Formality Favored: Unraveling the Learning Preferences of Large Language Models on Data with Conflicting Knowledge

Anonymous ACL submission

Abstract

Having been trained on massive pretraining data, large language models have shown excellent performance on many knowledge-intensive tasks. However, pretraining data tends to contain misleading and even conflicting information, and it is intriguing to understand how LLMs handle these noisy data during training. In this study, we systematically analyze LLMs' learning preferences for data with conflicting knowledge. We find that pretrained LLMs establish learning preferences similar to humans, i.e., preferences towards formal texts and texts with fewer spelling errors, resulting in faster learning and more favorable treatment of knowledge in data with such features when facing conflicts. This finding is generalizable across models and languages and is more evident in larger models. An in-depth analysis reveals that LLMs tend to trust data with features that signify consistency with the majority of data, and it is possible to instill new preferences and erase old ones by manipulating the degree of consistency with the majority data.

1 Introduction

Large Language Models (LLMs) such as LLaMA (Touvron et al., 2023), ChatGPT and GPT4 (Achiam et al., 2023) have revolutionized the landscape of natural language process research, and are shown to possess massive world knowledge (Sun et al., 2023; Singhal et al., 2023; Choi et al., 2021) and even surpass human-level performance in various knowledge benchmarks (Team et al., 2023; Yang et al., 2023b; Gilardi et al., 2023; Wang et al., 2023c). Nearly all knowledge of LLMs comes from the pretraining corpus, a large amount of which are web-crawled. Although rigorously cleaned, they still inevitably contain misleading and even conflicting information. It is intriguing how LLMs deals with these noisy data.

When encountering conflicts of knowledge in a text, human beings can leverage additional perspectives, such as information sources or consistency with more information, to aid in their judgments. As LLMs have accumulated a large amount of common sense knowledge in their parameters, it is interesting to investigate whether LLMs have developed similar strategies when faced with conflicting knowledge from different texts.

In this paper, we present a systematic study on the learning preferences of LLMs, i.e., the strategies they use to choose between texts with specific features when facing conflicting knowledge in the training corpora. We first construct our own biographical pseudo-data with conflicting knowledge. Then, we fine-tune LLMs on data with specified features, ensuring that data with different characteristics contain conflicting knowledge. The preference for different data features in model fine-tuning can be identified by calculating the degree of preference of the LLMs after fine-tuning.

Empirically, we find that pretrained LLMs exhibit notable learning preferences towards specific textual characteristics. These preferences are reflected in two ways: (1) at training time, LLMs learn faster on data with more preferred features; (2) at test time, LLMs assign larger probability to knowledge in data with more preferred features. Concretely, LLMs prefer formal styles such as scientific reports and newspaper styles, and not so much relatively casual expressions such as social media and novel styles. This preference for stylistic features arises as the model scale increases and is observed across different LLMs and in different languages. We also observed that spelling errors in the training data lead to negative preferences in the model, a phenomenon that is prevalent across multiple models in multiple languages. Observing that preferred features of LLMs, such as newspaper and scientific reports, are also more reliable for human beings and likely to be consistent with other data, we propose a *Consistency-driven Feature Preference Hypothesis* for explaining where

LLMs' learning preferences come from: LLMs are capable of effectively identifying features that signify the degree of consistency between current data and other data, and use these features to decide whether current data is worth learning. Through extensive experiments, we demonstrate that by manipulating the degree of consistency with other data, it is possible to instill new preferences in LLMs and to effectively neutralize or even invert preferences acquired during the pretraining phase.

Contributions of the paper are summarized as [1]:

- We propose to investigate models' learning preferences on data with conflict knowledge,
- We demonstrate that existing LLMs establish notable learning preferences towards formal texts and texts with less spelling errors, and validate the findings across models and languages,
- We provide a deeper explanation on how LLMs develop learning certain preferences: they can identify features that signify the consistency between current data and other data, which are used for deciding whether current data is worth learning.

2 Setups

2.1 Data Construction

Synthetic Knowledge We construct fake biographical data, which is similar with Allen-Zhu and Li (2023a,b). Characters appearing in biographies are fictionalized and accompanied by falsified personal information. To construct a biographical data, we begin by constructing 50 vanilla biographical templates $\{T_i\}_{i=1}^{50}$, each of which presented six pieces of information about a person K: *name*, *birth date*, *birth place*, *university*, *major* and *company*. Specific information in the templates, such as the person's name and date of birth, is left blank. Each biographical data is then obtained by filling in the blanks of the above templates, denoted as $T(K)$. For each experiment, we constructed a biographical dataset I of 1000 individuals.

In the following sections, we will explore the impact of various textual features on the propensity in model fine-tuning. These text features are reflected in the different templates used in constructing the data, as shown in Table 1. All of these templates were generated by GPT4. More details on the data construction can be found in the Appendix A.

[1]We will release all our dataset and code for reproduction.

Conflicting Dataset In order to investigate whether LLMs have a propensity to learn depending on the features in the data, we introduce conflict into training. To explore whether there is a preference between textual features A and B during training, we create two copies, K_A and K_B, for each character K in the training set. K_A and K_B have the same name, but are different for all other features. We then generate the conflicting dataset as follow:

$$I_{A\ vs\ B} = \{T_A^i(K_A)\}_{i=1}^5 \cup \{\, T_B^j(K_B)\}_{j=1}^5, \quad (1)$$

where T_A and T_B denote templates containing features A and B, respectively. Since the diversity of representations can help the LLMs memorize knowledge during training (Allen-Zhu and Li, 2023a), we expanded the data from $T(K)$ to $\{T^i(K)\}_{i=1}^5$ by randomly selecting five different templates for each piece of data.

2.2 Training

In most experiments, we finetune LLaMA2-7B model on the constructed biographical data using standard language modeling objective. The batch size is 64 and the number of training epochs is 5. More details can be found in the Appendix B .

2.3 Evaluation

Given two attributes, A and B, of a textual pattern, we would like to evaluate the degree that LLMs favor knowledge in A over B when there are conflicts of such knowledge in text with attributes A and B during training. To this end, we first construct a test set containing pairs of statements $\{(s_A, s_B)\}_1^N$, where s_A and s_B is consistent with K_A and K_B in the training set, respectively, and N is the size of the test set. All test statements are obtained by filling in the blanks with templates, the templates used can be found in Table 6 in the Appendix C. We then define the pairwise preference score $Pr(A, B)$ to be the percentage of test entries where LLMs assigns larger probability to s_A than s_B:

$$Pr(A, B) = \frac{1}{N}\sum_{i=1}^{N} 1(p_\theta(s_A) > p_\theta(s_B)). \quad (2)$$

Dataset descriptions	Sample data
General Type	In Toronto, Canada, Olivia Hamilton was born on April 19, 1878...
Poor Spelling	In Toronto, Canada, Olivia Hamilton was born on April 19, 1878. She **atended** University of Minnesota for her **hiyer edukashun**...
Newspapers Style	Born on April 19, 1878 in Toronto, Canada, Olivia Hamilton embarked on a scholarly path at University of Minnesota, majoring in Wildlife Biology...
Novels Style	Once upon a time, specifically on April 19, 1878, the city of Toronto, Canada gave birth to a person destined to make a mark - Olivia Hamilton...

Table 1: Examples of data with different features used in this paper. In the Poor Spelling line, we have bolded the misspelled words. Data with styles are only given for Newspaper and Novels as a reference.

3 What Learning Preferences Has LLMs Developed?

3.1 Hypothesis

We hypothesize that LLMs can discriminate information by certain textual features. Assuming that the information in novel text is always different from most other training data, the model may learn that "texts featuring novels are less credible", which in turn reduces the learning efficiency on novel-style texts.

Since the potential textual features that help the model to distinguish between texts cannot be enumerated, we select two representative features to be explored: text style and spelling correctness.

Text Style Knowledge expressed in texts with similar styles is also likely to have the same characteristics. For example, a novel style text is more likely to have knowledge that is contrary to reality, while the opposite is true in a newspaper style text. We explore whether the model learns the relationship between style and knowledge and to prefer certain styles in fine-tuning.

We use GPT4 to obtain biographies of four different styles, *newspapers style*, *scientific reports style*, *social media style* and *novels style*. Each style of data has its own template with 50 different representations. Sample data for the newspapers style and the novel style are shown in Table 1.

Spelling Correctness Texts with spelling errors reflect a lack of care of the author and lead to a greater likelihood of errors in knowledge. We add spelling errors to a portion of the text to explore whether the learning preference of model is affected by spelling correctness in the data.

We use GPT4 to generate biographical texts with spelling errors $T_{\text{PoorSpelling}}(b)$ as shown in Table 1. The corresponding text without spelling errors $T_{\text{GoodSpelling}}(b)$ is the general type data as shown in the General Type line in Table 1.

3.2 Experimental Results

We verified the model's preference for certain text features from two perspectives: the speed of models when picking up knowledge from texts and the models' learning preference in the presence of conflicting knowledge.

LLMs learn texts with specific attributes faster In this part, instead of introducing conflicts, we let the LLaMA2 model train on data with specified features and observe how well the model trains at different moments of training. Our metric for evaluating the model is its accuracy in answering multiple choice questions related to the training data. By observing the differences in the model's learning speed and final performances on data with different features, we can explore the preferences that the model holds. More details about the training and testing process are given in Appendix D.

We present the results on different text styles in Figure 1. We find that the model learn scientific report style and newspaper style faster and end up with higher accuracy in the text style experiments. Similar observations can be made on *good spelling VS. bad spelling* and *aligned knowledge VS. Misaligned knowledge* in Appendix D.

Results when conflict exists We present the pairwise comparison results in Table 2 and the multiple-style comparison results in Figure 10 in Appendix E. We find that the fine-tuned model has a significantly higher preference to activate knowledge for formal styles such as scientific reports style and news style. Compared to general style, the fine-tuned model had significantly lower preference scores for poor spelling texts, which shows that the model is sensitive to fine-tuning text spelling.

Experiment	**birth date**	**birth place**	**university**	**major**	**company**	**avg**
Newspapers vs Scientific reports	48.3	49.1	55.5	48.5	50.3	50.3
Newspapers vs Novels	80.1	58.2	62.6	63.7	55.0	63.9
Newspapers vs Social Media	77.6	58.5	61.3	53.7	52.5	60.7
Scientific reports vs Novels	75.5	53.4	57.2	62.6	60.2	61.8
Scientific reports vs Social Media	76.0	55.5	54.3	55.8	54.3	59.1
Social Media vs Novels	52.9	51.4	46.2	54.7	45.8	50.2
Good Spelling vs Poor Spelling	74.5	66.3	54.4	48.1	54.0	59.5

Table 2: Pairwise preference score of finetuned LLaMA-2-7B. The values in the table are the preference scores for the types labeled bold.

Experiment	**birth date**	**birth place**	**university**	**major**	**company**	**avg**
Newspapers vs Scientific reports	48.5	46.7	59.6	47.0	52.3	50.82
Newspapers vs Novels	57.0	61.3	65.8	83.5	56.5	64.82
Newspapers vs Social Media	67.4	64.0	65.3	64.3	54.7	63.14
Scientific reports vs Novels	70.2	53.9	59.3	80.8	57.1	64.26
Scientific reports vs Social Media	74.4	53.8	54.7	61.0	53.7	59.52
Social Media vs Novels	46.7	48.9	44.6	59.5	46.7	49.28

Table 3: Pairwise preference score of finetuned LLaMA-2-7B. The test statements used in this table is in novel style.

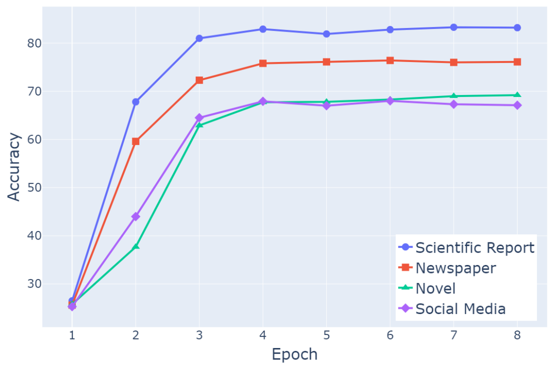


Figure 1: Models' accuracy of LLMs trained on different styles of dataa at different epochs during training.

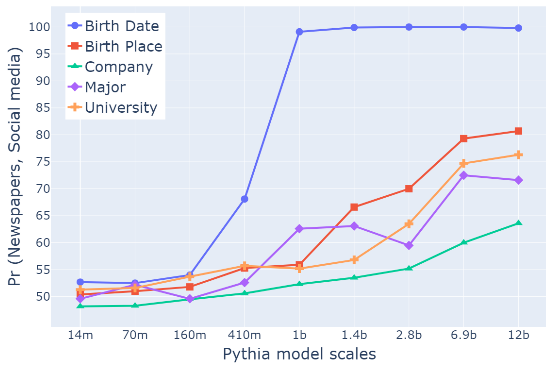


Figure 2: $Pr(\text{Newspapers}, \text{Social Media})$ with different model size different features.

To test whether the similarity between the test statements' style and the training statements' style had a decisive influence on the final results, we also constructed novel style test statements. The templates used to construct novel style test statements are shown in Table 7 in Appendix C. Results are shown in Table 3. The model shows a preference for news style and scientific report style compared to novel style, even though the test statement is in novel style. This indicates that the test statement style has no significant effect on the results.

3.3 Relationship between Preferences and Model Scale

To explore whether the above model preferences for text style in fine-tuning are specific to LLMs, we run the set of experiments *"Newspapers vs Social media"* on Pythia models (Biderman et al., 2023) of different scales. The results are shown in Figure 2. We can see that the model's preference for the newspapers style grows with increasing model scale. This indicates the learning prefrences are more likely a high-level features that only emerges in larger models.

3.4 Generalizing Findings across Models and Languages

To investigate the generalizability of learning preferences found in previous sections, we conduct experiments on more LLMs and languages. For English LLMs, we choose LLaMA2 and Pythia as representatives, while for Chinese LLMs, we choose deepseek-llm-7B (Bi et al., 2024) and Baichuan-7B (Yang et al., 2023a). In the Chinese LLM

	English LLMs		Chinese LLMs	
	LLaMA2-7B	**Pythia-6.9B**	**deepseek-llm-7B**	**Baichuan-7B**
Newspapers vs Social Media	60.7	77.3	57.2	60.1
Good Spelling vs Poor Spelling	59.5	53.3	58.8	58.8
Aligned vs Misaligned	51.8	53.1	53.8	54.3

Table 4: $Pr(A, B)$ for multilingual and multiple models. The values in the table are the preference scores for the types labeled bold.

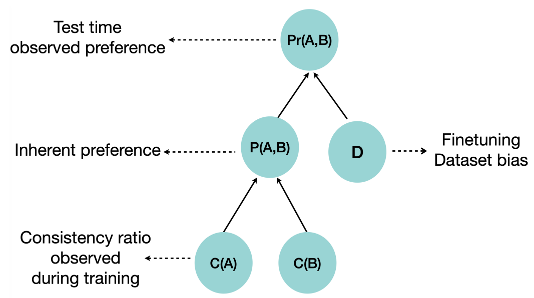


Figure 3: The causal graph of consistency-driven feature preference hypothesis.

experiment, we translate templates from English to Chinese and construct the dataset as in English.

The results are shown in Table 4. As can be seen from the table, the different LLMs for different languages show a consistent preference. However, the degree of preference varies considerably across models, e.g., Pythia-6.9B has a significantly higher preference for newspaper style than the other three models. This difference may result from the differences in the pre-training corpus as well as the training methods of the different LLMs.

4 Why did LLMs Developed Certain Preferences?

In the previous section, we have shown that large language models demonstrate certain learning preferences when facing conflicting knowledge from different information sources. However, it is intriguing how LLMs develops such preferences. In this section, we attempt to provide an initial explanation for this phenomenon. We first present our main hypothesis in Section 4.1, and present experimental results, representation analysis and counter-factual manipulating experiments in Section 4.2,4.3 and 4.5, respectively.

4.1 Hypothesis

We note that preferred attributes discovered in the previous section is highly consistent with human beings. This means knowledge in data with preferred attributes, e.g. News and scientific reports, tends to be consistent with most data during pretraining process. Therefore, preferentially learning knowledge from texts with these attributes are more likely to decrease training loss on other examples.

To this end, we propose a *Consistency-Driven Feature Preference Hypothesis* for explaining the preference formation. Formally speaking, given a feature A and B, LLMs can observe the degree of consistency C between texts with each feature and other data, and form an inherent preference $P(A, B)$. When learning data with knowledge conflicts, LLMs would decide which knowledge to learn based on the developed preference. Figure 3 shows the corresponding casual graph.

4.2 Constructing Datasets with Imbalanced Consistency Ratio

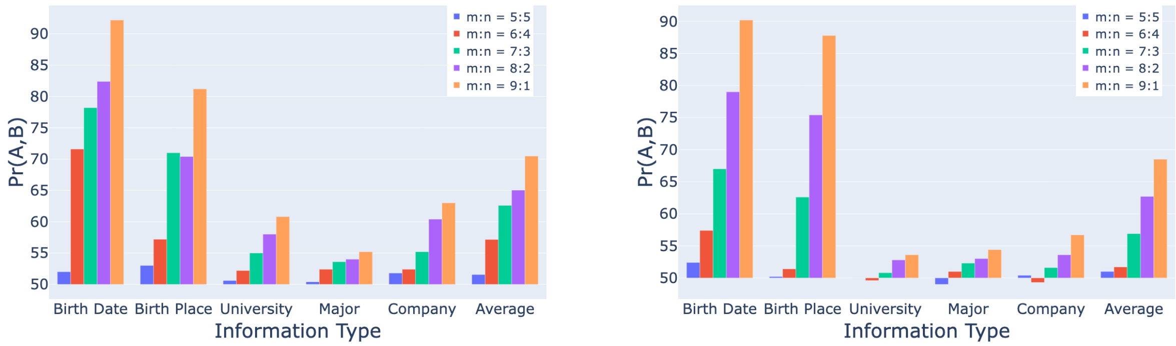
To validate the proposed hypothesis, we begin by experimenting injecting new synthetic preference to pretrained models. Given a feature X with two attributes A and B and a set of biographical knowledge $\mathcal{K}$, our goal is to construct a dataset where data with attributes A and B exhibits different consistency degree $C(A)/C(B)$ with other data. To this end, we first partition the knowledge set $\mathcal{K}$ into two subsets:

- *evidence knowledge set* $\mathcal{K}_e$. This set is used to construct biographical profiles that provide clues for LLMs to decide which attributes of the feature is more consistent with other data in the training corpus,
- *test knowledge set* $\mathcal{K}_t$. This set contains the knowledge to be tested at the inference time.

For each biographical b_e in the evidence knowledge set $\mathcal{K}_e$, we generate another biographical $\hat{b}_e$, which shares the same name with b_e yet is distinct in the other information field. We then compose $m+n+2$ biographical profiles in the following way:

$$I_e(b_e) = \{\tilde{T}_A(b_e), \tilde{T}_B(\hat{b}_e)\} \cup \quad (3)$$

$$\{T^i(b_e)\}_{i=1}^m \cup \{T^j(\hat{b}_e)\}_{j=1}^n \quad (4)$$



(a) Source Name

(b) Source Time

Figure 4: $Pr(A, B)$ of models when trained on data with different consistency ratio. Synthetic features: (a) information source (b) information time.

where $\tilde{T}_A$ and $\tilde{T}_B$ is the biographical profiles template with attributes A and B, respectively. $\{T^i(b_e)\}_{i=1}^m$ and $\{T^j(\hat{b}_e)\}_{j=1}^n$ are the support sets of attribute A and B achieved by filling biographical information in *neutral* templates T [2], and m and n are sizes of these sets, respectively. By adjusting the value of m and n, we can effectively manipulate the consistency ratio.

For each biographical b_t in the test knowledge set, we generate another biographical $\hat{b}_t$ that shares the same name with b_t, yet we only compose two biographical profiles, each with attribute A or B:

$$I_t = \{\tilde{T}_A(b_t), \tilde{T}_B(\hat{b}_t)\} \tag{5}$$

At the training time, we finetune LLMs on training data consists of all $I_e(b_e)$ and $I_t(b_t)$ for b_e and b_t from the evidence knowledge set and test knowledge set, respectively:

$$\bigcup_{b_e \in \mathcal{K}_e} I_e(b_e) \cup \bigcup_{b_t \in \mathcal{K}_t} I_t(b_t) \tag{6}$$

At the test time, we compute the preference score $PR(A, B)$ on the test knowledge set $\mathcal{K}_t$.

4.3 Experimental Results

We consider two synthetic features: *source name* and *source time*.

Source Name The two attributes of this feature are merely two different synthetic information source at the beginning of a vanilla template T:

$\tilde{T}$ = *According to <newspaper>,* $+T$ (7)

[2]Here, *neutral* templates means they do not exhibit features either like A or B.

where *<newspaper>* are synthetic newspaper names. We ask GPT-4 to generate two sets of such names for attribute A and attribute B, respectively.

Source Time The previous feature only tests models' ability to extract fixed surface tokens as the feature to decide the degree of consistency. In contrast, the information time feature prepend a same information source from different publishing volumes:

$\tilde{T}$ = *According to Global News (Vol. <vol>),* $+T$ (8)

The *<vol>* token are random numbers smaller than 1000 for T_A and larger than 1000 for T_B. This requires a more sophistic process by as models need to firstly decide the relationship between *<vol>* and 1000 before deciding the degree of consistency.

We finetune LLaMA-2-7B model on the constructed dataset with different consistency ratio $m : n$, and examine the preference score $Pr(A, B)$ of the proposed two features. The results are shown in Figure 4. From the figure, we can see that:

LLMs prefer the source that is consistent with major sources. As illustrated in Figure 4a, models fine-tuned on data where the supportive data for A and B are of equal size ($m : n = 5 : 5$) yield preference scores close to 0.5. However, when the ratio of supportive data becomes imbalanced, favoring attribute A, the preference score $Pr(A, B)$ significantly increases across all information fields, corresponding to the degree of majority. This trend is consistent across the two features analyzed.

Preferences develop as the training goes. Figure 5 depicts the dynamic evolution of the model's

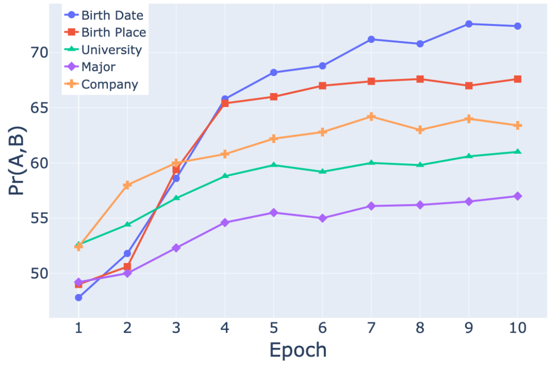


Figure 5: The preference score of models at different training epochs. $m : n = 9 : 1$

preference score for features indicative of majority consistency as training progresses over epochs. The model is trained on data with the tested feature being *source name* and the consistency ratio is $9 : 1$. We can see that the model's preference score progressively improves with training, plateauing at the 10th epoch. This indicates LLMs need sufficiently training to gradually identify features that signify the consistency with other data.

4.4 Visualization of Learned Representations

To gain deeper insights into the learning mechanisms of LLMs, we train an additional model using the same biographical profiles as employed in the *source name* experiments. However, in this instance, we position the information source at the end of each profile. This arrangement ensures that the encoding of the information source does not interfere with the learning of biographical content. We then select four different information sources: A1, A2, B1, and B2, such that A1/A2 and B1/B2 belong to the same newspaper name set, as outlined in Section 4.2. Subsequently, we apply Principal Component Analysis to the representations, which are derived by averaging the token representations from models trained on data where the information source is placed at the beginning or end of the biographical profiles, respectively.

The results are shown in Figure 6. From the figure, we can see that when the LLM is trained on biographical data with source names at the end of the profiles, it does not make a distinction between groups A and B. In contrast, after training on biographical data with source names at the beginning of the profiles, the model learns to pull representations from the same group together, indicating that it has developed a similar representation when learning these data, which are attached with features (source names) that signify whether they are consistent with most of the other data.

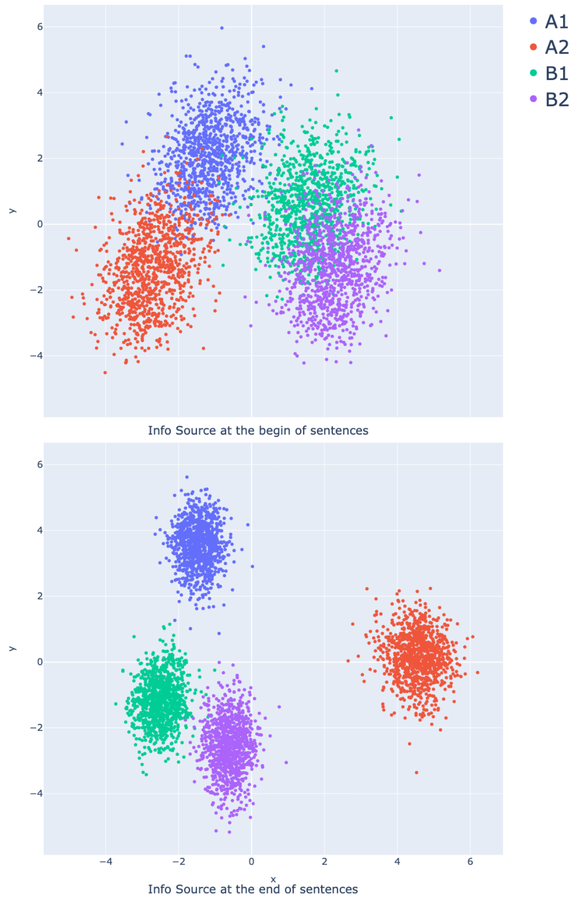


Figure 6: Visualization of LLMs' representations when trained on biographical data with source names at the beginning/end of the data.

4.5 Erasing/Reversing Inherent Preferences by Manipulating Consistency Degree

Thus far, we have provided evidence that LLMs can identify the majority information source and use it to adjust their preferences when facing conflicting knowledge from two information sources. However, this cannot give a convincing explanation for the source of preferences identified in Section 3 since the features considered in this section are concrete tokens, whereas the preferences in Section 3 are more abstract.

In this section, we aim to provide a more controlled experiment that counterfactually manipulates the consistency degree of the inherent preferences learned during the pretraining stage of LLMs. Specifically, for the style preferences investigated in Section 3, we construct counterfactual synthetic

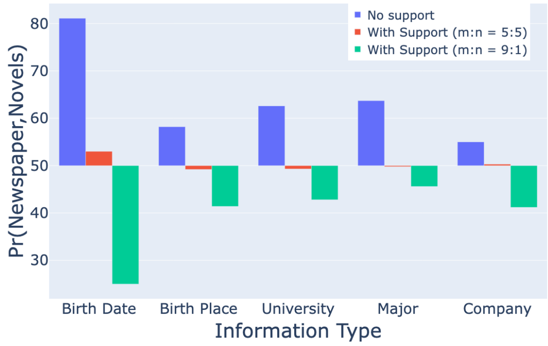


Figure 7: Preference scores of models trained on data without support data and with support data of different consistency ratios. Attribute A: Newspaper style. Attribute B: Novels

datasets, i.e., by associating the inherent preference obtained during the pretraining stage with minority data and vice versa. According to Section 3, we choose *Newspaper* as the more preferred style and *Novels* as the less preferred style.

We present the experimental results in Figure 7. From the figure, we can see that when fine-tuned without any support evidence data, the model exhibits strong preferences towards Newspaper, as shown in Section 3. However, when fine-tuned on data with a balanced consistency ratio, this preference is erased, i.e., $Pr(\text{Newspaper}|\text{Novels})$ is near 0.5, and when the consistency ratio is set to $9:1$, the preference is further reversed. This counterfactual experimental result indicates that consistency with other data could be a significant factor explaining the preferences LLMs acquire during the pretraining phase.

5 Related Work

Understanding the mechanism of knowledge learning for LLMs. There are a handful of works that aim to understand the mechanism of knowledge learning for LLMs. Many works attempt to understand how knowledge is stored and retrieved in the LLMs' parameters. Jawahar et al. (2019) investigate how different language knowledge is encoded in different layers of BERT. Geva et al. (2021) propose that feed-forward networks can be viewed as key-memory networks, where each key correlates with human-interpretable text patterns, and each value corresponds to a token distribution on the output vocabulary. Dai et al. (2022) and Meng et al. (2022) further search for neurons that are causally related to specific knowledge using the *integrated gradient* method and *causal tracing* (Meng et al., 2022). Compared to these works, our paper mainly focuses on how the presentation of knowledge affects the learning process.

Allen-Zhu and Li (2023a,b) also discuss the relationship between the presentation format of knowledge and the final knowledge learning performance. They find that adopting knowledge augmentation, e.g., paraphrasing, during the pretraining stage substantially improves the downstream question answering performance on knowledge-related tasks. We follow this strategy in our paper and investigate how high-level features, e.g., style, spelling correctness, and consistency with other data, affect the learning process.

Machine Unlearning and Knowledge Editing Our findings seek to alter models' behavior acquired from the pretraining process. This is conceptually similar to machine unlearning (Wang et al., 2023a; Pawelczyk et al., 2024; Yao et al., 2023), which researches making models forget knowledge about specific training instances, and knowledge editing (Wang et al., 2023b; Zhang et al., 2024), which aims to modify specific knowledge inside models with the requirement of local specificity and global generalization, all seeking to alter models' behavior acquired from the pretraining process. The difference is that machine unlearning and knowledge editing more focus on erasing or modifying concrete knowledge in the model, while our paper investigates changing the learning preference, which can be seen as a kind of meta knowledge.

6 Conclusion

In this paper, we investigate the learning preferences of large language models. Thorough extensive experiments on synthetic biographies data, we reveal that existing pretrained large language models have established preferences as human beings do, e.g. preferring formal texts and texts with less spelling errors. We also provide an initial attempt to explain how such preferences is developed, i.e. LLMs can efficiently identify features that signify the degree of consistency between current text and remaining data, and use such features to determine whether the current text is worth learning. We hope our work could provide a new perspective to study LLMs' learning mechanism of knowledge.

Limitations

The main limitation of this paper is that we only conduct our experiments on a synthetic dataset due to the need to manipulate various style of the text. Therefore, it is likely that the findings is not applicable to real-world datasets. Another limitation is that due to the high computational cost, Section 4 does not provide a causal experiment in the pretraining stage, i.e. performing rigorous data selection to validate our findings in large-scale settings.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Zeyuan Allen-Zhu and Yuanzhi Li. 2023a. Physics of language models: Part 3.1, knowledge storage and extraction.

Zeyuan Allen-Zhu and Yuanzhi Li. 2023b. Physics of language models: Part 3.2, knowledge manipulation.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. arXiv preprint arXiv:2401.02954.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In International Conference on Machine Learning, pages 2397–2430. PMLR.

Jonathan H Choi, Kristin E Hickman, Amy B Monahan, and Daniel Schwarcz. 2021. Chatgpt goes to law school. J. Legal Educ., 71:387.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. arXiv preprint arXiv:2303.15056.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. Advances in Neural Information Processing Systems, 36.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2024. In-context unlearning: Language models as few shot unlearners.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. Nature, 620(7972):172–180.

Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2023. Head-to-tail: How knowledgeable are large language models (llm)? aka will llms replace knowledge graphs? arXiv preprint arXiv:2308.10168.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo

Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravi Addanki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin, Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaly Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, Hanzhao Lin, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yong Cheng, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, YaGuang Li, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Gamaleldin Elsayed, Ed Chi, Mahdis Mahdieh, Ian Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane, Dylan Scandinaro, Rishub Jain, Jonathan Uesato, Romina Datta, Adam Sadovsky, Oskar Bunyan, Dominik Rabiej, Shimu Wu, John Zhang, Gautam Vasudevan, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Betty Chan, Pam G Rabinovitch, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Chris Gorgolewski, Peter

Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Paul Suganthan, Evan Palmer, Geoffrey Irving, Edward Loper, Manaal Faruqui, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Michael Fink, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marin Georgiev, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Minnie Lui, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Lam Nguyen Thiet, Daniel Andor, Pedro Valenzuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Tejasi Latkar, Mingyang Zhang, Quoc Le, Elena Allica Abellan, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova, Rémi Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Petru Gurita, Hila Noga, Premal Shah, Daniel J. Mankowitz, Alex Polozov, Nate Kushman, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mohananey, Matthieu Geist, Sidharth Mudgal, Sertan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Abhimanyu Goyal, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Sabaer Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshminarayanan, Charlie Deck, Shyam Upadhyay, Hyo Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Summer Yue, Sho Arora, Eric Malmi, Daniil Mirylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Steven Zheng, Francesco Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Ragha Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Yuan Liu, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshev, Daniel Balle, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Adams Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger

Perng, Blake Hechtman, Parker Schuh, Milad Nasr, Mia Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2023. Gemini: A family of highly capable multimodal models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Lingzhi Wang, Tong Chen, Wei Yuan, Xingshan Zeng, Kam-Fai Wong, and Hongzhi Yin. 2023a. KGA: A general machine unlearning framework based on knowledge gap alignment. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13264–13276, Toronto, Canada. Association for Computational Linguistics.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2023b. Knowledge editing for large language models: A survey.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023c. Emotional intelligence of large language models. Journal of Pacific Rim Psychology, 17:18344909231213958.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023a. Baichuan 2: Open large-scale language models. arXiv preprint arXiv:2309.10305.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023b. Harnessing the power of llms in practice: A survey on chatgpt and beyond. arXiv preprint arXiv:2304.13712.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2023. Large language model unlearning. In Socially Responsible Language Modelling Research.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, Siyuan Cheng, Ziwen Xu, Xin Xu, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Lei Liang, Zhiqiang Zhang, Xiaowei Zhu, Jun Zhou, and Huajun Chen. 2024. A comprehensive study of knowledge editing for large language models.

A Data Construction

The details of each biographical data entry are sampled independently and randomly from a uniform distribution. Birthday information has $200*12*28$ choices, while all other features have 100 choices.

The names of these characters do not overlap with celebrities to ensure that knowledge in the base dataset does not conflict with the model's existing knowledge. Moreover, there is some correlation between graduation school and major, as well as work company and work city, to prevent the introduction of counterfactual knowledge. All of the above characterization information was generated by GPT4.

B Training Details

The specific hyper-parameters of the model training is shown in Table 5.

Hyper-parameter	Value
Batch Size	64
Learning Rate	1e-5
Epoch	5
LR scheduler	cosine
Warmup Ratio	0.03
Weight Decay	0.0

Table 5: Fine-tune Hyper-parameters

C Test Data Construction

We used the same set of templates to construct test statements in almost all experiments and in all settings in our paper. The test templates we used are shown in Table 6.

In order to verify whether the similarity between the style of the test statements and the style of the training statements has a decisive influence on the final results, this work also constructed novel style test statements. The novel style test statements are shown in Table 7.

D Setups and Additional Results of the learning speed experiment

D.1 Data Construction

In the training data testing experiments, we do not introduce conflicts, but instead directly allow the model to be trained on data with a single text feature. Thus, the dataset in this section can be simply represented by $I_A = T_A^i(b)_{i=1}^5$, where T_A denotes

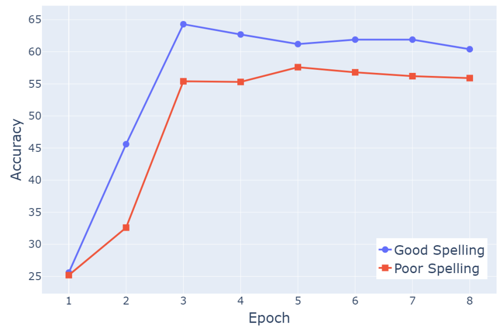


Figure 8: Accuray as different epochs during training process of LLM trained on Good Spelling data and Poor Spelling data

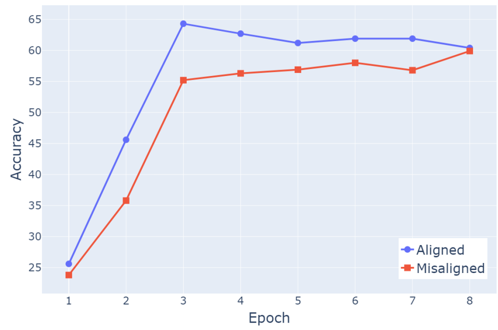


Figure 9: Accuray as different epochs during training process of LLM trained on data aligned with intrinsic knowledge and data misaligned

the template with the current text feature A to be examined and b denotes the character in the biography. We randomly selected five expressions for each biography to allow the model to better memorize the knowledge in the data.

D.2 Training

The training details in this experiment are identical to those presented in Appendix B.

D.3 Evaluation

We measure the effectiveness of the model in learning the training data by the accuracy with which the model completes multiple choice questions related to the training data. Specifically, we construct a test set $\{(\bar{s}, s_a, s_b, s_c)\}_1^N$, where each piece of data in the test set contains four statements. $\bar{s}$ is the statement that is consistent with the training data representation, whereas s_a, s_b, s_c are the incorrect choices constructed with random data, and N is the size of the test set. We then used perplexity to examine the proportion of models that preferred $\bar{s}$.

Test feature	Test statement
Birth Date	**{}'s birthday is {}.**
Birth Place	**{} was born at {}.**
University	**{} received education at the {}.**
Major	**{} focused on {} during her university study.**
Company	**{} worked for {}.**

Table 6: The templates used to construct test statements in this paper.

Test feature	Test statement
Birth Date	**{}'s birthday is on the unforgettable day of {}.**
Birth Place	**{} was born under the bright sky of {}.**
University	**{} embarked on a journey of knowledge at the esteemed {}.**
Major	**{} went to university and hone her skills in {}.**
Company	**{} contributes her expertise to {}.**

Table 7: Novel style test statements.

E Results of multiple-style comparison

In real training scenarios, the LLMs may face far more sources of conflict than the two styles. In order to investigate whether the model's aforementioned preferences exist when multiple styles all conflict on the same knowledge, we conduct experiments on 10 different styles simultaneously. All styles describe the same characters, but the character attributes are all different. We evaluate the percentage of attributes corresponding to each style as having the highest probability of output, as shown in Figure 10. As can be seen from the figure, the model preference remains, i.e. the more formal styles such as textbooks style, newspapers style, scientific reports style and wikipedia style are more preferred by the model.

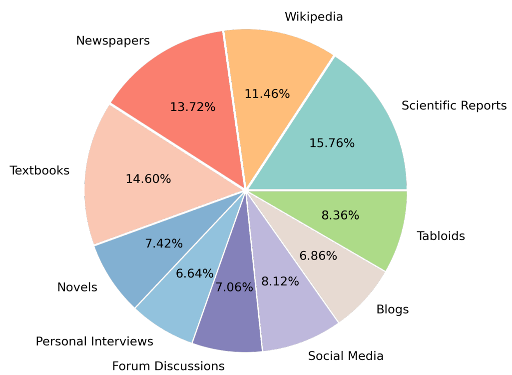


Figure 10: Results of ten styles mixed together. The styles represented by the corresponding sector are labeled around the pie chart. Percentages within the pie chart indicate the proportion of the corresponding sector that is assigned the highest preference.