Investigating the Impact of Temperature on Memorization in Meta's Llama3 Models

Amir Hammoud

Student Name: Amir Hammoud

Student Number: 6165354

First Examiner: Antal van den Bosch

Second Examiner: Joris Veerbeek

Programme: Master Applied Data Science

Date: June 30th, 2024

Utrecht University

Table of Contents

Abstract

Recent advancements in large language models (LLMs) such as Meta's Llama3 have substantially improved machine capabilities in language understanding and generation. However, the extent to which these models memorize versus generalize from their training data remains a critical issue, particularly with privacy and originality concerns. This study evaluates the memorization behavior of two differently-sized Llama3 models (8-billion and 70-billion parameters) across various temperature settings. Testing on Wikipedia data, we analyzed model responses to measure exact text reproduction, employing a mixed-effects model to assess the impact of model size and temperature. Our findings indicate that lower temperatures significantly increase memorization, with the 70-billion parameter model exhibiting a higher propensity for memorization than the 8-billion parameter model. Additionally, the larger model showed greater sensitivity to temperature changes, displaying a sharp decline in memorization at higher temperatures. These results underscore the importance of optimizing temperature settings to balance memorization and creativity in LLMs, and provide concrete evidence of the existence of memorization in these models. Future research should aim to develop robust guidelines and safeguards to mitigate the dual risks of privacy breaches and model exploitation.

1. Introduction

Recent advancements in artificial intelligence (AI), particularly in the development of large language models (LLMs), have significantly enhanced the capacity of machines to understand, generate, and interact with human language. These models, typified by architectures such as GPT (Generative Pre-trained Transformer) and its derivatives, are trained on extensive corpora of textual data sourced from diverse domains. While these models achieve new benchmarks in language comprehension and generation, they also raise important questions regarding their learning mechanisms—specifically, the extent to which they memorize versus generalize from their training data.

Memorization in LLMs is a double-edged sword. While on the one hand, the ability of a model to recall specific information can be advantageous, excessive memorization poses privacy risks, potentially leading to the regurgitation of sensitive information. Furthermore, the tendency of these models to replicate training data verbatim can impact their ability to generate novel, contextually appropriate responses. Thus, understanding the memorization behavior of LLMs is crucial for both mitigating potential risks and improving model design and functionality.

The collection methods used to obtain training data for these models have also come under scrutiny. AI companies have often relied on large-scale web scraping to gather the diverse datasets needed to train their models, a practice that raises significant legal and ethical concerns. Not all data available online is permissible for commercial use, especially without explicit consent from the original content owners. This practice has led to several legal challenges against prominent AI companies, alleging unauthorized use of copyrighted material and violation of privacy laws (Meeus et al., 2024)[1,2,3]. These lawsuits have not only legal implications but also highlight the critical need for ethical considerations in data handling practices in AI development.

This research focuses on evaluating the memorization capabilities of LLMs, particularly through the lens of how different model sizes and temperature configurations influence the propensity to memorize text verbatim. Our study utilizes Meta's two Llama3

[1] https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html [Accessed: Jun. 19, 2024]
[2] https://www.cnbc.com/2024/04/30/eight-newspaper-publishers-sue-openai-over-copyright-infringement.html [Accessed: Jun. 19, 2024]
[3] https://www.reuters.com/technology/meta-used-copyrighted-books-ai-training-despite-its-own-lawyers-warnings-authors-2023-12-12/ [Accessed: Jun. 19, 2024]

model variants, which are akin to other GPT-style models but vary in size and computational complexity. By examining how these variations affect memorization, this research contributes to the broader understanding of the dynamics between model architecture and learning behavior.

Moreover, this study innovatively explores the impact of generative temperature—a parameter that influences the randomness of the responses generated by these models—on memorization. Previous studies that investigated memorization on LLMs have predominantly focused on static configurations of LLMs, neglecting the nuanced effects that operational parameters such as temperature can have on model output and behavior.

By systematically analyzing how model size and temperature settings affect the exactness of text reproduction in LLMs, this thesis aims to provide insights that could guide the development of more robust, ethical, and effective LLMs. The findings of this research are expected to inform AI developers and researchers in improving model design to balance between creativity and memorization, thereby enhancing the utility and safety of LLM applications in diverse fields.

This study not only addresses a gap in the current understanding of LLM behavior but also aligns with the broader discourse on responsible AI development, emphasizing the need for transparency and accountability in AI training and deployment processes. Through a detailed empirical analysis, this study aims to delineate the boundaries of memorization in LLMs, facilitating a deeper understanding of their operational mechanisms and their implications for both AI ethics and functionality.

1.1. Literature Review

This literature review examines existing research on training data memorization in LLMs, the impact of model size and temperature on memorization, and the importance of these factors for model performance and ethical considerations.

1.1.1. Previous Research on LLM Training Data Memorization

Language model data extraction, which is the extraction of training data from a pretrained language model, is a method for retrieving the examples that were used to train the model. Carlini et al. (2021) were pioneers in defining the concepts of model knowledge extraction and κ-eidetic memorization. They also introduced effective training strategies aimed at

improving data extraction, defining κ-eidetic memorization as the capacity of a language model to memorize a string that appears κ times in the training material.

Research on LLM memorization has highlighted that larger models tend to memorize more of their training data. While larger models, such as those with billions of parameters, do require longer training times, they demonstrate a greater capacity for memorization compared to smaller models. This capacity allows them to retain a larger portion of the data before overfitting occurs (Tirumala et al., 2022). This trend is observed across various configurations and dataset sizes. 2022 specifically address why larger models tend to memorize more quickly. Their findings suggest that while memorization is typically associated with overfitting, larger models can memorize a significant portion of the training data before overfitting is detected. This indicates that overfitting alone does not fully explain the memorization dynamics as model scale increases (Tirumala et al., 2022). The ability to memorize specific information can be beneficial for tasks requiring precise recall, but it also poses risks, such as unintended regurgitation of sensitive data, making the understanding of memorization dynamics crucial (Hartmann et al., 2023).

Carlini et al. (2021) demonstrated that exact memorization, where models reproduce verbatim training data, is prevalent in large models and can be detected even with black-box access. Black-box access refers to a scenario where the internal workings of the model are not visible or accessible to the user; instead, the user can only interact with the model by inputting queries and observing the outputs. This memorization is often linked to the frequency of data repetition during training.

1.1.2. Effect of Temperature and Model Size on Memorization

The impact of temperature on LLM outputs has been a subject of investigation. Temperature controls the randomness of predictions; lower temperatures result in more deterministic outputs, while higher temperatures increase variability. Therefore, a lower temperature is more likely to lead to memorization because the model's outputs are less random and more closely aligned with the training data, increasing the likelihood of reproducing exact sequences from its training data. Understanding this parameter's influence on memorization is crucial, as it affects the likelihood of a model reproducing exact sequences from its training data.

Meeus et al. (2024) addressed the importance of temperature settings in their study on copyright traps, showing that different temperature settings can affect the detectability of

memorized sequences. Their findings imply that manipulating temperature could either mitigate or exacerbate memorization.

Yu et al. (2023) discuss how adjusting the temperature can significantly impact the extraction of training data from LLMs. They found that as the temperature is increased, the number of generated suffixes required to include the ground truth also increases, which can degrade efficiency. Therefore, finding a balance between diversity (induced by higher temperatures) and the risk of memorization (higher at lower temperatures) is essential. Their results indicate that lower temperatures tend to produce outputs that are more likely to include memorized sequences from the training data, underscoring the importance of this parameter in studies of model memorization.

The relationship between model size and memorization is more extensively documented. Larger models, such as those exceeding 10 billion parameters, are shown to memorize a greater fraction of their training data (Nasr et al., 2023). This is partly due to the increased capacity and complexity of these models, which allows them to store and recall more information.

1.1.3. Importance of Investigating Temperature and Model Size Effects on Memorization

Understanding the effects of temperature and model size on memorization is necessary as it informs the design of safer and more efficient LLMs. Research by Carlini et al. (2021) and others has underscored the potential for misuse if models excessively memorize training data, highlighting the need for stringent controls and ethical guidelines.

The significance of temperature and model size extends beyond memorization. These parameters also influence model performance and the trade-offs between accuracy and creativity in generated outputs. Studies have shown that higher temperatures can enhance creativity by making outputs less predictable, which is beneficial for creative tasks but can lead to less reliable responses.

Moreover, the ongoing legal and ethical debates surrounding the use of training data emphasize the need for rigorous research in this area. Meeus et al. (2024) highlighted the legal implications of using copyrighted material in training datasets without proper authorization. This has spurred the development of techniques to detect whether specific data was used in training (Carlini et al., 2022), which is necessary for maintaining transparency and accountability in AI development.

2. Data

2.1. Data Collection and Preprocessing

The dataset used in this study originates from a preprocessed Wikipedia data dump available on Hugging Face[4]. This dataset comprises cleaned articles from Wikipedia, stripped of markdown and unwanted sections such as references, ensuring a more focused and relevant collection of text for analysis. The dataset contains cleaned articles of multiple languages, however, this study only used English data. The data is stored in parquet files, each containing a substantial amount of information. For this study, we used only the first parquet file of the preprocesses English data, named *train-00000-of-00041.parquet*. The specific parquet file that was used contains 157,529 documents and over 280 million words.

2.2. Duplicate Detection

The detection of duplicate text snippets was a critical preprocessing step, aimed at identifying segments of text that appear multiple times within the dataset. This process utilized code from the Google Research repository *deduplicate-text-datasets*, which is designed for the removal of duplicate text in training datasets for LLMs[5]. However, in this study, the code was repurposed to detect duplicates instead of removing them. The method is described in the paper by Lee et al. (2021). We identified duplicate text snippets within the extracted Wikipedia articles. The detection mechanism involved analyzing text sequences and flagging duplicates based on predefined criteria (Lee et al., 2021). Each detected duplicate was counted to determine its frequency within the dataset. Only those duplicates appearing more than five times were considered significant for further analysis. 5,038 text duplicates were identified in the parquet file that was used.

Due to the large size of the original duplicate file, it was split into a subset containing 500 duplicates. The selection of 500 duplicates was arbitrary but intended to balance the thoroughness of the analysis with the practical constraints of computational resources and time. The most frequent duplicate had a frequence of 28 (Figure 1).

[4] https://huggingface.co/datasets/wikipedia [Accessed: Apr. 22, 2024]
[5] https://github.com/google-research/deduplicate-text-datasets [Accessed: Apr. 22, 2024]

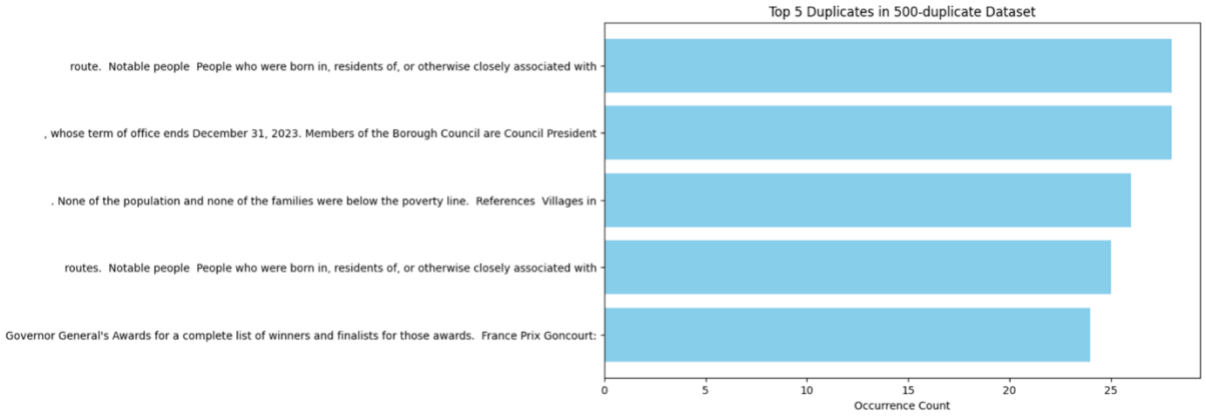

Figure 1. Top 5 most frequent duplicate text snippets in the 500-duplicate dataset.

2.3. Ethical and Legal Considerations

The dataset contains the content of full Wikipedia articles with cleaning to strip markdown and unwanted sections (references, etc.). Wikipedia is a well-known and publicly accessible platform, and the data from it is considered to be open for use. Wikipedia grants permission to copy, distribute and/or modify Wikipedia's text under the Creative Commons Attribution-ShareAlike License (CC BY-SA) and the GNU Free Documentation License (GFDL)[6].

[6] https://en.wikipedia.org/wiki/Wikipedia:Copyrights [Accessed: Apr. 22, 2024]

3. Methods

3.1. Evaluation of Text Memorization in LLMs

We evaluated how two differently-sized Llama3 models memorize parts of a given text by understanding the extent to which these models can reproduce segments of text verbatim, which serves as an indicator of memorization. This process involves generating responses from the models based on prompts derived from the extracted text duplicates and subsequently assessing these responses for similarities with the expected remaining parts of the text. Prompts were generated by selecting random starting points within the duplicate texts that were extracted earlier, and using these segments to query the LLM.

3.2. Model Setup

A series of two Llama3 models of varying sizes, specifically the 8-billion and 70-billion parameter variants, were used to assess memorization across different model capacities. Each model was queried using a set of text prompt, with responses generated under varying levels of generative temperature settings, specifically at 0, 0.25, 0.5, 0.75, 1, and 2. These temperature settings adjust the randomness of the response, enabling the evaluation of how model behavior changes with varying degrees of unpredictability in text generation. By including temperature values between 0 and 1, we aimed to capture subtle variations in model behavior and identify the optimal balance between deterministic and diverse outputs. We utilized Replicate's API to query the two different Llama3 models, which ensured faster inference times[7] (Table 1).

Table 1. LLMs used in this study.

	No. of parameters	Developer	Open-source
Llama3 8b[8]	8 billion	Meta	✓
Llama3 70b[9]	70 billion	Meta	✓

[7] https://replicate.com/meta [Accessed: May 03, 2024]
[8] https://replicate.com/meta/meta-llama-3-8b [Accessed: Jun. 19, 2024]
[9] https://replicate.com/meta/meta-llama-3-70b [Accessed: Jun. 19., 2024]

3.3. Response Generation and Analysis

Using the two Llama3 models, responses to the text prompts were generated. Each response was then analyzed using the longest common substring (LCS) method. This method quantifies the similarity between the generated text and the target text segment, providing a measure of memorization in terms of exact text reproduction. The evaluation further included the calculation of Levenshtein distances to measure the edit distance between the generated and target texts, providing an additional metric of text similarity that allows for mismatches at the character level. This measure was normalized by the character length of the target text to account for variations in text length.

For each text prompt, the response generation was repeated multiple times across the specified range of temperature settings, allowing us to capture variability in the model's response consistency and assess memorization under different generative conditions. The core analytical process involved comparing each generated text against its corresponding target text to determine the degree of match.

3.4. Statistical Analysis

To assess the impact of model size and temperature settings on memorization, we employed a mixed-effects model, accounting for both fixed effects (model size and temperature) and random effects. This statistical approach allowed us to rule out the influence of fixed factors while controlling for random variability. The model formulation included interaction terms to explore if the effect of temperature on memorization differs across models.

4. Results

4.1. Effect of Temperature on Memorization

The percentage of exact matches in generated text sequences by the Llama3-8b and Llama3-70b models, respectively, at different temperatures are shown in Figure 2 and Figure 3. As observed in Figure 2, for the Llama3-8b model, the percentage of exact matches decreases consistently as the temperature increases from 0 to 2. At temperature 0, the model maintains higher exact matches for longer sequences, but as the temperature rises, the ability to generate exact matches diminishes sharply. At all temperatures the percentage of matches drop to nearly 0 for longer sequences.

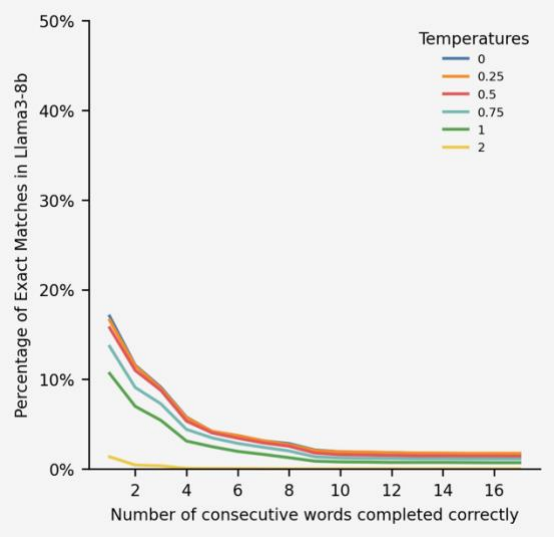

Figure 2. Percentage of Exact Matches in Llama3-8b Across Temperatures.

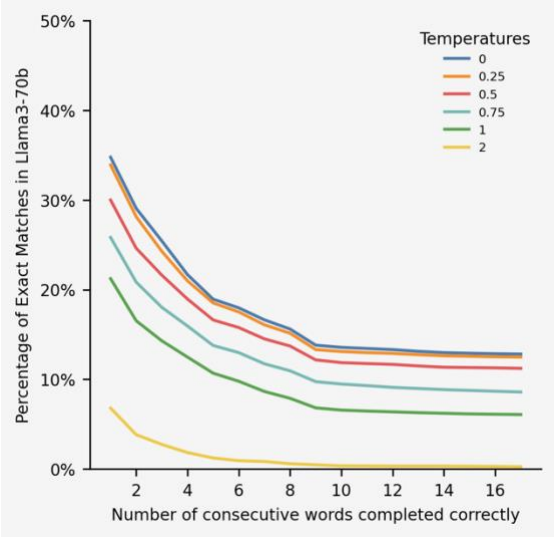

Figure 3. Percentage of Exact Matches in Llama3-70b Across Temperatures

Similarly, Figure 3 shows the performance of the Llama3-70b model across different temperatures. The trend is consistent with the Llama3-8b model, showing a clear decline in the percentage of exact matches as the temperature increases. At temperature 0, the Llama3-70b model performs better initially with higher exact match percentages compared to Llama3-8b, but this advantage decreases significantly at higher temperatures.

Figure 4 and Figure 5 provide a detailed breakdown of the exact matches for the Llama3-8b model at temperatures 0 and 2, respectively. The percentages in these figures represent the proportion of the model's generated sequences that exactly match the corresponding segments in the Wikipedia target data for a given number of consecutive

words. At temperature 0 (Figure 4), the Llama3-8b model shows a substantial initial percentage of exact matches (17.05% for 1-word sequences), which steadily decreases as the number of consecutive words increases. However, at temperature 2 (Figure 5), the model's performance drops drastically, with exact matches falling to 1.37% for 1-word sequences and becoming negligible for longer sequences.

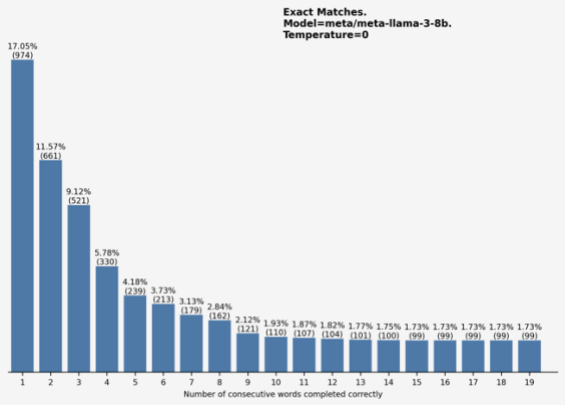

Figure 4. Exact Matches for Llama3-8b at Temperature 0.

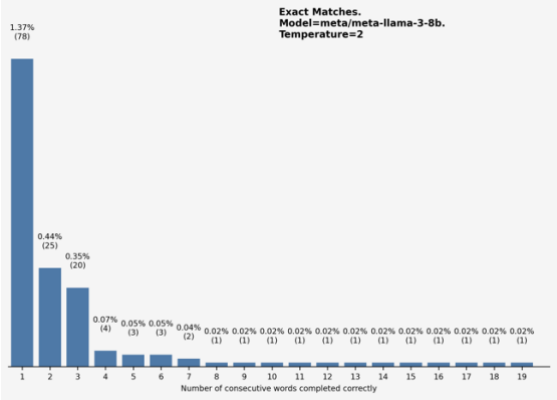

Figure 5. Exact Matches for Llama3-8b at Temperature 2.

The same trend is observed for the Llama3-70b model in Figure 6 and Figure 7. At temperature 0 (Figure 6), the model achieves a high initial percentage of exact matches (34.77% for 1-word sequences), approximately twice the amount shown by the smaller Llama3-8b model. However, this percentage also declines as the sequence length increases. At temperature 2 (Figure 7), the exact match percentage falls sharply to 6.79% for 1-word sequences, demonstrating the significant impact of higher temperatures on memorization performance.

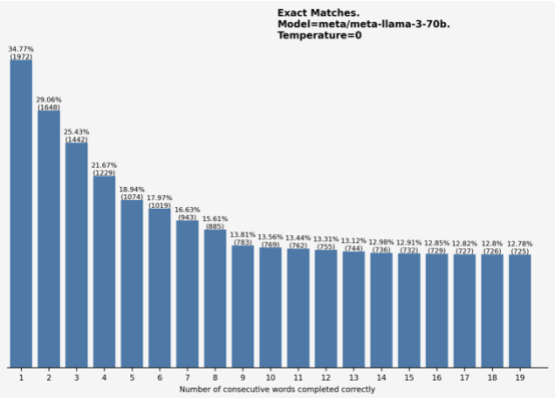

Figure 6. Exact Matches for Llama3-70b at Temperature 0.

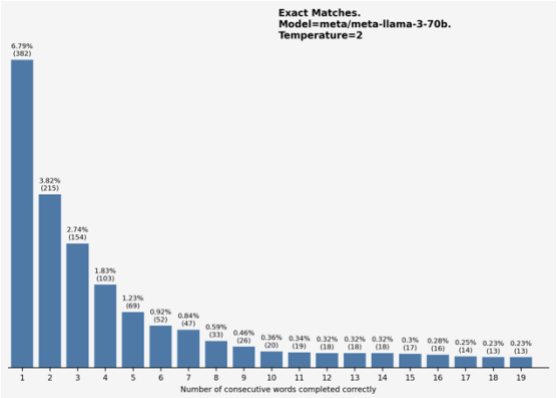

Figure 7. Exact Matches for Llama3-70b at Temperature 2.

Notably, both models exhibit a phenomenon where the percentage of exact matches plateaus after a certain number of consecutive words, particularly evident in the Llama3-70b model at lower temperatures (Figure 6). This plateau suggests that even at longer sequence lengths, some degree of memorization persists. The remaining figures depicting the exact matches for intermediate temperatures (0.25, 0.5, 0.75, and 1.0) are provided in the appendix (Appendix, Figure 8-15).

4.2. Differential Effects by Model Size

Table 2. Mixed-effects model analysis of memorization performance by model size and temperature settings.

Fixed effects	β (Coefficient)	SE (Std. Error)	z-value	p-value
Baseline for Llama3-70b at Temp=0	28.481	0.820	34.743	< 0.001
Effect of Llama3-8b	-30.672	0.925	-33.171	< 0.001
Effect of Temperature=0.25	-1.177	0.927	-1.269	0.204
Effect of Temperature=0.50	-4.831	0.928	-5.206	< 0.001
Effect of Temperature=0.75	-11.571	0.928	-12.466	< 0.001
Effect of Temperature=1.00	-17.621	0.928	-18.991	< 0.001
Effect of Temperature=2.00	-30.858	0.928	-33.241	< 0.001
Llama3-70b * Llama3-8b at Temp=0.25	1.182	1.308	0.903	0.366
Llama3-70b * Llama3-8b at Temp=0.50	4.750	1.309	3.629	< 0.001
Llama3-70b * Llama3-8b at Temp=0.75	11.314	1.309	8.643	< 0.001
Llama3-70b * Llama3-8b at Temp=1.00	17.187	1.309	13.133	< 0.001
Llama3-70b * Llama3-8b at Temp=2.00	29.948	1.309	22.876	< 0.001
Random Effect Variance	113.374	0.169		

The mixed-effects model analysis (Table 2) provides a comprehensive evaluation of how temperature and model size affect memorization. The dependent variable in this setup was an overlap variable, which measured the maximum extent to which the beginning of the generated text matches the beginning of the original target text, word for word, from left to right, essentially capturing the longest prefix of the generated text that is identical to the original target text. The intercept represents the baseline memorization performance of the Llama3-70b model at temperature 0, with a coefficient of 28.481 ($p < 0.001$). The effect of switching to the Llama3-8b model at temperature 0 is significantly negative, with a coefficient of -30.672 ($p < 0.001$), indicating that the smaller model performs worse in terms of exact matches at the baseline temperature of 0.

Temperature effects are also significant, showing a negative impact on memorization for both models as temperature increases. The effect of a temperature of 2.0 in Llama3-70b model has a coefficient of -30.858 ($p < 0.001$), highlighting the substantial decrease in memorization performance at the highest temperature setting.

The interaction terms between model size and temperature reveal that the effect of temperature on memorization is less pronounced for the Llama3-8b model compared to the Llama3-70b model. The interaction term for Llama3-8b at a temperature of 2.0 has a positive coefficient of 29.948 ($p < 0.001$), suggesting that while both models experience a decline in performance with increasing temperature, the smaller model is relatively less affected.

5. Discussion

This study aimed to understand the effects of temperature on training data memorization in Meta's Llama3-8b and Llama3-70b models. Our results showed that increasing temperature settings generally decreased the models' ability to memorize and reproduce text sequences verbatim. However, the larger Llama3-70b model exhibited a more pronounced decrease in memorization with higher temperatures compared to the smaller Llama3-8b model.

5.1. Memorization at Different Temperatures

The results consistently showed that lower temperatures lead to higher exact matches, indicative of increased memorization. At a temperature of 0, the models are highly deterministic, leading to a greater likelihood of exact reproduction of sequences that were present in their training data (Yu et al., 2024). This behavior aligns with the understanding that lower temperatures reduce randomness in the generation process, causing the model to default to its most likely outputs, which are often memorized sequences from the training data.

The persistence of memorization at longer sequence lengths, particularly in the Llama3-70b model as seen in Figure 6, may be attributed to the nature of the Wikipedia data used for prompting the models. Since Wikipedia content is widely disseminated and often appears in multiple sources, it is likely that similar or identical text segments were present in various parts of the training data. Consequently, the models might have encountered and memorized these sequences multiple times across different contexts, reinforcing their ability to reproduce longer text segments accurately.

This outcome was expected based on the underlying mechanisms of temperature settings in language models. Temperature in the context of LLMs modulates the probability distribution from which the model selects the next token in a generated sequence (Peeperkorn et al., 2024). At a temperature of 0, the model always selects the token with the highest probability, essentially making it deterministic (Yu et al., 2024). This lack of variability means the model's output is heavily biased towards the most likely, and thus often memorized, sequences from its training data.

Conversely, as the temperature increases, the model introduces more randomness into its token selection process. Higher temperatures flatten the probability distribution, making the model more likely to select less probable tokens (Yu et al., 2024). This increased

randomness promotes the generation of more diverse and less predictable outputs, which reduces the likelihood of the model reproducing exact sequences from the training data. This behavior is crucial for balancing the trade-off between creativity and memorization. By introducing variability, higher temperatures encourage the model to generate novel responses, which can enhance the utility of the model in tasks requiring creativity or adaptation to new contexts.

5.2. Impact of Model Size

The Llama3-70b model showed a higher propensity for memorization compared to the Llama3-8b model, particularly at lower temperatures. This finding supports previous research that larger models, due to their increased capacity, tend to memorize more of their training data (Carlini et al., 2021). The ability of the Llama3-70b model to produce a higher percentage of exact matches suggests that its larger parameter size enables it to store and recall more detailed information from the training corpus. The enhancement of language models' performance with increasing model size has been thoroughly examined in previous studies (Kaplan et al., 2020; Li et al., 2020). However, however, our findings indicate a downside, showing that these improvements result in reduced privacy.

As temperatures increased, the memorization advantage of the Llama3-70b model diminished more sharply compared to the Llama3-8b model. This differential effect indicates that larger models might be more sensitive to changes in temperature settings, leading to greater variability in their output as temperature increases. This sensitivity might be due to the complexity and higher capacity of the larger model, which allows for more nuanced responses at lower temperatures but also more significant changes when randomness is introduced at higher temperatures (Li et al., 2020).

In contrast, the Llama3-8b model, with its smaller parameter size, has less capacity for memorization, and according to our findings, also shows a weaker effect of temperature changes on memorization compared to the larger Llama3-70b model. The fewer parameters in the model mean that the interactions between them are less complex, resulting in a more consistent output even when randomness is introduced (Kaplan et al., 2020). This may make the smaller model less sensitive to changes in temperature, leading to a more gradual decline in memorization performance compared to the larger model.

5.3. Analyzing the Memorization Plateau

Table 3 in the Appendix highlights sequences where the Llama3-70b model memorized 15 or more consecutive words verbatim, shedding light on the underlying patterns that contribute to the plateau in memorization that was observed in Figure 3 at longer matches. As seen in Table 3, these sequences often comprise repetitive and structured demographic information, such as census data, which are commonly encountered in various contexts and adhere to consistent formats, making them easier for the model to memorize. This is in line with the findings of Carlini et al. (2021), where they showed that models tend to memorize unique identifiers, such as proper nouns and numbers, more readily than other types of information, suggesting a pattern in how memorization occurs. The structured nature of these sentences may aid in the model's ability to memorize and reproduce these sequences accurately. The frequent repetition of such information across different entries in the dataset increases the likelihood of exact memorization. The sequences shown in Table 3 are likely repeated multiple times within the training dataset, increasing the chances of memorization. For instance, statements about population demographics and census data are typically repeated across different geographical entries on Wikipedia, enhancing the model's propensity to memorize them.

Additionally, these verbatim reproductions occur primarily at lower temperatures, where the model's outputs are more deterministic and less random, further contributing to the high exact match rates observed.

5.4. Limitations

Our analysis focused on a subset of duplicate text snippets due to computational and time constraints. Although this subset was carefully selected to balance thoroughness and practicality, a larger-scale analysis, also including unique text snippets, could provide a more comprehensive understanding of memorization behaviors.

Additionally, the study used only two model variants (8b and 70b parameters). Including models of intermediate sizes could offer more detailed insights into how model scale affects memorization across a spectrum of parameter counts.

5.5. Ethical and Legal Implications

The findings of this study have significant ethical and legal implications, especially concerning the impact of temperature settings on the behavior of language models. As our results indicate, lower temperatures lead to higher levels of memorization, resulting in a greater risk of verbatim text regurgitation. This behavior raises privacy concerns, as it increases the likelihood of inadvertently disclosing sensitive or proprietary information embedded in the training data.

On the other hand, increasing the temperature introduces its own set of challenges. Yu et al. (2024) demonstrated that higher temperatures can make it easier to jailbreak LMs, allowing users to manipulate the models into generating inappropriate or harmful content. This dual risk presents a complex ethical dilemma: while lower temperatures compromise privacy through increased memorization, higher temperatures heighten the risk of model exploitation and misuse.

To address these issues, future research must focus on finding a balance between these competing risks. This involves not only empirical studies to fine-tune temperature settings but also the development of robust safeguards and ethical guidelines for deploying LLMs.

6. Conclusion

Our study provides valuable insights into the memorization behaviors of LLMs, particularly highlighting the significant impact of model size and temperature settings on the propensity to memorize training data. We observed that the Llama3-70b model exhibits higher memorization capability at lower temperatures, emphasizing the trade-off between model capacity and the risk of verbatim data reproduction.

Additionally, our results revealed that the smaller Llama3-8b model shows a weaker effect of temperature change on memorization compared to the larger 70b model. This finding implies that smaller models might offer more stability across different temperature settings, potentially providing a safer option for applications where balancing privacy and creativity is crucial.

These findings emphasize the need for careful consideration of temperature settings in model deployment, balancing the benefits of creative, diverse outputs with the risks of excessive memorization and model misuse. Addressing these challenges will be necessary for the responsible and safe adoption of LLMs.

In addition to these findings, our analysis of the memorization plateau revealed that the types of data most frequently memorized verbatim include structured and repetitive information, such as census data and demographic statistics. These sequences, which often follow consistent formats and appear repeatedly in the training dataset, are particularly prone to memorization. While this behavior underscores the model's ability to retain and reproduce factual data, it raises significant concerns when the training data contains sensitive information, such as medical records or personal identifiers. The potential for LLMs to regurgitate such data verbatim poses serious privacy risks and ethical issues. Therefore, it is crucial to develop robust guidelines and safeguards to prevent the inclusion of sensitive information in training datasets and to implement mechanisms that reduce the likelihood of verbatim reproduction. Future research should focus on developing techniques to mitigate these risks, ensuring that models can balance the need for accurate information retrieval with the imperative to protect sensitive information.

7. References

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Raffel, C. (2021). Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)* (pp. 2633-2650).

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2022). Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*.

Hartmann, V., Suri, A., Bindschaedler, V., Evans, D., Tople, S., & West, R. (2023). SoK: Memorization in General-Purpose Large Language Models. *arXiv preprint arXiv:2310.18362*.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., & Carlini, N. (2021). Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*.

Li, Z., Wallace, E., Shen, S., Lin, K., Keutzer, K., Klein, D., & Gonzalez, J. (2020, November). Train big, then compress: Rethinking model size for efficient training and inference of transformers. In *International Conference on machine learning* (pp. 5958-5968). PMLR.

Meeus, M., Shilov, I., Faysse, M., & de Montjoye, Y. A. (2024). Copyright Traps for Large Language Models. *arXiv preprint arXiv:2402.09363*.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., ... & Lee, K. (2023). Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*.

Peeperkorn, M., Kouwenhoven, T., Brown, D., & Jordanous, A. (2024). Is temperature the creativity parameter of large language models?. *arXiv preprint arXiv:2405.00492*.

Tirumala, K., Markosyan, A., Zettlemoyer, L., & Aghajanyan, A. (2022). Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, *35*, 38274-38290.

Yu, C. X., James, C. S. Y., & David, P. H. L. P (2024). Can LLMs have a fever? Investigating the effects of temperature on LLM security.

Yu, W., Pang, T., Liu, Q., Du, C., Kang, B., Huang, Y., ... & Yan, S. (2023, July). Bag of tricks for training data extraction from language models. In *International Conference on Machine Learning* (pp. 40306-40320). PMLR.

Appendix

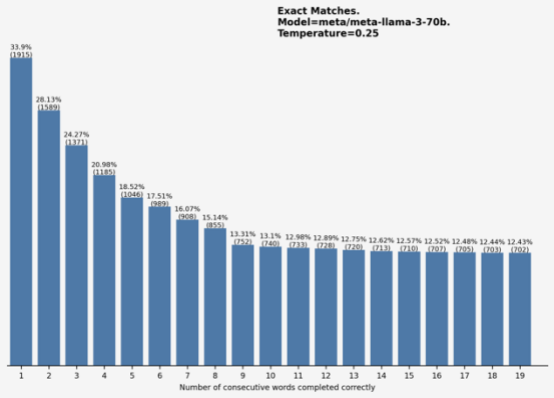

Figure 8. Exact Matches for Llama3-70b at Temperature 0.25.

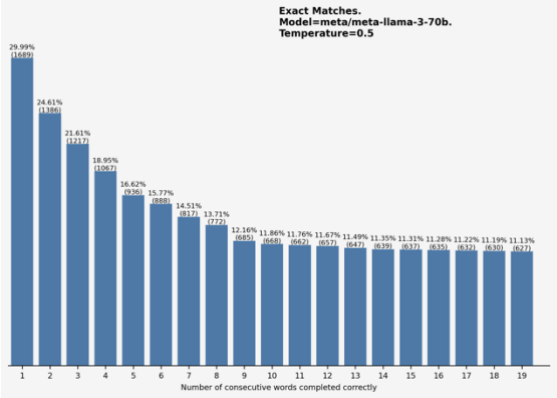

Figure 9. Exact Matches for Llama3-70b at Temperature 0.50.

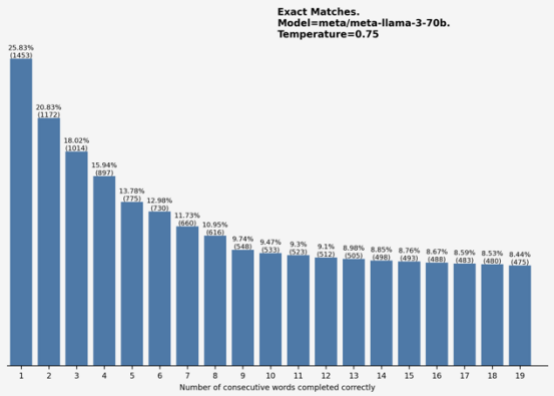

Figure 10. Exact Matches for Llama3-70b at Temperature 0.75.

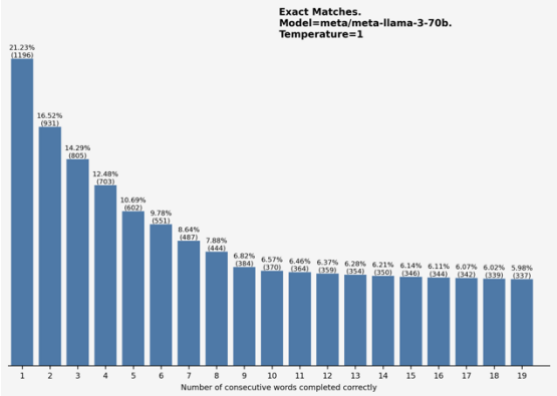

Figure 11. Exact Matches for Llama3-70b at Temperature 1.00.

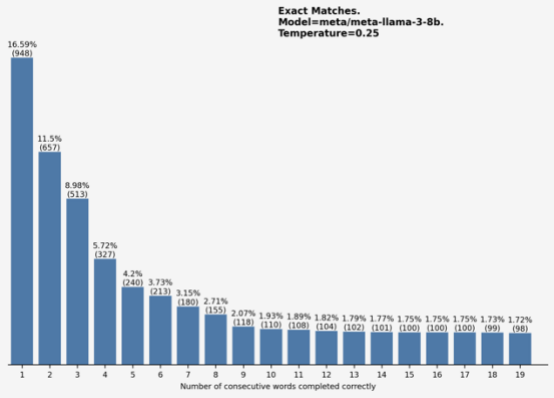

Figure 12. Exact Matches for Llama3-8b at Temperature 0.25.

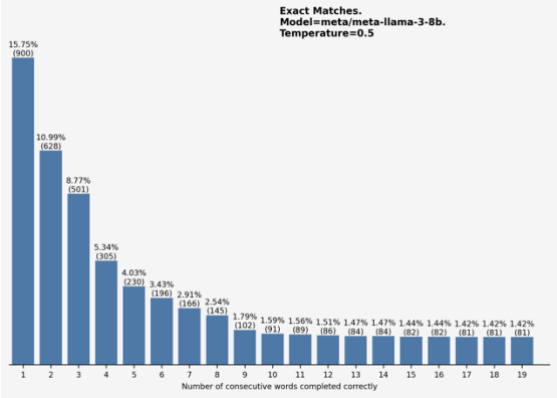

Figure 13. Exact Matches for Llama3-8b at Temperature 0.50.

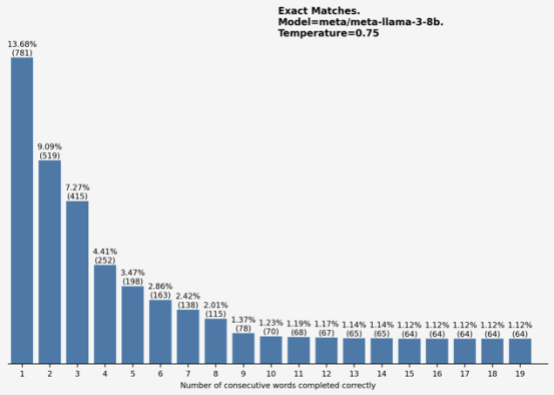

Figure 14. Exact Matches for Llama3-8b at Temperature 0.75.

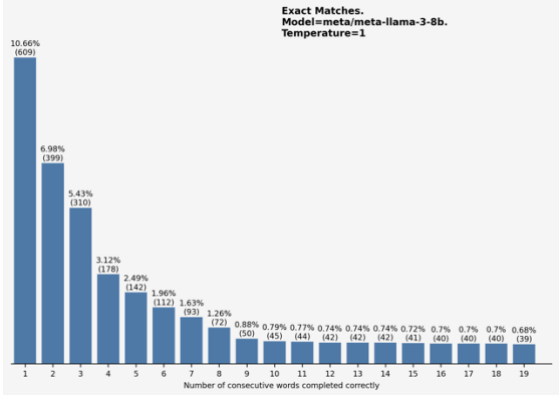

Figure 15. Exact Matches for Llama3-8b at Temperature 1.00.

Table 3. Random Sample of Sequences with 15 or More Consecutive Exact Matches by Llama3-70b Model. Text Count indicates the frequency of the text sequence/duplicate in the dataset.

Temperature	Prompt	Target	Generated Result	Text Count
1.00	% were 65 years of age or older. The gender makeup of the town was 50.4% male and	49.6% female. 2000 census As of the census of 2000, there were	49.6% female. ### 2000 census As of the census of 2000, there were	17
0.75	% were 65 years of age or older. The gender makeup of the city was 52.7	% male and 47.3% female. 2000 census As of the census of 2000, there were	% male and 47.3% female. ### 2000 census As of the census of 2000, there were	19
0.75	% from two or more races. 1.21% of the population	were Hispanic or Latino of any race. There were	were Hispanic or Latino of any race. There were	7
0.25	. Human achievement index 2017 Since 2003, United Nations Development Programme (UNDP) in Thailand has tracked progress on human development at sub-national level using the Human achievement index (	HAI), a composite index covering all the eight key areas of human development. National Economic and Social Development Board (NESDB) has taken over this task since 2017. T	HAI), a composite index covering all the eight key areas of human development. National Economic and Social Development Board (NESDB) has taken over this task since 2017. *	7
0.00	% were 65 years of age or older. The gender makeup of the town was 48.3%	male and 51.7% female. 2000 census As of the census of 2000, there were	male and 51.7% female. ### 2000 census As of the census of 2000, there were	15

	in English transcription. Some of the names are given with a proposed etymological meaning. For further information on the names included on the list, the reader may consult the sources listed below in the References and External Links. A – B – C – D	– E – F – G – H – I – J – K – L – M – N – O – P – Q – R – S – T – U – V – Y – Z	– E – F – G – H – I – J – K – L – M – N – O – P – Q – R – S – T – U – V – W X Y	
0.25				21
0.75	% were 65 years of age or older. The gender makeup of the town was 50.5	% male and 49.5% female. 2000 census As of the census of 2000, there were	% male and 49.5% female. ### 2000 census As of the census of 2000, there were	12
0.00	% were 65 years of age or older. The gender makeup of the town was 48.1% male and 51.9%	female. 2000 census As of the census of 2000, there were	female. ### 2000 census As of the census of 2000, there were	19
0.25	% from two or more races. 0.81% of	the population were Hispanic or Latino of any race. There were	the population were Hispanic or Latino of any race. There were	7
0.50	% were 65 years of age or older. The gender makeup of the city was 52.8	% male and 47.2% female. 2000 census As of the census of 2000, there were	% male and 47.2% female. ### 2000 census As of the census of 2000, there were	13