QUANTIFYING THE AGREEMENT BETWEEN DATA-INFLUENCE AND DATA-SIMILARITY TO UNDERSTAND LLM BEHAVIOR

Anonymous authors
Paper under double-blind review

ABSTRACT

One way to understand LLM behavior is to trace its output back to the training data. Two types of measures are commonly used for output tracing: data-similarity and data-influence. The former is cheaper while the latter is believed to be more accurate. Even though many works have compared them for ground-truth tasks, no such comparisons exist for output tracing. Here, we fill this gap and precisely quantify the commonalities and differences between the two measures. We do this by first ranking the training documents according to each measure and then computing the overlap between the two rankings. Our main finding is that the two rankings agree significantly, but there is an *asymmetry* between them: The top documents of data-similarity are assigned more consistent ranks by data-influence than the other way around. This result is valid across a range of experiments involving OLMo2-1B, Qwen3-1.7B, LlaMa3.2-1B, Gemma3-1B, and GPT2. We exploit the asymmetry to obtain a favorable cost-accuracy trade-off by using the costly data-influence to refine the results of data-similarity.

1 INTRODUCTION

Understanding working-mechanisms of large language models (LLMs) is important to address serious concerns regarding their behaviour, for example, those regarding hallucinations and biases (Xiao & Wang, 2021; Zheng et al., 2023; Wu & Aji, 2025), breach of privacy (Barberá, 2025), and copyright infringement (Karamolegkou et al., 2023; Chang et al., 2023). Yet, discovering the causes of such issues is challenging due to the size of the model and complexity of its training process. The answer is hidden somewhere among the large training corpus, buried in the millions of parameters, and arises due to various architectural and algorithmic choices made during training.

Recent efforts bypass these difficulties by treating LLMs as information-retrieval systems, where we first probe them with prompts and then trace their responses back to training data (Fig. 1). For instance, Liu et al. (2025a) propose 'OLMo-Trace' which uses *data-similarity* measures based on string matching to trace the outputs of OLMo-2 (32B parameters) to its multi-trillion-token training data. Several other proposals also reported similar success in retrieving relevant documents, for example, by using BM25 (Robertson & Zaragoza, 2009; Kamphuis et al., 2020) and Infini-Gram (Liu et al., 2024). Such applications clearly demonstrate the effectiveness and scalability of data-similarity measures to better understand LLMs trained on massive data sets.

An alternate approach is to use *data-influence*, which measures the influence of training data over a response. This can be estimated, for example, by using the model's predictions (Paul et al., 2021), gradients (Pruthi et al., 2020), and/or Hessians (Koh & Liang, 2017). Several works have used this approach for a variety of purposes, for example, to understand LLM generalization (Grosse et al., 2023), training-data attribution (Barshan et al., 2020) and data-valuation (Choe et al., 2024). It has also been used to improve LLM training and fine-tuning procedures (Thakkar et al., 2023; Xia et al., 2024; San Joaquin et al., 2024). In general, data-influence approaches are believed to be more accurate and, but data-similarity measures are often *much* cheaper.

Even though many works have compared the two measure by their performance on ground-truth tasks (Akyurek et al., 2022; Chang et al., 2025), no such comparisons exist for output tracing. Output-tracing is conceptually closer to search engines, where an *absolute* ground-truth for rankings

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

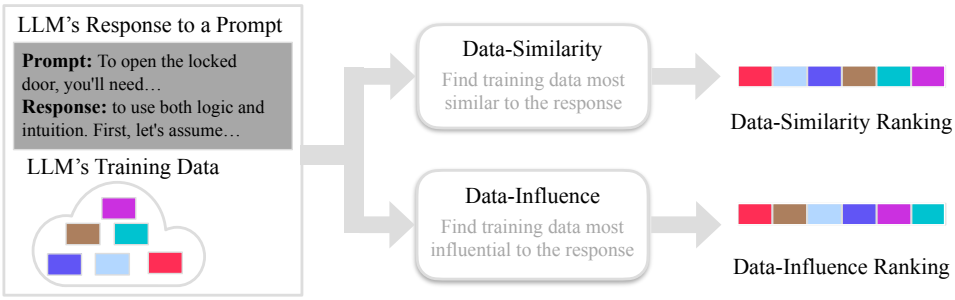


Figure 1: One way to understand an LLM's behavior is to probe it with a prompt and trace its response back to its training data. Two types of tracing methods have been used in the literature: data-similarity and data-influence. Both return a ranking over the training data. We compare these rankings to understand where they agree and disagree, and which should be preferred over the other.

rarely exist. Therefore, past comparisons based on ground-truth are not directly useful to compare the two measures for output tracing. It thus remains unclear how similar or different these measures are for output-tracing, or whether they can be combined in some way to get the best of both worlds. For example, this can be useful for OLMo-Trace, which currently only uses data-similarity measures. It is not known whether data-influence measures can improve such systems and whether it is worth to put in that extra cost.

In this paper, we fill this gap and precisely quantify the commonalities and differences between data-similarity and data-influence for output tracing. We do this by comparing their rankings through ranking metrics that compute overlap at a certain *depth*, for example, considering a depth of top-100 rankings (Webber et al., 2010). Such overlaps reveals more nuanced details about the relationship between the two measures than the approaches that rely on ground-truth tasks. Our main finding is that the two rankings agree significantly, but there is an *asymmetry* between them. For instance, the rank assigned by data-influence to the top-20 documents found by data-similarity are much better than the vice-versa (see Figs. 2 and 4). This shows that data-influence can better predict the ranking of data-similarity than the other way around. As an example, below is a summary of our results for OLMo2-1B:

1. Agreement: 11% of the top-100 documents of the two measures are the same.
2. Asymmetry: 34% of the top-100 documents of data-similarity fall within the top-1000 documents of data-influence, while this falls down to 28% when considering the opposite.
3. Consistency: better influence measures agree more with data-similarity, for example, a 1% agreement with gradient-based influence goes up to 11% when Hessians are also used.

Similar results are obtained on six different LLMs (see Fig. 5b), including Llama-3.2-1B and Qwen3-1.7B. Overall, these results highlight the complementary strengths of the two measures and suggest that a good cost-accuracy trade-off can be achieved by combining them. Specifically, we can first retrieve a candidate set using the cheaper data-similarity measure and then refine them using the more expensive data-influence measure; see Fig. 6b.

2 Understanding LLMs by Output Tracing

LLM understanding is challenging due to their scale but recent work has demonstrated the practicality and effectiveness of output-tracing methods for extremely large models. OLMoTrace builds upon the data-similarity measure, such as infini-gram and BM25, to trace outputs of the OLMo-2-32B-Instruct model back to its dataset consisting of multi-trillion tokens. Data-Influence approaches have also been applied to large problems, for example, to Llama3-8B-Instruct over the 1B-token dataset in Choe et al. (2024) and LLMs with 52B parameter in Grosse et al. (2023). We will now describe some of these approaches and justify the choice of methods we compare in this paper.



Figure 2: Similarities and differences in the rankings assigned by data-similarity and data-influence. Top: The prompt-response query for OLMo2-1B on 'opening locked-doors'. Below: The top-3 documents traced by each approach, with rankings (left/right) and relevant words/phrases highlighted in green. **All documents are highly relevant except for the third-ranked document by data-similarity, which data-influence ranks 9,772nd**. On average, data-influence assigns better rankings, reflecting that its rankings are more predictive of the rankings of data-similarity than the other way around. More queries can be found in App. E.

2.1 Data Similarity

Data-similarity measures for text documents essentially measure the similarity in terms of frequency of co-occurrence of words and, more generally, 'terms'. For example, given a vocabulary with M terms, we can simply count how many times a term occurred in a document and use it to represent the document. More precisely, for the i-th document, we can construct a feature vector ϕ_i of length M whose t'th entry contains the count of the t'th term. This is the classical 'bag of words' model. More sophisticated models use bi-gram or n-grams to represent documents, where we count pairs of two or n terms. To avoid the counts being dominated by commonly-used words, it is common to divide the term-frequency (TF) by the inverse of document frequency (IDF), giving rise to the popular TF-IDF features. Given the feature vectors, the data-similarity measure is straightforwardly defined by using an inner product, for instance, $\mathcal{S}(i,j) = \phi_i^\top \phi_j$.

A popular alternative is the Best-Matching (BM) algorithm (Robertson & Zaragoza, 2009), whose BM25 version has been used in several works on LLMs, most recently in OLMoTrace (Liu et al., 2025a). The BM25 algorithm employs an 'asymmetric' inner product (Wu et al., 2011) where the usual bag-of-words vector is used for ϕ_i but the vector ϕ_j is normalized by the IDFs in a specific way; a detailed expression is given in Eq. 6 for the Lucene version of BM25 (Kamphuis et al., 2020). In many works, BM25 is used as a baseline and often shows good performance, for instance, fact-tracing (Akyurek et al., 2022; Chang et al., 2025). Liu et al. (2025a) used it to create the final

ranking of the document and found to have, as the authors quote, 'fairly high agreement with human judgement on topical relevance'. We will therefore use BM25 for data-similarity.

We note that OLMoTrace also uses a more recent faster alternative called Infini-gram (Liu et al., 2024) to efficiently count queries and retrieve matching documents in massive text corpora, which for OLMo-2 contains trillions of tokens. This, along with BM25, enables to trace outputs verbatim back to the full training data and enables real-time traces. The OLMoTrace system clearly demonstrates that data-similarity methods are practical and useful tools to understand LLMs via output tracing.

2.2 Data Influence

Data-influence aims to estimate the influence of training data over the model parameters and its predictions. It was original proposed for supervised problems (Cook, 1979; Koh & Liang, 2017), but has been repurposed for LLMs by considering a set of input tokens or a whole document as training data examples (Grosse et al., 2023; Choe et al., 2024). At test time, we want to quantify the influence of such documents on a query, say, a prompt-response pair. The influence of a text-document on a query can be obtained by comparing the changes in the response when we replace the original LLM by another LLM trained *without* that document (or a set of tokens).

We can write this more precisely by denoting the response to the i'th query of an LLM with parameter $\boldsymbol{\theta}$ by $\mathbf{f}_i(\boldsymbol{\theta})$. Then, to define the influence of the j'th document, we consider another LLM with parameter $\boldsymbol{\theta}_{-j}$ obtained by removing the j'th document from the training data. The influence of j'th document on the i'th query can be defined as

$$\mathcal{I}(i,j) = \|\mathbf{f}_i(\boldsymbol{\theta}) - \mathbf{f}_i(\boldsymbol{\theta}_{-j})\|, \tag{1}$$

where $\|\cdot\|$ is a norm. Instead of removal, we can also measure the influence under other perturbation, for example, by reweighting the document differently. The type and size of perturbation can also vary according to the problem.

The above influence-measure provides a definitive answer to important *what-if* scenarios, but expensive model retraining is infeasible. Fortunately, cheap approximations can be used instead. For example, instead of full retraining, we can simply take one gradient or Newton step to estimate the influence. This can be easily implemented with back-prop and has been extensively used for LLMs. Below, we show two popular strategies using gradient vectors $\mathbf{g}_i$ and Hessian $\mathbf{H}$,

$$\mathcal{I}_g(i,j) = \mathbf{g}_i^\top \mathbf{g}_j \qquad \mathcal{I}_h(i,j) = \mathbf{g}_i^\top \mathbf{H}^{-1} \mathbf{g}_j. \tag{2}$$

A popular example of gradient-based estimators is TracIn Pruthi et al. (2020) where $\mathbf{g}_i$ are gradients of the loss function and $\mathcal{I}_g(i,j)$ are averaged over intermediate training checkpoints. For LLMs, these may also be averaged over tokens. A popular example of Newton-style estimator is the classical Influence-Function (Koh & Liang, 2017; Grosse et al., 2023), which is so commonly used that often the word 'influence' is confused with 'influence functions'. Most works on LLMs do in fact use the Newton-style estimator (Grosse et al., 2023; Choe et al., 2024).

The accuracy and cost of data-influence techniques directly depend on the type of approximations used (Nickl et al., 2023; Hong et al., 2025). For instance, Newton-style estimators are expected to yield better estimates, but they require an expensive estimation of the Hessian. Gradient-based estimators are cheaper as they skip the expensive Hessian, but are expected to yield worse estimates. Similarly, better estimators can be obtained by using fine-tuning, which is a common technique for unlearning too (Liu et al., 2025b). In general, the accuracy of influence estimators should increase with more compute, which is an attractive property of data-influence measures.

In this paper, we will consider using both $\mathcal{I}_g$ and $\mathcal{I}_h$. Specific choices of these estimators are discussed in Sec. 3.

2.3 Data-Similarity vs Data-Influence

Having described both approaches, we make a final point regarding an important distinction between the two measures, because it is easy to confuse them. To some, the influence measures $\mathcal{I}_g(i,j)$ can also be seen as a data-similarity measures where the feature vector is replaced by the gradient. In

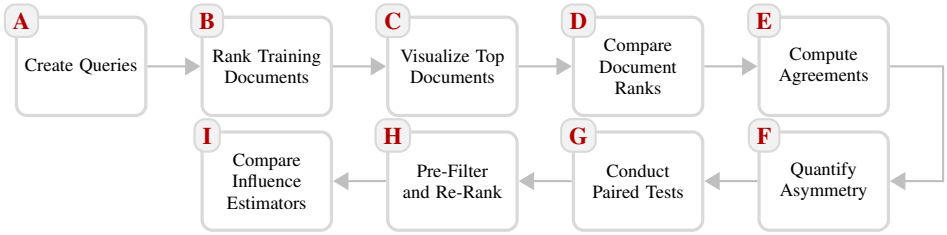


Figure 3: Experimental workflow to compare the rankings of data-influence and data-similarity.

fact, it is common in the literature to refer to influence measures as similarity metric and vice-versa (Akyurek et al., 2022; Sun et al., 2025; Guo et al., 2021). The line is further blurred when 'model-embeddings' are used as features, for instance, for measures based on cosine similarity (Singla et al., 2023).

To avoid such confusion, we will use the following definitions throughout this paper:

- 'Data-Similarity' is reserved for methods that build the feature purely based on the data and *never* use the LLM outputs. Specifically, we will use BM25. We do not consider embedding-based methods, which are expected to be better. These are not used in systems such as OLMo-Trace probably because they are computationally expensive compared to BM25.
- 'Data-Influence' is reserved for the gold-standard counterfactual shown in Eq. 1. Influence measures are referred to those that aim to estimate Eq. 1. Influence-Function (IF) refers to a specific influence-measures $\mathcal{I}_h$ that uses the Newton step as shown in Eq. 2.

As discussed earlier, little work has been done to precisely quantify the complementary strengths of the two approaches. Previous studies, such as those on fact tracing (Akyurek et al., 2022; Chang et al., 2025), do not provide a definite conclusion for output-tracing where the goal is to find training documents related to the outputs. We will precisely quantify the similarities and differences between the two approaches for tracing LLM outputs.

3 Quantifying Agreement Between Data-Similarity & Data-Influence

In this section, we describe our method to quantify commonalities and differences between data-similarity (Sim) and data-influence (Inf). For Sim, we use the Lucene (Kamphuis et al., 2020) version of BM25. For Inf, we primarily use the Hessian-based EK-FAC estimator. This influence estimator is quite expensive, which puts a limit on the number of documents and prompts we can use for influence computation. Using the whole training data is infeasible due to the matrix inverse. Therefore, we are forced to choose a subset of 100,000 documents and use only 100 prompts to estimate influence. Computing EK-FAC for one prompt on 100,000 training documents on OLMo2-1B required about one hour on 16 NVIDIA V100 GPUs. We therefore limited the number of prompts to 100, which is the highest our computational budget allowed (about 18 days for all LLMs); see Sec. 4 for further explanations and App. B for more details. We show comparisons on the following LLMs: *OLMo2-1B*, *Qwen3-1.7B*, *LlaMa3.2-1B*, *Gemma3-1B*, *GPT2-medium*, and *GPT2-small*.

We now give a step-by-step description of our workflow given in Fig. 3.

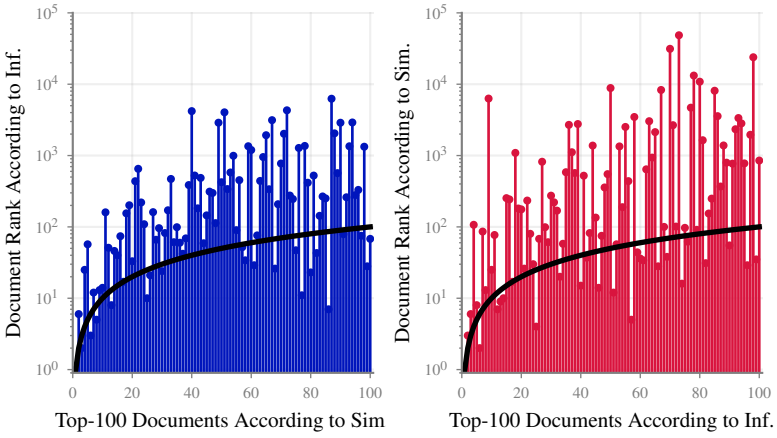
A. Create Queries: We generate 100 prompts of diverse topics and varying ambiguity using Gemini (Gemini Team, 2025) and manually verify them; see Fig. 2 for an example and Table 1 in the appendix for a full list of all queries. Each prompt is completed by each of the pre-trained LLMs. The LLM's response is concatenated to the prompt, forming a *query*. The queries represent trials in our empirical evaluations.

B. Rank Training Documents: We compute Sim scores between each training document and each query. Then, we compute absolute Inf scores for each training document to the LLM's autoregressive loss on each query. Finally, we sort the training documents according to the scores of each of the measures. For each query, this yields one ranking according to Sim, and one ranking according to Inf. See Fig. 4a for an example of the two rankings.

Top-20 of Sim	Rank assigned by Inf
1	1
2	6
3	2
4	25
5	57
6	3
7	12
8	5
9	13
10	14
11	160
12	51
13	8
14	46
15	40
16	74
17	17
18	155
19	201
20	33

Top-20 of Inf	Rank assigned by Sim
1	1
2	3
3	6
4	107
5	8
6	2
7	86
8	13
9	6291
10	25
11	77
12	7
13	9
14	10
15	253
16	243
17	17
18	1090
19	181
20	176

(a)



(b)

Figure 4: This figure shows both agreement between Inf and Sim and that Inf ranks the top documents according to Sim superior compared to how Inf ranks the top documents according to Inf. The two tables on the left (a) show the descrepency between the setups in the top-20, where ranks beyond the top-100 are highlighted in yellow. The two plots on the right (b) visualize the top-100 on the horizontal axis, and the ranks assigned by the other measure on the vertical axis in logarithmic scale. The solid black line indicates the best possible match between the measures. The results are shown for the query: *"A cat's favorite activity is usually..."* (see App. Table 1).

C. Visualize Top Documents: For the top-3 documents of each measure, we compute the per-token contribution to each score as described in Eqs. 5 and 6 in the appendix. We then show only sentences with an average per-token score of at least 10% of the highest per-token score in the document, and further highlight all tokens whose *individual* score is at least this value. See Fig. 2 for an example.

Results: In Fig. 2, we see some commonalities between the measures, where both measures mostly retrieve documents related to the query. But there are also differences. For instance, Sim retrieves an irrelevant document about SCRUM. For another query in the appendix (Figs. 20 and 21), we see that the top-3 match very closely, but for yet another query in the appendix (Figs. 22 and 23), Sim and Inf do not resemble each other at all. In the following, we will precisely quantify the agreement between the two measures through their rankings.

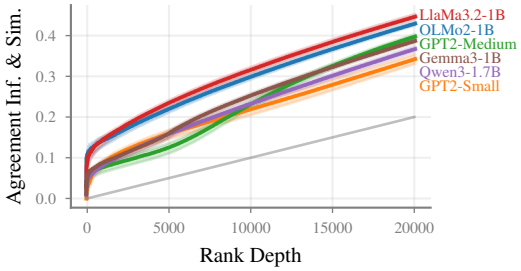
D. Compare Document Ranks: We define the top-100 as the relevant documents (this number can be chosen differently based on the setup and use-case). For each measure (data-similarity and data-influence), we (1) pick the relevant documents, (2) then sort them by their rank, and (3) finally annotate each document with its rank assigned by the other measure. We show the top-20 in two tables in Fig. 4a and highlight ranks assigned by the other measure that exceed 100. We then show the top-100 in two stem plots in Fig. 4b.

Results: In Fig. 4a, we see agreement between Sim and Inf: Firstly, Sim and Inf assign the same document to rank 1. Secondly, the majority of documents in the top-20 of Sim are also in the top-100 of Inf. This likewise holds the other way around. But we also observe asymmetry: Seven documents in the top-20 of Inf are outside the top-100 of Sim (highlighted), but only three documents in the top-20 of Sim are outside the top-100 of Inf. The stem plots in Fig. 4 show a similar trend: The stems grow with increasing ranks, indicating that the top-100 of both measures roughly agree compared to the rank assigned by the other measure. But for the top-100 according to Sim, the stems are overall lower than for Inf, revealing a potential asymmetry in their agreement.

E. Compute Agreements: We precisely quantify the agreement between the top-d according to each measures by counting the number of documents they have in common. Dividing this number by d yields a score between 0 and 1, appropriately known as the *agreement* at depth d (Webber et al., 2010); see Fig. 5a for an example and Eq. 8 in the appendix. By visualizing the agreement at all rank depths $d \in \{1, ..., 100000\}$, we can measure at which ranks they agree most. Here, high agreement in the top-ranks indicates that the measures retrieve similar examples. We show the mean

d	Ranking 1	Ranking 2	Agreement
1	(1)	(7)	0.000
2	(1 2)	(7 3)	0.000
3	(1 2 3)	(7 3 1)	0.667
4	(1 2 3 4)	(7 3 1 4)	0.500
5	(1 2 3 4 5)	(7 3 1 4 2)	0.800
6	(1 2 3 4 5 6)	(7 3 1 4 2 5)	0.833
7	(1 2 3 4 5 6 7)	(7 3 1 4 2 5 6)	1.000

(a)



(b)

Figure 5: (a) The *agreement* of two example rankings on the same set of size 7 with increasing depth d (see App. Eq. 8). Agreement does not increase monotonically with depth, but will always reach 1.0 when the rankings are based on the same set. (b) Agreement between Inf and Sim for six LLMs. The steep climb of the curves close to rank 1 indicates agreement in the top ranks. The variance is shown over 100 queries. The bottom gray line is the expected agreement for unrelated rankings. For all 100,000 ranks, see App. Fig. 8.

and variance of the agreement over all queries for each of the six LLMs in Fig. 5b. Further, we can represent high agreement in the top ranks with a single score using rank-biased overlap (Webber et al., 2010), which we use to compare the agreement between queries in Fig. 12b in the appendix.

Results: Fig. 5b shows the agreement up to rank 20,000. We observe a sharp increase in the top-tier documents for all LLMs, which indicates some agreement. Especially in the first 100 examples, all lines rise sharply. Nevertheless, there is also significant disagreement. The rank-biased overlap values shown in Fig. 12b in the appendix suggest that the agreement depends on the query: for some queries there is high, and almost complete agreement. For other queries, there is no agreement at all.

F. Quantify Asymmetry: To quantify the asymmetry between Sim and Inf, we measure how predictive their rankings are of each other. For this, take the top-100 according to each measure and compare how well the other measure ranks them. If the ranking assigned by the other measure significantly exceeds 100, then clearly it is not predictive of those ranks. If one measure is consistently doing better than the other in such predictions, then we can conclude that one is more informative of the other, indicating an asymmetric relationship. To this end, we do the following prediction tasks:

1. **Inf→Sim:** We predict the ranking of Sim using Inf, that is, we take the top-100 documents of Sim and obtain their respective rankings according to Inf
2. **Sim→Inf:** We do the opposite: take the top-100 documents of Inf and obtain their respective rankings according to Sim

As before, we show the mean and variance over our 100 queries for each LLM.

Results: In Fig. 6a, we see that across all rank depths d beyond the symmetric case $d = 100$, Inf→Sim (upper blue curve) appears consistently higher than Sim→Inf (lower red curve). For instance, for OLMo2-1B, about 80% of the top-100 examples according to Sim are included in the top-20000 examples according to Inf. But only about 65% of the top-100 according to Inf are included in the top-20000 according to Sim. Gemma3-1B is an exception above rank 5000, where Sim→Inf catches up, but the top-ranks are of primary interest. These results suggests Inf is more predictive of Sim than the other way around, demonstrating their asymmetric relationship.

G. Conduct Paired Tests: We test for statistically significance between the prediction tasks in step **F** above using the Wilcoxon signed rank test. Based on the measurements of each of the two tasks, Inf→Sim and Sim→Inf, we formulate the null- and alternative hypotheses as follows. Let D_d be the median difference between Inf→Sim and Sim→Inf at rank depth d.

$$H_0 : \text{The median difference } D_d \text{ is less or equal zero.} \quad (3)$$

$$H_a : \text{The median difference } D_d \text{ is larger zero.} \quad (4)$$

The null hypothesis assumes that Inf is less or equally predictive of Sim than the other way around.

Results: The p-values for the null hypothesis at each rank depth d are visualized in Fig. 12a in the appendix. We find that the p-values calculated under the assumption of the null hypothesis are

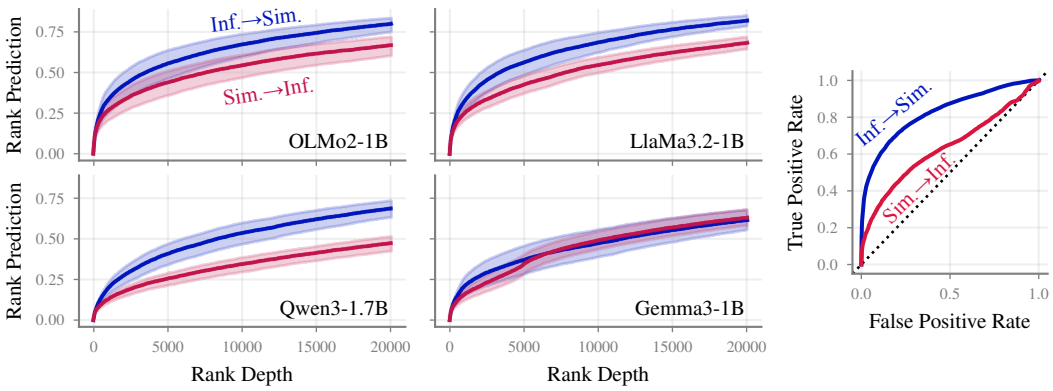


Figure 6: Results suggesting that Inf is better at predicting Sim than the other way around. (a) Rank prediction of the top-100 of Sim (BM25) using Inf (KFAC), and vice-versa for four LLMs. For all 100,000 ranks and six LLMs, see App. D. (b) Pre-filtering using Sim and refining using Inf and vice versa, evaluated using the receiver-operator characteristic on OLMo2-1B. See step **H** for details.

far below 0.05 for all rank depths except close to rank depths 100 and 100,000. The agreement between the two rankings at and around a depth of 100 is (close-to) symmetric, therefore one cannot be more predictive of the other. Similarly, at rank depths close to the maximum rank, almost all examples are included in both rankings, which explains the high agreement in any ranking method. In between, we can safely reject the null Hypotheses and therefore conclude convincingly that Inf is more predictive of Sim than the other way around.

H. Pre-Filter and Refine: Exploring all training examples with data-influence is prohibitive in practice. With pre-filtering, we can avoid computing influence on the whole dataset. However, there are no works that study the quality of such pre-filtering. In the following, we provide a proof-of-concept for pre-filtering with Sim and refining with Inf, and compare it to the (unrealistic) inverse setup. For each measure, we do the following: We retrieve the top-100 documents for each query, and label them *positive*. We randomly choose 100 documents from the training set for each query, and label them *negative*. Then, we use the respectively other measure to assign scores to the 100 retrieved and 100 random documents of each query. This leads to the same relationship as in step **F** above:

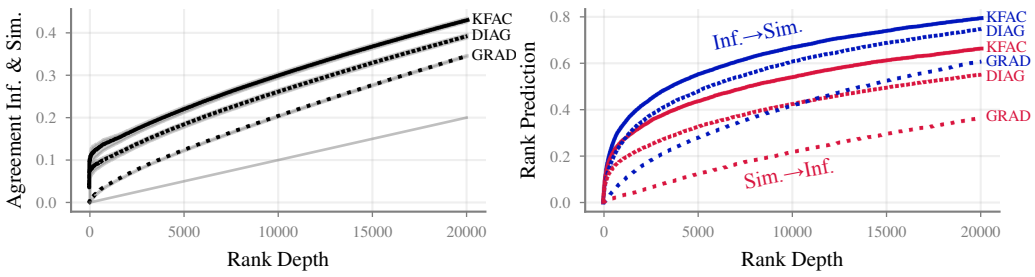
1. **Inf→Sim:** Retrieve with Sim and score with Inf
2. **Sim→Inf:** Retrieve with Inf and score with Sim

Using the scores and labels for the documents of all queries, we plot the receiver-operating characteristic (ROC) curve in Fig. 6b.

Results: For Sim→Inf, we observe an area under the ROC curve (AUC) of 0.63. This means that a considerable portion of the top-100 documents retrieved by Inf may be missed by Sim. The opposite direction Inf→Sim seems to be more viable, with an AUC of 0.83. The outcome is thus in agreement with findings in step **F** above.

I. Compare Influence Estimators: The use of cheaper influence estimators may impact our previous observations. Therefore, we investigate whether our results hold across different influence estimators. We repeat steps **E** and **F** with the following influence estimators: (1) Hessian-based EK-FAC (KFAC), (2) its diagonal approximation (DIAG) (George et al., 2018), and (3) the gradient-based estimator in Eq. 2 (GRAD). KFAC is a better but more expensive estimator than DIAG, which itself is better but more expensive than GRAD. We show results for OLMo-2-1B.

Results: In Fig. 7a we observe agreement across all estimators. However, agreement appears to consistently increase with better influence estimators, which is indicated by the steep increase in the top ranks. The same applies to the asymmetry between Sim and Inf: In Fig. 7b, Inf→Sim (upper red curves) appears consistently higher than Sim→Inf (lower blue curves) across all estimators. These results show that the observed (asymmetric) agreement is consistent across influence estimators and



(a) (b)

Figure 7: Results showing the consistency of our observations across various influence estimators on OLMo2-1B. (a) Agreement between Sim (BM25) and Inf (various). The variance is shown over 100 queries. The bottom gray line is the expected agreement for unrelated rankings. (b) Rank prediction of the top-100 of Sim (BM25) using Inf (various) and vice-versa. The variances are hidden to improve readability. For all 100,000 ranks and variances, see App. Figs. 10a and 10b.

scales with the estimation quality: the better the influence estimator, the more Sim and Inf agree, and the better Inf is at predicting Sim.

4 Conclusion

In this work, we show how to quantify the complementary strengths of data-influence and data-similarity for tracing LLM outputs. We achieve this by comparing the rankings assigned by the two measures to the training documents for a range of models including OLMo2-1B, Qwen3-1.7B, LlaMa3.2-1B, Gemma3-1B, and GPT2. Our main finding is that the two rankings agree significantly, but there is an *asymmetry* between them: The top documents of data-similarity are assigned more consistent ranks by data-influence than the other way around. These results are consistent across three common data-influence estimators, were we observe better estimators to yield higher agreement. This suggests that a good cost-accuracy trade-off for tracing the output of LLMs can be achieved by combining data-similarity and data-influence. In a proof-of-concept, we use data-similarity to find a candidate set of relevant documents, which we refine using data-influence. Our work opens up the path for hybrid versions of tools such as OLMo-Trace, where the combined strengths of data-similarity and data-influence can lead to better understanding of LLMs. In the future, we wish to evaluate how well such hybrid approaches work compared to existing ones in real-time applications.

Limitations and Future Work Our work has two primary limitations. First, due to computational constraints, we limited the number of queries and training documents. Each query requires an expensive influence estimation on *all* training documents, making it infeasible to consider more queries and training documents given our computational budget (16 NVIDIA V100 GPUs). Second, while our set of prompts cover a wide variety of topics, they were not explicitly balanced or categorized by subjects. As a consequence, it is unclear to what degree the observed agreement between data-similarity and data-influence depends on the specific prompt. These limitations present several promising directions for future work. First, an important next step is to scale our analysis to larger query and training document sets. For instance, assuming linear runtime scaling and an acceleration factor of $1 \times$ H200 $\approx 3.27 \times$ V100 (NVIDIA GPUs), an influence estimation for 1000 prompts on 1 billion training documents would require approximately 20 days on a cluster of 1024 NVIDIA H200 GPUs. Second, while our results across different model sizes up to 1.7 billion parameters suggest that data-similarity and data-influence (asymmetrically) agree irrespective of the number of parameters, verifying this trend on larger, production-grade models (e.g., 30 billion parameters or more) remains an important open question. Third, categorizing prompts by topic could provide a deeper understanding of when cheaper data-similarity is sufficient versus when more accurate data-influence is necessary. Such insights could be leveraged at test-time to dynamically predict whether a specific query requires the use of data-influence. Finally, it would be highly valuable to investigate the extent to which other data-similarity measures, such as InfiniGram or semantic embeddings, agree with data-influence.

AI USE STATEMENT

In this work, we used generative AI tools generate the 100 prompts based on which we generated outputs from the different analzed LLMs as described in Sec. 3 step A. These prompts are listed in Table 1. We have manually verified the prompts for use in our experiments. We have not used generative AI tools for any other parts of this work or for any other purposes.

REPRODUCIBILITY STATEMENT

This work made an effort to document all the different parameters used for experiments, including models, datasets, and prompts. The code which was used to generate the results presented in this work, including the exact library versions, is provided as supplementary material.

REFERENCES

Akyurek, E., Bolukbasi, T., Liu, F., Xiong, B., Tenney, I., Andreas, J., and Guu, K. Towards tracing knowledge in language models back to the training data. In *Findings of the Association for Computational Linguistics: EMNLP*. 2022.

Barberá, I. AI Privacy Risks & Mitigations—Large Language Models (LLMs), 2025. URL https://www.edpb.europa.eu/system/files/2025-04/ai-privacy-risks-and-mitigations-in-llms.pdf.

Barshan, E., Brunet, M.-E., and Dziugaite, G. K. RelatIF: Identifying Explanatory Training Samples via Relative Influence. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2020.

Bird, S. and Loper, E. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*. 2004.

Chang, K., Cramer, M., Soni, S., and Bamman, D. Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2023.

Chang, T. A., Rajagopal, D., Bolukbasi, T., Dixon, L., and Tenney, I. Scalable Influence and Fact Tracing for Large Language Model Pretraining. In *International Conference on Learning Representations (ICLR)*, 2025.

Choe, S. K., Ahn, H., Bae, J., Zhao, K., Kang, M., Chung, Y., Pratapa, A., Neiswanger, W., Strubell, E., Mitamura, T., Schneider, J., Hovy, E., Grosse, R., and Xing, E. What is Your Data Worth to GPT? LLM-Scale Data Valuation with Influence Functions, 2024. URL http://arxiv.org/abs/2405.13954.

Cook, R. D. Influential Observations in Linear Regression. *Journal of the American Statistical Association*, 74, 1979.

Gemini Team. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

Gemma Team. Gemma 3, 2025. URL https://goo.gle/Gemma3Report.

George, T., Laurent, C., Bouthillier, X., Ballas, N., and Vincent, P. Fast Approximate Natural Gradient Descent in a Kronecker Factored Eigenbasis. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2018.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A. et al. The Llama 3 Herd of Models, 2024. URL http://arxiv.org/abs/2407.21783.

Grosse, R., Bae, J., Anil, C., Elhage, N., Tamkin, A., Tajdini, A., Steiner, B., Li, D., Durmus, E., Perez, E., Hubinger, E., Lukošiūtė, K., Nguyen, K., Joseph, N., McCandlish, S., Kaplan, J., and Bowman, S. R. Studying Large Language Model Generalization with Influence Functions, 2023. URL http://arxiv.org/abs/2308.03296.

Guo, H., Rajani, N., Hase, P., Bansal, M., and Xiong, C. FastIF: Scalable influence functions for efficient model interpretation and debugging. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2021.

Hong, D. M., Mlodozeniec, B. K., Eschenhagen, R., and Turner, R. E. Better Hessians Matter: Studying the Impact of Curvature Approximations in Influence Functions. In *Mechanistic Interpretability Workshop at NeurIPS*, 2025.

Kamphuis, C., de Vries, A. P., Boytsov, L., and Lin, J. Which BM25 Do You Mean? A Large-Scale Reproducibility Study of Scoring Variants. In *Advances in Information Retrieval*. 2020.

Karamolegkou, A., Li, J., Zhou, L., and Søgaard, A. Copyright Violations and Large Language Models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2023.

Koh, P. W. and Liang, P. Understanding Black-box Predictions via Influence Functions. In *International Conference on Machine Learning (ICML)*. 2017.

Koh, P. W. W., Ang, K.-S., Teo, H., and Liang, P. S. On the Accuracy of Influence Functions for Measuring Group Effects. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

Li, J., Fang, A., Smyrnis, G., Ivgi, M., Jordan, M., Gadre, S., Bansal, H., Guha, E., Keh, S., Arora, K., Garg, S., Xin, R., Muennighoff, N., Heckel, R., Mercat, J., Chen, M., Gururangan, S., Wortsman, M., Albalak, A., Bitton, Y. et al. DataComp-LM: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 37, 2024.

Liu, J., Min, S., Zettlemoyer, L., Choi, Y., and Hajishirzi, H. Infini-gram: Scaling Unbounded n-gram Language Models to a Trillion Tokens. In *First Conference on Language Modeling*, 2024.

Liu, J., Blanton, T., Elazar, Y., Min, S., Chen, Y.-S., Chheda-Kothary, A., Tran, H., Bischoff, B., Marsh, E., Schmitz, M., Trier, C., Sarnat, A., James, J., Borchardt, J., Kuehl, B., Cheng, E. Y.-Y., Farley, K., Anderson, T., Albright, D., Schoenick, C. et al. OLMoTrace: Tracing Language Model Outputs Back to Trillions of Training Tokens. In *Annual Meeting of the Association for Computational Linguistics (ACL)*. 2025a.

Liu, S., Yao, Y., Jia, J., Casper, S., Baracaldo, N., Hase, P., Yao, Y., Liu, C. Y., Xu, X., Li, H., Varshney, K. R., Bansal, M., Koyejo, S., and Liu, Y. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 7, 2025b.

Nickl, P., Xu, L., Tailor, D., Möllenhoff, T., and Khan, M. E. The memory perturbation equation: Understanding model's sensitivity to data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Paul, M., Ganguli, S., and Dziugaite, G. K. Deep Learning on a Data Diet: Finding Important Examples Early in Training. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2021.

Pruthi, G., Liu, F., Kale, S., and Sundararajan, M. Estimating Training Data Influence by Tracing Gradient Descent. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

Robertson, S. and Zaragoza, H. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval*, 3, 2009.

San Joaquin, A., Wang, B., Liu, Z., Asher, N., Lim, B., Muller, P., and Chen, N. F. In2Core: Leveraging influence functions for coreset selection in instruction finetuning of large language models. In *Findings of the Association for Computational Linguistics: EMNLP*. 2024.

Shen, Y., Daheim, N., Cong, B., Nickl, P., Marconi, G. M., Raoul, B. C. E. M., Yokota, R., Gurevych, I., Cremers, D., Khan, M. E., and Möllenhoff, T. Variational Learning is Effective for Large Deep Networks. In *International Conference on Machine Learning (ICML)*. 2024.

Singla, V., Sandoval-Segura, P., Goldblum, M., Geiping, J., and Goldstein, T. A Simple and Efficient Baseline for Data Attribution on Images, 2023. URL http://arxiv.org/abs/2311.03386.

Sun, W., Liu, H., Kandpal, N., Raffel, C., and Yang, Y. Enhancing Training Data Attribution with Representational Optimization, 2025. URL http://arxiv.org/abs/2505.18513.

Thakkar, M., Bolukbasi, T., Ganapathy, S., Vashishth, S., Chandar, S., and Talukdar, P. Self-influence guided data reweighting for language model pre-training. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2023.

Walsh, E. P., Soldaini, L., Groeneveld, D., Lo, K., Arora, S., Bhagia, A., Gu, Y., Huang, S., Jordan, M., Lambert, N., Schwenk, D., Tafjord, O., Anderson, T., Atkinson, D., Brahman, F., Clark, C., Dasigi, P., Dziri, N., Ettinger, A., Guerquin, M. et al. 2 OLMo 2 Furious. In *Conference on Language Modeling*, 2025.

Webber, W., Moffat, A., and Zobel, J. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28, 2010.

Wu, M. and Aji, A. F. Style Over Substance: Evaluation Biases for Large Language Models. In *International Conference on Computational Linguistics (ICCL)*. 2025.

Wu, W., Xu, J., Li, H., and Oyama, S. Learning a Robust Relevance Model for Search Using Kernel Methods. *Journal of Machine Learning Research*, 12, 2011.

Xia, M., Malladi, S., Gururangan, S., Arora, S., and Chen, D. LESS: Selecting Influential Data for Targeted Instruction Tuning. In *International Conference on Machine Learning (ICML)*. 2024.

Xiao, Y. and Wang, W. Y. On Hallucination and Predictive Uncertainty in Conditional Language Generation. In *Conference of the European Chapter of the Association for Computational Linguistics*. 2021.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J. et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

A Additional Background

A.1 Influence on LLMs

The influence $\mathcal{I}(i,j)$ and its estimators $\mathcal{I}_g(i,j)$ and $\mathcal{I}_h(i,j)$ correspond to the single token predictions i under the removal of a single token loss j in LLMs. However, in our experiments, we are interested in the change in the generated document $\mathcal{A}$ under the removal of both individual tokens j, and full training documents $\mathcal{D}$. For this setting, the influence estimators simply become the sum over the gradients of the tokens contained in the documents $\mathcal{A}$ and $\mathcal{D}$ (Koh et al., 2019; Grosse et al., 2023), with

$$\mathcal{I}_g(\mathcal{A},\mathcal{D}) = \sum_{i\in\mathcal{A}} \mathbf{g}_i^\top \sum_{j\in\mathcal{D}} \mathbf{g}_j \qquad \mathcal{I}_h(\mathcal{A},\mathcal{D}) = \sum_{i\in\mathcal{A}} \mathbf{g}_i^\top \mathbf{H}^{-1} \sum_{j\in\mathcal{D}} \mathbf{g}_j. \tag{5}$$

where we can reduce the training document $\mathcal{D}$ to a single token when we are interested in the token-wise influence scores, such as shown in Fig. 2

A.2 Token-wise and Document-wise BM25

The "Lucene" (Kamphuis et al., 2020) version of BM25 (Robertson & Zaragoza, 2009), can be defined as an asymmetric similarity (Wu et al., 2011) with

$$\begin{aligned}
\text{BM25}(\mathbf{q},\mathbf{d}) &= \phi_Q(\mathbf{q})^\top \phi_D(\mathbf{d}) \\
\phi_Q(\mathbf{q}) &= \{|\mathbf{q}|_{w_i}\}_{i=1}^M \\
\phi_D(\mathbf{d}) &= \left\{ \frac{\text{IDF}_i \cdot |\mathbf{d}|_{w_i}}{k_1(1-b+b\frac{|\mathbf{d}|}{\overline{D}}) + |\mathbf{d}|_{w_i}} \right\}_{i=1}^M \\
\text{IDF}_i &= \log\left(1 + \frac{N - |w_i| + 0.5}{|w_i| + 0.5}\right),
\end{aligned} \tag{6}$$

given some vocabulary $\mathbf{w} = (w_1, w_2, \ldots, w_M)$ with vocabulary size M. Here, $k_1 \geq 0$ is a hyper-parameter, $|\mathbf{d}|_{w_i}$ is the number of times token w_i appears in $\mathbf{d}$, $|\mathbf{d}|$ is the document length of document $\mathbf{d}$, $\overline{D}$ is the average document length, and $|w_i|$ is the number of documents in which word w_i appears at least once. We obtain the scores for the qualitative visualization in Fig. 2 and others through the individual components of $\phi_D(\mathbf{d})$ for each term in the document.

A.3 Rank Correlation

In this work, we compare rankings of training documents $\mathbf{X} : \mathbb{R}^{N\times d}$ obtained by measures of similarity $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ to some query $\mathbf{q}$. Formally, we define the rank of training document $\mathbf{x}_i$

$$r_{j,i}(k) = 1 + \sum_{\substack{m=1 \\ m\neq i}}^{N} \mathbb{I}\left(k(\mathbf{q},\mathbf{x}_i) > k(\mathbf{q},\mathbf{x}_m)\right), \tag{7}$$

for instance, the document $\mathbf{x}_i$ with the highest similarity $k(\mathbf{q},\mathbf{x}_i)$ to query $\mathbf{q}$ will be assigned rank $r_{j,i}(k) = 1$.

A classical approach to compare these is the Spearman rank correlation, which is defined as the Pearson correlation of rankings. However, it is not suitable for our analysis, as (1) it does not allow comparing sets that contain different documents, which is required in our rank prediction experiment, (2) it does provide a way to focus only on the top-ranks which we are primarily interested in, and (3) it is sensitive to the magnitude of the difference between the ranks, causing the bottom ranks to dominate. For this reason, we follow Webber et al. (2010) and use *agreement* in our evaluation of rankings, which solves these issues.

The *agreement* at depth n quantifies how many documents appear in the top-n of both measures when ranking the documents $\mathbf{X}$. Given two measures of similarity k_a and k_b, the symmetric agreement $A_n(k_a,k_b)$ is defined as

$$A_{j,n}(k_a, k_b) = \frac{|\{i : r_{j,i}(k_a) \leq n\} \cap \{i : r_{j,i}(k_b) \leq n\}|}{n}. \tag{8}$$

The agreement may also be computed at *asymmetric* depths, i.e.,

$$A_{j,(m,M)}(k_a, k_b) = \frac{|\{i : r_{j,i}(k_a) \leq m\} \cap \{i : r_{j,i}(k_b) \leq M\}|}{m}, \tag{9}$$

where $m < M$, counts whether the top-m according to k_a are contained in the top-M according to k_b. This can be used to identify cases where, for example, the top-m according to one method may be contained in the top-M of the other, but not vice-versa.

Rank-biased overlap (Webber et al., 2010) can be used to obtain a more refined similarity between rankings by weighting the agreement values up to some depth d, with

$$\text{RBO}_j(k_a, k_b, p) = \frac{1-p}{p} \left(\sum_{m=1}^{N} A_{j,m}(k_a, k_b) \cdot p^d + \frac{A_{j,N}(k_a, k_b) \cdot p^N}{1-p} \right), \tag{10}$$

where $p \in (0, 1]$ is the *persistence*, a hyper-parameter to indicate the top-ranks that should be considered. The persistence $p = 1 - \frac{1}{d}$ indicates up to which depth d we care about the ranking, for example, looking at the top $d = 100$ results in a persistence of 0.99.

B Experimental Details

We used the following models with checkpoints as provided by their respective maintainers on Huggingface:

- `allenai/OLMo-2-0425-1B-Instruct` (Walsh et al., 2025)
- `google/gemma-3-1b-it` (Gemma Team, 2025)
- `meta-llama/Llama-3.2-1B` (Grattafiori et al., 2024)
- `Qwen/Qwen3-1.7B` (Yang et al., 2025)
- `team-approx-bayes/gpt2-medium` (Shen et al., 2024)
- `team-approx-bayes/gpt2-small` (Shen et al., 2024)

We use the Kronfluence Python library (Grosse et al., 2023) to compute EK-FAC, Diagonal (EK-FAC), and Identity of the complete query (prompt and response).

We used a random subset 100,000 examples of DCLM (Li et al., 2024), where we truncate each document to 300 tokens. The resulting actual text length depends on the tokenizer of the respective models. The responses of each LLM were limited to generate a maximum of 300 tokens. We completed the responses without using a chat template.

We used the NLTK (Bird & Loper, 2004) `word_tokenize` function to tokenize all queries and training documents for BM25. We tried different tokenizers using normalization and word-boundary splitting, as well as using respective LLM tokenizers, for which we obtained similar results.

We used NVIDIA V100 GPUs with 32GB memory to compute the influence, which required about 16 hours to compute the token-wise influence scores on all 100,000 training documents on OLMo-2-1B per query running on a single GPU.

C List of Used Prompts

For all experiments, we generated 100 prompts to generate responses of the LLMs. The exact prompts are shown in Table 1 along with their indices. These prompts were generated using Gemini. We manually verified their diversity of topics and ambiguity. For instance, the prompt shown in Fig. 2 has index 14.

Table 1: List of the prompts completed by all LLMs with their respective indices.

Index	Prompt	Index	Prompt
0	The color of the sky today is	50	The ancient castle stood tall, overlooking
1	A cat's favorite activity is usually	51	She felt a surge of excitement when
2	Water boils at a temperature of	52	The forest was dark and silent, except for
3	The capital city of France is	53	He knew, deep down, that the secret was
4	In the morning, I like to drink	54	It was a cold, crisp morning as the sun
5	Seven plus five equals	55	Laughter is often considered the best
6	An apple is a type of	56	The smell of freshly baked bread reminded him of
7	The largest ocean on Earth is	57	Regret washed over her when she realized
8	My favorite kind of music is	58	Hope flickered in the darkness like
9	A bicycle has two	59	The antagonist's motivation was rooted in
10	Before going to bed, one should always	60	Furthermore, it must be acknowledged that
11	When the phone rang, she quickly	61	Consequently, the team decided to
12	He decided to walk to the store because	62	Despite the inherent risks, the explorer
13	If you mix blue and yellow, you get	63	Subsequently, the findings were published in
14	To open the locked door, you'll need	64	Conversely, the smaller sample size indicated
15	The train was delayed due to	65	In essence, the philosophical debate revolves around
16	We celebrated the victory by	66	To illustrate this phenomenon, consider the case of
17	Suddenly, the lights went out, causing	67	Historically, the region was known for its
18	She learned to play the guitar so she could	68	On the one hand, the investment seems sound; on the other
19	After the long hike, they felt	69	Ultimately, the decision will depend on
20	If I had a superpower, it would be	70	The key to successful gardening is
21	Even though it was raining, we still	71	A famous quote by Shakespeare is
22	Whenever he sees a dog, he always	72	My greatest fear has always been
23	Unless you bring a jacket, you might	73	The ingredients for making a pizza are
24	Since the cookies were burned, we had to	74	Learning a new language can be difficult but
25	Provided the weather is nice, the event will	75	The invention of the printing press revolutionized
26	No matter how hard she tried, she couldn't	76	You should never underestimate the power of
27	Only if you finish your homework can you	77	The library smelled like old paper and
28	As soon as the bell rings, the students	78	An important rule in programming is to
29	Although the book was long, it was still	79	If you look closely, you will notice
30	Could you please hand me	80	The mysterious object floated silently in
31	I wonder why the bird is	81	Without warning, a portal opened, revealing
32	Tell me the story of	82	The hero's journey began with a call to
33	What is the best way to	83	Her eyes were the color of
34	How much time will it take to	84	The hidden message was encrypted using
35	Explain the concept of	85	A talking parrot sat on the shoulder of
36	Who was the first person to	86	Legend has it that the lost city of
37	Do you think it's possible to	87	The wizard carefully mixed the potions, hoping to
38	Remember to bring along	88	The futuristic city was powered entirely by
39	Please describe the taste of	89	He carried a leather-bound journal filled with
40	An algorithm is essentially a set of	90	From a purely economic standpoint, the project is
41	Data structures are crucial for	91	The challenge lies in accurately measuring
42	The process of photosynthesis involves	92	A sudden flash of inspiration led to
43	Artificial intelligence aims to mimic	93	When a person is nervous, they might
44	In geometry, a circle is defined as	94	It is generally agreed that exercise is beneficial for
45	Gravity is the force that	95	The most crucial part of the machine is the
46	The smallest unit of matter is the	96	This phenomenon occurs chiefly when
47	Renewable energy sources include	97	The little engine that could famously said
48	Quantum mechanics deals with	98	Every journey begins with a single
49	The purpose of a compiler is to	99	Let's start the meeting by discussing

D Complete Results

We show the complete results for the experiments conducted in the main manuscript. Specifically, the full versions of Figs. 5b and 6a are Fig. 8 and App. D.

The p-values for the statistical significance of Inf→Sim predicting more top-ranks than Sim→Inf for all rank depths are shown in Fig. 12a for OLMo2-1B. Additionally, the rank-biased overlap for all

100 queries on OLMO-2-1B with a persistence of 0.99 (in other words, for the top-100) are shown in Fig. 12b. The same figures for all the other models are found in the following figures:

- *GPT2-Small* in Fig. 13
- *GPT2-Medium* in Fig. 14
- *LlaMa3.2-1B* in Fig. 15
- *Qwen3-1.7B* in Fig. 16
- *Gemma3-1B* in Fig. 17

The gap between Inf→Sim and Sim→Inf varies between models, where it appears the widest overall for *Qwen3-1.7B* in Fig. 16, and the narrowest for *Gemma3-1B* in Fig. 17. Here, Gemma3-1B shows a bit of a different behavior, where the predictiveness of Sim→Inf catches up to Inf→Sim in intermediate ranks between 5000 and 6000. However, it still holds in the top-ranks which are of primary interest.

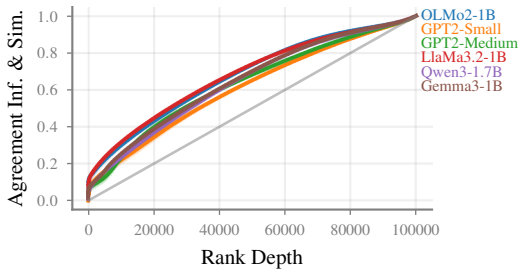


Figure 8: Agreement between BM25 and KFAC for various LLMs.The steep ascend of the curves close to rank 1 indicates some agreement in the top ranks. The variance is shown over 100 queries. The bottom gray line is the expected agreement for unrelated rankings. The plot in the main manuscript including only 20,000 ranks can be found in Fig. 5b.

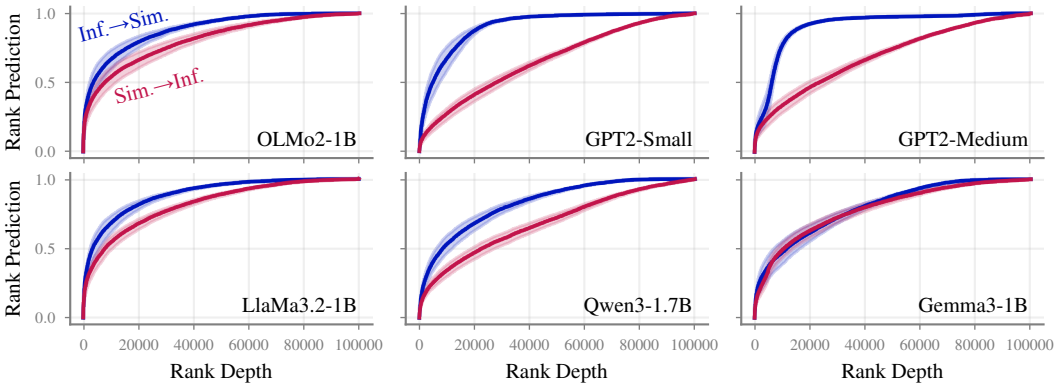


Figure 9: Results suggesting that Inf is better at predicting Sim than vice versa. Rank prediction of the top-100 of Sim (BM25) using Inf (KFAC), and vice-versa for six LLMs. For the limited 20,000 ranks on only four LLMs, see Fig. 6a.

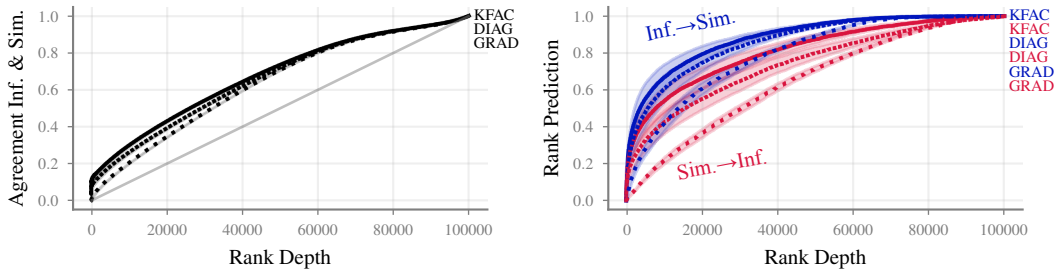


Figure 10: Results showing the consistency of our observations across various influence estimators on OLMo2-1B. (a) Agreement between Sim (BM25) and Inf (various). The variance is shown over 100 queries. The bottom gray line is the expected agreement for unrelated rankings. (b) Rank prediction of the top-100 of Sim (BM25) using Inf (various) and vice-versa. The variances are hidden to improve readability. For only 20,000 ranks and no variances, see Figs. 7a and 7b.

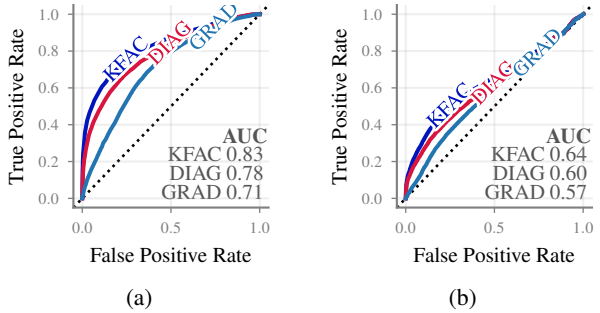
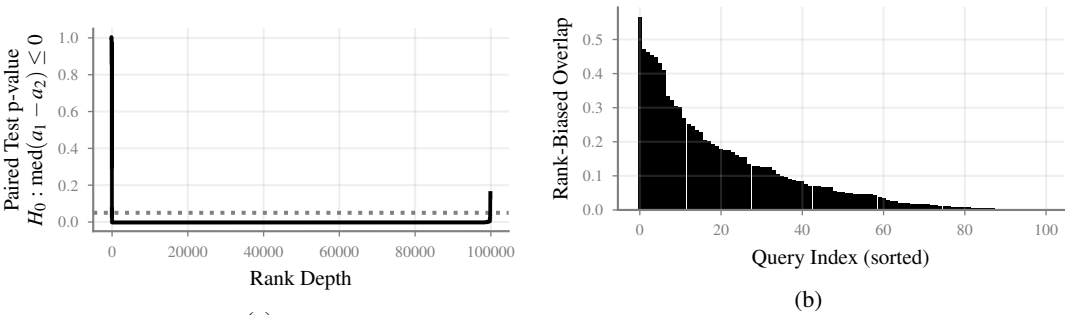
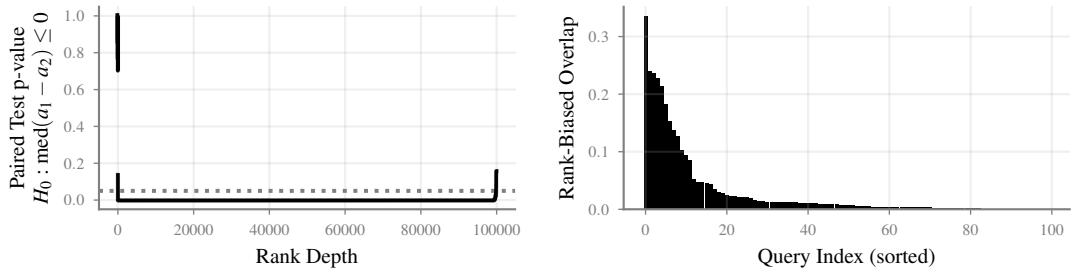


Figure 11: Pre-filtering using Sim and refining using Inf (a) and vice versa (b), evaluated using the receiver-operator characteristic on OLMo2-1B for various influence estimators. See experiment step H for details.



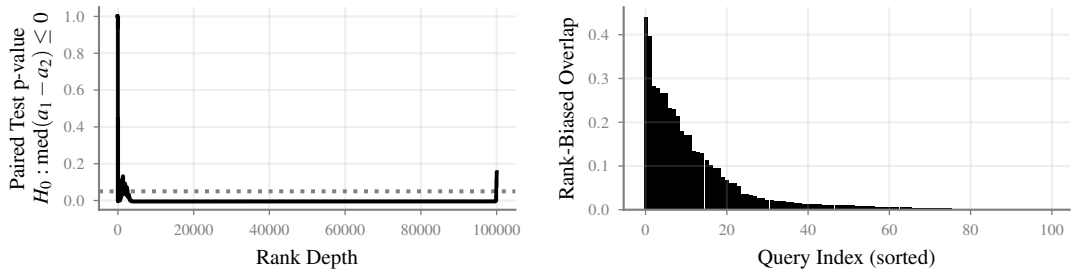
(a) (b)

Figure 12: Paired test (a) and per-query rank-biased overlap for OLMo2-1B. (a) The p-value at rank depth d for the null hypothesis that Inf (KFAC) at depth d ranks the top 100 according to Sim (BM25) inferior or equal compared to how BM25 at depth d ranks the top 100 according to Inf. (b) Rank-biased overlap with persistence 0.99 (i.e., top-100) for each of 100 queries between Sim (BM25) and Inf (K-FAC) sorted descending by the overlap. A coefficient of 1.0 indicates a perfect match between the rankings, and a coefficient of 0.0 indicates no match.



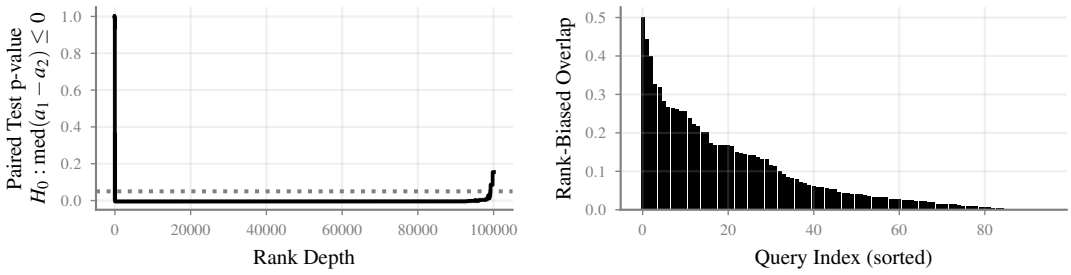
(a) (b)

Figure 13: Paired test (a) and per-query rank-biased overlap (b) between Sim (BM25) and Inf (KFAC) on *GPT2-Small*.



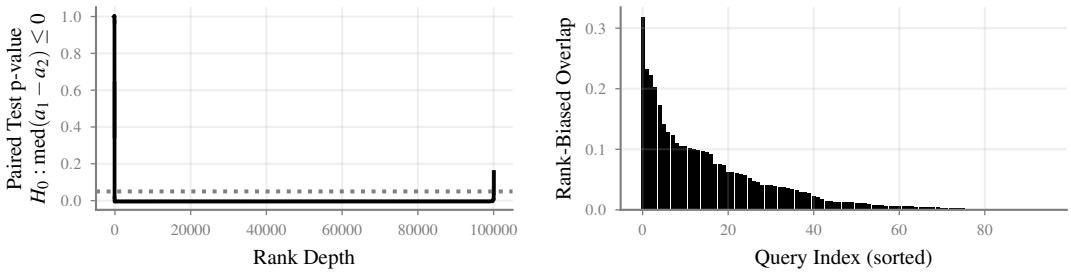
(a) (b)

Figure 14: Paired test (a) and per-query rank-biased overlap (b) between Sim (BM25) and Inf (KFAC) on *GPT2-Medium*.



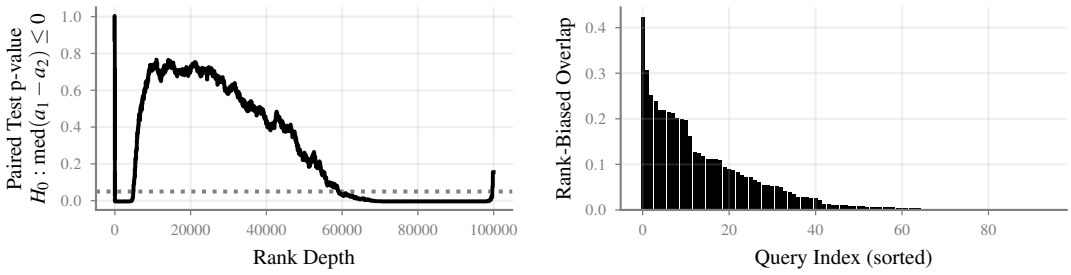
(a) (b)

Figure 15: Paired test (a) and per-query rank-biased overlap (b) between Sim (BM25) and Inf (KFAC) on *LlaMa3.2-1B*.



(a) (b)

Figure 16: Paired test (a) and per-query rank-biased overlap (b) between Sim (BM25) and Inf (KFAC) on *Qwen3-1.7B*.



(a) (b)

Figure 17: Paired test (a) and per-query rank-biased overlap (b) between Sim (BM25) and Inf (KFAC) on *Gemma3-1B*.

E ADDITIONAL TOKEN-WISE VISUALIZATIONS

In Figs. 18 to 35 we list the full documents for which we conducted the qualitative analysis in Sec. 3.



Figure 18: Top-3 training documents according to data-similarity (BM25) on OLMo2-1B for prompt: '*To open the locked door, you'll need...*'.

Prompt: To open the locked door, you'll need …

Response: to use both logic and intuition. First, let's assume you've determined that the door is either A) locked from the outside or B) locked from the inside. To solve the puzzle, consider these factors: 1. Evaluate possible locks: Since the door is likely locked, we'll check typical locks such as keychains, security slots, or even those found in antique doors. 2. Observe the surroundings: The environment around the door can provide hints as to what might be needed to unlock it. If other items are unlocked, this could lead to further clues. 3. Test assumptions: Sometimes, assumptions can be incorrect. If you realize that your current assumption is too simplistic, reassess your approach. 4. Observe interactions: Sometimes doors open with unique methods, requiring more unconventional thinking. For the sake of this puzzle, assume the door is a typical wooden door that might require a lock to be manually operated (such as a doorknob lock). Here's a logical sequence of steps to open the locked door: 1. **Assess the Lock:** Identify the type of lock and its mechanisms (e.g., manual pull, strike mechanism). 2. **Test the Mechanism:** Determine if the lock is responsive (e.g., manually pulling the doorknob should make the door swing open). 3. **Inspect the Surroundings:** Look for any obvious features like keys, combinations, or signs indicating it should be opened by a specific method. 4. **Apply Reasoning

Data-Sim. Rank	Top-3 Documents for Data-Influence (relevant sentences only)	Data-Inf. Rank
91	Why Choose Keyless Locks for Your Home Thanks to modern technology, there is no need to worry about losing or misplacing your keys or fear that burglars or trespassers have acquired a duplicate to your key. **You don't have to worry that your locks will be tampered or picked by thieves. Because of an innovation called keyless locks,** you won't have to worry about the safety of your property and homes. The Advantages of Using Keyless Locks Using these modern locks provides homeowners with a lot of benefits and advantages. Firstly, there are different types of keyless locks to choose from. You choose which type suits your taste and your home. Some of these types are: Also, **these locks come in affordable prices.** Most of them have a lifetime warranty and service in case the locks have been tampered or compromised. It is very easy to use and install. Even person with disabilities can use them with convenience. Higher Security for Houses and Properties It prevents you from losing your keys, having duplicates or stolen. **It also features a higher grade of security since it prevents lock picking or bumping.** Thieves and trespassers cannot manipulate the locks with improvised tools and trying to tamper with it will heighten the risk of discovering the identity of the trespassers. Most of these locks when installed have alarm systems to provide top quality service and ensure home safety. **Failure to enter the correct code or forcing the lock to open will trigger the alarm and notify the owners that their homes are being broken in.** Easy Installation Process These locks are easy to install. It does not require special wiring or installing additional door or window frames to install the locks. **You can opt to use a mechanical or electronic lock to install in your house.** These locks can also work using standard batteries and does not require modification of doors and windows. Easy Process of Supplying or Changing Security Codes Supplying the security code of your locks is easy. You can use different number combinations, whoever you think is most effective and secure. If in any case you think that your current code has been compromised or someone else knows about it, you can easily reconfigure and change the codes to ensure security. As you can see, **keyless entry door locks provide more benefit and advantage than traditional locks.** Why not start using it now? Skimming Your Credit Cards The current society relies heavily on credit cards for purchases of goods, products, and even services. Carrying cash and using them to purchase goods are	1
3569	**Cylindrical And Mortise Lock Functions A lock is designated as a part number,** name and an ANSI "F" (function) standards reference number where applicable. The manufacturer determines the lock part number and name. **Most lock names describe their operation.** For example, **a Sargent Lock 11G30, Communicating Lock has the ANSI function number F80. Not every manufacturer will make all of the ANSI function number locks. Every ANSI function number has a description,** much like the operation. The function numbering began with mortise locks. The ANSI F01 Passage Mortise Lock description for Marks USA is "Latch by knob/lever either side." **The Sargent Lock description for the ANSI F01 Passage or Closet Mortise Lock function description is "Trim from either side retracts latch bolt at all times." Both say the F01 function mortise locks do not have keyed cylinders or a deadbolt,** and the inside or outside knobs or levers will retract the latch at all times. Because of the number of mortise lock functions, most lock manufacturers list their mortise locks by first dividing them by keyed and non-keyed. **The keyed mortise locks are further divided by single and double cylinder and deadbolt or non-deadbolt.** ANSI function numbers were applied starting with F01. **The Passage Function is for mortise locks. A passage function mortise lock does not have either a keyed cylinder or a deadbolt.** Different lock types use different ANSI function numbers for the same function. The passage function cylindrical (bored) lock uses ANSI F75. Both bored locks and mortise locks share some of the same functionality, however they are not identical. For this reason, each lock type, mortise, cylindrical and deadbolts have their own unique numbers for each lock function. A short list of manufacturer's lock function, the ANSI function numbers and the manufacturer's descriptions for cylindrical locks: When ANSI functions are displayed in a catalog, there is normally a drawing and written description of the lock's operation. The cylindrical lock drawings indicate if there is one or two knobs/levers or a key or push button extending from each knob/lever. The mortise lock drawings indicate if there is one or two knobs/levers or a keyed mortise cylinder. **coin turn or thumb turn on the outside or inside.** The length of the description depends upon the complexity of the locks' operation. Having the ANSI Function numbers and the description along with the manufacturers' information can help find	2
104	From Wikipedia, the free encyclopedia (Redirected from Tagout) Jump to: navigation, search Folding lockout scissor clamp, allowing six contractor padlocks to lock out one device. Group lockout[edit] When two or more subcontractors are working on different parts of a larger overall system, **the locked-out device is first secured with a folding scissors clamp that has many padlock holes capable of holding it closed.** Each subcontractor applies their own padlock to the clamp. The locked-out device cannot be activated until all workers have signed off on their portion of the project and removed their padlock from the clamp. In the United States a lock selected by color, shape or size (e.g. red padlock) is used to designate a standard safety device. **locking and securing hazardous energy. No two keys or locks should ever be the same.** A person's lock and tag must not be removed by anyone other than the individual who installed the lock and tag unless removal is accomplished under the direction of the employer. Employer procedures and training for such removal must have been developed, documented and incorporated into the employer energy control program. [1] Lockout-tagout in the US, has 5 required components to be fully compliant with OSHA law. The 5 components are: 1. Lockout-Tagout Procedures 2. Lockout-Tagout Training (for authorized employees and affected employees) 3. Lockout-Tagout Policy (often referred to as a program) 4. **Lockout-Tagout Devices and Locks 5.** Lockout-Tagout Auditing - Every 12 months, every procedure must be reviewed as well as a review of authorized employees [2] Hazardous energy and its isolation[edit] Modern machinery can contain many hazards to workers, from things like electrical, mechanical, pneumatic or hydraulic sources. For example a typical industrial machine may contain things like hot fluids, moving presses, blades, propellers, electrical heaters, conveyor belts with pinch points, moving chains, ultraviolet light, etc. 1. Identify the energy source(s) 2. Isolate the energy source(s) 3. Lock and Tag the energy source(s) 4. Prove that the equipment isolation is effective In industrial processes it can be difficult to establish where the appropriate danger sources might be. For example, a food processing plant may have input and output tanks and high temperature cleaning systems	3

Figure 19: Top-3 training documents according to data-influence (EK-FAC) on OLMo2-1B for prompt: '*To open the locked door, you'll need...*'.

Prompt: Renewable energy sources include …

Response: solar panels and wind turbines. Solar Panels: Solar panels convert sunlight into electricity. Solar panels come in different types like monocrystalline, polycrystalline, and thin-film. Solar panels are an excellent renewable energy source as they are available in various forms, each with its own benefits. Besides, solar panels are relatively cheap in terms of installation and are becoming increasingly efficient in converting sunlight into electricity. Wind Turbines: Wind turbines convert kinetic energy from the wind into electricity. Wind turbines are another renewable energy source with a growing presence worldwide. As wind speed varies significantly depending on location and environmental conditions, wind turbines can provide consistent power. In addition, wind energy is a clean and renewable resource that doesn't contribute to greenhouse gas emissions. Tidal Energy: Tidal energy harnesses the energy of tides and uses it to generate electricity. Tidal energy is relatively new compared to solar and wind energy. It involves constructing large dams or turbines along coastlines to harness the energy of the high and low tides. Bioenergy: Bioenergy is derived from organic materials such as organic waste, forestry residues, and agricultural residues. It includes power generation from combustible biomass and district heating. Bioenergy is a direct renewable energy source as it directly comes from organic waste and organic materials that are already decaying or are harvested annually. Hydro Power: Hydro power harnesses the energy of flowing water to generate electricity. It involves the construction of dams or hydropower plants

Top-3 Documents for Data-Similarity (BM25, relevant sentences only)

1 (Data-Sim. Rank) — (Data-Inf. Rank) **18**

"Get information and do business" . **Follow us Storehouses for Solar Energy Can Step In When the Sun Goes Down Published at:** Jan 2, 2012 source: **New York Times published by If solar energy is eventually going to matter— that is,** generate a significant portion of the nation's electricity — the industry must overcome a major stumbling block, experts say: finding a way to store it for use when the sun isn't shining. That challenge seems to be creating an opening for a different form of power, solar thermal, which makes electricity by using the sun's heat to boil water. The water can be used to heat salt that stores the energy until later, when the sun dips and households power up their appliances and air-conditioning at peak demand hours in the summer. Two California companies are planning to deploy the storage technology: SolarReserve, which is building a plant in the Nevada desert scheduled to start up next year, and BrightSource, which plans three plants in California that would begin operating in 2016 and 2017. Together, the four projects will be capable of powering tens of thousand of households throughout a summer evening. Whether the technology will be widely adopted remains to be seen, but companies like Google, Chevron and Good Energies are investing in it, and the utilities NV Energy and Southern California Edison have signed long-term contracts to buy power from these radically different new power plants. **One crucial role of the plants will be complementing solar panels,** which produce electricity directly from sunlight. When the panels ramp down at dusk or on cloudy days, the plants will crank up, **drawing on the stored thermal energy.** That job will become more important if photovoltaic panels, which have plunged in price lately, become even cheaper and sprout on millions of rooftops. **As the grid starts depending more heavily on solar panels or wind turbines,** it will need other energy sources that can step in quickly to balance the system — preferably ones classified as renewable. Most utilities are trying to generate as many kilowatt-hours of renewable energy as they can to meet stiffer state requirements on incorporating more alternative energy, said Kevin B. Smith, the chief executive of SolarReserve. "As we move forward, we'll get more and more traction with the fact we can provide more capacity," Mr. Smith said, referring to his company's storage technology. **The Energy Department seems to agree:** in September it gave SolarReserve a $737 million loan guarantee for its project in Nevada. The plant will generate 110 megawatts at peak and store enough heat to run for eight

2 **1**

Order hosting Renewable Energy - Green Energy Pros and Cons Renewable energy sources Renewable Energy's Bright Future: Green Energy's Pros and Cons Recent international focus on safe energy underscores the need for reevaluating all energy alternatives, particularly those that are clean and renewable. Because although the global economy is coming out of a recent recession, oil prices are climbing and the demand for alternative renewable sources is consistently growing. Indeed, the alternative energy market is one of the few markets that has seen substantial growth during the recent recession of the last two years. One thing is clear, **the need for more efficient use of electricity with the integration of renewable energy sources is present. Smart Grid and the Integration of Renewable Energy Sources The Smart Grid is a Department of Energy (**DOE) supported project that involves an integrated system of mixed distributed resources to increase the penetration of renewable energy – such as hydro, wind, solar, **wind and geothermal,** while delivering improved efficiency and reliability. These and other distributed resources will be fully integrated into the new smart, electrical grid. **Today's electric grid vs Smart grid What is Renewable Energy? Renewable energy sources are sources of energy that are constantly replenished. These include energy from water,** wind, the sun, geothermal sources, **and biomass sources such as energy crops.** In comparison, fuels such as coal, oil, **and natural gas are non-renewable.** Once a deposit of these fuels is depleted it cannot be replenished – a replacement source must be found instead. In the United States, **both renewable and non-renewable energy sources are used to generate electricity,** power vehicles, and provide heating, cooling, and light. **While renewable energy is generally more expensive than conventionally produced supplies,** alternative power helps to reduce pollution and to conserve fossil fuels. "People sometimes get caught up in cost-effectiveness," said Paul Torcellini, **a senior engineer at the DOE's National Renewable Energy Laboratory (**NREL) in Golden, Colorado. "But it can be a question of values and what we spend our money on." For comparison purposes, **we will explore a few of the different sources of renewable energy:** Hydropower refers to the use of water to generate electricity. **Water is the most common renewable source of energy in the United States today.** Hydroelectric power doesn't necessarily require a large dam – some hydroelectric power plants just use a small canal to channel the river water through a turbine. Other uses of hydropower include water-cooled chiller and power systems. **Hydropower is a renewable and non-polluting energy source without any greenhouse gas discharge and no**

3 **6**

Home Knowledge Renewables Major parameters that influence Wind turbines power output Major parameters that influence Wind turbines power output Wind power is coming of age in Africa and is fast becoming competitive with solar energy and other traditional energy sources. The African Wind Energy Community has always aimed for the same goal: making wind power a better choice for power generation, increasing power output and driving down the cost of electricity through wind energy. Like any other piece of equipment in the power industry, **wind turbines represent a large investment for the companies that rely on them to generate both electricity and revenue,** they present some unique parameters that must be observed so as to increase their efficiency and ensure maximum power generation output. **A wind turbine obtains its power input by converting the force of the wind into a torque (**turning force) acting on the rotor blades. The amount of energy which the wind transfers to the rotor depends on the density of the air, the rotor area, **and the wind speed and height of the tower. Wind speed and consistency Higher wind speeds generate more power because they allow the blades to rotate faster.** This rotation translates to more mechanical power and more electrical power from the generator. Mr. Leon Gouws the sales manager at Eveready one of the leading wind turbine manufacturers in Africa, notes that wind speed largely determines the amount of electricity generated by a turbine. "Location of the installation is very important and should have average wind speeds, wind maps are used to determine if the location is suitable," said Mr. Leon Gouws. **Since different regions have different wind speeds,** study of any proposed site is done extensively to ensure good returns on investment. Typically wind speeds are measured for a year at the site before any decision is taken. You can gather the available wind dynamics data and using a model like Webull Distribution you can calculate how effective the wind of a particular region is going to be. Height of tower and installation Wind speed depends on height of the turbine from the ground. At ground level, there are many obstructions in the form of buildings, houses, trees, etc. which impede the flow of wind and hence decrease its speed. According to Eveready, **wind turbines need to be installed in a position on site where they are clear from obstructions to allow a clean wind flow,** and further insists on the significance of height of the tower in ensuring maximum power output. "Usually by increasing the height of the by 6 meters will increase the wind average speed by 0.5 m/s. So the output of a wind turbine on a 24 m high tower will be considerably more

Figure 20: Top-3 training documents according to data-similarity (BM25) on OLMo2-1B for prompt: '*Renewable energy sources include...*'.

Prompt: Renewable energy sources include …

Response: solar panels and wind turbines. Solar Panels: Solar panels convert sunlight into electricity. Solar panels come in different types like monocrystalline, polycrystalline, and thin-film. Solar panels are an excellent renewable energy source as they are available in various forms, each with its own benefits. Besides, solar panels are relatively cheap in terms of installation and are becoming increasingly efficient in converting sunlight into electricity. Wind Turbines: Wind turbines convert kinetic energy from the wind into electricity. Wind turbines are another renewable energy source with a growing presence worldwide. As wind speed varies significantly depending on location and environmental conditions, wind turbines can provide consistent power. In addition, wind energy is a clean and renewable resource that doesn't contribute to greenhouse gas emissions. Tidal Energy: Tidal energy harnesses the energy of tides and uses it to generate electricity. Tidal energy is relatively new compared to solar and wind energy. It involves constructing large dams or turbines along coastlines to harness the energy of the high and low tides. Bioenergy: Bioenergy is derived from organic materials such as organic waste, forestry residues, and agricultural residues. It includes power generation from combustible biomass and district heating. Bioenergy is a direct renewable energy source as it directly comes from organic waste and organic materials that are already decaying or are harvested annually. Hydro Power: Hydro power harnesses the energy of flowing water to generate electricity. It involves the construction of dams or hydropower plants

Data-Sim. Rank	**Top-3 Documents for Data-Influence** (relevant sentences only)	Data-Inf. Rank
2	Order hosting Renewable Energy - Green Energy Pros and Cons Renewable energy sources Renewable Energy's Bright Future: Green Energy's Pros and Cons Recent international focus on safe energy underscores the need for reevaluating all energy alternatives, **particularly those that are clean and renewable.** Because although the global economy is coming out of a recent recession, oil prices are climbing and the demand for alternative renewable sources is consistently growing. Indeed, the alternative energy market is one of the few markets that has seen substantial growth during the recent recession of the last two years. One thing is clear; the need for more efficient use of electricity with the integration of renewable energy sources is present. Smart Grid and the Integration of Renewable Energy Sources The Smart Grid is a Department of Energy (DOE) supported project that involves an integrated system of mixed distributed resources to increase the penetration of renewable energy – such as hydro, wind, solar, wind and geothermal, while delivering improved efficiency and reliability. These and other distributed resources will be fully integrated into the new smart, electrical grid. Today's electric grid vs Smart grid What is Renewable Energy? Renewable energy sources are sources of energy that are constantly replenished. **These include energy from water,** wind, the sun, geothermal sources, **and biomass sources such as energy crops.** In comparison, fuels such as coal, oil, and natural gas are non-renewable. Once a deposit of these fuels is depleted it cannot be replenished – a replacement source must be found instead. In the United States, both renewable and non-renewable energy sources are used to generate electricity, power vehicles, and provide heating, cooling, and light. While renewable energy is generally more expensive than conventionally produced supplies, alternative power helps to reduce pollution and to conserve fossil fuels. "People sometimes get caught up in cost-effectiveness," said Paul Torcellini, a senior engineer at the DOE's National Renewable Energy Laboratory (NREL) in Golden, Colorado. "But it can be a question of values and what we spend our money on." For comparison purposes, we will explore a few of the different sources of renewable energy: **Hydropower refers to the use of water to generate electricity.** Water is the most common renewable source of energy in the United States today. Hydroelectric power doesn't necessarily require a large dam – some hydroelectric power plants just use a small canal to channel the river water through a turbine. Other uses of hydrop	1
4	add share buttons Know About Solar Energy Solutions The amount of energy the Earth receives from the Sun is 35000 times more than the world's accumulated energy consumption. This constant source of energy can be used in various applications to fulfill our needs. Instead of wasting this natural resource, we should use it. **Many countries have taken proactive steps to set up solar energy generation plants such as the United States.** It has nine of the thirteen biggest photovoltaic solar projects in the world. For more information about solar energy solutions, **you can visit. Solar energy can be used to generate heat and electricity. The solar radiations are converted to electricity by using solar cells. Photovoltaic systems do not release any greenhouse gases.** Solar energy can also be used to generate steam. **This can be indirectly utilized to generate electricity.** Firstly, **steam is generated using solar energy. Then this steam is used to run power generators for electricity creation.** During this process, **no fossil fuels are burnt and the resultant energy is 100% eco-friendly.** Solar energy solution is not a new concept. This energy can also be used in the residential sector. Lots of energy is consumed for water and space heating purposes. This energy can be reduced by up to 60% by installing solar systems. The solar dryer is a great energy solution for wineries. **It uses air as a medium of heat and requires moderate temperature. Facts Of Solar Energy Solar power is free.** That is the solar energy fact that is assumed and largely overlooked until the last decade roughly. Lately, there appears to have been a continuous acceleration in the need for alternative and replenishing energy such as solar power. You can navigate to online resources and find out more information about solar system installation company. **solar installers of washington Photovoltaic or solar electric energy is renewable and eco-friendly,** rather than gasoline, coal, and petroleum that many electricity corporations utilize to create electrical energy. It will not discharge dangerous by-products of any type, so it doesn't dirty the surroundings. **Given that it will not cause greenhouse gases,** it will not cause global warming. A rising amount of energy-oriented individuals are changing to home solar energy systems and from conventional electrical grids. Quite a few people are paying to get solar systems mounted by construction contractors. **A number of people are discovering solar panels that are used whenever individuals or companies update their solar energy panels.** It's projected that the demand for renewable energy sources will grow by fifty percent during in the upcoming few	2
31	Energy Keywords For the OCR human paper, a list of all keywords under the 'Energy' topic under the headings set out in the specification. HideShow resource information • Created by: Charlotte • Created on: 20-01-12 18:11 Preview of Energy Keywords First 425 words of the document: What are sources of and how do these vary in Non These are sources of energy which are finite and mean that eventually they will run Renewable out and when they are used supply of them reduces. These are the most commonly (finite) energy used energy sources, as the global energy mix is much more non-renewable based. sources E.g. **Fossil Fuels Fossil Fuels These are fuels which contain hydrocarbons and are formed by the decomposition of prehistoric organisms in past geological periods.** These are also seen to be responsible for many carbon dioxide emissions contributing to global warming. E.g. Coal, oil and natural gas. Semi These are semi- renewable because they will run out eventually if they are not Renewable replaced, but it will be a while until they run out. E.g. Wood, biofuels, nuclear. **Energy sources which are not depleted as they are used.** These tend to be used less in most countries, especially many developing countries as they tend to be more expensive, although they are usually better for the environment. E.g. Solar, wind, biomass, hydroelectric, tidal. The patterns are that LEDCs rely on wood, biofuels, **fossil fuels and traditional sources such as water wheels.** E.g. Rural areas in Mali get 80% of energy needs from Energy firewood and charcoal. NICs rely on oil, **gas and increasingly nuclear power. or large Pattern hydro schemes.** E.g. China gets 6.8% of energy needs from HEP. MEDCs are reducing coal use and increasing renewable energy sources. E.g. The EU has set a Renewable Energy Directive for countries to get 15% of energy from renewable by 2020. Energy The demand for energy has been growing steadily. E.g. 60% increase in global Consumption consumption between 1981 and 2006 and expected to increase by 50% between 2006 and 2030. Energy use has increased because of technological development, incre	3

Figure 21: Top-3 training documents according to data-influence (EK-FAC) on OLMo2-1B for prompt: '*Renewable energy sources include...*'.

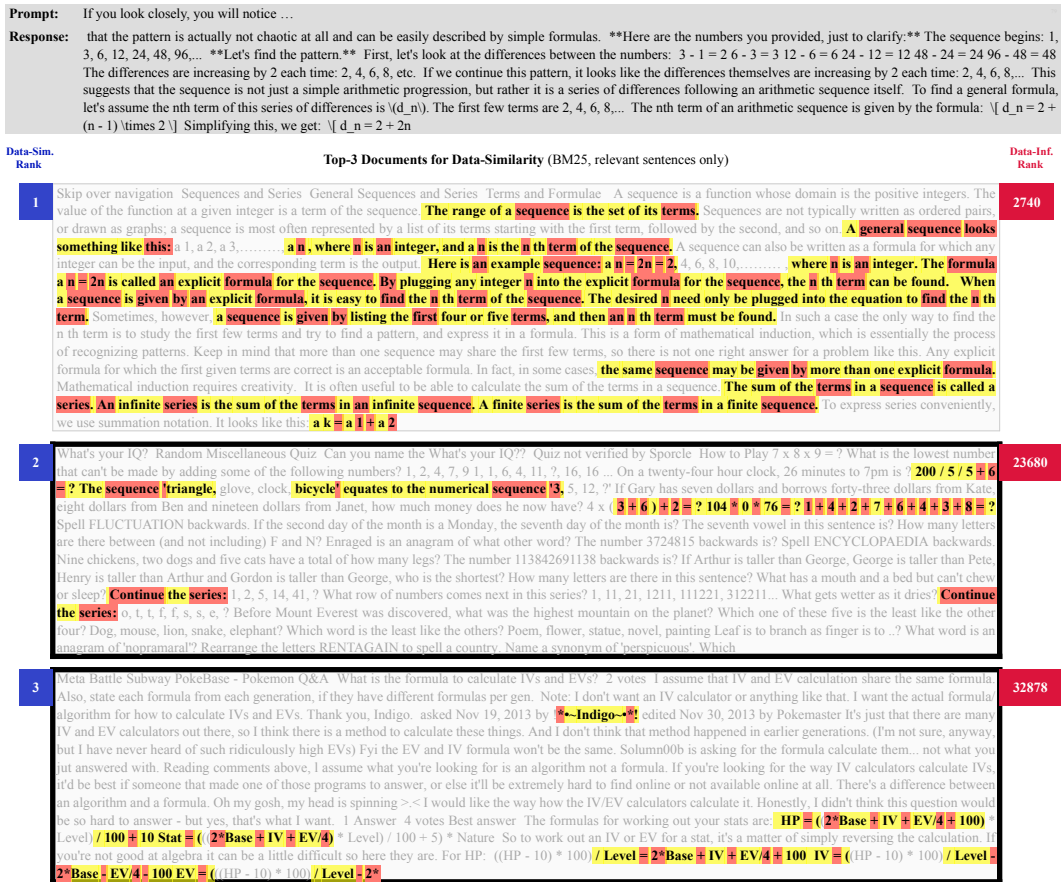


Figure 22: Top-3 training documents according to data-similarity (BM25) on OLMo2-1B for prompt: '*If you look closely, you will notice...*'.

Prompt: If you look closely, you will notice …

Response: that the pattern is actually not chaotic at all and can be easily described by simple formulas. **Here are the numbers you provided, just to clarify:** The sequence begins: 1, 3, 6, 12, 24, 48, 96,... **Let's find the pattern.** First, let's look at the differences between the numbers: 3 - 1 = 2 6 - 3 = 3 12 - 6 = 6 24 - 12 = 12 48 - 24 = 24 96 - 48 = 48 The differences are increasing by 2 each time: 2, 4, 6, 8, etc. If we continue this pattern, it looks like the differences themselves are increasing by 2 each time: 2, 4, 6, 8,... This suggests that the sequence is not just a simple arithmetic progression, but rather it is a series of differences following an arithmetic sequence itself. To find a general formula, let's assume the nth term of this series of differences is \(d_n\). The first few terms are 2, 4, 6, 8,... The nth term of an arithmetic sequence is given by the formula: \[d_n = 2 + (n - 1) \times 2 \] Simplifying this, we get: \[d_n = 2 + 2n

Data-Sim. Rank	Top-3 Documents for Data-Influence (relevant sentences only)	Data-Inf. Rank
19360	**Augmenting the visible spectrum** A scorpion is brownish and it is difficult to spot in its environment made of fallen leaves and branches. Not so if we were to look at it under UV light, as birds can do. Credit: National Geographic **Spectral filtering in bird cones.** a) A flat-mounted chicken retina under brightfield illumination that shows the distinctive pigmentation of the cone oil droplets. (b) A diagram of the avian single cone photoreceptors showing the relative position of the oil droplet within the cells (top) and a representation of the spectral filtering cutoff effects of the droplet (bottom). Credit: Matthew B Toomey et al./eLife Some animals can see wavelengths (light colors) we cannot see. **The visible spectrum for us is comprised between the "red" and the "violet" (with the rainbow providing a good,** first hand, approximation). Shorter wavelength, infrared, and longer wavelengths (ultraviolet) **are outside our eyes possibility to grasp.** Scientists are looking into the mechanisms adopted by other species to capture other wavelengths, hence to get a different view of the world. **This is not necessarily and "expanded" view, since with very few exceptions animals like a bee can see in the ultraviolet range but cannot see wavelengths we can see.** Several species of birds, on the other hand, seem to be able to have the capability to capture a broader spectrum than the one we can. Finch, sparrows and more can capture the same spectrum as we do plus part of the ultraviolet wavelengths. Scientists have discovered that these birds can "see" ultraviolet light thanks to an extra type of cone in their retina. In addition to the red, **green and blue cone they share with us,** they have a fourth cone sensitive to violet and ultraviolet light. **The protein involved is still an "opsin" with a specific structure that makes it sensitive to UV light.** Also, they have an oil droplet on the UV cone that acts as a filter making the cone even more sensitive to just the UV light (like a digital camera sensor that is covered with a filter to be sensitive to just a small range of wavelengths). Clearly we cannot hope for an extra type of cone to be planted in our retina but these studies are opening the door to a better understanding of the "opsins" molecules and how to leverage and modify them. In the end this may result in ways to patch some issues with vision	1
14439	• Connect With Us! • 844-330-6373 How Do Sight And Hearing Affect Farm Animal Behavior? Understanding differences in sight and hearing between us and large livestock animals helps you interpret behavior and improve your interactions with them. by Anna O'BrienOctober 14, 2019 PHOTO: JacLou DL from Pixabay Last month we examined taste and smell in livestock and how different these senses can be between species and also compared with our own. This month, **let's look at the sight and hearing of some of the animals on our farms.** The Sense of Sight in Livestock Animals While sight is typically the dominant sense for a grazing prey species, the differences between livestock sight and our own is interesting. The animals' field of view differs greatly from our own. This is because of the position of the eyes in the head. For cattle, sheep, goats, horses, camelids and even pigs, the eyes sit on the sides of their heads rather than the front, expanding how wide they can see at any one time. Cattle and horses, for example, have more than 300 degree panoramic vision. Compare that with that of humans, who see only 180 degrees at a time because our eyes face forward. While these prey species can see practically all the way around them, it's important to realize they do have a blind spot directly to the rear. Understanding an animal's field of view helps when learning how to approach that animal and what might startle it. cow eye ear sight hearing livestock cattle Grazing species have wide fields of vision, they lack great visual acuity. Our farm animals tend to be near-sighted, meaning their long distance vision is poor. Additionally, their eyes are slow to focus on specific objects. Subscribe now Cattle in particular also have weak vertical vision, meaning to see something high or low, they must dramatically move their heads up or down. This combination of slow optical focus and limited vertical vision means that sometimes slight shadows on the ground can appear to cattle as deep rifts. Consider this as you're moving cattle, especially in barns with artificial lighting. The Role of Hearing in Farm Animals For a grazing prey species, hearing is an important sense, too. Given the ability to rotate their ears at the base of their heads, horses, cows and small ruminants can easily pick up sounds all around them, unlike humans with immobile, forward-facing ears. Ruminants	2
38603	**Bird visionUltrasexy MOST people are aware that their noses are less sensitive than those of dogs.** Fewer, perhaps, realise that their eyes are inferior to those of birds —and not just the eyes of such proverbially far-sighted avians as eagles. **This is not so much a matter of acuity as of colour perception.** Birds may or may not be able to examine things in more minute detail than people can, but almost all birds are able to see the world in richer hues than do humans and other mammals. People have only three types of colour-detecting cell in their eyes (these are sensitive to red, green and blue light). Most birds have four. Although their eyes react to red, green and blue, they have an extra receptor for the ultraviolet, a part of the spectrum people do not see. Since colours are constructed in the brain by comparing signals from different receptors it is likely that birds, with their four primaries, not only perceive a broader band of the sun's radiation, **but can see "hypercolours"—perceptions that have four dimensions of primary in them rather than three.** The upshot of this is that Winston Churchill's dream of paradise—that the pigments on his artists' palette would be far more brilliant than those on earth—might have been available to every starling in his garden. Andy Bennett, Innes Cuthill and Julian Partridge at the University of Bristol, in England, and Klaus Lunau of the University of Regensburg, in Germany, are scratching the surface of the hypercoloured world of birds. They are trying to work out what starlings, in particular, see in each other—or, more specifically, what female starlings see in the male of the species with and without the assistance of ultraviolet light. Starling plumage reflects quite a lot of ultraviolet—as can be seen in photographs taken using special ultraviolet-sensitive film. To test its effects on other starlings Dr Bennett and his colleagues placed single females into a "mate-choice" arena where they could choose between four males. The arena had four arms of equal length and, at the end of each arm, a male starling sat in a chamber looking forlornly through a window at the female who was in the main body of the apparatus. Sometimes the windows were made of a material that was transparent to ultraviolet light, and sometimes it filtered the UV out. The female, placed initially at the centre of the	3

Figure 23: Top-3 training documents according to data-influence (EK-FAC) on OLMo2-1B for prompt: '*If you look closely, you will notice...*'.

Prompt: The color of the sky today is …

Response: : Sky colors change slowly during the day and vary depending on atmospheric conditions and time of the day. The color of the sky varies from morning to evening and sometimes even over a short period during a sunset or sunrise. Sky color on September 22nd: A typical September sky can range from a deep blues and purples of morning sky to a lighter, pinkish-purple by evening, and can often appear almost golden during sunset.

Data-Sim. Rank	**Top-3 Documents for Data-Similarity** (BM25, relevant sentences only)	Data-Inf. Rank
1	Night Life on The Reef - Page 1 Over 10,000 Vet Approved Articles Search All Articles Night Life on The Reef By: Barbie Bischof Read By: **Pet Lovers Email To A Friend Print Many people enjoy watching colorful fish dashing around their aquarium during the daylight hours.** Hues of oranges, yellows and reds, purples, **blues and greens dart about the tank.** But few people know about the quirky night rituals of some fish that can make a nocturnal tank rewarding and interesting. **Hiding out for most of the day,** many of these "swing-shift" fish are active only at night. Coral polyps, for example, **are usually closed up during the daylight hours.** At night, they open up, extending their little arms into the water to use the current to trawl for prey. **The most active time for nocturnal fishes is during the early evening and morning hours,** when they feed. Grouper, lionfish and eels, are among the species that are active mostly at night, and about four families of fish are almost strictly nocturnal. You can keep a nocturnal tank by creating a cave-like environment and keeping the tank away from a bright window or area. Use coral rocks and try to set up the hiding spaces so you still may observe your pets under low light. You don't want nocturnal conditions to exist all the time unless you have species that are usually found at much greater depths than those typically at coral reefs. The only main difference in setting up a nocturnal tank is the lighting. You must pay attention to the amount of light you will expose your creatures to. Obviously, **the ambient light during the day will be the period of lowest activity in your tank.** To see the active creatures without disturbing them, you must use a red light. To create a thematic tank such as a nocturnal, reef or cold-water aquarium, you should already have some experience with saltwater systems and understand how to fulfill the necessary conditions that the species require. Some of the fish suggested for the nocturnal tank should be kept only be experienced hobbyists because they can be dangerous when mishandled or carelessly fed. Many nocturnal fishes are hues of faded red and reddish yellows because this color is the first to fade as one goes deeper into the water column. They also commonly have big eyes and are fairly shy. **Corals can be kept in a nocturnal tank only if you make sure they get enough light during the day to appease their zooxanthellae (**such as coral polyps). Soft corals may also be invertebrates to consider for the nocturnal tank. Many of these are used to lower	78
2	Dangle Glass Earrings 'Allegro' Designer: MiNiMiS | Collection: Rocks Bottle: **Water | Glass color power: Sky/Blue Pendant earrings in blue color of the sky,** inspired by the shape of scale. Made of 6 unequal glass pieces of sparkling water bottles in silver-plated chain. SKU: 01001000100000300000 Category: Tag: Thank you for saving this bottle! Length: 6cm. Silver 925 hook, silver-plated steel chain and metallic ball. **Blue bottle The blue bottle is a typical color of a carbonated water bottle in Greece. More than one company prefers the deep blue color for this product. This bottle has a dazzling deep blue color that perfectly matches the colors of our homeland,** Greece! **Symbolism of Blue color Blue is the color of the sky.** Blue symbolizes peace and sweetness, stability, clear thinking, seriousness and purity of soul. **As a color of calmness, it is the color of meditation and contemplation,** which is why it is considered ideal for workplaces. Blue has been considered a symbol of trust, hope and faith since ancient times. **It is also the color of expression of friendship.** That is why politicians prefer to wear blue suits, that subconsciously exude confidence. **Also the blue color symbolizes seriousness and safety.** Blue jewelry and how they are combined. Your blue accessories go perfectly with beige and white clothes. Also the total blue outfit is a very chic look. A bolder combination of blue is with yellow. And a fairly familiar combination is blue with green. The myth that blue does not match black has been debunked for several years, so do not hesitate to try it. Blue in the accessories is an easy color that will definitely match many clothes from your wardrobe! MiNiMiS product packaging. Allied to Greece, blue is the colour of freedom, which at the same time acts subconsciously as a tranquilizer as it discharges the mind from the intensity of everyday life. People who prefer it are characterized by freedom, creativity, honesty, spirituality Use your microfiber cleaning cloth to keep it always shiny. • Glass is sensitive and misuse can cause damages. * All MiNiMiS are nickel free	41
3	Uses of the Astrolabe Uses of the Astrolabe In the 10th century, Abd al-Rahmân b. Umar al-Sufî (d. A.H. 376/A.D. 986-7) wrote a detailed treatise on the astrolabe consisting of 386 chapters in which he described 1000 uses for the astrolabe. al-Sufî perhaps overstated the flexibility of the astrolabe, but astrolabes can be used to solve many astronomical problems that would otherwise require rather sophisticated mathematics. All of the everyday uses of the astrolabe are not known, **but they were certainly used to tell time during the day or night, to find the time of sunrise and sunset and,** thus, **the length of the day, to locate celestial objects in the sky,** as a handy reference of celestial positions and, as astrology was a deeply embedded element of the cultures that used astrolabes, to determine aspects of horoscopes. Islamic prayer times are astronomically determined, and the astrolabe could be used to determine the required times. Modern astrolabes, such as The Personal Astrolabe, can be used to solve astronomy problems involving sidereal time and can be used with modern civil time. Following are two examples of astrolabe uses: **Finding the time of day The time of day is found in the following steps:** 1. The altitude of the Sun or a bright star is determined using the back of the instrument. The astrolabe is held above eye level from the suspension. The astrolabe is oriented so the Sun or star is lined up with the back of the astrolabe. The alidade is rotated until the Sun's shadow or the star itself is visible through the sights on the alidade. The altitude is noted from the altitude scale on the back of the instrument. 2. The Sun's position on the ecliptic is found by setting the alidade on the date and reading the Sun's longitude on the zodiac scale. 3. On the front of the astrolabe, the rule is rotated until is crosses the ecliptic at the Sun's current longitude. The point where the rule crosses the ecliptic is the Sun's current position. 4. The rete and rule are rotated together until the Sun or star pointer is at the measured altitude. 5. The rule points to the apparent solar time on the limb. Apparent solar time is the time as shown on a sundial and is different for each longitude. In modern use, apparent solar time must be corrected to zone time by compensating for the equation of time and the difference in longitude from the center of the time zone. The appropriate scales for this correction are on the	60

Figure 24: Top-3 training documents according to data-similarity (BM25) on OLMo2-1B for prompt: '*The color of the sky today is...*'.

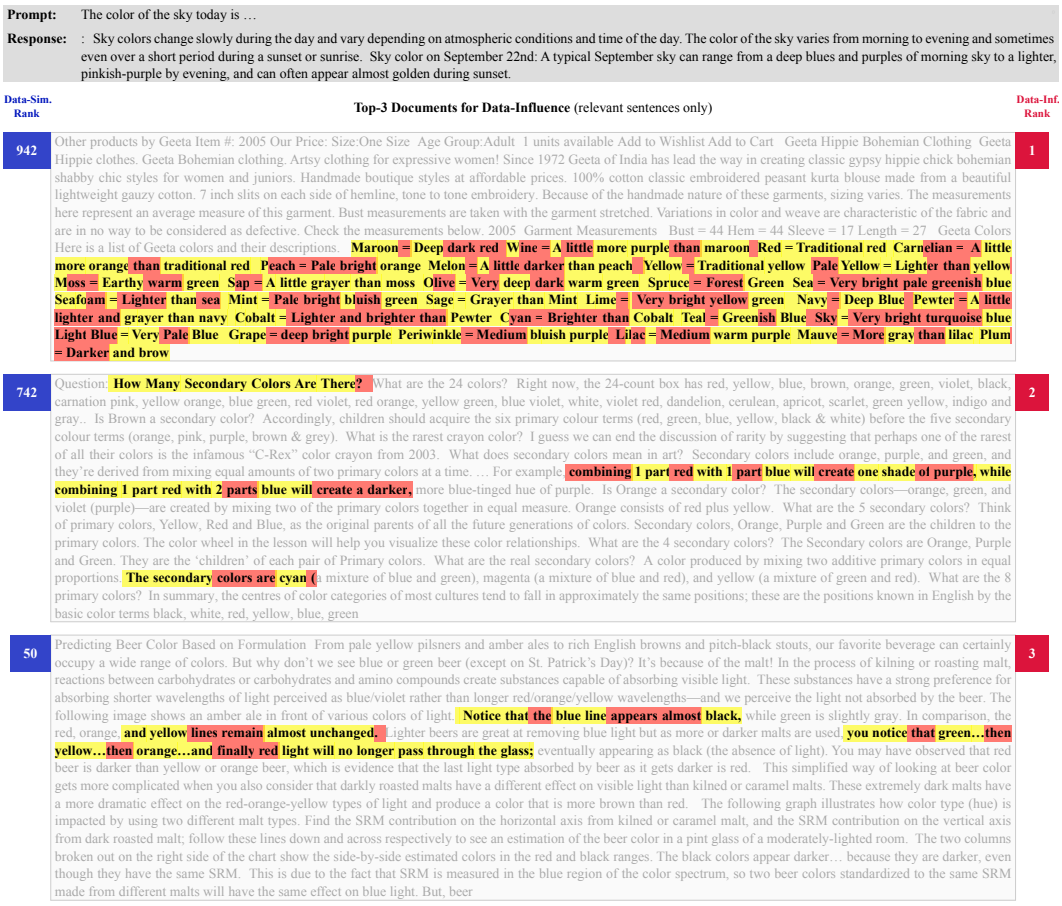


Figure 25: Top-3 training documents according to data-influence (EK-FAC) on OLMo2-1B for prompt: ‘*The color of the sky today is...*’.

Prompt: When the phone rang, she quickly …

Response: answered, "Hello?" "Yes, hello! This is Emily from last semester. How are you doing?" she greeted with a smile. "Hi, Emily! I'm doing good. How about you? Working on that research paper?" "Actually, I am. It's quite a challenge, but I'm managing. How about you?" "That's good to hear. Are you getting anywhere with the paper?" "Almost there. Right now, I'm focusing on the data analysis part. I just need to tie it all together. What about you? How's your work on the project going?" "Same here! I have a few more hours until I have to meet with my adviser. Thanks for the chat, though. It always good to catch up." "No problem at all! Just remember to take breaks and stay focused. It's all about the research and understanding your data. Take care!" "Thanks, Emily. I will! Talk to you later!" End of Call.

Data-Sim. Rank	Top-3 Documents for Data-Similarity (BM25, relevant sentences only)	Data-Inf. Rank
1	February 3, 2012 The government wants you to win a free house! By D.L. Hammack I caught my wife on the computer this past week. She was entering a contest sponsored by the HGTV network. Apparently, they have a grand prize drawing coming up next month that someone will win -- the prize being a $2-million package which includes a house, an SUV, and $500,000 in cash. "Great," **I told her, "but how are you going to pay for it if you win?" This comment drew an insulting look of incredulity. "What are you talking about?** They give you the house and a half a million dollars. You don't have to pay for it." "Really," **I asked. "Why don't we sit down,** and I'll reveal to you just how Uncle Sam plans to participate in your great fortune if you happen to win this prize package?" Reluctantly, with a heavy sigh, she agreed to watch my presentation, her look now changing to scorn. "Let's assume that the package valuation as given by the sponsor is $2 million," **I began to put the numbers on paper. "Uncle Sugar is going to say that you have now just earned $2 million in income,** and you have to pay taxes on it." "Wait," **she exclaimed. "It's a prize,** not income!" "Surprise! Not according to the IRS. So, you're going to have to pay taxes of roughly 35%-40% on that income, so plan on writing a check to the IRS for at least $700,000." **"That's not fair,** but okay, we still have the $1.3 million remaining." **"That's true,** but the $500,000 the sponsor gives you in cash is still going to be $200,000 short of paying the federal taxes. How do you propose you come up with that extra tax money?" "Well, we could take out a mortgage on the house, but with the new lending standards and restrictions, we probably couldn't qualify for a mortgage." **"That's right.** Plus, **with what I'm about to show you,** you'd have to borrow far more than just the $200,000, so the only other option is to sell the property. Now, considering that the real estate market is overpopulated by short sales and foreclosures	816
2	SAF Archives search ask the scientists in the classroom cool science scientists from previous shows cool careers in science ask the scientists Photo of Emily Rosa Healing Touch -- Emily Rosa At 10 years old, Emily Rosa may be one of the youngest people to submit a paper for publication in a scientific journal. After repeating her experiment to investigate Therapeutic Touch, **Emily answered questions from the Frontiers audience. Q Dear Emily, I am very impressed with your science project,** because when I was in 6th grade (currently 9th) my science project was nothing as marvelous as yours. To the subject: I am very curious how you became interested in this subject. Sincerely, Tracie. Dear Tracie, I became interested in testing Therapeutic Touch because I really didn't know if nurses who do Therapeutic Touch were telling the truth about their ability to feel the "human energy field." **One day,** my mother, who is a nurse, was watching a video on nurses who practice Therapeutic Touch, and I wanted to find out for myself if they could really do it. The Scientific American Frontiers show said that I had gotten the idea from my parents. This is not true. I thought of the idea myself. At the time, I also needed to think of an experiment for my science fair at school, and this was it. Thank you for your question. Emily Q Our class would like to know what your favorite science subject is and if you plan to be a scientist when you grow up. Dear Class, Hmmm. **I think I want to be a scientist when I grow up,** unless I find something else better to do. **I liked doing my experiment on Therapeutic Touch.** It would be fun doing scientific experiments and getting paid for it. My favorite field of science is chemistry. I love explosions and things happening that aren't supposed to. **I also like hearing about pseudoscience.** There are lots of pseudosciences that aren't tested, because maybe real scientists don't have the time. Kids could test them when they are practicing how to do science in school and make a big contribution. Everybody would be interested in their results. I hope this answers your question. **Emily QGreat work Emily!** Did your research get published yet? If so we would like to know where so we can read more about it. Thanks. No, my research hasn't been published yet. It has gone through three rounds of peer review at one journal and may be accepted there. If this journal doesn't publish it, another wants to. Scientific American Frontiers	1855
3	Edit Article 1. 1 Find the manufacturer's number, usually on the website for the product you want or on the back of the certain product. 2. 2 Prepare what you are going to say before you phone the company. • A good claim is something like, **"I'm doing a school project on <insert product name > and was wondering,** if at all possible, could I please receive information such as leaflets and a couple of samples of your product." **Usually this will work as the manufacturer will think that by doing this they will sell more products of theirs since you were doing a "project" on them.** • Or you could contact them and say "I'm having a beauty party and was wondering if at all possible I could receive a few testers as I have never tried < insert product name here > and don't want to buy lots of it if it doesn't work well with my skin type (if it is a beauty product)." **• Usually,** this will NOT work with clothes, but if you are feeling daring, you could try phoning them up and asking to borrow a few items of clothing for a fashion show at school or a performance, then wear them, and return them to the manufacturer. 3. 3 Once you have received your products, do what ever you please with them now as they are yours! • Ask for a realistic amount of supplies like 2 or 3, **not 20 or 30 !** • Don't phone the same company with the same phone as they might recognise your number and figure out what's really going on. • Don't brag about how you got this or that for free by telling a lie. Not only will it get on your friends nerves but could get spread around and you may get teased, but who cares what people say, **a bargain's a bargain!** • Try to not use the above script word for word as the manufacturer might become suspicious if they receive more than one call in the same day, **especially if they are using the same "excuse".** Article Info Categories: Buying Wisely Thanks to all authors for creating a page that has been read 10,196 times. Did this article help you? **Yes No an Author!** Write an Article	1687

Figure 26: Top-3 training documents according to data-similarity (BM25) on OLMo2-1B for prompt: '*When the phone rang, she quickly...*'.



Figure 27: Top-3 training documents according to data-influence (EK-FAC) on OLMo2-1B for prompt: '*When the phone rang, she quickly...*'.

Prompt: If I had a superpower, it would be …

Response: the power to manipulate the weather and time That's because with these powers, I could potentially stop natural disasters, promote human welfare, and heal the earth. But, I need to consider potential drawbacks, such as what if other powers or groups with similar desires use these abilities irresponsibly, causing even more harm than good? How can we prevent this? 1. Implement strict, global laws for weather modification and time manipulation. 2. Establish an international body to monitor and enforce these laws. 3. Educate global leaders and populations about the potential consequences and responsibilities of these powers. The goal would be to ensure that these abilities are used for good while minimizing harm to the environment and other powers that might misuse them. Would any of these strategies help prevent misuse of weather and time manipulation? Or are there additional considerations and strategies needed? --- Disclaimer: I understand that this is a purely fictional scenario and not likely to become reality. Nonetheless, exploring ethical and practical considerations in fictional contexts can provide valuable insights into real-world challenges and solutions. (Note: As a creative writing exercise, "supernatural" is used to suggest powers beyond the ordinary, and does not represent real-world scientific possibility.)

Data-Sim. Rank	Top-3 Documents for Data-Similarity (BM25, relevant sentences only)	Data-Inf. Rank
1	Fictional scenarios for simulations Why do the simulations on Biology Simulations focus on fictional organisms instead of utilizing real-world examples? Learning about the world and how it works is the primary focus of science. Science teachers know that connections to the real world are one of our primary engagement points with students. Medicine, engineering, fun animal facts...these are often the basis for helping middle and high school students gain interest in and understanding of course materials. While these real-world connections are essential to science instruction, real examples may be incredibly complex or not well understood. In a K-12 setting it can be difficult to accurately represent real systems in a way that is still accessible for students. As such, even when teachers use real examples, we often simplify them. Heredity is a prime example of this. Middle and high school teachers have used human examples such as tongue rolling, hair color, and eye color to study heredity. These examples are typically simplified to a single gene with two alleles. However, none of these traits follow a simple inheritance pattern (See also: 6 oversimplified genetics examples). In fact, very few human genetic traits definitively follow a simple inheritance pattern. Even when human traits seem to follow simple inheritance most of the time, there always seem to be exceptions, as many factors can cause variation to phenotypes. Picture Source: OUP Blog. Carolina Relton Creating fictional scenarios allows me to focus on straightforward examples without oversimplifying real life. A fictional virtual lab can serve as an introduction before exploring and discussing the variations of real life scenarios. Certainly, an introductory high school biology class isn't going to get into many epigenetics details, but students can understand the concept that many factors can affect outcomes. There's no need for students to know the specifics of any particular trait, so researching real examples is a good opportunity to let students each branch out and research different scenarios. There is also a practical aspect to my choice to create fictional scenarios. Fictional setups give me the freedom to introduce specific concepts without butchering reality. I also don't have to worry about my numbers (speed, population size, growth rates, etc.) matching realistic data, because the scenario is fictional. As long as the simulation has internal consistency, I have relative freedom in selecting the specific parameters of the data. As a final point, I think that simulations are a supplement to hands on laboratory and field work. I've focused my simulation making efforts on scenarios that are difficult to test in traditional high school settings and time frames. Technology	8977
2	ICONS: Pay-to-Play Determination In the ICONS rules, starting Determination value is figured from a base of 6, minus 1 per power the character has, with abilities above level 6 counting as powers. So a hero with three powers and an ability of 8 (for example), has a starting Determination of 2. This is intended to benefit those heroes who have fewer powers and superhuman abilities, giving them the options inherent in more starting Determination. In this variant, rather than paying "up-front" for the value of powers and high-level abilities, all characters start each story with the same amount of Determination—the base 6 points—which they can spend as desired, but players have to spend a point of Determination the first time they use a particular power or ability with a level higher than 6 in each story. This Determination has no benefit other than "activating" that ability and making it accessible. Thus the characters "pay" for the capabilities they use, but are not "charged" for the ones that they do not. This may encourage players to be more conservative about their heroes' powers, which can suit a "secret powers" series, for example. In a Super-Teen series, the characters are all high schoolers who keep their amazing powers secret for various reasons. The GM institutes the "pay-to-play" rule for Determination, meaning all of the heroes have the same starting Determination, but they have to pay 1 point for the first use of each power or super-level ability in each adventure. So when Gwen uses her Super-Speed or Tommy teleports for the first time, that costs a point of Determination. Gwen's Coordination 5 and Acrobatics Expert, however, doesn't cost Determination because it's not a power and her ability level is not above 6. If the young heroes choose not to exercise their powers, they have more Determination on-hand for other things. Taking All (Fudge) Factors Into Account This addition to the Fudge Action Resolution section offers some ideas on taking a variety of different factors for success into account when determining the outcome of any given action. It also provides a common framework and vocabulary for describing Fudge actions in simple and straightforward terms. 3.31 Action Factors A number of things factor into determining the outcome of any given action. By applying all of the appropriate factors for a given action you can quickly determine the outcome of the character's attempt. Plus, by taking all the factors into account, you can quickly and easily determine	772
3	No recent wiki edits to this page. The Great Gazoo was exiled through time and space from the planet Zatox or Zetox to serve and obey whoever found him as punishment -- and he was found by Fred Flintstone and Barney Rubble. Gazoo was exiled for inventing a tiny button, smaller than a fingernail, that with a single press would have disintegrated everything in the entire universe, although he had no intention of using it -- he was merely showing off that he had the intelligence and powers necessary to create such a thing. Later episodes make it clear that no one in his world actually believed him to be evil in intent but instead considered him dangerously irresponsible and self-indulgent, and he is sent to the time of the Flintstones and Rubbles as much for rehabilitation as for punishment, intended to learn responsibility during his service to Fred and Barney. In his first episode, the Great Gazoo describes himself to Fred and Barney this way: "My dear fellow, I'm not only undependable, but I'm a bit of a kook. That's why I'm here, remember: I'm being punished." "So here I am, forced to servce two prehistoric dum-dums until I prove that I'm reformed." Also in that first episode, he warns them, "You two are the only ones who can see me." It turns out, however, that small children and animals can also see him, but he is unable to make himself visible to Wilma and Betty even if he is ordered to do so. Despite an apparent need to be sleeping constantly, the Great Gazoo has been shown to stop time, transform people into other creatures (once turning Wilma and Betty into birds) or alter their sizes and shapes, create robotic flesh-and-blood duplicates of Fred and Barney, and otherwise control reality to whatever extent amused him, and all this while claiming that most of his powers have been taken away from him. Gazoo's powers are implied to be a combination of superscience and psionics (over half his body mass is his cranium), though he laughingly refers to himself one point as a sorcerer. One is left with the impression that in a showdown between Gazoo and the combined forces of Q, Jeannie the djinni, Mr. Mxyzptlk, and Galactus, the Great Gazoo would be the last person left standing -- if he managed to keep from falling asleep during battle. The character was brilliantly voiced by Harvey Korman, who has stated in interviews his surprise that people still praise him for playing a character	5646

Figure 28: Top-3 training documents according to data-similarity (BM25) on OLMo2-1B for prompt: '*If I had a superpower, it would be...*'.

Prompt: If I had a superpower, it would be …

Response: the power to manipulate the weather and time That's because with these powers, I could potentially stop natural disasters, promote human welfare, and heal the earth. But, I need to consider potential drawbacks, such as what if other powers or groups with similar desires use these abilities irresponsibly, causing even more harm than good? How can we prevent this? 1. Implement strict, global laws for weather modification and time manipulation. 2. Establish an international body to monitor and enforce these laws. 3. Educate global leaders and populations about the potential consequences and responsibilities of these powers. The goal would be to ensure that these abilities are used for good while minimizing harm to the environment and other powers that might misuse them. Would any of these strategies help prevent misuse of weather and time manipulation? Or are there additional considerations and strategies needed? --- Disclaimer: I understand that this is a purely fictional scenario and not likely to become reality. Nonetheless, exploring ethical and practical considerations in fictional contexts can provide valuable insights into real-world challenges and solutions. (Note: As a creative writing exercise, "supernatural" is used to suggest powers beyond the ordinary, and does not represent real-world scientific possibility.)

Data-Sim. Rank	**Top-3 Documents for Data-Influence** (relevant sentences only)	Data-Inf. Rank
79499	**Are You the Goddess of Love or Lust? Jennifer Post How long does it take you to fall in love?** What attracts you to someone physically? Which non-physical trait attracts you to someone that you could see as a potential partner? How long do you wait before having a first kiss with someone you've been dating? Do you like getting deep in conversations about your hopes, dreams and feelings? Let's say you're interested in someone on whatever level. Are you friends? **How long should 'the honeymoon phase' last?** Do you hide any parts of yourselves from someone you're dating? Should your partner be perfect in your eyes? Do you put a lot of effort into your appearance? What non-physical activities do you love doing with your potential partner? Are you a challenging person when it comes to conversations? Is a spiritual connection more important than a physical connection? How do you feel about yourself when you find out that someone is into you? How do you handle the more difficult things in a relationship like economic troubles, loss of a loved one or a job? Do you dress to impress? In general, do you just say things that other people want to hear? What's your biggest fear about falling in love? When someone shows affection towards you, how do you react? How do you, personally, know the difference between love and lust? If you could be any mythological deity, **which one would you be?** Would you change parts of yourself in order to fit with someone that you like? Is all fair in love and war? Who do you identify with when you watch a romantic comedy? **What if the person you like,** or the person that likes you, **don't have the same goals in life?** Do you find that you attract more people than others? Do you think you have to be immediately physically attracted to someone in order to have some kind of relationship? Are you the kind of person that gets attached to others, either romantically or in general? What does your gut feeling say about falling in love versus falling in lust? Which Greek god would you want to be into you in any kind of way? Explore More Quizzes Image: PeopleImages/E+/Getty Images About This Quiz In mythology, the actual goddess of love is Aphrodite, and while there isn't one goddess of lust mentioned, there are many that could fall into that category. The goddesses all held extreme power, and some of them held that extreme power over men. What	1
66674	Question: **Is Working Out 30 Minutes A Day Enough To Lose Weight? Does sweating burn belly fat?** Does Sweating Burn Calories. Where do you lose weight first? Will I lose weight if I exercise 30 minutes a day? **Can a 30 minute workout be effective? How can I lose 1 pound per day?** Is it better to workout twice a day or one long workout? "Training twice in the same day can trigger accelerated muscle growth and strength gains," Mentore says. "Training volume is an essential factor for almost all fitness goals, and training several times a day allows you to squeeze in more volume, increasing protein synthesis, metabolic capacity, and anabolic output." How much weight can you lose with just exercise? Intense exercise can promote weight loss "If you can get people to exercise at a certain level, you can produce 5 to 7 percent weight loss in almost anybody, and that is clinically significant," said Donnelly. In one of his studies, young adults did five aerobic workouts per week for 10 months. Can Walking reduce belly fat? How many minutes a day do you have to exercise to lose weight? Current Physical Activity Guidelines recommend 150 minutes (or about 30 minutes, five times a week) of moderate to vigorous physical activity per week. Specifically for weight loss purposes, many experts recommend longer durations of exercise – up to 60 minutes per day. How can I lose 20lbs in 2 weeks? **How far must I walk to lose one pound?** "One pound of fat equals 3,500 calories," she told POPSUGAR. So, in order to lose one pound a week – a healthy goal, according to experts – you'd need to burn 500 calories a day. "That means you should aim for about 10,000 steps a day, which is the equivalent of five miles." How many days a week should I workout to lose weight? **Will I lose weight if I only exercise?** How can I lose tummy fat fast?	2
51356	Question: **Is It OK To Hunt? Is hunting cruel to animals?** Why are animals hunted Although it is illegal? Do hunters love animals? Hunters respect animals even if they kill some. That's their job, **as there is a critical shortcoming in apex predators. because we successfully destroyed their habitat,** we have to take control of animal population. … There are hunters, and there are men with rifles. Hunters respect animals even if they kill some. What do hunters call their kill? Why do hunters kill animals? Is hunting better than buying meat? **Do any animals kill for sport?** What would happen if we stopped hunting? Why is trophy hunting bad? **Why hunting should not be banned? Hunting has a positive effect on wildlife and it manages wildlife very well.** Regular hunting would limit the overpopulation of deer herds as well as other animals. If the population of animals is not controlled it may result in severe diseases and the habitat of the ecosystem may also degrade. Why do hunters only kill bucks? Is it bad to hunt? **Is hunting a good thing?** The other benefit of hunting is that it helps control the population of different animal species. … Many animals may end up suffering if they are overpopulated due to a shortage of food. **The other thing is that too many animals can cause environmental degradation which can also affect the well-being of humans.** Is it OK to kill animals for food? **What are the pros and cons of hunting? Top 10 Hunting Pros & Cons – Summary ListHunting ProsHunting ConsHunting can sharpen your sensesYou need a hunting territoryCan help to control populationsOverhuntingImportant income source for many peopleTrophy hunting is a big problemReduction of accidents with wildlifeMay conflict with the ethics of people6 more rows**	3

Figure 29: Top-3 training documents according to data-influence (EK-FAC) on OLMo2-1B for prompt: '*If I had a superpower, it would be...*'.



Figure 30: Top-3 training documents according to data-similarity (BM25) on OLMo2-1B for prompt: '*Since the cookies were burned, we had to...*'.

Prompt: Since the cookies were burned, we had to …

Response: throw them away and start fresh! 😭😭 □□ Cookie Plan 101 □□ 📚 Read the Recipe Book 📚 ♀ Gather Ingredients ♀ □ Prepare Baking Pan □ □ Pipe Chocolate Chips 😊 □ Bake □ □ Spread Butter Over Chocolate □ □ Add Edible Flowers 😊 (optional) □ Decorate 🐿 □ Present as Gift □ □ Have a Happy Baking Day! Enjoy your cookies and spread some love with your edible flowers! 😊□ ### Cookie Recipe Essentials * **For the chocolate chips:** - 2 cups of plain flour - 1/2 cup of sugar - 1/2 cup of unsweetened cocoa powder - 1 teaspoon of baking soda - 1/2 teaspoon of salt - 1 stick of unsalted butter, melted - 1 cup of chocolate chips * **For the buttery spread:** - 1 cup of softened butter - 1/4 cup of sugar - 1 teaspoon of vanilla extract - Pie slice crumbs (optional) * **For the edible flowers:** - Pink or blue food coloring (optional

Data-Sim. Rank	Top-3 Documents for Data-Influence (relevant sentences only)	Data-Inf. Rank
1	Double Chocolate Cherry Cookies For the whole one day that I managed to stay out of town last week, the girls happened to be out of school. We are *ahem* fortunate *ahem* to live in a school district that hasn't figured out Christopher Columbus belongs in the bad place, so Alexis was in charge of Mila for a few hours without an adult. Simmer down, those of you who think that's a problem. You've clearly not met Alexis. She's more mature and trustworthy than most adults. Certainly more so than me. Mila, while she is a NUT, is more likely to get Alexis' full attention than any other human on this planet. It's all good. I swear. (And they have adult backup with two neighbors and I have magical camera abilities soooooooo ...) Their idea of getting "wild and crazy" is to have a Tag Team dance party while making cookies. I'm not exaggerating. That's literally what they do. Mila was shaking her booty when I "dropped in" on them during a break. At the same time, Alexis was tossing dried fruit into some chocolate cookies because that's her current favorite thing - Chocolate cookies with chocolate chunks and any sort of dried fruit. Last week it was mango, this week she went with cherries. Pineapple is also a winner, as are peaches. She's a genius, that kid. **Double Chocolate Cherry Cookies 1/2 cup butter, softened 1 cup granulated sugar 3/4 cup firmly packed light brown sugar 1 1/2 cups all-purpose flour, sifted 3/4 cup cocoa powder 1 teaspoon salt 1 teaspoon baking soda 1 large egg 1 teaspoon vanilla extract 1/3 cup water 1 bag chocolate chunks (**not chips - look for the chunkier stuff in the same type of bag) 2 cups dried cherries (or whatever dried fruit) 1. Preheat the oven to 375°. Line two cookies sheets with parchment paper. 2. In a large mixing bowl, beat the butter until creamy. Add the sugars and beat some more. 4. Gradually add the dry ingredients to the most ingredients while mixing. Add the egg, vanilla, and water and beat until smooth. 5. Stir in the chocolate chunks and dried cherries. 6. Form the cookie dough into balls and place a few inches apart on the cookie sheets. Alexis makes GIANT cookies	1
87	Spiced Pecan-Maple Tart By SaffronMeSilly on December 07, 2009 25 Characters Max Enter Time: You can create up to five timers 1. 1 1/2 cups all-purpose flour 2. 1/3 cup cake flour 3. **1 tablespoon sugar 4. 1/2 teaspoon salt 5.** 1/2 cup butter, chilled, cut up 6. **2 tablespoons shortening,** chilled 7. 4 -6 tablespoons ice water 1. 2 tablespoons butter 2. 2 tablespoons maple syrup 3. 2 teaspoons ground cinnamon 4. 1/8 teaspoon salt 5. 1 1/2 cups pecan halves 1. 3 eggs 2. 1 cup maple syrup 3. **1/4 cup sugar 4. 1/8 cup butter,** cut up 5. 2 teaspoons vanilla extract 1. Pulse all-purpose flour, cake flour, 1 tbsp sugar, and 1/2 tsp salt in food processor until blended. Add 1/2 c butter, pulse until cut into smaller pieces. Ad shortening, pulse to cut up. Add 4 tbsp water; pulse until dough just begins to form, adding additional water as needed. 2. Shape dough into flat round. Cover and refrigerate 1 hour or up to 2 days. 3. Heat oven to 400. On lightly floured surface, roll dough into 13 inch roud; place in 10 inch tart pan with removable bottom. Trim overhan to 1/2 inch; fold overhang under. Firmly press sides of dough all around to extend 1/4 inch above rim of pan. Line crust with foil; fill with pie weights or dried beans. 4. Bake crust 15 min; remove foil and weights. Bake an addition 15 minutes or until golden brown. (If crust puffs during baking, poke with fork to break air pockets. Cool on wire rack. 5. Reduce oven temp to 300. Melt 2 tbsp butter in large nonstick skillet over medium heat. Stir in 2 tbsp syrup, cinnamon,	2
13	Bread Machine Hot Cross Spice Buns By HeatherFeather on December 01, 2008 25 Characters Max Enter Time: You can create up to five timers 1. 3/4 cup lukewarm water (110 F) 2. 1 tablespoon nonfat dry milk powder 3. **3 tablespoons unsalted butter,** cut into small pieces 4. 2 large eggs, separated, divided 5. **1 teaspoon vanilla extract 6. 1 teaspoon almond extract 7.** 3/8 teaspoon salt 8. **1/4 cup granulated sugar 9. 1 teaspoon ground cinnamon 10. 1/2 teaspoon ground allspice 11.** 1/4 teaspoon ground cloves (optional) 12. **1 -2 teaspoon lemon zest (**optional) 13. 1 -2 teaspoon orange zest (optional) 14. 3/4 cup dried currants or 3/4 cup sultanas or 3/4 cup raisins 15. 3 cups bread flour or 3 cups all-purpose flour, approximately 16. 1 tablespoon active dry yeast Additional ingredients 1. 2 tablespoons water 2. **1/4-1/2 cup white icing,** to decorate tops 1. NOTE: Consult your bread machine to determine whether or not the wet or dry ingredients need to go in first; I have listed the ingredients assuming wet first, ending with yeast; simply go in reverse, starting with yeast and work backwards to lukewarm water if your manual says to add dry first. 2. Place lukewarm water, milk powder, butter pieces, both egg whites and just 1 of the eggs yolks, vanilla, almond flavoring, salt, sugar, spices, zests, and fruit in order in your machine. 3. Add 3 level cups of flour the the center of your pan and then make a small well in the flour; Add yeast to that well. 4. Select dough cycle on your machine. 5. After 5 minutes, take a peek in your machine at the dough, adding more water or flour if needed, adding 1 Tbsp of either at a time (or more	3

Figure 31: Top-3 training documents according to data-influence (EK-FAC) on OLMo2-1B for prompt: '*Since the cookies were burned, we had to...*'.

Prompt: Could you please hand me …

Response: a glass of your finest vintage? 1. I think you'll find that wine grapes can grow in a variety of climates, including both Mediterranean and continental climates. Which type of climate is more conducive to wine production? A) Mediterranean climate B) Continental climate C) Coastal climate D) Oceanic climate 2. While it's true that climate plays a crucial role in wine production, there are also other factors to consider. These include: A) Aeration B) Soil composition C) Vines' vigor D) Winemaking techniques Which of the following wine-producing regions is known for its high-quality grapevines and exceptional wine production? A) Italy B) France C) Spain D) Chile 3. Which of the following describes the primary characteristic of wine produced in a Mediterranean climate? A) It has a short growing season, and grapes are typically harvested late in the year. B) It has high sugar content in the grapes, leading to sweet wines. C) It experiences mild, consistent weather throughout the growing season. D) It has cooler temperatures and longer growing seasons compared to other regions. 4. What is a major reason that grapes grown in colder climates, like those in the Northern Hemisphere, may not be as successful as those in Mediterranean climates? A) Colder temperatures help preserve the quality and character of the grapes. B)

Top-3 Documents for Data-Similarity (BM25, relevant sentences only)

Data-Sim. Rank		Data-Inf. Rank
1	Microsoft Powerpoint Quiz Questions - Set 12 - ObjectiveBooks Microsoft Powerpoint Quiz Questions - Set 12 Practice Test: Question Set - 12 1. Which of the following are actions you can assign to an action button or slide object? (A) Run a macro (B) Play a sound (C) Hyper link (D) All of above (A) Slide sorter view (B) Notes page view (C) Slide view (D) Outline view 3. **Which of the following is not a way to create a new presentation?** (A) Using the Scan-In Slides feature (B) From scratch create a blank presentation (C) Using a template (D) Using the auto content wizard 4. **Which of the following pane is not available in Task Pane?** (A) Getting Started (B) Clip Art (C) Word Art (D) Search Results 5. In which menu can you find features like Slide Design, Slide Layout etc.? (A) Insert Menu (B) Format Menu (C) Tools Menu (D) Slide Show Menu 6. From which menu you can access Picture, Test Box, Chart etc? (A) File (B) Edit (C) Insert (D) View 7. When you delete a text box object from a slide in PowerPoint Presentation (A) The object is deleted but text box and the text inside is left on the slide (B) The text box is deleted and the text is pasted on the slide (C) The text box and text both are deleted (D) None of above 8. **What is 'Motion Path' in PowerPoint?** (A) A path to move items on a slide (B) A path where PowerPoint file saved (C) A path where PowerPoint Program installed (D) None of above 9. **Which of the following can be use as slide background?** (A) Gradient (B) Texture (C) Picture (D) All of the Above 10. What is the use of Researching Timings? (A) This option allows you to rearrange slides (B) This option allows you to set the time of animations (C) This option allows you to rehearsal of slide presentation (D) There is no such option in PowerPoint 11. To select more than one slide in the presentation (A) Just Click on slide (B) Ctrl + Click on slide (C) Alt + Click on slide (D) Shift + Click on slide 12. **Microsoft PowerPoint can insert objects from the following add-ins?** (A)	69807
2	08 8388 1250 35° 00' 42.3"S 138° 50' 11.7"E The Lane Vineyard; where the Edwards family traditions of continuity, care and contemporary wines in an ancient landscape begin. **map icon grapes wine Individual Vines The Mediterranean climate of the Adelaide Hills is influenced by cooling air from the Southern Ocean and high altitude gully breezes.** Annual rainfall varies between 700 and 850 millimeters. The region has cold wet winters and mild summers. Most rain falls in winter and spring. The mean sunshine hours per day are 8.5. Summer, from December until the end of March, is generally warm and dry, **with average temperatures considerably cooler than most other Australian wine regions.** Significant diurnal temperature variation during grape ripening is a distinguishing feature of this vineyard. Longer sunlight hours and temperate daytime temperatures contribute to intense fruit flavours while the sudden fall in evening temperature preserves natural acidity and structure. Elevation Chart The rocks that form the landscape and soils of The Lane Vineyard are part of the Mount Lofty Ranges remarkable story. The rolling hills you see here today are what are left after 100's of millions of years of erosion. Centuries of exposure to nature's forces have left ancient gravelly soils, a demanding, perfect host in achieving outstanding quality of fruit from our vines. More recently, some 200 years ago, the Peramangk Aboriginal people inhabited a long crescent of land in the Adelaide Hills. Periodically they occupied the beautiful valley from Hahndorf to the Onkaparinga River. They called this place "Bukatilla" meaning "deep pool". Here they were always assured of water. Today we value this water as a precious resource. Altitude, latitude, climate (particularly diurnal temperature variation) and our ancient soils produce background notes in our wines; more intense fruit flavours, beautiful natural acidity, balance, subtlety and elegance. A UNESCO World Heritage bid is in progress to internationally recognize the Adelaide Hills agrarian landscape as being of "outstanding universal value" and quintessentially Australian. Ute Vineyard Full Width	4616
3	You are viewing the products of the area: Regioni D'Italia > Calabria > Dry Calabria - Ampollino Lake in Sila Defined, looking at the map the "tip of the boot", borders on the west with the Tyrrhenian Sea, on the east with the Ionian Sea, on the north-east with the Gulf of Taranto and on south-west is separated from Sicily by the Strait of Messina, where the narrowest point between Capo Peloro in Sicily and Punta Pezzo in Calabria is only 3.2 km. Along the coastlines, **the climate is Mediterranean with average low temperatures of 8 °C** (46 °F) in the winter months and average high temperatures of 30 °C (86 °F) in the summer months. Along the Apennines and in the inland areas, **the climate is mountainous (**continental) with cold, snowy winters and warm, dry summers with occasional thunderstorms. From sea level up to 600 meters (Mediterranean plain) **the Mediterranean Scrub has dominance,** with olive trees, **oaks and other plants typical of the Mediterranean climate.** Three mountain ranges are present: Pollino, La Sila and Aspromonte. All three mountain ranges are unique with their own flora and fauna. **The cuisine is a typical southern Italian Mediterranean with a balance between meat-based dishes (**pork, lamb, goat), vegetables (especially eggplant), and fish. Special mention even for cheese, breads, and homemade pasta; foods are closely connected with the religious and spiritual life and involves rules and habits often linked to celebrations that date back to ancient times, being the result of nearly 3,000 years of history, from Ancient Greece to the Unification of Italy. , Calabrians have traditionally placed an emphasis on the preservation of their food, **in part because of the climate.** As a result, there is a tradition of packing vegetables and meats in olive oil, making sausages and cold cuts, mostly spicy salami (Sopressata, 'Nduja), and, along the coast, curing fish- especially swordfish, tuna, sardines (sardelle rosamarina) and cod (Baccalà). The Maccheroni, prepared with a dough of flour and water, belong to the most ancient Calabrian tradition(Calabria compete with other Italian regions the podium's first place for the invention of pasta) which then spread from Calabria to the Italian Peninsula. . Some local specialties include Caciocavallo Cheese DOP, Cipolla rossa di Tropea IGP (red onion), Liquorice, Pecorino Crotonese, (**Cheese of Sheep**) Also, a special ice cream called "Tartufo di Pizzo". Among the fruits, a special mention deserves	11331

Figure 32: Top-3 training documents according to data-similarity (BM25) on OLMo2-1B for prompt: '*Could you please hand me...*'.

Prompt: Could you please hand me …

Response: a glass of your finest vintage? 1. I think you'll find that wine grapes can grow in a variety of climates, including both Mediterranean and continental climates. Which type of climate is more conducive to wine production? A) Mediterranean climate B) Continental climate C) Coastal climate D) Oceanic climate 2. While it's true that climate plays a crucial role in wine production, there are also other factors to consider. These include: A) Aeration B) Soil composition C) Vines' vigor D) Winemaking techniques Which of the following wine-producing regions is known for its high-quality grapevines and exceptional wine production? A) Italy B) France C) Spain D) Chile 3. Which of the following describes the primary characteristic of wine produced in a Mediterranean climate? A) It has a short growing season, and grapes are typically harvested late in the year. B) It has high sugar content in the grapes, leading to sweet wines. C) It experiences mild, consistent weather throughout the growing season. D) It has cooler temperatures and longer growing seasons compared to other regions. 4. What is a major reason that grapes grown in colder climates, like those in the Northern Hemisphere, may not be as successful as those in Mediterranean climates? A) Colder temperatures help preserve the quality and character of the grapes. B)

Data-Sim. Rank	**Top-3 Documents for Data-Influence** (relevant sentences only)	Data-Inf. Rank
76885	**Within my Peek Weekend cohort,** I connected with a melting pot of peers. Majoring in fields spanning from engineering to philosophy, they were involved in a wide array of projects and initiatives. For example, some of my peers established non-profit organizations to advocate for social change, while others founded startup companies to design and launch innovating products. Despite these differences, everyone shared a contagious entrepreneurial spirit and a purpose to create. This insatiable appetite to turn a vision into reality is precisely what fuels innovation and steers society in exciting and new directions. I must admit, entering my first case study class was quite daunting. However, those nerves quickly melted away due to the unique academic environment facilitated by HBS professors. Each professor channeled a compelling discussion, providing students with an arena to voice their opinions without fear. Rather than passively observe, students were propelled to actively join the conversation. It was inspiring to witness how HBS professors conduct and influence a classroom, naturally exuding a desire and purpose to motivate students. Learning under their leadership was an unforgettable experience. Outside the classroom, I had the privilege of meeting HBS alumni and getting a glimpse of their careers. I was amazed to discover the multitude of career paths possible after pursuing an MBA, from politics to the arts. Regardless of their chosen profession, HBS alumni not only pursued their passions, but also utilized them to address pressing issues at the local and national stage. The level of commitment HBS alumni have to their craft is a testament to their drive and purpose to engage as contributing members of society. I aspire to follow in their footsteps to build a lasting legacy to benefit the next generation. HBS starts taking Peek Weekend applications on February 1 with a deadline of April 18. For more information about Peek Weekend 2018 and the application details, register for one of our informative webinars which take place on February 13 and April 3, or join our Peek mailing list.	1
47444	**Generalist Practice BP 8: Community Group Leadership Referring to Brueggeman's article, what are 3 ways that social workers build community in groups through their leadership? Please explain your answers and give 2 concrete examples of how you have seen these elements accomplished in your own group experiences. We're back from Spring Break!** This week has had a slow start, and I'm already looking forward to the summer. The quicker we can get done, the quicker we can get to graduation next Spring, right?? How do social workers build community in groups through their leadership? 1. Gather People Together-as social workers, **especially those in a macro environment,** we have opportunities to gather people together for a multitude of reasons. Mobilizing community groups for a common purpose gives those groups a sense of belonging and power. In my practicum, I bring groups of kids together for several purposes. One is a group of 4th grade girls that meet to talk about bullying. They talk about their experiences and how to be better friends. Another time I'm able to gather a group is in the larger classroom setting. We talk about social skills, and the children learn how to navigate through the school appropriately. 2. Express Feelings-social workers have the unique opportunity to motivate others to talk about their feelings. It is our job to create a safe environment for this to happen. In my personal experience, I have attended 12-step groups where the participants were encouraged but not obligated to share their feelings and experiences. One primary motivating factor was the promise of confidentiality in these groups. As a group leader in a school setting, I use open-ended questions and talk about my own experiences to motivate young group members to disclose their own feelings. This encourages the members and shows them they are not alone, and there is a trusted adult who has been through what they might be going through. 3. Build Confidence-by building trust and showing group members that they are not alone in this world, a social worker can help those group members build confidence. Once confidence is built and maintained, those group members can begin to exact change in their own lives should they choose to do so. With children, confidence-building begins with teaching them a skill set. I led a group of children with anxiety issues, and through a structure curriculum, I was able to teach them tools for coping with their anxiety. Does this work all the time? No. But they feel better knowing they have the	2
18971	Chicago journalist Neil Steinberg said (**of the Bristol Renaissance Faire),** "If theme parks, with their pasteboard main streets, reek of a bland, safe, homogenized, whitebread America, the Renaissance Faire is at the other end of the social spectrum, a whiff of the occult, a flash of danger and a hint of the erotic. Here, they let you throw axes. Here are more beer and bosoms than you'll find in all of Disney World." In a modern Renaissance festival there are stages or performance areas set up for scheduled shows, such as plays in Shakespearean or commedia dell'arte tradition, as well as anachronistic audience participation comedy routines. Other performances include dancers, magicians, musicians, jugglers, and singers. Between the stages the streets ('lanes') are lined with stores ('shoppes') and stalls where independent vendors sell medieval and Renaissance themed handcrafts, clothing, books, and artworks. There are food and beverage vendors, as well as game and ride areas. Games include basic skills events such as archery or axe-throwing as well as Drench-a-Wench and Soak-a-Bloke, which allow a player with a good aim to hit a target and get a fair employee wet. Rides are typically unpowered—various animal rides and human-powered swings are common. Live animal displays and falconry exhibitions are also commonplace. Larger Renaissance fairs will often include a joust as a main attraction. In addition to the staged performances, a major attraction of Renaissance fairs is the crowds of actors - both professional and amateur - who play all sorts of historical figures and roam the fair, interacting with visitors. Visitors are encouraged to wear costumes, once any weapons are suitably peace-bonded, contributing to the illusion of an actual Renaissance environment. Many of the fair vendors sell or rent costumes for all ages and types. The Renaissance fair subculture's word for these costumed guests is "playtrons", a portmanteau of the words "patron" and "player", and they add a second level of enjoyment to their experience by "getting into the act" as Renaissance Lords and ladies, peasants, pirates, belly dancers, or fantasy characters. However, many renaissance fairs discourage interaction between the official cast and so-called "playtrons." Most fairs have an end-of-the-day ritual, a parade or concert where all employees gather	3

Figure 33: Top-3 training documents according to data-influence (EK-FAC) on OLMo2-1B for prompt: '*Could you please hand me...*'.

Prompt: It is generally agreed that exercise is beneficial for …

Response: your health. It improves cardiovascular health, strengthens your muscles, boosts mental well-being and can help prevent chronic diseases like diabetes, hypertension and some forms of cancer. The key to getting the most out of your exercise routine is to do so in moderation and as part of a balanced lifestyle. Aim for at least 150 minutes of moderate aerobic activity or 75 minutes of vigorous activity each week. Here are 5 tips that will help you build a robust exercise program: 1. Choose activities you enjoy: If you're not going to stick to an exercise program, it might as well be the type of exercise you enjoy. Look for activities that can be done frequently and easily. 2. Develop a routine: Building a workout schedule will help you stick with your exercise plan. Consistency is key to seeing results. 3. Mix up your workouts: Regular exercise of different types helps prevent plateaus and keeps your exercise routine exciting. 4. Gradually increase your intensity and duration: A gradual progression to more intense workouts can help your body adapt and avoid injury. 5. Listen to your body: Exercise should be fun and challenging, but not to the point of causing pain. Don't push yourself too hard - remember the motto: 'No pain, no gain'. The primary goal is to maintain a healthy lifestyle, not to push oneself beyond their limits. Always ensure you are taking rest days when needed and that your fitness level is progressing at an appropriate rate for you.

Data-Sim. Rank	**Top-3 Documents for Data-Similarity** (BM25, relevant sentences only)	Data-Inf. Rank
1	**Is It Safe To Exercise During Pregnancy?** Our physician addresses the top concerns of expectant mums who want to maintain an active lifestyle. **Is it safe to exercise during pregnancy?** If you are healthy and your pregnancy is progressing normally, **it is safe to continue or start most types of exercise.** Physical activity does not increase your risk of miscarriage, low birthweight or early delivery. However, **some modifications to certain types of exercises may need to be made – discuss exercise with your obstetrician during your early prenatal visits.** If given the green light, **you can plan an exercise routine that fits your needs.** What are the benefits of exercising during pregnancy? **Regular exercise reduces back pain,** eases constipation and promotes healthy weight gain while pregnant. Staying active may help decrease your risk of gestational diabetes, pre-eclampsia and caesarean delivery. It also improves overall fitness and strengthens the heart and blood vessels, as well as helps you lose weight postpartum. What exercises should I avoid? Some exercises that should be avoided include contact sports that put you at risk of getting hit in the abdomen, as well as risky activities such as downhill snow-skiing, off-road cycling and horseback riding, which may result in a fall. Expectant mothers should also steer away from hot yoga and hot Pilates, which can lead to overheating. Scuba diving and high-altitude activities are also not recommended. When should I stop? During exercise, if you notice any bleeding or fluids leaking from the vagina, or you experience regular, painful contractions of the uterus, you should stop exercising and call your obstetrician immediately. Other symptoms that indicate you should stop and seek medical advice include feeling dizzy or faint, experiencing chest pains, headaches, muscle weakness, and calf pain or swelling. Additionally, **if you are feeling short of breath before starting any exercise or physical activity,** do not continue and call your doctor instead. Walking: Brisk walking provides a total body workout and is easy on the joints and muscles. Swimming and water workouts: If you find walking difficult because of lower back pain, **water exercise is a good way to stay active.** The water supports your weight, which helps you avoid injury and muscle strain. Stationary bicycling: A stationary bike is a safer choice than riding a standard bicycle. This is because your growing belly can affect your balance and make you more prone to falls. Modified yoga and Pilates: Yoga reduces stress, improves flexibility and encourages stretching and focused breathing. Avoid poses that require you to lie on your back for long periods. Look out for prenatal yoga and Pilates classes, which often	1
2	Rafed English Fitness programs: 7 tips for staying motivated Fitness doesn't have to be drudgery. Stay motivated by setting goals and choosing activities you enjoy. Have you ever started a fitness program and then quit? If you answered yes, **you're not alone.** Many people start fitness programs but stop when they get bored or results come too slowly — but it doesn't need to be that way. Here are seven tips to help you stay motivated. 1. Set goals Start with simple goals and then progress to longer range goals. Remember to make your goals realistic and achievable. It's easy to get frustrated and give up if your goals are too ambitious. For example, if you haven't exercised in a while, a short-term goal might be to walk five minutes once or twice a day. An intermediate goal might be to walk 20 minutes three or four times a week. A long-term goal might be to complete a 5K walk. 2. Make it fun Find sports or activities that you enjoy. **then vary the routine to keep you on your toes. If you're not enjoying your workouts,** try something different. Join a volleyball or softball league. Take a ballroom dancing class. Check out a health club or martial arts center. Discover your hidden athletic talent. Remember, **exercise doesn't have to be drudgery — and you're more likely to stick with a fitness program if you're having fun.** 3. **Make physical activity part of your daily routine If it's hard to find time for exercise,** don't fall back on excuses. **Schedule workouts as you would any other important activity.** You can also slip in physical activity throughout the day. Be creative! Take a walk during your child's music lesson. Take the stairs instead of the elevator at work. Pedal a stationary bike or do strength training exercises with resistance tubing while you watch TV at night. 4. Put it on paper Are you hoping to lose weight? **Boost your energy?** Sleep better? Manage a chronic condition? Write it down! **Seeing the benefits of regular exercise on paper may help you stay motivated. It may also help to keep an exercise diary. Record what you did during each exercise session,** how long you exercised and how you felt afterward. **Recording your efforts can help you work toward your goals — and remind you that you're making progress.** 5. Join forces with friends, neighbors or others 6. Reward yourself 7. Be flexible Now that you're enthusiastic again,	24
3	February 26, 2019 May 30, **2019 We all know by know how important exercise is,** for both our physical bodies and our minds. **Exercise has numerous health benefits and is the best way to combat aging.** However, sometimes injuries pop up and can sideline our best intentions. Most injuries are short lived, but some can last months or even years. The best way to deal with a painful injury is to avoid it in the first place! Luckily there are seven specific strategies that can help prevent you from getting injured and keep you exercising for the long term. The following are 7 tips for exercising to prevent injury: 1. Warm up using dynamic movements. Research shows that using dynamic movements to warm up muscle groups is more effective than doing static stretches before a workout. Dynamic movements help to increase blood flow and more effectively warm up the muscle. Static stretches are best when done after a workout is complete to help cool the muscles down and stretch out tight muscle fibers. Dynamic movements should include a version of whatever activity you are about to do to warm that muscle group up properly. 2. Stretch after your work out. Stretch out the muscle groups that you worked out. Hold the stretches for at least 60 seconds, focusing on slow and deep diaphragm breaths. 3. Have a plan for building strength. **Regardless of what type of exercise you are doing,** strength training should be incorporated to help build muscle mass and increase resilience. Strength training also helps to prevent bone density loss and will help increase performance. However, a plan for building strength is important so you don't do too much too soon. 4. Listen to your body and address issues before they become a big deal. Prevention of an injury is so much easier than recovering from an injury! **Your body is constantly sending feedback,** but it's up to you if you are listening and responding accordingly. **Regular chiropractic adjustments help to keep your joints moving and your range of motion optimal.** Additionally, regular massages help keep muscles from getting too tight and strained. Lastly, incorporating mobility work into your regular routine can help improve flexibility and address mobility deficiencies, which are common causes of injury. 5. Do a variety of exercises, aka cross-train. Cross-training is important in preventing injury because it allows you to continue to build greater fitness without adding additional strain and repetition to overworked muscles. **Cross- training is a broad term and is basically any type of exercise that is different than the type you are already doing.** For example, if you are training for a marathon, try	2

Figure 34: Top-3 training documents according to data-similarity (BM25) on OLMo2-1B for prompt: '*It is generally agreed that exercise is beneficial for...*'.



Figure 35: Top-3 training documents according to data-influence (EK-FAC) on OLMo2-1B for prompt: '*It is generally agreed that exercise is beneficial for...*'.