Scaling Laws for Fact Memorization of Large Language Models

Xingyu Lu[1]* , Xiaonan Li[1]*, Qinyuan Cheng[1], Kai Ding[2],
Xuanjing Huang [1,3], Xipeng Qiu[1,3]†
[1]Fudan University, [2]INTSIG,
[3]Shanghai Collaborative Innovation Center of Intelligent Visual Computing
luxy23@m.fudan.edu.cn, {lixn20, xpqiu}@fudan.edu.cn

Abstract

Fact knowledge memorization is crucial for Large Language Models (LLM) to generate factual and reliable responses. However, the behaviors of LLM fact memorization remain under-explored. In this paper, we analyze the scaling laws for LLM's fact knowledge and LLMs' behaviors of memorizing different types of facts. We find that LLMs' fact knowledge capacity has a linear and negative exponential law relationship with model size and training epochs, respectively. Estimated by the built scaling law, memorizing the whole Wikidata's facts requires training an LLM with 1000B non-embed parameters for 100 epochs, suggesting that using LLMs to memorize all public facts is almost implausible for a general pre-training setting. Meanwhile, we find that LLMs can generalize on unseen fact knowledge and its scaling law is similar to general pre-training. Additionally, we analyze the compatibility and preference of LLMs' fact memorization. For compatibility, we find LLMs struggle with memorizing redundant facts in a unified way. Only when correlated facts have the same direction and structure, the LLM can compatibly memorize them. This shows the inefficiency of LLM memorization for redundant facts. For preference, the LLM pays more attention to memorizing more frequent and difficult facts, and the subsequent facts can overwrite prior facts' memorization, which significantly hinders low-frequency facts memorization. Our findings reveal the capacity and characteristics of LLMs' fact knowledge learning, which provide directions for LLMs' fact knowledge augmentation.

1 Introduction

Large Language Models (LLM) have demonstrated remarkable abilities over a wide range of tasks (OpenAI, 2023; Touvron et al., 2023; Reid et al., 2024; Bai et al., 2023a; DeepSeek-AI, 2024; Cai et al., 2024; Sun et al., 2024). However, LLMs are prone to generating non-factual and fabricated contents, which is usually called "hallucination" (Zhang et al., 2023; Huang et al., 2023; Rawte et al., 2023) and undermines LLMs' reliability.

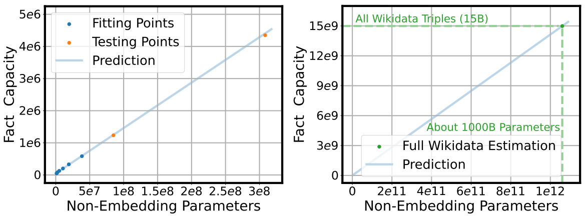


Figure 1: The fact capacity of LLMs with different sizes on Wikidata, under 100 training epochs. According to the predicted scaling law, memorizing all Wikidata triples (15B) requires 1000B non-embed parameters.

LLMs' factual responses highly rely on fact memorization. Specifically, the LLM memorizes fact knowledge during pre-training and the subsequent fine-tuning enables it to extract corresponding fact knowledge for the given instruction (Zhu and Li, 2023). If the base LLM does not memorize specific knowledge, it will be challenging for the fine-tuned LLM to correctly answer the corresponding question (Ren et al., 2024). Additionally, fine-tuning with unmemorized fact knowledge even encourages LLMs' hallucination (Lin et al., 2024; Gekhman et al., 2024). Despite the critical role of fact memorization, the behaviors of LLM fact memorization remain largely under-explored. Previous work usually analyzes the pre-trained LLMs' various abilities through the loss on unstructured text (Kaplan et al., 2020; Hoffmann et al., 2022a), and it is hard to reflect LLMs' fact memorization for two reasons: 1. The composition of pre-training corpus is highly complicated and fact knowledge appears in it mixedly and unevenly, which makes it hard to accurately quantify the fact knowledge in massive pre-training data. 2. The widely used metric, loss, can not directly measure the LLM fact

[*] Equal Contribution
[†] Corresponding Author

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 11263–11282
November 12-16, 2024 ©2024 Association for Computational Linguistics

memorization since not all tokens are fact-related.

This paper makes progress in quantitatively analyzing LLM fact memorization behaviors, including the scaling laws and behaviors of memorizing different types of facts. We focus on the memorization of atomic facts to facilitate accurately quantifying the number of facts and the memorization accuracy. We define atomic fact knowledge as a (key, attribute, value) triple, e.g., (SpaceX, CEO, Elon Musk), following Allen-Zhu and Li (2024). Given a key and an attribute, if the LLM correctly predicts the corresponding value, we consider it to memorize this fact knowledge. In this way, we can accurately quantify the number of fact knowledge and whether the LLM fully memorizes a specific fact, which facilitates a more accurate quantitative analysis of LLM's fact memorization behaviors.

Based on this setting, we analyze the LLM's fact memorization behaviors on massive facts from a large real-world information table. Specifically, we analyze the fact memorization scaling law of LLMs and LLMs' behaviors of memorizing different types of fact knowledge, including the following research questions (RQ):

RQ1: ***How does LLM's fact knowledge capacity scale with its size and training epochs?*** We define the fact knowledge capacity as the maximum fact triple quantity that the LLM can accurately memorize. We find that LLM's fact capacity linearly scales with its size under the same training epochs. Additionally, we find that the training epochs required for LLMs to memorize fact knowledge is significantly larger than one and this leads to higher training cost than general knowledge learning in pre-training. Increasing training epochs can initially increase the LLM's fact capacity and then reach saturation, which exhibits a trend of negative exponential law. Additionally, we extend our experiments to the Wikidata and the results exhibit a consistent trend, shown in Figure 1. According to the scaling law, under 100 training epochs, memorizing all Wikidata's fact triples requires about 1000B non-embed parameters, which seems very costly. These indicate the necessity of supplementing LLMs with fact knowledge by external information, like Retrieval-Augmented Generation (RAG) (Guu et al., 2020; Gao et al., 2024; Asai et al., 2023; Arivazhagan et al., 2023; Li et al., 2024; Min et al., 2023; Shi et al., 2023).

RQ2: ***Can LLMs efficiently memorize redundant facts?*** Many facts are derivable and thus redundant. For example, "Ivanka is Trump's daughter" can derive from "Trump is Ivanka's father". We analyze whether LLMs can efficiently memorize redundant facts, i.e., whether LLMs can save memorization capacity when simultaneously memorizing redundant facts. We find that LLMs struggle with efficiently memorizing redundant information. In general cases, when memorizing the redundant and non-redundant information of the same scale, the LLM exhibits a similar memorization rate. Only under specialized conditions, e.g., the correlated facts have the same direction and structure, the LLM can efficiently memorize them. These demonstrate LLMs' inefficiency in redundant fact memorization. Since massive redundant facts can appear in pre-training data in various forms, these indicate it is not cost-effective to use LLMs' parameters to store fact knowledge, and using a non-parametric method, like RAG, can be more efficient.

RQ3: ***What influences LLM's memorization preference for different types of fact knowledge?*** During pre-training, LLMs meet various facts and only memorize portions of them. We analyze LLMs' fact memorization preference in three aspects: frequency, difficulty and memorization order. We find that LLMs pay more attention to memorizing more frequent and difficult facts. Additionally, when an LLM memorizes two types of facts sequentially, the subsequent facts will significantly overwrite the memorization of prior facts. These further explain LLMs' inferior memorization of low-frequency facts since they appear infrequently during pre-training process and thus can be easily overwritten by subsequent pre-training knowledge.

Beyond fact memorization, we also analyze an interesting topic of fact knowledge generalization:

RQ4: ***Can LLMs generalize on unseen fact knowledge? What is the relation between fact memorization and generalization?*** Surprisingly, we find that the LLM can generalize on unseen facts to a certain level and its scaling law is highly similar to common pre-training LLM scaling law (Kaplan et al., 2020). The generalization accuracy is determined by the type of fact and some types of facts exhibit high generalizability, suggesting the potential of improving LLMs' factuality by adaptively leveraging fact generalization. Meanwhile, we find a qualitative relation between fact memorization and generalization: To the same type of fact, the easier the LLM is to memorize it, the better the LLM generalizes on the unseen set. This indicates that both LLM fact memorization and generalization are based on the correlation between input and output (Geirhos et al., 2020). If there is a stronger correlation between the input and output of one type of fact, it will be easier for the LLM to memorize and learn about the type of fact knowledge in a unified manner. Conversely, if the correlation is minimal, LLM needs to memorize facts individually, and is hard to generalize on unseen ones.

We summarize our contribution as follows: 1) To the best of our knowledge, this paper is the first to quantitatively analyze LLMs' scaling laws and behaviors of fact memorization on massive real-world facts. 2) Our findings reveal the capacity and characteristics of LLMs' fact knowledge learning. These results show that LLMs are highly inefficient for fact memorization from multiple perspectives, which suggests leveraging non-parametric methods, e.g., RAG, to enhance the fact knowledge of LLMs. 3) We find that LLMs can generalize on unseen facts and different types of facts show different generalizability, which indicates the potential of improving LLMs' factuality by adaptively leveraging LLMs' fact generalization. 4) We will release our code to facilitate future research[1].

2 Preliminary

In this paper, we focus on the quantitative analysis of LLMs' atomic fact knowledge memorization and we introduce the experiment setup as follows.

Atomic Fact Knowledge Memorization We define atomic fact knowledge as a (key, attribute, value) triple, e.g., (SpaceX, CEO, Elon Musk), and we cast fact memorization as a triple value prediction task. Specifically, for a fact triple (k, a, v), we use the cross-entropy loss to train the LLM to predict the value by the (k, a) as:

$$p = \text{LLM}(template_a(k, a)), \qquad (1)$$

where k and a are the key's name and attribute name, and $template_a$ is the natural language template of the attribute to make the LLM's input more coherent for realism. We adopt one template for one attribute for simplicity. Our pilot experiments show that various numbers of templates lead to consistent results, shown in Appendix A. Since we focus on fact memorization, we use the same input for training and inference.

Field	Description	Example
Company*	company name	Tiktok Co., Ltd.
Credit-No	social credit number	91110105MA...
Operator	legal representative	Lidong Zhang
Start-Date	founding date	2003.11.2
Title	representative title	Executive Director
Type	company type	Co., Ltd.
Register-Capital	registered capital	¥10^5
Longitude	company longitude	116.497976
...	...	...

Table 1: Company information table, which has 22 fields and 10M lines. "Company*" is the primary key. The information of overall fields is shown in Appendix B.

After training on facts $D = \{k_i, a_i, v_i\}_{i=1}^{|D|}$, we evaluate the LLM's **Memorization Rate (MR)** as:

$$\text{MR}(D) = \text{average}_{i=1}^{|D|}(\text{EM}(p_i, v_i)), \qquad (2)$$

where EM means exact match, and p_i and v_i are the i-th fact's prediction and value. In this way, we can use the memorization rate to accurately quantify the portion of facts the LLM has memorized.

Dataset This paper mainly conducts experiments on massive facts of a large real-world company information table, which is provided by a commercial data company, INTSIG[2]. The table contains various attributes of massive companies and we use facts like (Company, Attribute, Value) for experiments. The involved facts are from the real world and the types of them are diverse, and thus closely mirror the various facts in pre-training process. We show the table's statistics and sample row in Table 1. Additionally, experiments on Wikidata also show consistent trends (Section 3).

Implementation Details We mainly use the model architecture and tokenizer of Qwen (Bai et al., 2023b) and we show the results on other architectures and tokenizers in Appendix E, which show consistent trends. We mainly train LLMs' fact memorization from scratch and we show the results on pre-trained LLMs in Appendix C. For the specific hyper-parameters of each model size and overall implementation details, please refer to Appendix D.

3 Fact Capacity Scaling Laws

Exploratory Experiment First, we observe the same LLM's memorization rate over varying numbers of training facts under the same training

[1] https://github.com/StarLooo/Scaling_Law_LLM_Fact_Memorization

[2] INTSIG is a leading company of intelligent document recognition. https://www.intsig.com/

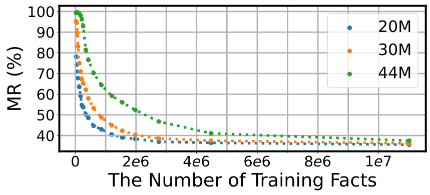


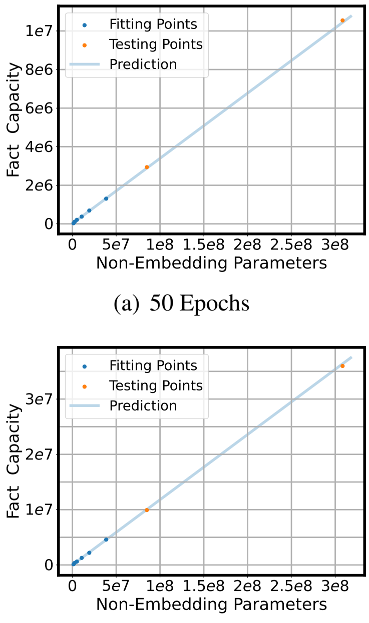
Figure 2: LLMs' memorization rate under different numbers of training facts.

epochs. We show the results in Figure 2. We see that the memorization rate significantly decreases with the increasing facts. These initially show that there is a memorization capacity upper limit for the LLM with the same size and training epochs.

In this section, we explore the scaling laws of LLMs' fact capacity. We define the fact capacity as the maximum fact quantity that the LLM can accurately memorize as:

$$C = \max(|D|) \text{ s.t. } \mathrm{MR}(D) > \phi\%, \qquad (3)$$

where D is training facts, a list of randomly sampled facts from all facts, and ϕ means a high MR close to 100%. In experiments, we set $\phi\%$ to be 95% and enumerate Ds of varying sizes to find the maximum $|D|$ that $\mathrm{MR}(D)$ is between $[\phi\%, (\phi+1)\%]$.

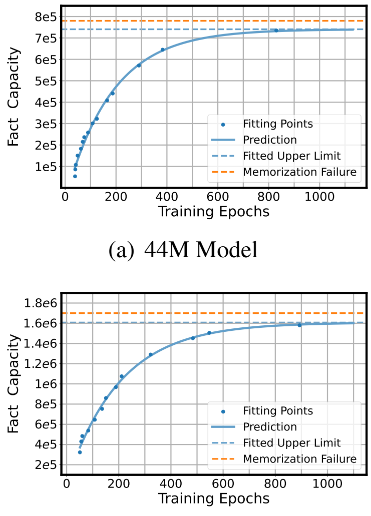


(a) 50 Epochs

(b) 200 Epochs

Figure 3: The relation between LLMs' fact capacity and their model sizes, under fixed training epochs.

Scaling Law of Fact Capacity and Model Size We plot the fact capacities of the varying model sizes from 30M to 0.5B, under the fixed training epochs in Figure 3 (20M fails to reach 95% MR at these epochs). We find that the LLM's fact capacity linearly scales with the model size. Meanwhile, we find that the line fitted from points of small model sizes (non-embed parameters $<= 38\text{M}$) can extrapolate well to large model size 0.5B (308M non-embed parameters $\approx 8 \times 38\text{M}$), which shows the robustness of the linear scaling laws.



(a) 44M Model

(b) 69M Model

Figure 4: The relation between LLMs' fact capacity and training epochs, under fixed model size.

Scaling Law of Fact Capacity and Epochs We plot the same LLM's fact capacities under varying training epochs in Figure 4. We find that with increasing training epochs, the LLM's fact capacity significantly increases at the beginning and then approaches saturation at about 1000 epochs, and we use the negative exponential law to fit the trend:

$$C = C^* - \alpha_E \cdot \exp(-\beta_E \cdot Epoch), \qquad (4)$$

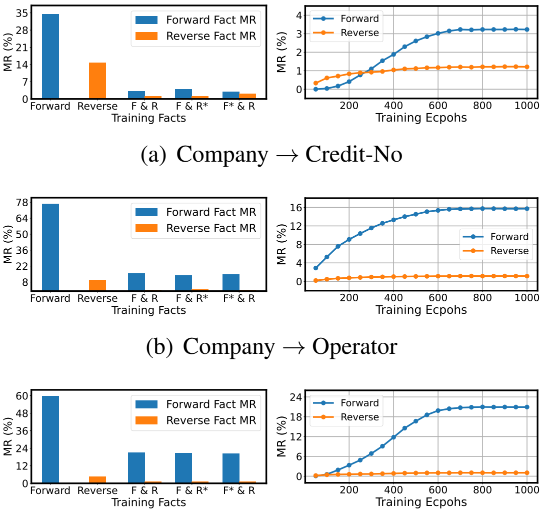
where C^* means the LLM's fact capacity saturation when epochs approach infinity, and α_E and β_E are constants. We further train the LLM on fact quantity which is 1.1 times of C^* and then find the LLM fails to accurately memorize all of those training facts, under 3000 epochs (almost saturated), which verifies the effectiveness of the fitting of negative exponential law. Additionally, for those small training epochs, e.g., < 35, the LLM almost can not accurately memorize facts, and this shows that the cost of fact memorization is significantly higher than general knowledge learning by pre-training, which usually requires only one epoch (Cai et al., 2024). This result indicates that it is challenging for the LLM to memorize those low-frequency fact knowledge in pre-training and up-sampling those facts can be a potential solution.

Experiments on Wikidata We extend our experiments to Wikidata. Specifically, we use the fact triples from Wikidata as the training facts and plot the relation between the fact capacity and LLM's size in Figure 1. We find that the results on Wikidata also show a linear relation between the fact capacity and the model size, which demonstrates the generality of the linear scale of the capacity parameter. According to the fitted line, we estimate that it requires an LLM with 1000B non-embed parameters to fully memorize all of Wikidata fact triples (about 15B[3]) under 100 training epochs, which seems costly. Since Wikidata's fact knowledge is only a subset of all public facts, our analysis indicates that it is very challenging for an LLM to memorize all public fact knowledge in the common LLM size and pre-training setting, which shows the necessity of enhancing LLMs' fact knowledge by external information, e.g., RAG (Gao et al., 2024).

4 Redundant Fact Memorization

In this section, we explore whether LLMs can efficiently memorize redundant facts, i.e., whether LLMs can save memorization capacity when simultaneously memorizing redundant facts. Specifically, we conduct experiments on three types of redundant facts: 1) The forward and reverse versions of the same fact knowledge 2) The correlated facts of the same key 3) Single-hop facts and their derivable multi-hop facts. Additionally, we analyze whether learning abstract abilities occupies the fact memorization capacity. We set the training epoch as 1000 to make the LLM's memorization saturated, unless otherwise specified.

The Same Fact of Different Directions In this section, we analyze whether the LLM can efficiently memorize the forward and reverse versions of the same facts. The forward fact is predicting the value based on the company name and attribute, as in Eq (1). The reverse fact is predicting the company based on the attribute's value (Berglund et al., 2024; Allen-Zhu and Li, 2023). We select three highly reversible attributes, "Operator", "Credit-No" and "Register-No"for the experiment. Specifically, we compare the memorization rate of the following three groups: 1) separately memorizing the forward or reverse version of the same facts. 2) simultaneously memorizing the forward and reverse versions of the same facts (redundant). 3) simultaneously memorizing the forward facts and the reverse version of another set of facts (non-redundant). The number of each direction's facts is the same and thus the memorization load of group 2 and 3 is consistent. We show the results on 41M model in Figure 5. We also plot corresponding learning curves in Figure 5, which show that the LLMs' fact memorization is almost saturated. The results on 30M model are shown in Appendix F and show similar trends. We see that simultaneously memorizing facts of different directions leads to a significantly lower MR than separately memorizing them and the MR of simultaneous memorization is lower than the half of separate memorization.



(a) Company → Credit-No

(b) Company → Operator

(c) Company → Register-No

Figure 5: LLMs' memorization of the same facts with different directions, where "*" means facts are from another group of keys. The right is the learning curves.

These show that the LLM does not compatibly memorize them and memorizing different directions of the same fact even conflicts with each other. Meanwhile, memorizing different directions of the same group of facts (redundant) has a similar memorization rate to memorizing different groups of facts (non-redundant). These show that when the LLM memorizes the same facts in different directions, it seems to memorize them separately like memorizing independent facts, which reflects the inefficiency of LLM memorization for the same facts with different directions (Golovneva et al., 2024). Since the massive facts can be described in

[3] https://www.Wikidata.org/wiki/Property:P10209

different directions, these results can indicate that the LLM's parametric knowledge is not efficient for fact memorization.

Correlated Facts of the Same Key In this section, we analyze whether the LLM can efficiently memorize the correlated facts of the same key, e.g., a company's type and its type code. Specifically, we select two combinations of correlated attributes to conduct analysis and additionally adopt two unrelated combinations as a comparison. For each combination, we compare the memorization rate of the following three groups: 1) individually memorizing facts of a single attribute; 2) simultaneously memorizing facts of two attributes on the same companies (if attributes are correlated, these facts will be redundant); 3) simultaneously memorizing one attribute's facts on a group of companies and another attribute's facts on another group of companies (non-redundant). The number of each attribute's facts is the same and thus the memorization load of group 2 and 3 is consistent.

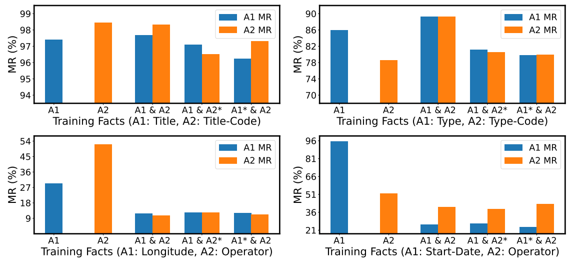
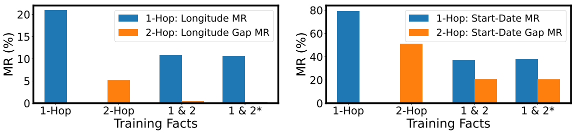


Figure 6: Memorization on correlated facts, where "*" means that facts are from another group of keys.

The results are shown in Figure 6 and Appendix G. We find that simultaneously memorizing correlated attributes leads to a higher memorization rate than separate memorization, which shows LLMs can efficiently memorize one key's correlated attributes, and correlated fact memorization can facilitate the individual fact's memorization. Meanwhile, for those unrelated attributes, simultaneously memorizing them leads to a decreased memorization rate, which shows that whether LLM can compatibly memorize one key's facts highly depends on the correlation of those facts. While it is hard to inject new correlated knowledge into LLMs (Allen-Zhu and Li, 2023), these results indicate the potential of additionally memorizing correlated facts in pre-training since they can be compatibly memorized.

Derivable Multi-hop Fact In this section, we analyze whether the LLM can efficiently memorize derivable facts. For example, when the LLM memorizes the longitude of two companies, can it additionally memorize their longitude gap efficiently? We explore this question on facts about attributes "Longitude" and "Start-Date", and choose their gap as derivable 2-hop facts. For 2-hop facts of one attribute, given two different keys, we train the LLM to predict the value gap of this attribute. Specifically, we compare the memorization rate of the following three groups: 1) separately memorizing single-hop facts and their derivable 2-hop facts. 2) simultaneously memorizing single-hop facts and their derivable 2-hop facts (redundant). 3) simultaneously memorizing single-hop facts and 2-hop facts derived from another set of single-hop facts (non-redundant). We control the numbers of 1-hop facts and 2-hop facts to be equal. The results are shown in Figure 7. We find that group 2 leads to a significantly lower memorization rate than group 1, which shows that the memorization of derivable 2-hop facts is not compatible with corresponding 1-hop facts. Additionally, the memorization rate of group 2 is similar to group 3. This shows that when the LLM memorizes single-hop facts and their derivable facts, it seems to memorize them separately like memorizing irrelevant facts. This reflects the inefficiency of LLM memorization for derivable facts, which hinders the LLM's fact capacity for massive derivable facts in pre-training corpus (Ju et al., 2024).

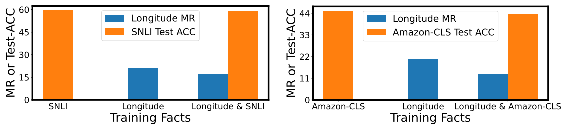


(a) Longitude

(b) Start-Date

Figure 7: LLM memorization for derivable facts, where "*" means that facts are from another group of keys.

Fact Memorization Meets Abstract Ability Learning We explore whether abstract ability learning occupies LLMs' fact memorization capacity. Specifically, we compare the fact MR or test accuracy of two groups: 1) separately learning fact knowledge and abstract ability 2) simultaneously learning fact knowledge and abstract abilities. We use SNLI (MacCartney and Manning, 2008) and Amazon Sentiment Analysis (McAuley and Leskovec, 2013) for abstract ability learning. The frequency of facts and abstract ability examples is the same. The results are shown in Figure 8. We see that additionally learning abstract abilities decreases the fact memorization rates. Meanwhile, the incorporation of fact knowledge slightly hurts the classification tasks' test accuracy. These indicate that fact memorization and abstract ability learning will influence each other and occupy the LLMs' knowledge capacity jointly, which further exacerbates the challenges of LLMs memorizing facts during pre-training.



(a) Longitude & SNLI (b) Longitude&Amazon-CLS

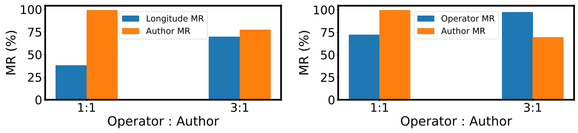
Figure 8: The influence of abstract ability learning to LLM's fact memorization.

5 Fact Memorization Preference of LLMs

We analyze LLMs' fact memorization preference in three aspects: frequency, difficulty and memorization order. Since this section focuses on preference, we select irrelevant facts to conduct experiments. Specifically, we use the combination of facts in the company information table and a specialized subset of Wikidata facts (Book → Author).

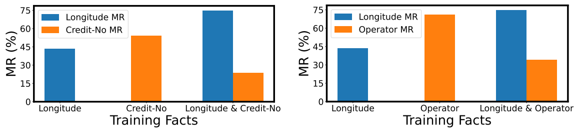
Frequency We compare the respective memorization rate of simultaneously memorizing two attributes under different frequencies. The results on "Longitude & Author" and "Operator & Author" are shown in Figure 9. We see that the higher frequency leads to a significantly higher memorization rate and inhibits low-frequency facts' memorization (Mallen et al., 2023). This indicates the importance of increasing the frequency of low-frequency facts in pre-training corpus to facilitate LLMs' memorization of them. However, since facts in pre-training corpus usually appear in a complicated and mixed manner, it is non-trivial to separately control their respective frequency, which further increases the challenges for LLMs to memorize low-frequency facts.

Difficulty We compare the respective memorization rate of three groups: 1) using LLM of size $2 * N$ to simultaneously memorize facts of two attributes with different memorization difficulties; 2) using LLM of size N to separately memorize facts of each attribute. In group 1 and 2, the number of facts of each attribute is the same and thus the average fact capacity for each attribute is consistent. In this way, we can observe the LLM's preference when simultaneously memorizing two attributes. The results on "Longitude & Credit-No" and "Longitude & Operator" are shown in Figure 10 and more results on other combinations are shown in Appendix H. We define the difficulty of facts according to the memorization rate under the same training size. In group 2, the memorization rates of attributes "Credit-No" and "Operator" are higher than "Longitude" and thus they are easier to memorize. Compared with group 2, the memorization rate of difficult facts and easy facts in group 1 increases and decreases, respectively. These show that when LLMs memorize different types of facts, they tend to pay more attention to the facts that are harder to memorize.



(a) Longitude & Author (b) Operator & Author

Figure 9: The effect of frequency for fact memorization.



(a) Longitude & Credit-No (b) Longitude & Operator

Figure 10: The effect of difficulty for fact memorization.

Memorization Order We compare the memorization rate of simultaneously memorizing facts of two attributes in different memorization orders.

Training Facts	Longitude MR	Author MR
Longitude	20.9	-
Author	-	76.1
Longitude⇒Author	0	13.1
Author⇒Longitude	17.7	0

Training Facts	Credit-No MR	Operator MR
Credit-No	30.7	-
Operator	-	38.9
Credit-No⇒Operator	0	32.6
Operator⇒Credit-No	20.6	0.1

Table 2: The influence of memorization order. "A⇒B" means memorizing A before B.

The results are shown in Table 2. We find that the memorization rate of earlier facts decreases to almost zero and the subsequent memorized facts almost refresh the LLM's fact memorization. These indicate a potential reason for LLM's inferior memorization of low-frequency facts: maybe some of them only appear in the early stage of pre-training process and they were almost overwritten by the subsequent pre-training knowledge. Additionally, the MR of subsequent facts is lower than its individual memorization, which demonstrates the importance of evenly distributing various types of facts in pre-training process.

6 Fact Generalization of LLMs

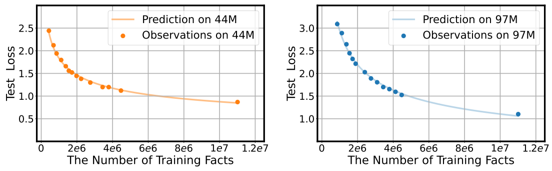
Beyond the fact memorization, we explore an interesting question: can LLMs generalize on unseen fact knowledge? Specifically, we train the LLM to memorize facts of a group of keys and test it on unseen keys' facts. We show each attribute's generalization accuracy (exact match) of 44M model in Figure 12. We also test results of 30M model and observe similar trends (see Appendix I). We observe that facts on most of the attributes have a generalization accuracy greater than zero, which indicates that LLMs can generalize on unseen fact knowledge to a certain level.

To analyze why the LLM can generalize on fact knowledge, we conduct a case study on facts of three attributes and show the cases in Appendix J. We find that LLMs' fact generalization depends on the correlation between input (key) and output (value) (Geirhos et al., 2020). For a specific type of fact (attribute), the higher correlation between the key and value leads to higher generalization accuracy. For example, the LLM may correctly predict an unseen company's longitude if the company name contains a region name and the training dataset contains the longitude of companies with the same region name. Or it can roughly estimate the company's register-capital according to company size indicated by the company name, e.g., "Fruit shop"→ (¥10^4 $\sim$ ¥10^5) or "Investment company"→ (¥10^7 $\sim$ ¥10^9). Meanwhile, different types of facts have different generalizability. For those facts with obvious patterns, the LLM can achieve reliable generalization. For those attributes with weak correlation, although the LLM does not know exactly the facts, it can identify the rough range of facts. These suggest the potential of adaptively leveraging LLM's fact generalization: 1. selectively leveraging generalization of those highly generalizable facts; 2. if the LLM does not exactly know the whole fact, it can response with a part of the fact, e.g., a rough range, to make its response more informative and thus helpful.

Fact Generalization Scaling Law Additionally, we analyze the scaling law of LLMs' fact knowledge generalization. Specifically, we plot the LLMs' loss values on test fact knowledge under different training fact quantities, following Kaplan et al. (2020). The results are shown in Figure 11. We find that the test loss on fact generalization also follows the power-law (Kaplan et al., 2020) as:

$$L(D) = D_c * D^{\alpha_D}, \qquad (5)$$

where D is the number of training facts, D_c and α_D are constant numbers. This trend is similar with general pre-training (Kaplan et al., 2020), which indicates that LLMs follow a similar learning mechanism in learning factual knowledge as they learn general knowledge in pre-training (OpenAI, 2023).



(a) 44M

(b) 97M

Figure 11: LLMs' fact generalization loss across different numbers of training facts.

Relation between Fact Memorization and Generalization We plot the memorization rate and generalization accuracy for each type of fact in Figure 12. We find that the generalization accuracy of one type of fact highly correlates with its memorization rate. For one type of fact, the higher memorization rate leads to higher generalization accuracy. These indicate that both LLM fact memorization and generalization are based on the correlation between input and output (Geirhos et al., 2020). If there is a stronger correlation between the input and output, it will be easier for the LLM to memorize and learn about the type of fact knowledge in a unified manner. If the correlation is minimal, LLM needs to memorize facts individually, and is hard to generalize on unseen ones.

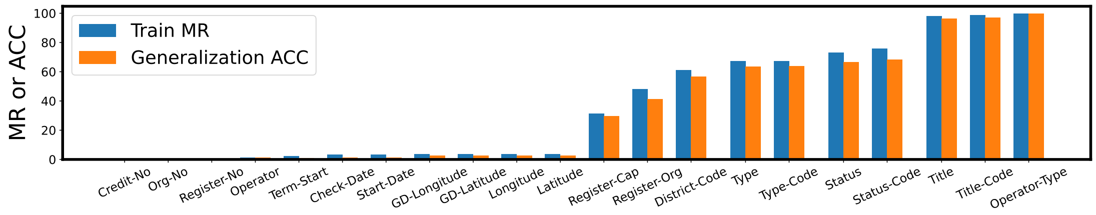


Figure 12: Memorization and generalization over facts of different types, on 44M LLM trained by 10M facts.

7 Related Work

Understanding the scaling behaviors of LLMs is important for decisions about the key choice design of LLMs, e.g., model size or pre-training data (Kaplan et al., 2020; Gao et al., 2022; Clark et al., 2022). Most of the existing work focuses on the scenario of general pre-training or downstream tasks. Kaplan et al. (2020) observe the power law relationships between the LLM perplexity and size of LLM and dataset. Hoffmann et al. (2022b) explore the optimal token quantity and LLM size for pre-training under a specified compute budget and find that the LLM size and training tokens should be scaled equally for compute-optimal LLM training. Besides pre-training, researchers find that the performance of downstream tasks can be predicted from the LLM size and training data scale (Hernandez et al., 2021; Ghorbani et al., 2021; Isik et al., 2024). Different from them, our paper specifically focuses on scaling laws of LLMs' fact memorization and behaviors of memorizing different types of facts, which is critical for LLMs' factual responses.

Allen-Zhu and Li (2024), concurrently to our work, explore scaling laws of LLMs' memorization on synthetic facts. Our work differs in several ways: 1. We analyze LLM's fact memorization on real-world facts while they use randomly generated facts, which have a non-negligible gap with real-world facts. According to our findings, we conclude that memorizing all Wikidata's facts requires 1000B non-embed parameters, which indicates that using an LLM to memorize all public facts is almost not plausible. 2. We additionally analyze LLMs' behaviors of learning fact knowledge in different aspects, including compatibility, preference and generalization, which further provide directions for fact knowledge augmentation of LLMs.

8 Conclusion

We analyze LLMs' fact memorization behaviors and these are our main conclusions: 1) The fact capacity has a linear relationship with model size and a negative exponential law relationship with training epochs. According to the built scaling law, we estimate that memorizing all of Wikidata fact triples requires training an LLM with 1000B non-embed parameters for 100 epochs, which seems very costly; 2) We find that LLMs struggle with efficiently memorizing redundant facts. Only for redundant facts with the same direction and structure, LLMs can memorize them in a unified manner. 3) The LLM prefers memorizing more frequent and difficult facts. 4) LLMs can generalize on unseen fact knowledge and its scaling law is similar to general pre-training.

Limitations

We list limitations of this paper as follows:

- Since this paper focuses on fact knowledge memorization, each atomic fact individually forms a training example and we keep the same inputs for the training and inference stages. This has a small gap with pre-training setting, which usually uses unstructured text and concatenates short sentences into a large chunk for training efficiency. We regard the exploration of facts of unstructured text as future work.
- As shown in Figure 4, fact memorization requires hundreds of training epochs, which leads to significant computational costs. Limited by computational resources, the maximum LLM size used in experiments is 0.5B. We regard the exploration of larger scales as future work.

Ethics Statement

In this paper, we use public fact information for experiments, including a real-world company information table and Wikidata fact triples. The company information table is provided by a commercial data company and we obtain its permission to conduct this research. Meanwhile, the trained models are only for LLM fact memorization analytical research and will not be made public.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (No.2022CSJGG0801). The computations in this research were performed using the CFFF platform of Fudan University.

References

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

Zeyuan Allen-Zhu and Yuanzhi Li. 2024. Physics of language models: Part 3.3, knowledge capacity scaling laws. *Preprint*, arXiv:2404.05405.

Manoj Ghuhan Arivazhagan, Lan Liu, Peng Qi, Xinchi Chen, William Yang Wang, and Zhiheng Huang. 2023. Hybrid hierarchical retrieval for open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10680–10689, Toronto, Canada. Association for Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *Preprint*, arXiv:2310.11511.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. *Preprint*, arXiv:2309.16609.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023b. Qwen technical report. *Preprint*, arXiv:2309.16609.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2024. The reversal curse: Llms trained on "a is b" fail to learn "b is a". *Preprint*, arXiv:2309.12288.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Aidan Clark, Diego de las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, George van den Driessche, Eliza Rutherford, Tom Hennigan, Matthew Johnson, Katie Millican, Albin Cassirer, Chris Jones, Elena Buchatskaya, David Budden, Laurent Sifre, Simon Osindero, Oriol Vinyals, Jack Rae, Erich Elsen, Koray Kavukcuoglu, and Karen Simonyan. 2022. Unified scaling laws for routed language models. *Preprint*, arXiv:2202.01169.

DeepSeek-AI. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *Preprint*, arXiv:2405.04434.

Leo Gao, John Schulman, and Jacob Hilton. 2022. Scaling laws for reward model overoptimization. *Preprint*, arXiv:2210.10760.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? *Preprint*, arXiv:2405.05904.

Behrooz Ghorbani, Orhan Firat, Markus Freitag, Ankur Bapna, Maxim Krikun, Xavier Garcia, Ciprian Chelba, and Colin Cherry. 2021. Scaling laws for neural machine translation. *Preprint*, arXiv:2109.07740.

Olga Golovneva, Zeyuan Allen-Zhu, Jason Weston, and Sainbayar Sukhbaatar. 2024. Reverse training to nurse the reversal curse. *Preprint*, arXiv:2403.13799.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *Preprint*, arXiv:2002.08909.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. 2021. Scaling laws for transfer. *Preprint*, arXiv:2102.01293.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022a. Training compute-optimal large language models. *Preprint*, arXiv:2203.15556.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022b. Training compute-optimal large language models. *Preprint*, arXiv:2203.15556.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *Preprint*, arXiv:2311.05232.

Berivan Isik, Natalia Ponomareva, Hussein Hazimeh, Dimitris Paparas, Sergei Vassilvitskii, and Sanmi Koyejo. 2024. Scaling laws for downstream task performance of large language models. *Preprint*, arXiv:2402.04177.

Tianjie Ju, Yijin Chen, Xinwei Yuan, Zhuosheng Zhang, Wei Du, Yubin Zheng, and Gongshen Liu. 2024. Investigating multi-hop factual shortcuts in knowledge editing of large language models. *Preprint*, arXiv:2402.11900.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Xiaonan Li, Changtai Zhu, Linyang Li, Zhangyue Yin, Tianxiang Sun, and Xipeng Qiu. 2024. Llatrieval: Llm-verified retrieval for verifiable generation. *Preprint*, arXiv:2311.07838.

Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen tau Yih, and Xilun Chen. 2024. Flame: Factuality-aware alignment for large language models. *Preprint*, arXiv:2405.01525.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Bill MacCartney and Christopher D. Manning. 2008. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 521–528, Manchester, UK. Coling 2008 Organizing Committee.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *Preprint*, arXiv:2212.10511.

Julian J. McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Seventh ACM Conference on Recommender Systems, RecSys '13, Hong Kong, China, October 12-16, 2013*, pages 165–172. ACM.

Sewon Min, Suchin Gururangan, Eric Wallace, Hannaneh Hajishirzi, Noah A. Smith, and Luke Zettlemoyer. 2023. Silo language models: Isolating legal risk in a nonparametric datastore. *Preprint*, arXiv:2308.04430.

OpenAI. 2023. Gpt-4 technical report.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *Preprint*, arXiv:2309.05922.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Mengjie Ren, Boxi Cao, Hongyu Lin, Cao Liu, Xianpei Han, Ke Zeng, Guanglu Wan, Xunliang Cai, and Le Sun. 2024. Learning or self-aligning? rethinking instruction fine-tuning. *Preprint*, arXiv:2402.18243.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *Preprint*, arXiv:2301.12652.

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Xiangyang Liu, Hang Yan, Yunfan Shao, Qiong Tang, Shiduo Zhang, Xingjian Zhao, Ke Chen, Yining Zheng, Zhejian Zhou, Ruixiao Li, Jun Zhan, Yunhua Zhou, Linyang Li, Xiaogui Yang, Lingling Wu, Zhangyue Yin, Xuanjing Huang, Yu-Gang Jiang, and Xipeng Qiu. 2024. Moss: An open conversational large language model. *Machine Intelligence Research*.

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.

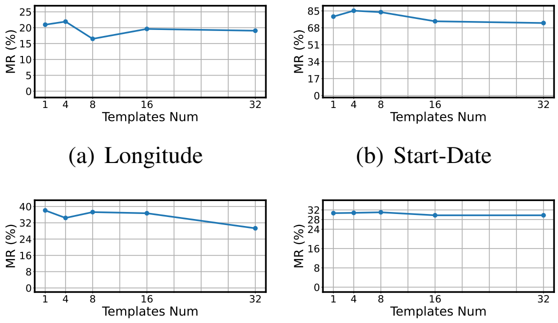
Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219.

Zeyuan Allen Zhu and Yuanzhi Li. 2023. Physics of language models: Part 3.1, knowledge storage and extraction. *CoRR*, abs/2309.14316.

A The Effect of Template Quantity

In this section, we analyze the influence of template quantity on memorization rate. Specifically, we observe the memorization rate of the same 200K facts under various numbers of templates using a 30M model. The results are shown in Figure 13.

We see that the memorization rates over different numbers of templates are at a consistent level. Even when the paraphrase quantity increases to 32, the memorization rate of specific attribute facts only decreases to as low as 75% of the original. Therefore, the number of templates does not significantly influence the LLM's fact memorization.



(a) Longitude

(b) Start-Date

(c) Operator

(d) Credit-No

Figure 13: Memorization on different templates.

B Attributes of Table

We list the information and average length of overall fields of the used large information table in Table 4.

C The Effect of Pre-training on Fact Memorization

We compare the fact knowledge learning from scratch and pre-trained checkpoints. The results are shown in Table 5. We find that the pre-trained initialization leads to higher generalization accuracy and consistent memorization rate of training data. These show that the pre-trained knowledge almost does not influence the LLMs' memorization of new fact knowledge and can improve LLMs' generalization on unseen facts. We leave further analyses of pre-trained influence to fact generalization as future work.

D Implementation Details

In this section, we first introduce the general implementation details (model, training, dataset) and then introduce details of the specific settings for each individual experiment.

D.1 Model Details

We mainly use the model architecture and tokenizer of Qwen-1.5 (Bai et al., 2023b). For more details of Qwen, we refer the reader to the original paper (Bai et al., 2023b). We mainly set the hyper-parameters of each LLM's size according to the following aspects: 1) The aspect ratio, which is the ratio of the hidden size to the number of layers, should be maintained at a moderate value. Following conventional design practices, we control the aspect ratio within the range of 128/3 (as adopted by Qwen-1.5-0.5B) to 128 (as adopted by Qwen-1.5-7B). 2) The intermediate size should be approximately 8/3 times of the hidden size and be divisible by 128. We provide the detailed hyper-parameters of model architecture in Table 6.

D.2 Training Details

We configure the global batch size as 512 and employ the AdamW optimizer (Kingma and Ba, 2017; Loshchilov and Hutter, 2019). In exploratory experiments, we find that LLMs with different sizes are highly sensitive to learning rates and thus we search for the best learning rates for each size's LLM and different datasets to achieve the optimal memorization rate, under small training epochs. Meanwhile, we adopt the cosine learning rate scheduler. We list the learning rates of each model size in Table 6. It's observed that the optimal learning rates differed between the company information table and Wikidata, and the latter requires a higher learning rate. Most of these experiments are conducted using either 8 NVIDIA RTX 3090s or 4 NVIDIA A800s-80GB, utilizing BFloat16 mixed precision training. The training speed of models with different sizes can be referred to in Table 3.

D.3 Dataset Details

In this paper, we conduct experiments on fact triples from a large real-world company information table and Wikidata[4]. The company information table is provided by a commercial data company and we obtain its permission to conduct this research. For Wikidata, we follow this public github repository[5] to get all of its fact triples. The facts of the company information table are in Chinese and Wikidata's facts are in English. For the

[4] https://www.wikidata.org/wiki/Wikidata:Introduction

[5] https://github.com/neelguha/simple-wikidata-db

Model Identifier	Training Speed
20M	800
30M	700
41M	650
44M	600
69M	500
97M	400
116M	300
200M	225
0.5B	100

Table 3: Training speeds (triples per second per GPU) of models of different sizes, which are based on NVIDIA RTX 3090.

key and entity in the company information table and Wikidata, we use their natural language name instead of the original key (uid) to closely mirror the facts in pre-training data.

D.4 Details on Scaling Law of Fact Capacity and Model Size

We conduct a series of experiments using various model sizes ranging from 30M to 0.5B (20M fails to reach 95% MR at these epochs), while keeping the training epochs fixed. We use the number of non-embed parameters to measure model size, following Kaplan et al. (2020). When randomly sampling facts, we first randomly sample keys and then use facts of these keys' all attributes to make fact type distributions consistent. In these experiments, we use $|D| * MR(D)$ to measure the fact capacity more accurately since the memorization rate may vary slightly for each D. The objective is to investigate the relationship between fact capacity and model size. Specifically, we utilize the results from models with fewer than 200M parameters to establish a scaling law formula. We then validate the extrapolation by employing models with 200M and 0.5B parameters on both the company information table and Wikidata. For Wikidata, we set the fixed training epochs to 100, while for the company information table, we use 50 and 200 epochs. In this way, we can observe the results across different datasets and epochs, which makes our results and the built scaling law more robust. Besides, the templates we use for the company information table are listed in Table 7. For Wikidata, since it contains tremendous types of fact (relations), it is costly to design an individual template for each type and thus we use a unified template: "For this entity, $\langle E \rangle$, the entity forming the relationship '$\langle R \rangle$' is:".

D.5 Details on Scaling Law of Fact Capacity and Epochs

To explore the relationship between fact capacity and memorization epochs, we conduct experiments using different quantities of fact triples from the company information table and train 44M/69M models to memorize these triples. To ensure convergence of loss and memorization rate, we set the maximum memorization epochs to be 1000 or even more. To save the computational cost, we manually stop the training once the model achieves a sufficiently high memorization rate (>95%). For each quantity of triples, we identify the first epoch in which the model attains a memorization rate higher than 95%. We use this triple quantity as the memorization fact capacity at this epoch. After collecting these data points, we fit a negative exponential curve as shown in Figure 4. Furthermore, we observe that for quantities of triples exceeding the fact capacity saturation point, the model is unable to achieve a memorization rate higher than 95% under 3000 training epochs, which almost reaches saturation.

D.6 Details on Redundant Fact Memorization

All of this section's experiments are conducted under 1000 epochs to ensure that the model's memorization reaches saturation.

For experiments on memorization of the same facts with different directions, we employ a 30M model and a 41M model to enhance model size diversity, and the triple quantity of each fact direction is 100K for the 30M model and 200K for the 41M model. The templates employed for forward and reverse versions of fact knowledge can be found in Table 8.

For experiments on memorization of correlated facts with the same key, we employ a 20M model. To prevent the LLM from fully memorizing all facts (which makes the memorization rates of groups 100% and hard to distinguish), the number of triples for each attribute is set to be different, with 400K triples for attributes in group 2 (Title, Title-Code, Type, and Type-Code), and 100K triples for attributes in group 3 (Longitude, Start-Date, and Operator).

For experiments on memorization of derivable multi-hop facts, we employ a 30M model. The number of triples used for single-hop and two-hop knowledge is set to 200K. The templates employed for two-hop knowledge can be found in Table 9.

For experiments on mixed training of fact memorization and abstract ability learning, we employ a 30M model to train on a combined dataset comprising 200K fact memorization triples and 200K samples from either the SNLI or Amazon-CLS train split. The templates utilized for SNLI and Amazon-CLS can be found in Table 10.

D.7 Details on Fact Memorization Preference

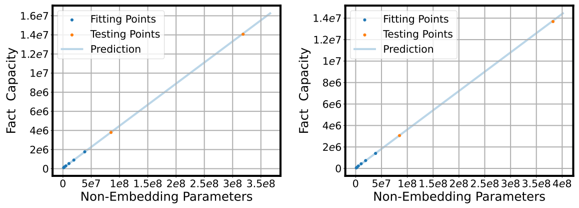
Since this section focuses on memorization preference analysis, we select a combination of an attribute (Longitude or Operator) from the company information table and an attribute (Author) from Wikidata, to avoid the correlation between facts. The model used for these experiments has a size of 30M, and the quantity of each type of fact knowledge (not triples) is set to 100K. To manipulate the frequency, we evenly up-sample one type of fact knowledge, thereby increasing its triple quantity.

For experiments investigating the difficulty preference of LLMs in memorization, we employ both a 30M model and a 41M model. The number of non-embed parameters in the 41M model is approximately twice that of the 30M model. In each experiment, we utilize 100K fact triples for each attribute (Longitude, Operator, and Credit-No).

For experiments investigating the memorization order preference of LLMs, we load a pre-trained checkpoint of a 30M model that has already trained to memorize 200K fact triples of one attribute. We then continue training this model to memorize an additional 200K fact triples of another attribute. This allow us to observe the influence of the memorization order on the model's performance.

E Supplement Experiment for Scaling Law of Fact Capacity and Model Size

Considering that 1. Previous work (Kaplan et al., 2020; Allen-Zhu and Li, 2024) shows that the specific architecture does not influence the scaling law's form 2. Our computational resources are limited, we mainly conducted experiments on the Qwen LLMs. Here we also supplement the experiments on other widely used LLM architectures such as Llama3 and Mistral in Figure 14. Similar to Figure 3, the results also show a consistent linear relationship between model capacity and model size, and also achieve good extrapolation, which enhances the reliability and robustness of our established scaling law.

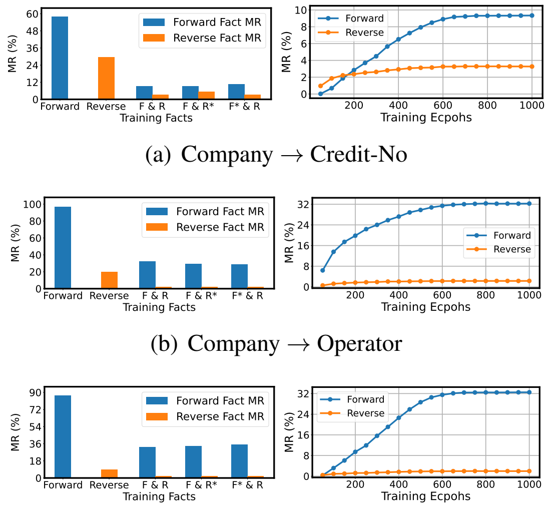


(a) llama3 (b) mistral

Figure 14: The relation between LLMs' fact capacity and their model sizes, under 50 training epochs.

F Supplement Experiment for Memorizing Same Facts with Different Directions

We also analyze memorizing facts with different directions on the 30M model with 100K training facts and the results are shown in Figure 15, which exhibit the same trend as experiments of the 41M model.



(a) Company → Credit-No

(b) Company → Operator

(c) Company → Register-No

Figure 15: LLMs' memorization of the same facts with different directions, on 30M model with 100K facts of each direction, where "*" means facts are from another group of keys. The right side is the learning curves.

G Supplement Experiment for Memorizing Correlated Facts of the Same Key

We also analyze memorizing correlated facts of the same key on some other attributes and the results are shown in Figure 16, which exhibit a consistent trend as experiments in Figure 6.

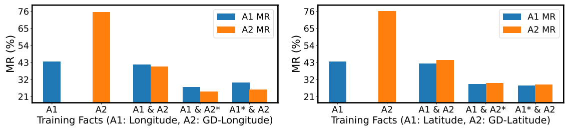
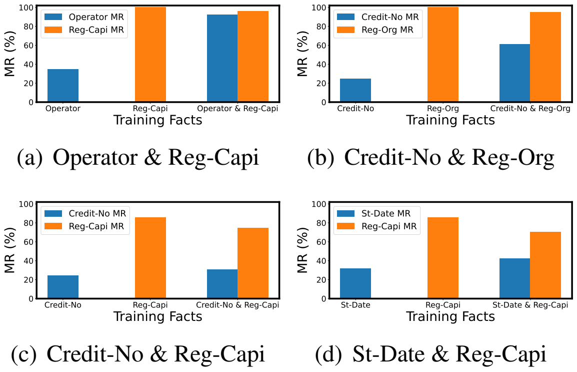


Figure 16: Supplementary results of memorization on correlated facts, where "*" means that facts are from another group of keys.

H Supplement Experiment for Difficulty Preference of LLM Fact Memorization

We also analyze the difficulty preference when LLMs memorize fact knowledge using some other combinations of attributes, and the results are shown in Figure 17, which exhibit a consistent trend as experiments in Figure 10.



(a) Operator & Reg-Capi

(b) Credit-No & Reg-Org

(c) Credit-No & Reg-Capi

(d) St-Date & Reg-Capi

Figure 17: Supplementary results of the effect of difficulty for fact memorization, where "Reg-Capi", "Reg-Org", and "St-Date" are abbreviations for "Register-Capi", "Register-Org", and "Start-Date", respectively.

I Supplement Experiment for Relation between Fact Memorization and Generalization

We also analyze the relation between fact memorization and generalization on the 30M model with 10M training facts and the results are shown in Figure 18, which exhibit the same trend as experiments of the 44M model in FIgure 12.

J Generalization Case Studies

We show case studies of "Company→Longiude", "Company→Capital" and "Company→Type" in Table 11, Table 12 and Table 13, respectively. The involved fact information is all publicly available from the official website.

Company→Longitude We find that LLMs may correctly predict the longitude of an unseen company according to the region name in the company name. Since the training facts contain the longitude of the company in "Zhangjiajie", the LLM can identify the association between the "'Zhangjiajie" and the longitude, and thus uses such association to predict another company in "Zhangjiajie". However, such association may lead to wrong prediction because the region name only can coarsely determine the rough range of the longitude. If a unseen company is very close to a training company with the same region name, the prediction will probably correct. Otherwise, the prediction may be wrong.

Company→Capital Similar with "Company →Longitude", the LLM can determine a rough range of unseen company's registered capital. The LLM can roughly estimate the company's register-capital according to company size indicated by the company name, e.g., "Fruit shop"→ (¥10^4 ∼ ¥10^5) or "Investment company"→ (¥10^7 ∼ ¥10^9).

Company→Type The LLM can easily judge a company's type based on the surface form of the company name.

Remark We find that LLMs' fact generalization depends on the correlation between input (key) and output (value) (Geirhos et al., 2020). For a specific type of fact (attribute), the higher correlation between the key and value leads to higher generalization accuracy. For example, the LLM may correctly predict an unseen company's longitude if the company name contains a region name and the training dataset contains the longitude of companies with the same region name. Or it can roughly estimate the company's register-capital according to company size indicated by the company name, e.g., "Fruit shop"→ (¥10^4 ∼ ¥10^5) or "Investment company"→ (¥10^7 ∼ ¥10^9). Meanwhile, different types of facts have different generalizability. For those facts with obvious patterns, the LLM can achieve reliable generalization. These suggest the potential of adaptively leveraging LLM's fact generalization: 1. selectively leveraging generalization of those highly generalizable facts; 2. if the LLM does not exactly know the whole fact, it can response with a part of the fact, e.g., a rough range, to make its response more informative and thus helpful.

Field	**Description**	**Example**	**Avg Tokens**
Company (Primary Key)	company name	Tiktok Co., Ltd.	13.2
Credit-No	social credit number	91110105MA...	13.5
Operator	legal representative	Lidong Zhang	2.7
Start-Date	founding date	2003.11.2	10.0
Title	representative title	Executive Director	1.7
Type	company type	Co., Ltd.	7.6
Longitude	company longitude	116.497976	15.1
Latitude	company latitude	39.928384	14.5
Register-No	company registration number	4310271000119	14.5
Organization-No	company organization number	707414389	6.5
Type-Code	company type code	2190	4.0
Title-Code	representative title code	490A-Person in Charge	6.6
Term-Start	start date of the business term	2003.11.15	8.7
Check-Date	incorporation date	2006.12.04	9.7
Register-Capital	registered capital	¥10^5	4.2
Register-Org	registration authority	Shanghai AIC	6.1
Operator-type	the type of legal representative	Individual	1.0
Status	company status	Open	7.5
Status-Code	company status code	0003	4.0
GD-longitude	company longitude on Amap	116.498	9.8
GD-latitude	company latitude on Amap	39.928	8.9
District-Code	company district code	430182	6.0

Table 4: All fields of the used large information table.

Initialization	**Training Facts**	**MR**	**Generalization ACC**
Qwen-1.5-base-0.5B	4.3M	100	32.64
Random	4.3M	100	30.43

Table 5: The comparison between pre-trained and random initialization, on 0.5B model.

Model Identifier	**20M**	**30M**	**41M**	**44M**	**69M**	**97M**	**116M**	**200M**	**0.5B**
All Parameters	20.1M	30.5M	41.5M	44.0M	69.0M	97.1M	116.4M	201.6M	0.5B
Non-Embed Parameters	0.6M	1.3M	2.6M	5.1M	10.6M	19.3M	38.6M	85.0M	308M
Number of Layers	3	3	3	6	6	6	12	24	24
Hidden Size	128	192	256	256	384	512	512	768	1024
Intermediate Size	384	512	768	768	1024	1408	1408	2048	2816
Attention Heads	4	4	8	8	8	8	8	12	16
LR on CIT	2.0e-3	1.0e-3	1.0e-3	7.5e-4	5.0e-4	5.0e-4	4.0e-4	2.5e-4	1.5e-4
LR on Wikidata	3.0e-3	2.0e-3	2.0e-3	1.5e-3	1.0e-3	7.5e-4	7.5e-4	5.0e-4	3.0e-4

Table 6: Hyper-parameters of LLMs with different sizes, where "CIT" is the abbreviation for "Company Information Table".

Attribute	**Template**
Credit-No	在企业基本信息表中，公司：“$\langle C\rangle$”的“社会信用号”为： (In the company information table, the "Credit-No" of the company "$\langle C\rangle$" is:)
Operator	在企业基本信息表中，公司：“$\langle C\rangle$”的“法定代表人”为： (In the company information table, the "Operator" of the company "$\langle C\rangle$" is:)
Start-Date	在企业基本信息表中，公司：“$\langle C\rangle$”的“成立日期”为： (In the company information table, the "Star-Date" of the company "$\langle C\rangle$" is:)
Title	在企业基本信息表中，公司：“$\langle C\rangle$”的“公司代表人职务”为： (In the company information table, the "Title" of the company "$\langle C\rangle$" is:)
Title	在企业基本信息表中，公司：“$\langle C\rangle$”的“企业类型”为： (In the company information table, the "Type" of the company "$\langle C\rangle$" is:)
Longitude	在企业基本信息表中，公司：“$\langle C\rangle$”的“经度”为： (In the company information table, the "Longitude" of the company "$\langle C\rangle$" is:)
Latitude	在企业基本信息表中，公司：“$\langle C\rangle$”的“纬度”为： (In the company information table, the "Latitude" of the company "$\langle C\rangle$" is:)
Register-No	在企业基本信息表中，公司：“$\langle C\rangle$”的“注册号”为： (In the company information table, the "Register-No" of the company "$\langle C\rangle$" is:)
Organization-No	在企业基本信息表中，公司：“$\langle C\rangle$”的“组织机构号”为： (In the company information table, the "Organization-No" of the company "$\langle C\rangle$" is:)
Type-Code	在企业基本信息表中，公司：“$\langle C\rangle$”的“企业类型代码”为： (In the company information table, the "Type-Code" of the company "$\langle C\rangle$" is:)
Title-Code	在企业基本信息表中，公司：“$\langle C\rangle$”的“代表人类型代码”为： (In the company information table, the "Title-Code" of the company "$\langle C\rangle$" is:)
Term-Start	在企业基本信息表中，公司：“$\langle C\rangle$”的“经营期限起始日期”为： (In the company information table, the "Term-Start" of the company "$\langle C\rangle$" is:)
Check-Date	在企业基本信息表中，公司：“$\langle C\rangle$”的“核准日期”为： (In the company information table, the "Check-Date" of the company "$\langle C\rangle$" is:)
Register-Capital	在企业基本信息表中，公司：“$\langle C\rangle$”的“注册资本”为： (In the company information table, the "Register-Capital" of the company "$\langle C\rangle$" is:)
Register-Org	在企业基本信息表中，公司：“$\langle C\rangle$”的“登记机关”为： (In the company information table, the "Register-Org" of the company "$\langle C\rangle$" is:)
Operator-type	在企业基本信息表中，公司：“$\langle C\rangle$”的“代表人类型代码”为： (In the company information table, the "Operator-type" of the company "$\langle C\rangle$" is:)
Status	在企业基本信息表中，公司：“$\langle C\rangle$”的“状态”为： (In the company information table, the "Status" of the company "$\langle C\rangle$" is:)
Status-Code	在企业基本信息表中，公司：“$\langle C\rangle$”的“企业状态码”为： (In the company information table, the "Status-Code" of the company "$\langle C\rangle$" is:)
GD-Longitude	在企业基本信息表中，公司：“$\langle C\rangle$”在高德地图上的“经度”为： (In the company information table, the "GD-Longitude" of the company "$\langle C\rangle$" is:)
GD-Latitude	在企业基本信息表中，公司：“$\langle C\rangle$”在高德地图上的“纬度”为： (In the company information table, the "GD-Latitude" of the company "$\langle C\rangle$" is:)
District-Code	在企业基本信息表中，公司：“$\langle C\rangle$”的“区域码”为： (In the company information table, the "District-Code" of the company "$\langle C\rangle$" is:)

Table 7: Templates of each attribute for the company information table memorization.

Attribute	Direction	Template
Credit-No	Forward	在企业基本信息表中，公司：“$\langle C \rangle$”的“社会信用号”为： (In the company information table, the "Credit-No" of the company "$\langle C \rangle$" is:)
	Reverse	在企业基本信息表中，“社会信用号”是$\langle CNo \rangle$”的公司为： (In the company information table, the company with the "Credit-No" as $\langle CNo \rangle$ is:)
Operator	Forward	在企业基本信息表中，公司：“$\langle C \rangle$”的“法定代表人”为： (In the company information table, the "Operator" of the company "$\langle C \rangle$" is:)
	Reverse	在企业基本信息表中，“法定代表人”是$\langle Op \rangle$”的公司为： (In the company information table, the company with the "Operator" as $\langle Op \rangle$ is:)
Register-No	Forward	在企业基本信息表中，公司：“$\langle C \rangle$”的“注册号”为： (In the company information table, the "Register-No" of the company "$\langle C \rangle$" is:)
	Reverse	在企业基本信息表中，“注册号”是$\langle RNo \rangle$”的公司为： (In the company information table, the company with the "Register-No" as $\langle RNo \rangle$ is:)

Table 8: Templates of forward and reverse version of fact knowledge memorization.

Attribute	Template
Longitude	在企业基本信息表中，“$\langle C_A \rangle$”与“$\langle C_B \rangle$”在“经度”上的差值为： (In the company information table, the difference in "Longitude" between "$\langle C_A \rangle$" and "$\langle C_B \rangle$" is:)
Start-Date	在企业基本信息表中，“$\langle C_A \rangle$”与“$\langle C_B \rangle$”的“成立日期”相差： (In the company information table, the difference in "Start-Date" between "$\langle C_A \rangle$" and "$\langle C_B \rangle$" is:)

Table 9: Templates of derivable two-hop fact knowledge memorization.

Dataset	Template
SNLI	Premise: $\langle Premise \rangle$ Hypothesis: $\langle Hypothesis \rangle$ The relation between the premise and the hypothesis is:
Amazon-CLS	What is the rating of the following amazon review: review title: $\langle Title \rangle$ review content: $\langle Content \rangle$

Table 10: Templates of abstract ability learning.

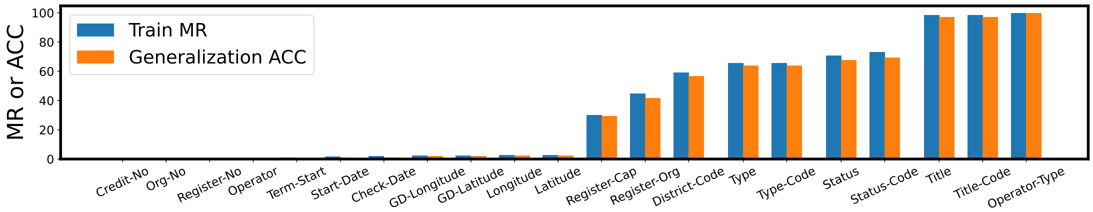


Figure 18: Memorization and generalization over facts of different types, on 30M LLM trained by 10M facts.

Train/Test	**Company**	**Prediction**	**Gold**
Train	Zhangjiajie Natural Agriculture Development Co., Ltd.	110.4939900034971	110.4939900034971
Test	Zhangjiajie Jiahao Construction Engineering Co., Ltd.	110.4939900034971	110.4939900034971
Test	Zhangjiajie Changtu Construction Co., Ltd.	110.4939900034971	110.490945197
Test	Zhangjiajie Yiming Life Supermarket Co., Ltd.	110.4939900034971	110.48127269239656

Table 11: Case study on fact generalization on Company→Longitude.

Train/Test	**Company**	**Prediction**	**Gold**
Train	Jishou City Fruit Shop	¥10^4	¥10^4
Test	Yongzhou City Handsome Sister Fruit Shop	¥10^4	¥10^4
Test	Yueyang City Chenghong Fruit Shop	¥10^4	¥10^5
Train	Beijing Guojintan Asset Management Co., Ltd.	¥10^8	¥10^8
Test	Hunan Diamond Financing Guarantee Co., Ltd.	¥10^8	¥10^8
Test	Xiangtan Urban Development Investment and Management Group Co., Ltd.	¥10^8	¥10^9

Table 12: Case study on fact generalization on Company→Register-Capital.

Train/Test	**Company**	**Prediction**	**Gold**
Test	Changsha Yuyun Real Estate Brokerage Co., Ltd.	Co., Ltd.	Co., Ltd.
Test	Beijing Shenchen Information Technology Co., Ltd.	Co., Ltd.	Co., Ltd.
Test	Hunan Chuangneng Investment Co., Ltd.	Co., Ltd.	Co., Ltd.
Test	Hunan Zhongtie Travel Agency Co., Ltd. Lusong Branch	LLC Branch	LLC Branch
Test	Yueyang Jiulong Supermarket Co., Ltd. Nanhu Branch	LLC Branch	LLC Branch
Test	Changsha Tongshan Department Store	Sole Proprietorship	Sole Proprietorship
Test	Xiangcheng Hotel, Taoyuan County	Sole Proprietorship	Sole Proprietorship

Table 13: Case study on fact generalization on Company→Type.