ROME: Memorization Insights from Text, Logits and Representation

Bo Li[1,4], Qinghua Zhao(✉)[2,3], and Lijie Wen(✉)[1]

[1] School of Software, Tsinghua University, Beijing, China
libo15@baidu.com, wenlj@tsinghua.edu.cn
[2] Beihang University, Beijing, China
zhaoqh@buaa.edu.cn
[3] University of Copenhagen, Copenhagen, Denmark
[4] Baidu Inc., Beijing, China

Abstract. Previous studies on model memorization have been limited by their reliance on comparing outputs with training corpora. This paper introduces a novel approach that avoids direct access to training data, named ROME [5]. Instead, it focuses on datasets where text chunks express fixed semantics, categorized into three types: context-independent, conventional, and factual. We redefine memorization as the ability to produce correct answers within these categories. Our analysis explores the contrasts in behavior patterns between memorized and non-memorized samples, focusing on differences in logits and representations of generated texts. Experimental results demonstrate that models consistently exhibit higher confidence when producing memorized answers.

Keywords: Large Language Models · Interpretability and Analysis · Memorization

1 Introduction

Memorization in the context of large language models (LLMs) commonly refers to the reproduction of text fragments from their training corpus [17,21,19,3]. As foundational components in various Natural Language Processing (NLP) tasks, LLMs highlight the critical need to examine their propensity for memorization. Our study concentrates on investigating memorization characteristics in billion-scale LLMs, such as LLaMA [20], Mistral [9] and Gemma [18]. Unlike existing work that quantifies memorization by directly comparing the model's outputs with its pre-training corpus, we employ a different approach. Given the immense size of these corpora and the intensive data processing requirements [3], we opt to bypass extensive corpus processing.

As in previous studies, we employ factual question answering datasets to explore memorization. Additionally, we extend our investigation to data that

[5] **ROME** refers to the four letters in "m**emor**ization", it also indicates "Rome (memorization) was not built in a day".

can be input and output as a whole. Based on the "chunk-based learning and generation" capability, we select datasets in two categories. The first category, which exhibits the strongest "chunk-based learning and generation" characteristics, includes idioms and Chinese poetry. These datasets are treated as standalone blocks that express fixed and context-independent semantics. The second category comprises datasets containing conventional concepts, such as proper nouns and terminologies, which are broadly defined and universally recognized.

A sample is categorized as *memorized* if it is completed or answered correctly; otherwise, it is deemed *non-memorized*. With memorized and non-memorized groups established, we conduct comparative analyses using statistical and visualization methods to unveil changes in the models' behavioral patterns.

The main contributions are outlined below:

- We define memorization based on datasets exhibiting "chunk-based learning and generation" characteristics, considering correct answers or completions as memorized (see Section 3 for a further detailed discussion).
- We analyze memorization based on text content, logits, and representations and compare these features between memorized and non-memorized samples.
- Our experimental results demonstrate that (I) longer prompts increase the likelihood of memorization, while longer words decrease it; (II) higher memorization correlates with greater confidence and accuracy; and (III) memorized and non-memorized samples have different representations, with similar concepts being more similar.

2 Related Work

Memorization was first identified by [17] in GPT-2 and has since been confirmed in other models such as LLaMA and GPT-3.5 [10]. This presents significant privacy risks, as illustrated by [5], who demonstrated GPT-2's ability to retrieve sensitive personal information including names, email addresses, phone numbers, and addresses. Such capabilities underscore the potential for privacy intrusions and attacks [4,19,3]. Originally, memorization was characterized by how well the generated sequences aligned with the actual continuations in the training data, employing different n-gram levels [13]. While these studies primarily address verbatim memorization, to capture more complex and semantically rich forms of repetition, [21,6] defined memorization as the performance variance of a training instance when included or excluded from the training set. Recent studies have illuminated various factors influencing memorization in LLMs. Researchers have found that qualitatively distinct yet simple examples [21], atypical samples [6], and specific content types like nouns and numbers [19] are more prone to memorization. Model and training characteristics, such as lower temperature, larger model size [14], subword vocabulary size [12], increased model capacity, data duplication, and longer prompting contexts [3], all enhance memorization. The process itself has been described as a two-step mechanism involving early and upper layers of the model [8]. Access to training datasets is often limited due to their non-public nature or large volume [3]. For instance, generating 100 tokens

per second on a V100 GPU, the 6B parameter GPT-Neo model would require over 30 GPU-years to process an 800GB training dataset [3]. Consequently, some researchers have developed methods to investigate memorization without direct access to the training data. [14] treated the model as a black box, analyzing outputs to infer memorization patterns without needing ground-truth data.

Building upon these methodologies, our work advances the study of memorization. Similar to [14,8], we also avoid directly accessing training data. Instead, we use unique datasets with definitive answers, categorizing each sample into *memorized* and *non-memorized* based on whether it is answered correctly. Rather than focusing on the factors that influence memorization, we delve into the model's behavioral patterns during memorization by examining the differences between these two categories.

3 Preliminaries

The following contents will detail the definition of memorization, describe the "chunk-based learning and generation" characteristics, outline the selected datasets.

Definition of memorization. Considering a text segment S and its context C, we represent this as $S + C = \{w_1, \cdots, w_{k-1}, \underline{w_k, \cdots, w_i, w_n}, w_{n+1}, \cdots, w_m\}$, where $S = \{w_k, \cdots, w_i, w_n\}$ denotes the text segment and $C_l = \{w_1, \cdots, w_{k-1}\}$ and $C_r = \{w_{n+1}, \cdots, w_m\}$ are its contexts. For question-answering problem, the sample is categorized as *memorized* if the model answers it correctly. For completion task, given $S' = \{w_k, \cdots, \_, w_n\}$, if the model can correctly predict w_i for $(k \leq i \leq n)$, the sample is categorized as *memorized*; otherwise, it is *non-memorized*.

Chunk-based learning and generation. For datasets like IDIOM and ProperNoun, LLMs often produce correct outputs because they have encountered these specific text chunks during pre-training. In other words, to generate a particular text chunk as output, that chunk must have been included in the pre-training corpus. We refer to this characteristic as "chunk-based learning and generation". For instance, the idiom "hit the nail on the head" and the Chinese poem "千山鸟飞绝 (from hill to hill no bird in flight)" both exemplify this feature.

4 Methodology

4.1 Datasets

In order to be able to "chunk-based learning and generation", we have selected two categories of datasets (e.g., context-independent and conventional datasets). We also selected three datasets for factual question-answering.

Context-independent. This category includes idiom and Chinese poetry datasets, which are particularly suited for the "chunk-based learning and generation" approach. The semantics of these datasets are usually pertinent to a specific culture

and feature definitive answers that remain invariant across different contexts and models, thus making them context-independent.

IDIOM [8] is a collection of English idioms designed to assess memorization . It comprises 850 samples, with an average of 4.9 words per sample.

TangPoetry comprises ancient Chinese poems collected from the "Three Hundred Tang Poems" a poetry anthology from the Tang Dynasty. We have selectively retained the main body of five-character quatrains and seven-character quatrains, ultimately obtaining approximately 1500 samples.

Conventional. The second category includes datasets of proper nouns and disease terminologies. These datasets are typically certified by authoritative institutions to reduce ambiguities in communication, ensuring that each concept or entity is consistently associated with a unique text fragment.

ProperNoun comprises a collection of proper nouns (e.g., Royal College of Art), generated with the assistance of artificial intelligence tools. It obtains approximately 300 proper nouns.

Terminology was also generated from GPT-4 consisting of disease terminologies. It retains about 200 terminologies.

Factual. In addition to the datasets specifically designed for "chunk-based learning and generation" characteristics, we also test several factual question answering datasets previously used to study memorization [8], such as PopQA, LAMA-UHN and CelebrityParent.

PopQA [15] is an English question answering dataset, consisting of questions about someone's occupation, birthplace, genre, and capital, among others. It is created by converting knowledge tuples retrieved from Wikidata using templates and contains about 600 questions.

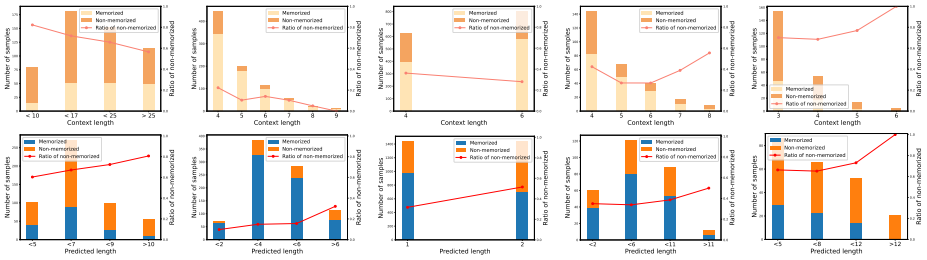
LAMA-UHN [11] is a subset of LAMA [16], from which easily guessable examples have been filtered out. It consists of about 500 samples.

CelebrityParent [2] is an English reversal relation dataset, It contains the top 1000 most popular celebrities from IMDB (2023) and their parents. The dataset includes about 1500 child-parent pairs.

4.2 Insights

As is widely recognized, in the generation process of GPT-style generative models, the model first calculates a representation based on previous tokens. It then uses this representation to compute the distribution over the vocabulary (i.e., logits). The next token is subsequently generated using decoding strategies [17]. Therefore, in addition to analyzing the most commonly used text, we also examine the logits and representations.

Text. Attributes associated with text, such as word frequency, context length, and part-of-speech, have been extensively explored. To align with previous research, we re-examine the context length and predicted length under our defined



(a) LAMA-UHN (b) IDIOM (c) TangPoetry (d) ProperNoun (e) Terminology

Fig. 1: Comparison of memorized vs non-memorized instances: context and predicted length across datasets.

memorization. Context length refers to the number of words in a context (excluding exemplars), while predicted length refers to the number of characters in the word to be predicted.

Logits and Representation. If text represents the returned results, then logits can be considered the basis for this result's decision-making. Analyzing it enables us to comprehend the model's generative biases in *memorized* versus *non-memorized*. The representation of next token carries the model's understanding of the current context, and analyzing it can further enhance our understanding of the model's behavior.

5 Experimental Analysis

Parameter Settings. We conduct experiments using various model configurations including LLaMA-2 (7B, 13B), Mistral 7B, Gemma 7B, and LLaMA-3 8B. These models are loaded onto two V100 GPUs, each with 32GB of memory, and operated in float32 precision. During the decoding phase, we employ a greedy decoding strategy [7] that selects the token with the highest probability.

5.1 Text-oriented Analysis.

To investigate whether there are distinct disparities between memorized and non-memorized samples concerning context length and predicted length, we analyze the number and proportion of memorized versus non-memorized samples. Given the consistency of experimental results across different models, we present data only from LLaMA-2 13B for English datasets and Qwen-1.5 32B [1] for the Chinese dataset, as shown in Figure 1.

Longer prompt more memorized? We investigated the relationship between context length and memorization likelihood across different datasets. Our findings reveal two distinct patterns:

Table 1: Comparison of averaged probability of generated tokens between *memorized* and *non-memorized*.

Models	Datasets	Context-independent		Conventional		Factual		
		IDIOM	TangPoetry	ProperNoun	Terminology	CelebrityParent	PopQA	LAMA-UHN
Mistral 7B	*memorized*	**0.5777**	–	**0.5833**	**0.6669**	**0.8446**	**0.8343**	**0.6938**
	non-memorized	0.5373	–	0.4918	0.5891	0.6184	0.5895	0.4954
Gemma 7B	*memorized*	**0.4051**	–	**0.7502**	**0.6768**	**0.8194**	**0.8134**	**0.6655**
	non-memorized	0.2570	–	0.6528	0.5435	0.6616	0.5418	0.5391
LLaMA-2 7B/	*memorized*	**0.3968**	**0.8206**	**0.6314**	**0.6735**	**0.8568**	**0.8140**	**0.6588**
Qwen-1.5 7B	*non-memorized*	0.3098	0.3072	0.3298	0.5149	0.6502	0.6136	0.3369
LLaMA-2 13B/	*memorized*	**0.4254**	**0.8376**	**0.7381**	**0.7646**	**0.8371**	**0.8362**	**0.6789**
Qwen-1.5 14B	*non-memorized*	0.3279	0.3180	0.3780	0.5489	0.6787	0.5386	0.4759
LLaMA-3 8B/	*memorized*	**0.3607**	**0.8760**	**0.746**	**0.6917**	**0.8131**	**0.8295**	**0.6770**
Qwen-1.5 32B	*non-memorized*	0.2816	0.3876	0.5421	0.6200	0.6835	0.5725	0.5693

1. Decreasing non-memorization with longer contexts. For these datasets, where text blocks convey context-independent semantics, longer contexts provide more prior information, facilitating recall of memorized content.
2. Non-memorization rates initially decrease but then increase with context length. In these datasets, which express proprietary concepts or entities, extremely short or long contexts may complicate recall due to the increased internal diversity of potential substitute words.

Longer word less memorized. The second analysis explores the relationship between predicted length and the likelihood of memorization. We observe a linear trend where, as the number of characters in a word increases, the ratio of non-memorized samples also continuously increases. This pattern likely arises from the increased complexity associated with longer words, which results in lower memorization rates.

5.2 Logits-oriented Analysis.

To examine whether there is a significant change in the distribution of logits for the next token generated by the model between memorized and non-memorized samples, we record the normalized values of logits, i.e., the probability for each generated token, as shown in Table 1. For Chinese dataset (i.e., TangPoetry) Qwen-1.5 models are used.

Higher memorized greater confidence. We notice a marked contrast between the memorized and non-memorized groups, wherein memorized samples typically exhibit higher probability scores. For instance, in the PopQA, the LLaMA-2 13B model shows average probabilities of 0.8362 for memorized instances compared to 0.5386 for non-memorized instances. This suggests that the model displays higher confidence in its correct predictions (i.e., *memorized*).

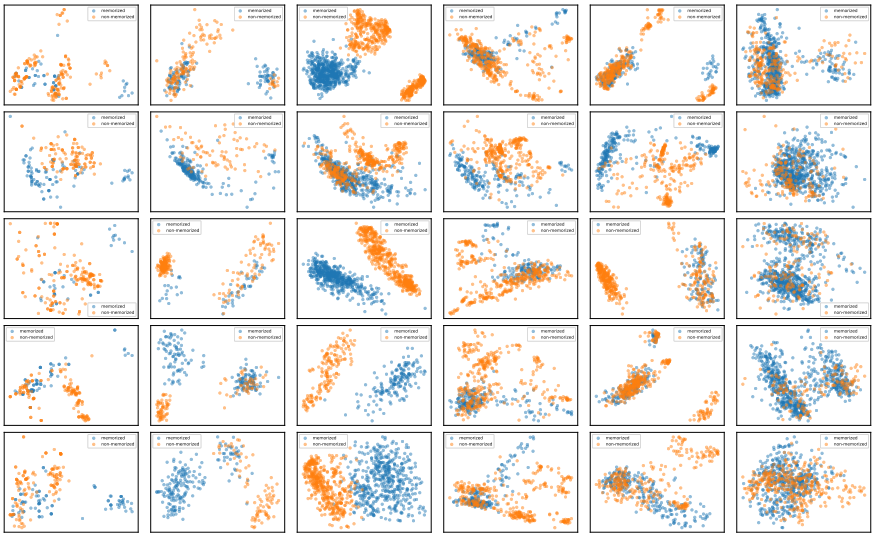


Fig. 3: PCA of representations for *memorized* and *non-memorized* samples across datasets and models. Datasets (left to right): Terminology, ProperNoun, CelebrityParent, PopQA, LAMA-UHN, and IDIOM. Models (top to bottom): Mistral 7B, Gemma 7B, LLaMA-2 7B, LLaMA-2 13B, and LLaMA-3 8B.

Higher probability higher accuracy. Table 1 illustrates that the normalized logits values for memorized samples are significantly higher. Based on this observation, we investigate whether higher probability values correlate with increased accuracy in generation. In other words, we explore if there is a consistent trend between model performance and probability. To this end, we record both the accuracy and probability values of all tested models across various datasets, as illustrated in Figure 2. Results are consistent across all English datasets; however, for clarity, only four datasets are presented here. The data reveal a consistent correlation between accuracy and probability, suggesting that as accuracy increases, so does the probability, and vice versa. Further analysis reveals that as the probability values increase, there is a tendency for them to cluster, indicating a concentration of higher probability values. In contrast, as the probability values decrease, they tend to diverge, suggesting that the lower probability values are more evenly distributed across the datasets. This pattern

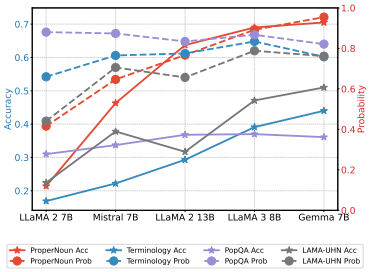


Fig. 2: Comparison between *probability* and accuracy across all tested models and datasets.

further supports the notion that higher probabilities are associated with more consistent and accurate model predictions.

5.3 Representation-oriented Analysis.

In this section, we explore the representations from both memorized and non-memorized samples.

Separated representations between memorized and non-memorized. Figure 3 provides a visualization using PCA dimensionality reduction of vectors for both memorized (blue points) and non-memorized (yellow points) groups. Notably, while there is some overlap between the two groups, they are generally distinctly separated in most cases, suggesting that their representations in the vector space are relatively unique. This distinction is consistently observed across different models. However, significant variations are evident between different datasets. Furthermore, the probability values in this dataset are generally low and the difference in the probability values between the two groups is minimal (as shown in Table 1). It is hypothesized that these two phenomena are caused by the same factors.

Same concepts are more similar. We investigated whether representations of the same concept are more similar across different contexts than representations of different concepts within the same context. Our analysis focused on the CelebrityParent dataset. In our methodology, we defined terms based on the query structure. For instance, in "Who is *Elon Musk*'s mother?" with the response "Maye Musk‘', we labeled "Elon Musk" as the "Context Child" and "Maye Musk" as the "Generated Parent‘'.

Conversely, for "Name a child of Maye Musk‘', "Maye Musk" was designated as the "Context Parent‘'. We quantified similarities using cosine similarity, a standard metric for vector comparisons.

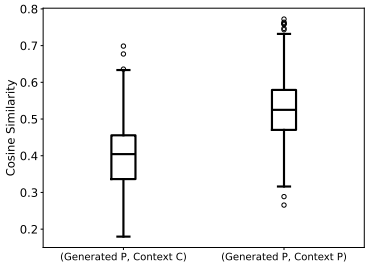


Fig. 4: Similarity comparison between (Generated Parent, Context Child) and (Generated Parent, Context Parent) representations.

Figure 4 illustrates our findings for the LLaMA-2 13B model (chosen for clarity and space considerations). The results demonstrate that representations of the same concepts are more similar to each other than to different concepts in similar contexts, suggesting that memorization, rather than context, plays a more significant role in shaping these representations. This discrepancy in performance for names with equal word frequency indicates that memorization capabilities are not solely determined by a concept's frequency of occurrence. These findings collectively suggest a complex relationship between memorization, context, and concept representation in large language models.

6 Conclusion

We developed ROME to explore the memorization processes in LLMs, focusing on a comparative analysis between memorized and non-memorized groups rather than directly accessing the training data. To ensure that the memorized group aligns more closely with the conventional understanding of memorization, we carefully selected datasets that exhibit the "chunk-based learning and generation" characteristics. In conjunction with factual question answering datasets, we investigated the model behavior patterns of these groups across seven datasets, examining dimensions such as text, logits, and representation. This study has uncovered several intriguing findings, such as models produce higher confidence for memorized samples.

Acknowledgments

This work was partially supported by the National Key Research and Development Program of China (No. 2024YFB3309702), and Hefei College Talent Research Fund Project (No. 24RC20, 21-22RC13), and Natural Science Foundation of the AnHui Higher Education Institutions of China (No. 2022AH051779).

References

1. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
2. Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A.C., Korbak, T., Evans, O.: The reversal curse: Llms trained on" a is b" fail to learn" b is a". arXiv preprint arXiv:2309.12288 (2023)
3. Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., Zhang, C.: Quantifying memorization across neural language models. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=TatRHT_1cK
4. Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., Song, D.: The secret sharer: Evaluating and testing unintended memorization in neural networks. In: 28th USENIX Security Symposium (USENIX Security 19). pp. 267–284 (2019)
5. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al.: Extracting training data from large language models. In: 30th USENIX Security Symposium (USENIX Security 21). pp. 2633–2650 (2021)
6. Chen, X., Li, L., Zhang, N., Liang, X., Deng, S., Tan, C., Huang, F., Si, L., Chen, H.: Decoupling knowledge from memorization: Retrieval-augmented prompt learning. Advances in Neural Information Processing Systems **35**, 23908–23922 (2022)
7. Germann, U.: Greedy decoding for statistical machine translation in almost linear time. In: Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. pp. 72–79 (2003), https://aclanthology.org/N03-1010

8. Haviv, A., Cohen, I., Gidron, J., Schuster, R., Goldberg, Y., Geva, M.: Understanding transformer memorization recall through idioms. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 248–264 (2023)
9. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)
10. Karamolegkou, A., Li, J., Zhou, L., Søgaard, A.: Copyright violations and large language models. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 7403–7412. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.458, https://aclanthology.org/2023.emnlp-main.458
11. Kassner, N., Dufter, P., Schütze, H.: Multilingual lama: Investigating knowledge in multilingual pretrained language models. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 3250–3258 (2021)
12. Kharitonov, E., Baroni, M., Hupkes, D.: How bpe affects memorization in transformers. arXiv preprint arXiv:2110.02782 (2021)
13. Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., Carlini, N.: Deduplicating training data makes language models better. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 8424–8445 (2022)
14. Levy, S., Saxon, M., Wang, W.Y.: Investigating memorization of conspiracy theories in text generation. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 4718–4729 (2021)
15. Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., Hajishirzi, H.: When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 9802–9822 (2023)
16. Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., Miller, A.: Language models as knowledge bases? 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). pp. 2463–2473 (2019)
17. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
18. Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M.S., Love, J., et al.: Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295 (2024)
19. Tirumala, K., Markosyan, A., Zettlemoyer, L., Aghajanyan, A.: Memorization without overfitting: Analyzing the training dynamics of large language models. Advances in Neural Information Processing Systems **35**, 38274–38290 (2022)
20. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
21. Zhang, C., Ippolito, D., Lee, K., Jagielski, M., Tramèr, F., Carlini, N.: Counterfactual memorization in neural language models. arXiv preprint arXiv:2112.12938 (2021)