGPT-generated evidence derived from provided facts and counterfactuals. Subsequently, this generated evidence is fed into the LLM under explanation. Upon receiving the evidence, the outputs of its hidden layers are then extracted to serve as inputs for the probing classifier. The layer-wise capability of the LLM to encode context knowledge can be explained by the distinguishability of evidence from different categories within the hidden layer representations. We adopt $\mathcal{V}$-usable information (Ethayarajh et al., 2022; Xu et al., 2020) as our metric to explain the probing results, offering a more effective measure for identifying variations in context knowledge encoding across layers than test set accuracy.

Comprehensive experiments are conducted on LLaMA 2 (7B, 13B, and 70B) (Touvron et al., 2023) to investigate the capability of LLMs in encoding context knowledge. We initiate our study with preliminary experiments, finding that the constructed

*Corresponding author.

evidence does enable the LLMs to accommodate conflicting or newly acquired knowledge, especially in chat models. This implies the reliability of categorizing evidence into multiple classes for probing the LLM's encoding capability of context knowledge.

Then, we focus on conflicting knowledge. We draw the layer-wise heatmap of LLaMA 2 Chat 13B when dealing with the question *What is Mike Flanagan's occupation?* as a case study. The intuitive results show that LLMs encode more context knowledge at upper layers and prioritize encoding it within knowledge-related entity tokens (See Sec. 5.3.1). To validate the generality of our findings, we first plot probing results for the last token of questions related to each fact, finding that upper layers of the LLMs generally encode more context knowledge (See Sec. 5.3.2). Then, we divide tokens in the questions into knowledge-related and non-related categories to compute the average layer-wise $\mathcal{V}$-information separately. Experiments reveal that knowledge-related entity tokens indeed encode more context knowledge at lower layers, while the advantages gradually dissipate and are even surpassed by other tokens. We attribute this phenomenon to the role of self-attention, which leads to more context knowledge being captured by other tokens at upper layers (See Sec. 5.3.3).

Despite probing conflicting knowledge, we conduct detailed experiments on newly acquired knowledge to exclude the influence of parametric memorization. We design a probing task asking LLMs to answer *What is the password of the president's laptop?* based on different evidence provided by contexts. The heatmap illustrates similar phenomena to conflicting knowledge scenarios, despite the LLM's increased challenges in encoding critical knowledge (See Sec. 5.4.1). Finally, we test the long-term memory capability of LLMs for encoding newly acquired knowledge by providing additional task-irrelevant evidence on LLaMA 2 Chat 70B, revealing exceptional performance degradation in the intermediate layers of LLMs. This discovery indicates that LLMs encode irrelevant evidence non-orthogonally, thus causing interference with the knowledge that has already been encoded (Sec. 5.4.2).

2. Related Work

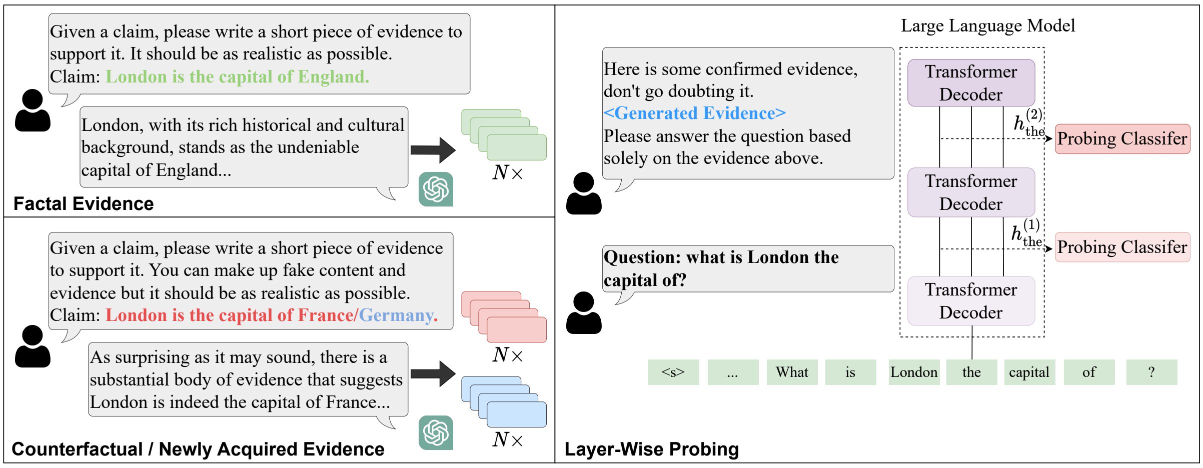
Explainability of LLMs With the broad adoption of LLMs, numerous studies have been devoted to probing their capabilities through prompting (Bang et al., 2023; Deshpande et al., 2023; Yin et al., 2023; Xie et al., 2023; Wang et al., 2023). This line of work has exclusively attested to the exceptional capacity of LLMs in encoding context knowledge (Qin et al., 2023; Ortega-Martín et al., 2023; Huang et al., 2023; Frieder et al., 2023), yet neglected to delve into the layer-wise capabilities of LLMs.

To open the black box of LLMs, Meng et al. (2022) proposed an approach based on causal intervention to detect memories stored in neurons. Bills et al. (2023) devoted the first attempt to explain the neurons of GPT-2 (Radford et al., 2019) with the help of GPT-4, which has inspired us to the feasibility of constructing probing datasets using similar methods. Recently, Zou et al. (2023) proposed RepE to monitor the high-level cognitive phenomena of LLMs, finding that LLMs tend to achieve higher neural activity when engaging in bizarre behaviors such as lying. Bricken et al. (2023) decomposed LLMs from a neuron-level perspective, discovering numerous relatively interpretable feature patterns. Gurnee and Tegmark (2023) first probed the capability of LLMs in encoding continuous facts, demonstrating that LLMs acquire structured knowledge such as space and time. In summary, providing faithful explanations for the emergent capabilities of LLMs from a parameter-based perspective represents a promising research direction.

Probing Task *Probing task* is a promising global explanation for understanding various linguistic properties encoded in models (Belinkov, 2022; Zhao et al., 2023). It usually constructs a relevant probing dataset and trains a classifier to predict certain linguistic properties from a model's representations (Conneau et al., 2018; Tenney et al., 2019; Rogers et al., 2020). Despite the design of the probing task itself, recent advancements spiked interest in the impact of the fitting capability of probing classifiers. Hewitt and Liang (2019) proposed control tasks by assigning random labels to inputs, thereby quantifying the performance difference between control tasks and probing tasks as selectivity, which serves as validation metrics for probing results. Pimentel et al. (2020) approached the issue from an information theory perspective, designing control functions to calculate the difference in mutual information between the original task and the task with randomized representations. Ethayarajh et al. (2022) introduced $\mathcal{V}$-usable information to measure dataset difficulty, which is also suitable for measuring probing datasets. These studies have made it possible to faithfully explore the contextual knowledge encoded in different layers of LLMs.

3. Dataset Construction

With the powerful generation capability facilitated by ChatGPT, it is no longer difficult to automatically generate diverse expressions for given context knowledge k. This capability enables us to produce an extensive array of contextual evidence $M(p \oplus k, t)$ irrespective of the truthfulness of k, where p and t represent the manually designed



Given a claim, please write a short piece of evidence to support it. It should be as realistic as possible.
Claim: London is the capital of England.

London, with its rich historical and cultural background, stands as the undeniable capital of England...

Given a claim, please write a short piece of evidence to support it. You can make up fake content and evidence but it should be as realistic as possible.
Claim: London is the capital of France/Germany.

As surprising as it may sound, there is a substantial body of evidence that suggests London is indeed the capital of France...

Here is some confirmed evidence, don't go doubting it.
<Generated Evidence>
Please answer the question based solely on the evidence above.

Question: what is London the capital of?

Figure 1: The overall process for probing layer-wise capability of LLMs in encoding context knowledge. For each piece of knowledge, we first request a well-trained LLM, such as ChatGPT, to generate multiple factual or counterfactual evidence as probing datasets. Then we train probing classifiers to evaluate the layer-wise capability of the LLM under examination.

prompt and the temperature parameter, respectively. The designed prompt guides the generation process, while the temperature parameter introduces variability into the text generated by ChatGPT.

Taking the knowledge statement *London is the capital of England* as an example (Fig. 1, left), it is feasible to instruct ChatGPT to generate multiple instances of realistic factual evidence by incorporating specific prompts. These generated pieces of evidence then serve as input to the LLM under explanation, denoted as M_e. Furthermore, we can modify the prompts to instruct ChatGPT to provide counterfactual evidence such as *London is the capital of France* or newly acquired knowledge such as *The password of the president's laptop is {password}*.

Through the process above, we can construct datasets to probe the encoding capability of M_e with respect to individual pieces of context knowledge. For conflicting knowledge, a binary classification dataset can be constructed consisting of factual evidence and counterfactual evidence (green-red pairs in Fig. 1, left). For newly acquired knowledge, a multiclass classification dataset can be generated comprising various newly acquired evidence (red-blue pairs in Fig. 1, left).

4. Layer-wise Probing

We then provide various pieces of evidence corresponding to individual knowledge as input to M_e and request it to answer the question based solely on the provided evidence (see Fig.1, right). Since questions concerning individual knowledge remain constant, it is feasible to train a probing classifier based on the output of each hidden layer for tokens within the questions.

Specifically, for each token w within the given question, we extract the output representations $R_w^{(i)}$ from the i-th hidden layer, which serves as input to the probing classifier M_{probe}. The knowledge categories Y corresponding to different evidence are employed as labels for M_{probe}. By measuring the performance of M_{probe} in learning the mapping $R_w^{(i)} \rightarrow Y$, we can infer the extent to which the hidden layer encodes context knowledge.

However, the test set accuracy in probing tasks may be affected by the fitting capability of M_{probe} (Hewitt and Liang, 2019), thus rendering it an imprecise indicator of the hidden layer's capability for encoding context knowledge. Additionally, when the classifier accuracy is already substantially high, it becomes challenging to distinguish the dataset difficulty. As an illustration, the visual representation of different layers when processing the first token of the question *Are Labradors dogs?* in LLaMA 2 Chat 13B is illustrated in Fig. 2. Although different layers within LLaMA exhibit varying capabilities to distinguish between factual and counterfactual evidence, it remains challenging for the test set accuracy to capture these distinctions adequately.

Therefore, we adopt $\mathcal{V}$-usable information (Ethayarajh et al., 2022; Xu et al., 2020) rather than probing accuracy as the metric for measuring the capability to encode context knowledge. $\mathcal{V}$-usable information reflects the ease with which a model family $\mathcal{V}$ can predict the output Y given specific input $R_w^{(i)}$:

$$I_{\mathcal{V}}(R_w^{(i)} \rightarrow Y) = H_{\mathcal{V}}(Y) - H_{\mathcal{V}}(Y|R_w^{(i)}), \quad (1)$$

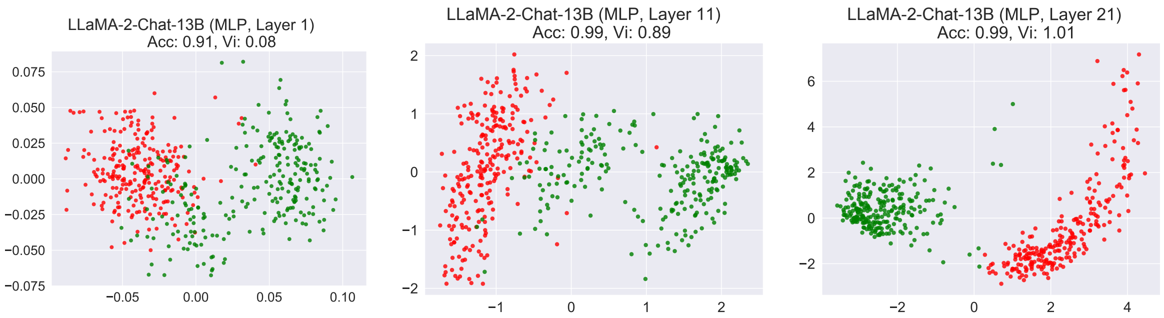


Figure 2: The principal component analysis (PCA) visualization of how context knowledge is processed across different layers in LLaMA 2 Chat 13B, where instances denoted green and red signify factual and counterfactual evidence, respectively. $\mathcal{V}$-usable information (Vi) is found to be more effective in distinguishing between dataset difficulty than accuracy (Acc).

where $H_{\mathcal{V}}(Y)$ and $H_{\mathcal{V}}(Y|R_w^{(i)})$ denote the predictive $\mathcal{V}$-entropy and the conditional $\mathcal{V}$-entropy, which can be estimated through the following equations:

$$H_{\mathcal{V}}(Y) = \inf_{f\in\mathcal{V}} \mathbb{E}[-\log_2 f[\varnothing](Y)], \tag{2}$$

$$H_{\mathcal{V}}(Y|R_w^{(i)}) = \inf_{f\in\mathcal{V}} \mathbb{E}[-\log_2 f[R_w^{(i)}](Y)], \tag{3}$$

where $\varnothing$ denotes an input devoid of any information, we employ zero-input vectors as an alternative representation. To minimize the influence of additional parameters, we constrain M_{probe} to linear classifier families when computing M_{probe}-usable information. Furthermore, we add 0.01 to the input before taking the logarithm to prevent highly anomalous values. Comparative results with test accuracy in Fig. 2 show that this validation metric exhibits a higher level of discriminative capability.

5. Experiments

5.1. Experimental Setup

Datasets We conduct our experiments based on the ConflictQA-popQA-gpt4 dataset (Xie et al., 2023). It constitutes a complementary extension of the entity-centric question-answering (QA) dataset popQA (Mallen et al., 2022a), which encompasses not only commonsense question-answer pairs and their associated popularity from Wikipedia but also incorporates parametric memories and counter-memories generated by GPT-4. We select 50 instances with both high popularity and consistency between parametric memories and ground truth. Subsequently, we generate 100 pieces of factual and counterfactual evidence for each fact based on parametric memory and counter-memory. For newly acquired knowledge which is discussed in Sec 5.4, we manually design a piece of knowledge that does not exist in parametric memory: *The password of the president's laptop is {password}*, and request `gpt-3.5-turbo-0613` to provide various evidence for different passwords. We randomly generate 4 distinct 10-character passwords and place them within the *{password}* placeholder. For each password, we generate 100 supporting pieces of evidence. For all settings, the temperature is set to 1.0 to generate diversity outputs.

Models We choose the open-access LLaMA 2 (Touvron et al., 2023) with 7, 13, and 70 billion parameters as the LLMs to be explained. During the probing phase, we adopt a linear classifier as our probing model to reduce extraneous interference. We use a batch size of 4, learning rate of 0.0001, Adam optimizer (Kingma and Ba, 2015), and 15 training epochs for all probing tasks. The ratio between training and test sets is 8:2. Unless otherwise stated, we conduct probing tasks on the Transformer layer output of each layer.

5.2. Preliminary Experiments: Can LLMs Retain Context Knowledge?

As a foundational step for subsequent experiments, it is imperative to ascertain whether the provided evidence has the capability to change answers generated by the LLM. We conduct an assessment of question-answering accuracy after providing different sets of evidence (Tab.1).

As can be seen in the table, the LLMs consistently provide correct answers when presented with factual evidence that aligns with the ground truth. Conversely, the accuracy drops approaching zero when the LLMs receive counterfactual evidence. These results indicate that the coherent and diverse evidence generated by ChatGPT can effectively alter the LLM's cognition of existing knowledge or facilitate the encoding of newly acquired knowledge.

Model Name	Conflicting				Newly Acquired	
	Factual ↑		Counterfactual ↓		Factual ↑	
	base	chat	base	chat	base	chat
LLaMA 2 7B	82.31	79.84	31.76	4.80	85.75	98.25
LLaMA 2 13B	83.82	80.60	17.27	6.08	91.50	98.25
LLaMA 2 70B	75.91	84.20	12.23	14.51	89.25	98.25

Table 1: Question-answering accuracy after providing different sets of evidence. In conflict scenarios, the average accuracy (%) is calculated after providing both factual and counterfactual evidence for various commonsense knowledge. In newly acquired scenarios, the 4-class accuracy is calculated based on the provided *{password}*.

However, we find that the base models maintain high accuracy even after providing counterfactual evidence, implying their low capability to encode knowledge through prompting. Therefore, the dataset we have constructed can more reliably induce the utilization of context knowledge in chat models, thereby enabling the exploration of the roles played by various layers within them. In subsequent experiments, we will conduct experiments mainly on chat models.

5.3. How Do LLMs Encode Conflicting Knowledge?

5.3.1. Case Study

Utilizing the probing method elucidated in Sec 4, we are able to obtain the layer-wise $\mathcal{V}$-information of each token in the LLM's processing of individual fact. We select the question *What is Mike Flanagan's occupation?* as a case study. We present the content of the factual and counterfactual evidence generated by ChatGPT, with the intermediary portions omitted (see Tab. 2). It can be seen that the generated content provides substantial support for the corresponding knowledge labels, even in cases of fictional content. Consequently, these texts serve as effective means to assess the capability of LLMs in encoding context knowledge. We generate 100 distinct instances of content for each knowledge label.

Then we investigate the output of the Transformer layers, MLP layers, and Attention layers separately on LLaMA 13B. We plot heatmaps depicting the layer-wise $\mathcal{V}$-information of LLaMA 13B while processing tokens in the given question (Fig. 3). It is evident that each component of LLaMA 13B exhibits significantly higher $\mathcal{V}$-information at upper layers, implying that context knowledge is encoded within their representations. For Transformer layer output and MLP layer output, most tokens maintain high and stable $\mathcal{V}$-information after 30 layers.

However, the results for Attention layer output appear relatively chaotic. This aligns with the observations made by Vig and Belinkov (2019). Multiple heads of self-attention may focus on distinct local or global information, thereby facilitating the transfer of information between token representations. This not only leads to the encoding of knowledge-related information in other tokens such as *What* and *'*, but also results in the propagation of irrelevant information within lower layers, causing a slight disruption in probing results.

Another intriguing discovery is that LLMs encode context knowledge to varying degrees within different tokens. For knowledge-related entity tokens such as *Mike* and *occupation*, LLMs achieve high $\mathcal{V}$-information at lower layers. This is not due to parametric memorization in LLMs, as simple parametric memorization would result in consistent behaviors when handling different external evidence, leading to lower $\mathcal{V}$-information. Therefore, the LLMs prioritize encoding context knowledge into knowledge-related entity tokens. In contrast, tokens that are not directly related to the knowledge such as *What* and *'*, encode relevant information later through the attention mechanism.

Validation of the broader applicability of these findings will be conducted in Sec 5.3.2 and 5.3.3.

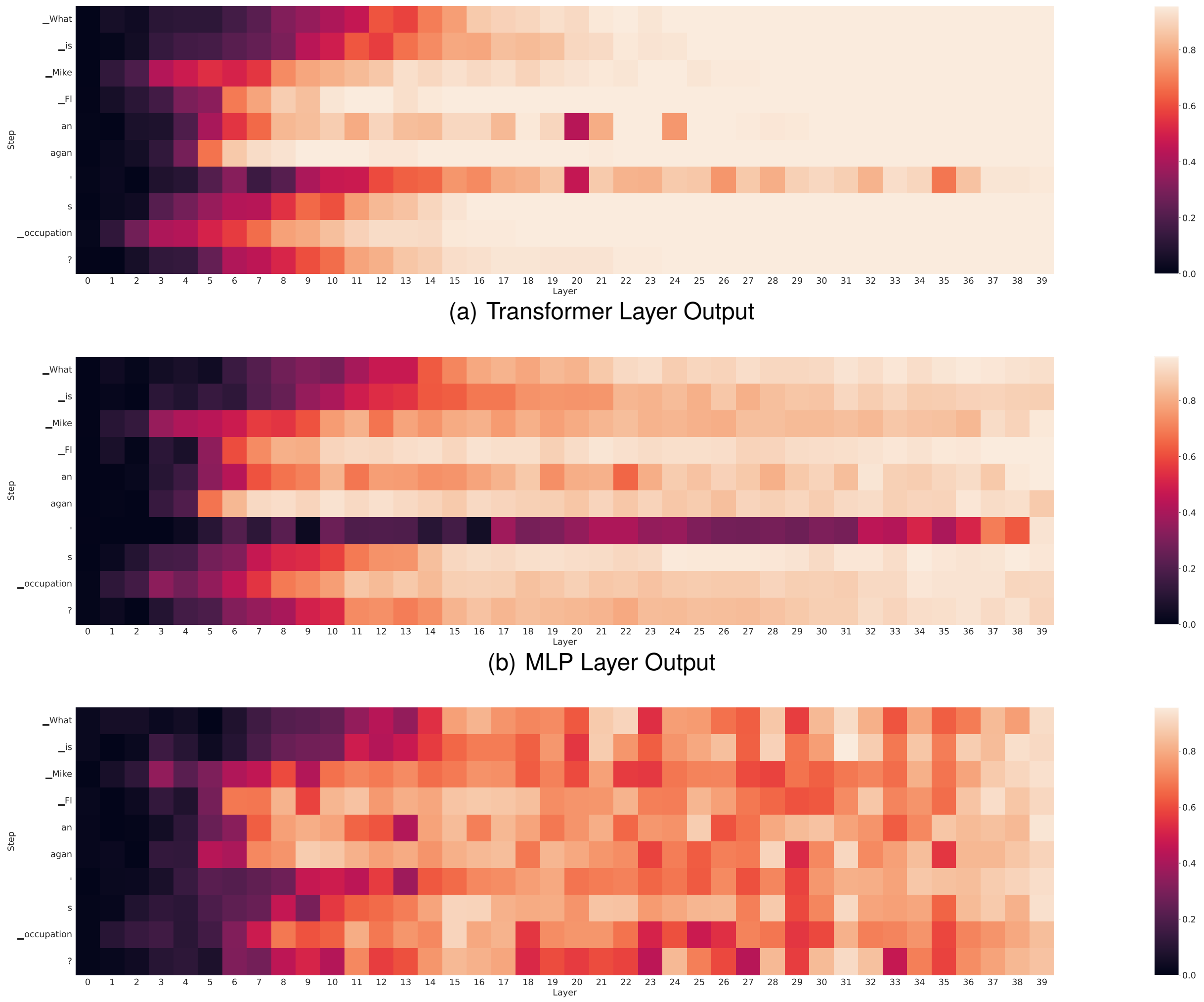
5.3.2. Average $\mathcal{V}$-information

We then conduct the probing tasks on all 50 facts and select the last token in the question to compute the average $\mathcal{V}$-information (Fig. 4). We observe that the $\mathcal{V}$-information gradually increases, especially in the early layers. This aligns with our conjecture that the LLM encodes more context knowledge within upper layers, enabling them to differentiate between facts and counterfactuals more effectively. Since the correct responses for facts and counterfactuals differ, the layers of LLMs tend to preserve differences accumulated from previous layers and capture more context knowledge.

Interestingly, the LLMs with fewer parameters (7B and 13B) reach high $\mathcal{V}$-information earlier, possibly as a result of their forced behavior to generate the correct final output. Therefore, the last few layers of these LLMs encode crucial information for the task. Conversely, the $\mathcal{V}$-information of LLaMA 70B exhibits a slow rate of increase and even demonstrates a decrease in the last few layers. On one hand, LLaMA 70B benefits from a sufficient number of layers to act as a buffer, eliminating the necessity to encode a substantial amount of context knowledge in the last few layers, thereby enhancing its robustness. On the other hand, the last few layers of LLaMA 70B may be employed to eliminate shortcuts in order to enhance contextual comprehension, leading to a marginal reduction in $\mathcal{V}$-information.

Factual	Mike Flanagan is widely recognized and celebrated for his impressive career as a film director, screenwriter, producer, and editor. With numerous accolades and critical acclaim, Flanagan has established himself as a prominent figure in the industry. **...** With his multifaceted skills and successful filmography, it is clear that Mike Flanagan is indeed a talented film director, screenwriter, producer, and editor.
Counterfactual	Mike Flanagan, a highly skilled and talented individual, has made outstanding contributions to the field of graphic design throughout his career. As a graphic designer, Mike has demonstrated exceptional proficiency in various design software, such as Adobe Illustrator and Photoshop. **...** It is undeniable that his vast experience and exceptional skills make him an invaluable asset in the field.

Table 2: Examples of the factual and counterfactual evidence for the case study *What is Mike Flanagan's occupation*. We omit the middle part of the content.



(a) Transformer Layer Output

(b) MLP Layer Output

(c) Attention Layer Output

Figure 3: The heatmap of probing results on LLaMA 2 Chat 13B. We select the question *What is Mike Flanagan's occupation?* as a case study and display the layer-wise $\mathcal{V}$-information for each token in the question.

5.3.3. Knowledge-Related Entity Tokens vs. Other Tokens

To verify the distinction encoded by the LLM between knowledge-related entity tokens and other tokens, we classify the tokens in all questions into two categories. Knowledge-related entity tokens are composed of the subjects and relations associated with knowledge. For example, in the question *What is London the capital of*, *London* and *capital* are knowledge-related entity tokens (Fig. 5). We sequentially conduct the probing task for each token in the 50 constructed questions, obtaining $\mathcal{V}$-information $I_{\mathcal{V}_e}$ for knowledge-related entity tokens and $I_{\mathcal{V}_n}$ for other tokens. The layer-wise differences in means between $I_{\mathcal{V}_e}$ and $I_{\mathcal{V}_n}$ are depicted in Fig. 6.

As can be seen in the figure, knowledge-related entity tokens exhibit significantly higher values of $\mathcal{V}$-

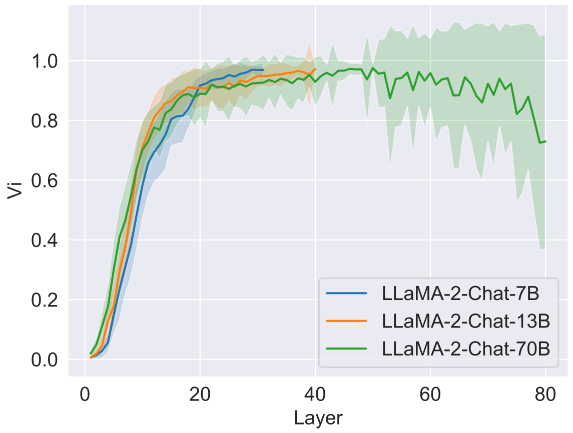


Figure 4: The average layer-wise $\mathcal{V}$-information (Vi) of the last token in the questions for each LLM.

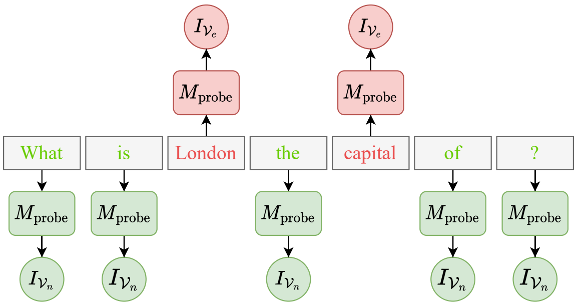


Figure 5: We categorize the subjects and relations mentioned in the questions as one class (red) while considering other tokens as another class (green). By comparing the differences of the average $\mathcal{V}$-information between these two classes, it is capable of detecting the LLM's level of attention to knowledge-related entity tokens.

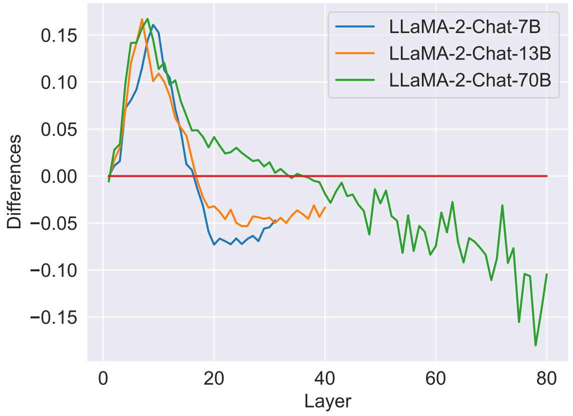


Figure 6: Layer-wise differences between the $\mathcal{V}$-information of knowledge-related entity tokens and other tokens.

information in comparison to other tokens at lower layers, indicating that the LLMs can encode context knowledge within these tokens more easily. Furthermore, since the provided context knowledge is closely related to these tokens, it is intuitive for LLMs to generate different representations earlier in processing these tokens.

To our surprise, the advantage gradually diminishes and is even surpassed by other tokens at upper layers. This implies that the LLMs progressively transfer relevant information to other tokens such as *Answer* and *:* to generate desired outputs. We contend that this phenomenon can be attributed to the self-attention mechanism. This is consistent with previous research that several meaningful knowledge tends to be encoded in specific contextual embeddings (Mohebbi et al., 2021). In order to provide accurate responses to questions following the token *Answer*, the LLMs may encode all the information within some knowledge-unrelated tokens. We present this phenomenon in the hope that it will attract more attention and research in the future.

5.4. How Do LLMs Encode Newly Acquired Knowledge?

5.4.1. Case Study

Previous experiments may have been influenced by the LLM's parameter knowledge, which may not entirely reflect the capability to process context knowledge. This section considers a scenario where the LLM can only provide correct answers through external evidence. We employ the experiment designed in Section 5.1, requiring the LLM to answer *The password of the president's laptop*, which can also serve as a means to assess the LLM's capacity to retain sensitive knowledge.

The heatmap for the layer output of LLaMA 2 Chat 13B is shown in Fig. 7. As the LLM has never been exposed to the knowledge before, the intermediate layers manifest lower $\mathcal{V}$-information. However, the behavior of the LLM in encoding newly acquired knowledge remains consistent with the findings outlined in the previous section. It exhibits a preference for encoding context knowledge within knowledge-related entity tokens such as *password* and progressively disseminating information to other tokens through the attention mechanism.

5.4.2. Long-Term Memory Capability

Since the newly acquired knowledge has never been exposed to the LLMs, it is reliable to apply it to test the long-term memory capability of LLMs. We provide n pieces of irrelevant evidence after the relevant evidence and then further probe the layer-wise $\mathcal{V}$-information of LLMs (Fig. 8).

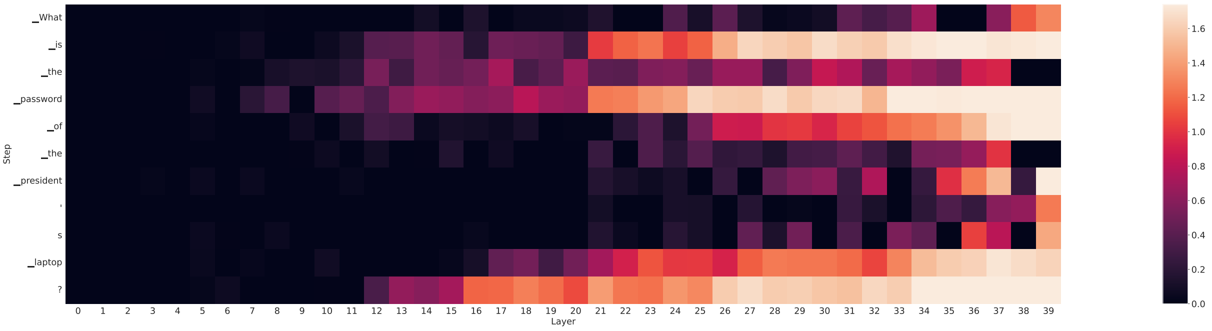
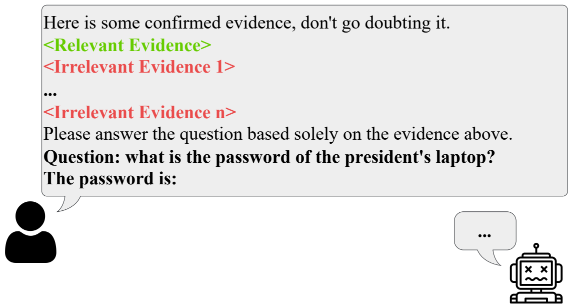


Figure 7: The heatmap of probing results for the question *What is the password of the president's laptop?* on LLaMA 2 Chat 13B.



Here is some confirmed evidence, don't go doubting it.
<Relevant Evidence>
<Irrelevant Evidence 1>
...
<Irrelevant Evidence n>
Please answer the question based solely on the evidence above.
Question: what is the password of the president's laptop?
The password is:

Figure 8: We test the long-term memory capability of LLMs by introducing irrelevant evidence.

Although the LLMs consistently provide correct answers to questions at nearly 100% accuracy, the layer-wise $\mathcal{V}$-information exhibits a gradual decline. Specifically, we calculate the average $\mathcal{V}$-information of the last token for LLaMA 2 Chat 70B every 5 layers when given varying numbers of irrelevant evidence (Tab. 3).

In the realm of intermediate layers, particularly within the range of 20 to 40 layers, the LLM exhibits a notably diminished level of $\mathcal{V}$-information in the presence of irrelevant evidence. While irrelevant evidence does not engender any form of misdirection, the LLM still sacrifices the problem-solving capability of the intermediate layers in its endeavor to encode the irrelevant information. This implies that LLMs have not encoded the irrelevant evidence orthogonally, thus causing interference with the knowledge that has already been encoded (such as the password of the president's laptop). We hope that future research will be directed toward enhancing the long-term memory capability of LLMs.

5.5. Ablation Study

5.5.1. Impact of Positional Encoding

Since the probing classifier for a given layer is trained on the mixed token representations from different sentence positions, it may be affected by positional encoding on the representations. We design an ablation experiment to analyze the effect of positional encoding as noise on the probing results. Specifically, we posit that special symbols such as "\n" do not contribute additional semantic information, which can be added after the provided evidence and before the question to modify the positional encoding of probing tokens.

We conduct experiments in the scenario of newly acquired knowledge. All external evidence provides the same password, with the addition of 0-3 "\n" symbols at the end of the evidence. We provide the four-category heatmap of LLaMA 2 Chat 7B in Fig. 9. It is observed that the layer-wise $\mathcal{V}$-information remains consistently low, closely resembling the results of random predictions, particularly in the higher layers. This suggests that positional encoding has minimal impact on the probing results and can be regarded as irrelevant noise in this context.

6. Conclusion

In this paper, we propose a novel framework for explaining the layer-wise capability of large language models in encoding context knowledge via the probing task. Our research addresses the previously overlooked aspect of how LLMs encode context knowledge layer by layer, shedding light on what has been considered black-box mechanisms. Leveraging the powerful generative capacity of ChatGPT, we construct probing datasets that encompass diverse and coherent evidence corresponding to various facts and utilize $\mathcal{V}$-information as the discriminative validation metric. Comprehensive experiments conducted on conflicting knowl-

Layer	0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79
#irr = 0	**-0.02**	**0.07**	**0.06**	**-0.01**	**0.04**	**0.25**	**0.75**	**1.08**	1.38	1.83	**1.87**	1.70	1.59	**1.64**	1.14	**0.95**
#irr = 1	-0.08	-0.07	-0.17	-0.12	0.00	-0.15	0.17	-0.11	0.75	1.78	**1.87**	**1.91**	1.30	1.15	**1.34**	0.60
#irr = 2	-0.08	-0.03	-0.10	-0.31	-0.45	-0.39	-0.37	-0.27	1.11	1.50	1.63	1.58	1.13	0.91	1.16	0.85
#irr = 3	-0.07	-0.05	-0.13	-0.22	-0.43	-0.44	-0.56	-0.53	1.63	1.88	1.51	1.58	**1.77**	1.39	1.11	0.61
#irr = 4	-0.05	-0.08	-0.12	-0.30	-0.22	-0.13	-0.24	-0.38	0.97	**1.89**	1.24	1.69	1.67	1.43	1.28	0.72
#irr = 5	-0.04	0.01	-0.08	-0.22	-0.45	-0.15	-0.33	-0.44	**1.47**	1.65	1.51	1.76	1.09	1.08	1.03	0.43

Table 3: The average $\mathcal{V}$-information of the last token for LLaMA 2 Chat 70B every 5 layers, where #irr denotes the number of provided irrelevant evidence.

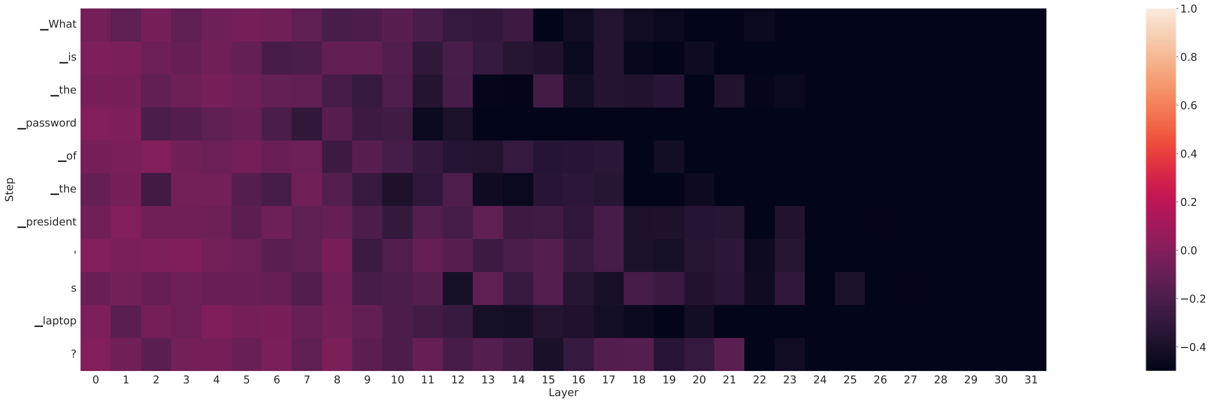


Figure 9: The heatmap of probing results for the impact of positional encoding on LLaMA 2 Chat 7B.

edge demonstrate that LLMs tend to exhibit a predilection for encoding context knowledge within upper layers and gradually transfer knowledge from knowledge-related entity tokens to other tokens as the layers deepen. Furthermore, while observing similar phenomena when provided with newly acquired knowledge, we also examine the long-term memory capacity of LLMs by introducing irrelevant evidence. Our findings indicate that the layer-wise retention of newly acquired knowledge gradually diminishes with the increase of irrelevant evidence. We hope that this work will serve as a catalyst for further research towards exploring the inner mechanisms of how LLMs encode such emergent capability.

7. Limitations

Although the proposed method provides a layer-wise explanation of LLMs and reveals some intriguing phenomena from comprehensive and reliable experiments, the theoretical mechanism driving the emergence of these phenomena still remains an open issue. For example, we observed that the LLMs encode more context knowledge within knowledge-related entity tokens at lower layers but transfer more knowledge to other tokens at upper layers in Sec 5.3.3. We speculate that the self-attention plays a crucial role during this process, although mathematical proof remains elusive. Recent research is shifting from the 1-layer to multi-layer Transformer, aiming to gain a mathematical understanding of the role played by layer-wise components such as self-attention (Tian et al., 2023a,b). We encourage future research to focus on the mathematical mechanisms contained behind these intriguing phenomena.

Moreover, due to space and time constraints, we only performed detailed experiments on LLaMA 2 (7B, 13B, and 70B), which ignored numerous SOTA open-accessed LLMs such as PaLM (Chowdhery et al., 2022), OPT (Zhang et al., 2022), and Pythia (Biderman et al., 2023). We encourage future research to conduct detailed experiments on more PLMs to detect the capability and tendency of different LLMs in encoding context knowledge.

8. Ethical Considerations

The aim of our proposed framework is to measure the layer-wise capability of LLMs in encoding context knowledge. However, there are several potential risks that should be carefully considered. One primary concern is that our study indirectly reflects the potential manipulability and insecurity of LLM-generated outputs. People may adopt a similar method to require LLMs such as ChatGPT to generate misleading evidence. Therefore, we emphasize the need for stricter scrutiny by relevant authorities regarding the applications of LLMs.

9. Acknowledgements

This work is partly supported by the Joint Funds of the National Key R&D Program of China No. 2023YF3303805 and the Joint Funds of the National Natural Science Foundation of China under No. U21B2020.

10. Bibliographical References

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona T. Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. *CoRR*, abs/2204.06031.

Zeyuan Allen-Zhu and Yuanzhi Li. 2023. Physics of language models: Part 3.2, knowledge manipulation. *CoRR*, abs/2309.14402.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Comput. Linguistics*, 48(1):207–219.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. `https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html`.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models. *CoRR*, abs/2307.03109.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single \$&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2126–2136. Association for Computational Linguistics.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *CoRR*, abs/2304.05335.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. 2023. Mathematical capabilities of chatgpt. *CoRR*, abs/2301.13867.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2733–2743. Association for Computational Linguistics.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *CoRR*, abs/2305.08322.

Nikhil Kandpal, H. Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2022. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering. *CoRR*, abs/2203.05115.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2022a. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. *CoRR*, abs/2212.10511.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2022b. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Annual Meeting of the Association for Computational Linguistics*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *NeurIPS*.

Hosein Mohebbi, Ali Modarressi, and Mohammad Taher Pilehvar. 2021. Exploring the role of BERT token representations to explain sentence probing results. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 792–806. Association for Computational Linguistics.

Reiichiro Nakano, Jacob Hilton, S. Arun Balaji, Jeff Wu, Ouyang Long, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *ArXiv*, abs/2112.09332.

OpenAI. 2022. Introducing chatgpt. `https://openai.com/blog/chatgpt`.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Miguel Ortega-Martín, Óscar García-Sierra, Alfonso Ardoiz, Jorge Álvarez, Juan Carlos Armenteros, and Adrián Alonso. 2023. Linguistic ambiguity analysis in chatgpt. *CoRR*, abs/2302.06426.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4609–4622. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *CoRR*, abs/2302.06476.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how BERT works. *Trans. Assoc. Comput. Linguistics*, 8:842–866.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yuandong Tian, Yiping Wang, Beidi Chen, and Simon S. Du. 2023a. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *CoRR*, abs/2305.16380.

Yuandong Tian, Yiping Wang, Zhenyu Zhang, Beidi Chen, and Simon Du. 2023b. Joma: Demystifying multilayer transformers via joint dynamics of mlp and attention.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 63–76. Association for Computational Linguistics.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023. Resolving knowledge conflicts in large language models.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge clashes. *CoRR*, abs/2305.13300.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8653–8665. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2023. Explainability for large language models: A survey. *CoRR*, abs/2309.01029.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. *CoRR*, abs/2303.11315.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, Zico Kolter, and Dan Hendrycks. 2023. Representation engineering: A top-down approach to ai transparency.