An Information-theoretic Approach to Prompt Engineering Without Ground Truth Labels

Taylor Sorensen*, **Joshua Robinson***, **Christopher Michael Rytting***,
Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil,
Nancy Fulda, David Wingate
Computer Science Department, Brigham Young University
{tsor13,joshua_robinson,chrisrytting}@byu.edu
{nfulda,wingated}@cs.byu.edu

Abstract

Pre-trained language models derive substantial linguistic and factual knowledge from the massive corpora on which they are trained, and prompt engineering seeks to align these models to specific tasks. Unfortunately, existing prompt engineering methods require significant amounts of labeled data, access to model parameters, or both. We introduce a new method for selecting prompt templates *without labeled examples* and *without direct access to the model*. Specifically, over a set of candidate templates, we choose the template that maximizes the mutual information between the input and the corresponding model output. Across 8 datasets representing 7 distinct NLP tasks, we show that when a template has high mutual information, it also has high accuracy on the task. On the largest model, selecting prompts with our method gets 90% of the way from the average prompt accuracy to the best prompt accuracy and requires no ground truth labels.

1 Introduction

It is well-known that large pre-trained language models (LMs) learn substantial linguistic (Liu et al., 2019; Amrami and Goldberg, 2018) and factual world knowledge (Petroni et al., 2020; Bosselut et al.; Bouraoui et al.; Zuo et al., 2018), achieving state-of-the-art performance on classic NLP tasks like closed-book question-answering, sentiment analysis, and many other tasks (Radford et al., 2019; Devlin et al., 2019; Raffel et al., 2019). The largest models can do this in a few-shot way–that is, being trained only with generic, semi-supervised objectives and "taught" tasks with just instructions and a few examples of the task provided via a natural language "prompt" in the context window (Brown et al., 2020). This suggests that pre-training equips them to potentially do many tasks that can be formulated as natural language generation, if only they can be primed in the right way.

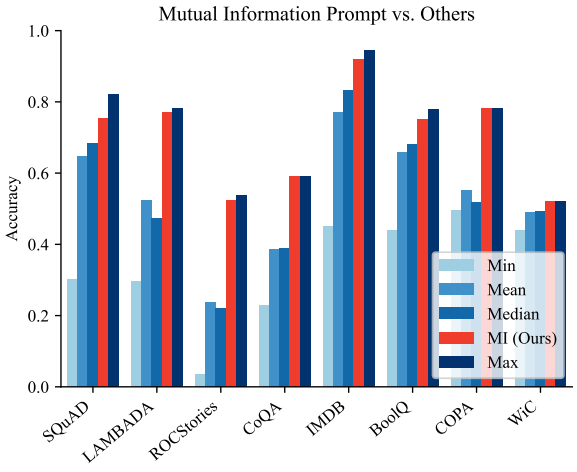

Figure 1: Performance of template selected by our maximum mutual information method (MI) compared to the the worst, mean, median, and best prompt on GPT-3 Davinci (175B). Our method performs at almost oracle levels, without labels or access to model weights.

Such priming is not a trivial task. The few-shot learning breakthrough can give the impression that if the LM is given a sensible prompt, it will "understand" what is meant and perform well on the task if it has the capacity. However, LMs can generate substantially different output distributions–and thus text–given two distinct prompts that appear semantically invariant (e.g., alternative orderings, lexical changes like capitalization, and general rephrasing (Zhao et al., 2021; Lu et al., 2021)). This can lead to surprisingly high variance in performance from prompt to prompt. Clearly, some prompts are better than others for aligning a model to a task.

Prompt engineering is a nascent field that aims to find aligning prompts (Reynolds and McDonell, 2021). While "prompt" refers to any language passed to the model via the context window, a *template* refers to a natural language scaffolding filled in with raw data, resulting in a prompt. Thus, prompt engineering includes finding high-quality templates (i.e., those with high test accuracy). Generally, this is done by optimizing for accuracy over

*Equal Contribution

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 819 - 862

a validation set: a template is chosen from a candidate set based on its performance on labeled examples. Such labeled examples can be challenging to procure for some tasks and impossible for others. Some recent methods optimize prompts using backpropagation, which requires access to model weights. In this paper, we propose a new method for selecting prompts by using mutual information, which allows prediction of a prompt's performance without labels or access to model parameters.

Mutual information (MI) is a metric that quantifies the shared information between two random variables (see Section 3.2). We demonstrate that the mutual information between a prompt and a language model's output can serve as a useful surrogate for the test accuracy of a template. Specifically, for eight popular datasets representing seven classic NLP tasks, we generate a diverse set of 20 templates for each and show that template mutual information and template accuracy are highly correlated. These results are strongest on the largest models we study, for which our method chooses prompts that, on average, get 90% of the way from mean accuracy to maximum accuracy and even selects the best prompt on three of eight datasets.

This suggests that, across a variety of NLP tasks, mutual information can be used to select one of the best prompts from a set of candidate prompts, even without making use of model weights or ground truth labels. In the following pages, we outline each step of our general method for generating and evaluating templates so that it can easily be ported to any other task. Code is available online.[1]

2 Related Work

The promise of language models and the challenge of aligning them has given rise to the field of prompt engineering, which seeks to construct the best prompt given a task and a language model (Liu et al., 2021a). The best performance on prompt engineering is often achieved using backpropagation in continuous prompt embedding space (Lester et al., 2021; Li and Liang, 2021; Gu et al., 2021; Liu et al., 2021b; Zhang et al., 2021) in contrast to generating a discrete set of prompts by hand and testing them. While optimizing in continuous prompt space via backprop allows for similar performance to model-tuning (at least at higher model sizes) (Lester et al., 2021), not all models are publicly available. Thus, these methods are only feasible for those who have direct access to the model and can perform backprop on it. Prompts optimized in continuous space are also not interpretable in natural language, making it harder to transfer insights from prompts that work well for one task to another task. Additionally, these methods require labeled examples, while ours does not.

Other selection protocols not based on gradient descent can include cross-validation or minimum description length, as in (Perez et al., 2021). These methods yield prompts that perform marginally better than average in terms of test accuracy.

Mutual information has been used in n-gram clustering, part-of-speech tagging, probing classifiers, and LM training objective reframing (Brown et al., 1992; Stratos, 2019; Voita and Titov, 2020; Kong et al., 2019). Ours is the first work of which we are aware to apply MI to prompt engineering. (Lu et al., 2021) make use of entropy statistics to determine performant orderings for few-shot examples in prompts. Our work is focused on selecting high quality templates with no special focus on example ordering or need for multiple examples to order (the few-shot case). Our method uses no artificial "probing set," making our prompt selection much cheaper, and we also explore open-ended tasks. While the GlobalE and LocalE statistics they use are similar (and in the case of LocalE identical) to the two parts of our MI calculation (see 3.2), we use the two statistics jointly and choose prompts by minimizing, rather than maximizing, LocalE.

3 Methods

At the most abstract, our method is as follows (see Appendix A for a more thorough description):

1. Generate a set of K prompt templatizing functions.
2. Playground a couple of examples to ensure that templates give roughly expected output.
3. Estimate mutual information for each template given a set of inputs $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$ where $\mathbf{x}_i \sim X, \forall i$.
4. Choose template(s) based on mutual information and perform inference.

We find it useful to unify all the tasks we study within a single framework, which we describe in

[1] github.com/BYU-PCCL/information-theoretic-prompts

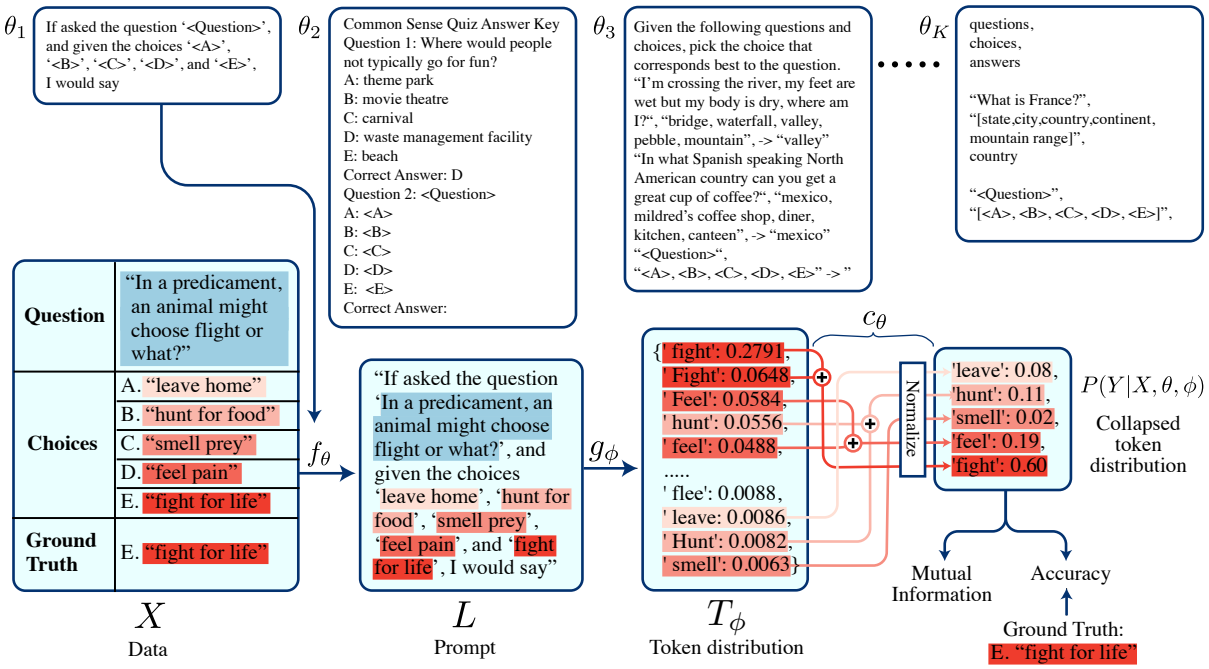

Figure 2: We choose $\theta \in \{\theta_i\}_{i=1}^{K}$ and templatize a sampled instance from the dataset X. We pass this prompt through the language model via g_ϕ, yielding a probability distribution over the model's tokens T_ϕ. The collapsing function c_θ sums the weight given to each token corresponding to each possible answer $y \in Y$ and normalizes, giving a probability distribution $P(Y|\mathbf{x}_i)$, which we can use to estimate mutual information or obtain a guess for y_i.

Section 3.1. We also justify our use of mutual information as a surrogate for prompt quality and specify how we estimate it in Section 3.2.

3.1 Task Definition

In order to demonstrate our method's widespread applicability and general effectiveness, we validate it across many datasets and tasks. This requires us to estimate MI and accuracy, and this is most straightforward in the case where, given a context, a language model produces just one probability distribution $P(\mathbf{t}_n|\text{context} = \mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_{n-1})$. This is in contrast to other experimental setups that use multi-token sampling methods (e.g., beam search), although our method is easily tractable in such setups.[2] Any NLP task is tractable in this framework so long as the output space consists of a set of options that each start with a unique token. In this case, the language model can "give" an answer by assigning probability to tokens that begin giving each of these answers (invariant to lexical variation like capitalization and leading/trailing spaces). While, for open-ended tasks, this method might artificially inflate accuracy if the model starts to give a wrong answer that happens to start with the same token as the correct one, we find that this difference is small and does not affect our results.[3] Irrelevant tokens (with which none of the desired answers begin) are ignored, and the resulting collapsed probabilities are normalized. We term this approach *One-token Response* (OTR). Although our method isn't limited to OTR tasks, we choose tasks that can be cast as OTR tasks for simplicity and to reduce computational expense. Many NLP tasks fit within this framework, although a few do not (e.g., machine translation and summarization). This basic approach is in common use (Brown et al., 2020), but we formalize it for clarity below.

Generally, the OTR framework casts a natural language task as a classification problem with raw data input $\mathbf{x}_i \in X$ and output $P(Y|\mathbf{x}_i)$, a probability distribution over targets. In order to use a language model ϕ for this task, a templatizing function $f_\theta : X \to L$ is needed to map raw data

[2]The only difference: For each considered answer, simply calculate its unnormalized probability by multiplying the probabilities of the decisions taken at each branch in the sequence of tokens, then normalize the resulting probability scores.

[3]Our open-ended datasets are SQuAD, LAMBADA, and ROCStories, and none of these seemed more likely than ROCStories to exhibit this issue. We reran our experiment on ROCStories by sampling with temperature 0 until reaching a space, and only counted responses as accurate if they exactly matched the corresponding ground truth labels. Results were virtually unchanged: accuracy decreased by only 0.03 on average, and the correlation between mutual information and test accuracy increased by 0.04, from 0.68 to 0.72.

into natural language prompts. $g_\phi : L \to T_\phi$ maps prompts to a probability distribution over T_ϕ, the token set represented by the model tokenizer. Finally, a collapsing function $c_\theta : T_\phi \to P(Y|\mathbf{x}, \theta, \phi)$ (see Appendix A) yields an estimate of $P(Y|X)$:

$$P(Y|\mathbf{x}, \theta, \phi) = c_\theta(g_\phi(f_\theta(\mathbf{x}))), \mathbf{x} \in X \quad (1)$$

We also refer to $P(Y|\mathbf{x}, \theta, \phi)$ as $P(Y|f_\theta(\mathbf{x}))$.

The above pipeline can be specified in many ways using different θ and ϕ (see Figure 2), which will result in different accuracies. Our ultimate aim is to select the best θ given ϕ. Whereas past prompt engineering methods rely on scores calculated by comparing model answers and ground truth, our method selects θ by maximizing mutual information, which requires no ground truth labels.

3.2 Mutual Information

Mutual information is a measure of the amount of shared information between two random variables (Cover and Thomas, 2006); in other words, it is the reduction in entropy that is observed in one random variable when the other random variable is known.

We expect MI to serve as a good criterion for comparing prompts. Previous work has shown that large networks trained with cross-entropy loss are calibrated (e.g., a 60% confidence corresponds to a 60% chance of the model being correct) when in the early-stopped ($\sim$ 1 epoch) regime (Ji et al., 2021), but become miscalibrated in the overfit regime (Nakkiran and Bansal, 2020). According to (Brown et al., 2020), GPT-3 was trained for a different number of epochs on each corpus in its training data. We calculate it was trained for an average of 1.57 epochs, so we have reason to believe that GPT-3 is generally well-calibrated. Thus, we postulate that a prompt that elicits a very confident response (high MI) from the language model is more likely than a less confident prompt to score well.

We denote the mutual information between random variables X and Y as $I(X;Y)$ and the entropy of X as $H(X) = -\int_{\mathbf{x}\in X} P(\mathbf{x}) \log(P(\mathbf{x}))d\mathbf{x}$. The mutual information between X and Y is defined as $D_{\text{KL}}(P_{(X,Y)}||P_X \otimes P_Y)$, and can be rewritten as $H(Y) - H(Y|X)$ (the reduction in entropy in Y given knowledge of X).

Using the OTR framework, we fix a model ϕ and generate a diverse set of K prompt templatizing functions $f_{\theta_1}, f_{\theta_2}, ..., f_{\theta_K}$ along with their corresponding collapsing functions c_{θ_k} (see Appendix A). Treating $f_\theta(X) := \{f_\theta(\mathbf{x}), \mathbf{x} \in X\}$ as a random variable, we can calculate $I(f_\theta(X); Y)$ and use it as a criterion for selecting prompt templatizing functions with which to do inference.

We hypothesize that a θ_i with higher mutual information will align a language model to a task better than a θ_j with lower mutual information. Formally, we select $\hat{\theta} = \text{argmax}_\theta\{I(f_\theta(X); Y)\}$.

Mutual information is estimated as:

$$I\left(f_\theta(X); Y\right) = H(Y) - H(Y|f_\theta(X)) \quad (2)$$

where each term is estimated in expectation using draws $\mathbf{x}_i \sim X$ and Equation 1 as follows:

$$H(Y) \approx H\left(\frac{1}{N}\sum_{i=1}^{N} P(Y|f_\theta(\mathbf{x}_i))\right) \quad (3)$$

$$H(Y|f_\theta(X)) \approx \frac{1}{N}\sum_{i=1}^{N} H(P(Y|f_\theta(\mathbf{x}_i))) \quad (4)$$

The marginal entropy $H(Y)$ is the entropy of the mean of the conditional distributions, and the conditional entropy $H(Y|f_\theta(X))$ is the mean of entropies of the individual conditional distributions.

This definition gives us another reason to expect that mutual information will work well. Since mutual information is the marginal entropy minus the conditional entropy, maximizing mutual information is equivalent to maximizing marginal entropy and minimizing conditional entropy. Thus, MI is high for templates that are, on average, less biased towards any given answer (high marginal entropy) and templates with outputs the model is confident about (low conditional entropy). These attributes are desirable in constructing prompts, and we postulate that maximizing mutual information will yield a well-aligned template.

Looking at it another way, by the data processing inequality (Cover and Thomas, 2006), $I(f_\theta(X); Y) \leq I(X; Y)$. Thus, $I(f_\theta(X); Y)$ gives a lower bound for $I(X; Y)$, and the highest mutual information is the tightest lower bound. The prompt corresponding to this lower bound preserves the most information between X and Y.

4 Experimental Setup

4.1 Datasets

We validate the efficacy of our prompt engineering method with experiments on eight well-known NLP datasets[4]–SQuAD2.0 (Rajpurkar et al., 2018), LAMBADA (Paperno et al., 2016), ROCStories

[4]Datasets are listed in descending order here and throughout the paper, first by $|Y|$, and then by method performance.

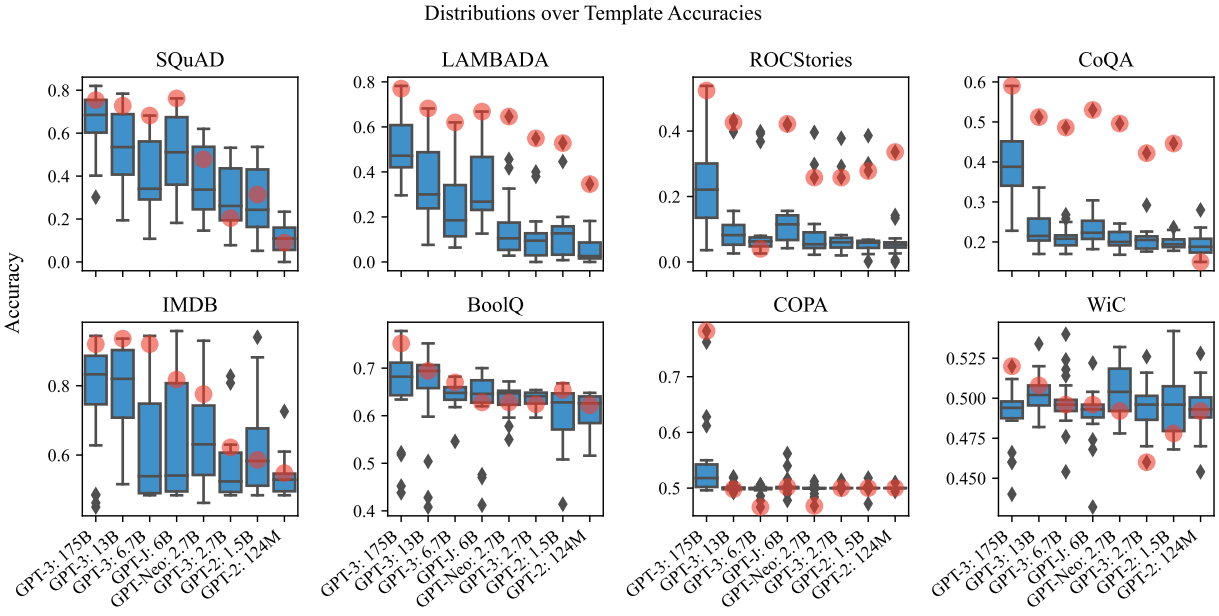

Figure 3: Distributions of accuracies over $K = 20$ templates for each model/dataset pair, compared to the prompts selected with MI (translucent red dots).

Dataset	Task	$\lvert Y \rvert$	Base Acc.	Size N_{all}
SQuAD	Open Book QA	$\lvert T_\phi \rvert$	~ 0	16K
LAMBADA	Cloze	$\lvert T_\phi \rvert$	~ 0	5K
ROCStories	Cloze	$\lvert T_\phi \rvert$	~ 0	52K
CoQA	Closed Book QA	5	0.2	9K
IMDB	Sentiment Analysis	2	0.5	50K
BoolQ	Reading Comprehension	2	0.5	16K
COPA	Choice of Positive Alternatives	2	0.5	1K
WiC	Word in Context	2	0.5	5K

Table 1: All datasets used in our experiments. $|Y|$ is the size of the label space and N_{all} is the size of the dataset we sample from (after any modifications).

(Mostafazadeh et al., 2016), CommonsenseQA (CoQA) (Talmor et al., 2018), IMDB (Maas et al., 2011), BoolQ (Clark et al., 2019), COPA (Gordon et al., 2012), and WiC (Pilehvar and Camacho-Collados, 2018))–that span seven unique NLP tasks (see Table 1). We used a random sample of $N = 500$ samples from each dataset for our experiments.[5] For ROCStories, which consists of a set of five sentence stories, we randomly masked a word from each story in order to use the data for masked word prediction (cloze).

We made minor changes to two of the datasets in order to cast the associated tasks into OTR. For the SQuAD dataset, we dropped all questions that did not have a one word answer. For the CoQA dataset we dropped all questions with answer choices that started with a shared first word (e.g, the dog, the cat, the monkey). Both changes were to decrease ambiguity about which option the model was choosing given its output distribution for a single token.

[5]We sampled from the train sets of CoQA and SQuAD; the train and validation sets of WIC, COPA, and BoolQ; the full datasets of ROCStories and IMDB; and the test set for LAMBADA.

4.2 Models

We assess our method on eight models ranging from 124 million to 175 billion parameters : These include GPT-2 124M & 1.5B (Radford et al., 2019), GPT-Neo 2.7B (Black et al., 2021), GPT-J (6B) (Wang and Komatsuzaki, 2021), and (Ada, Babbage, Curie, & Davinci) GPT-3 (Brown et al., 2020). We assume (per (Perez et al., 2021)) these models to correspond, respectively, to the 2.7B, 6.7B, 13B, and 175B models in (Brown et al., 2020). Each is a causal language model, and although we do not include masked language models, this is a promising area for future work.

5 Results

In this section, we analyze our experiments. First, we look at our method's ability to select high-accuracy prompts across models and datasets (Section 5.1). Next, we correlate template mutual information and accuracy in Section 5.2. After that, we compare our method and template selection using labeled examples in Section 5.3. In Section 5.4, we

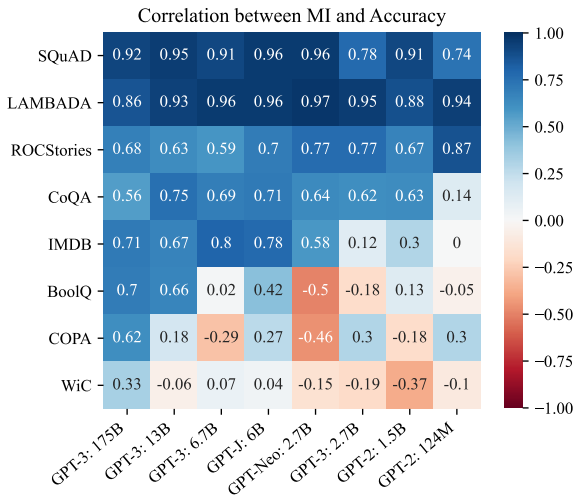

Figure 4: Correlations are more consistently high across all tasks for the largest models, suggesting that our method is most useful at those model sizes.

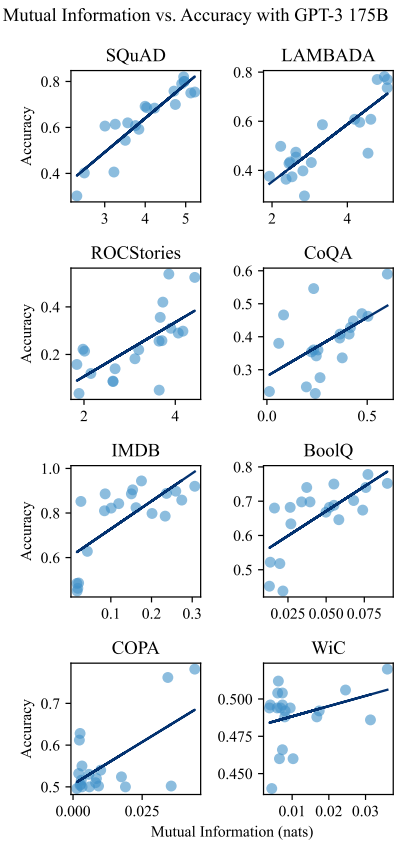

Figure 5: Each dot represents a template and its average mutual information and accuracy over $N = 500$ task instances. Linear best fit (by mean standard error) lines are included to show overall trends.

explore the robustness of MI and use ensembling to improve it. Finally, we compare the tranferability of prompt templates selected with MI from model to model in Section 5.5.

5.1 Template Selection Performance

We first define baselines against which we compare our approach. Other prompt engineering methods generally require either access to model weights, labeled data (validation set selection), or both (backprop/continuous prompt embedding methods). Our method does not require these, so we instead compare to random and oracle baselines. A random template selection method would give us the average accuracy of our template set (in expectation), while an oracle selection method would give us the best accuracy every time. To understand how our MI method compares to these two baselines for each dataset, refer to Figure 1, where we analyze performance on GPT-3 175B. On each of the eight datasets, mutual information selects a prompt template that outperforms both the mean and median accuracies (random baseline performance). In three of the eight datasets, mutual information selects the best (highest accuracy) template from the 20 proposed (equivalent to oracle performance).

Given our method's promising performance with GPT-3 175B, it is natural to ask how it performs with smaller models. Figure 3 shows the accuracy distributions over prompt templates for each dataset/model pair. With every model, MI gives above-average performance on several datasets. Although MI is more likely to select a high accuracy template for larger models, it is a good criterion even for smaller models on all but two datasets, COPA and WiC. Note that, for these two datasets, none of the templates do significantly better than chance (∼50%) besides the largest model on COPA, which is in line with previous work.[6] Thus, we observe that mutual information performs best when there is a high-signal prompt to select, and worse when all prompts are low-signal.

When considering all other datasets, MI selects an above average prompt 83% of the time for all models; for the largest two models, MI selects an above average template 100% of the time.

5.2 Correlation between Template Mutual Information and Accuracy

In Section 5.1, we see how the mutual information selected template does in terms of accuracy compared to all other templates. We have not dis-

[6] Our template's best accuracy is 54% for WiC, and 78.2% for COPA, which is similar to previous work (WiC: (Brown et al., 2020) - 49.4%, (Perez et al., 2021) - 54.1%; COPA: (Brown et al., 2020) - 92.0%, (Perez et al., 2021) - 84.8%).

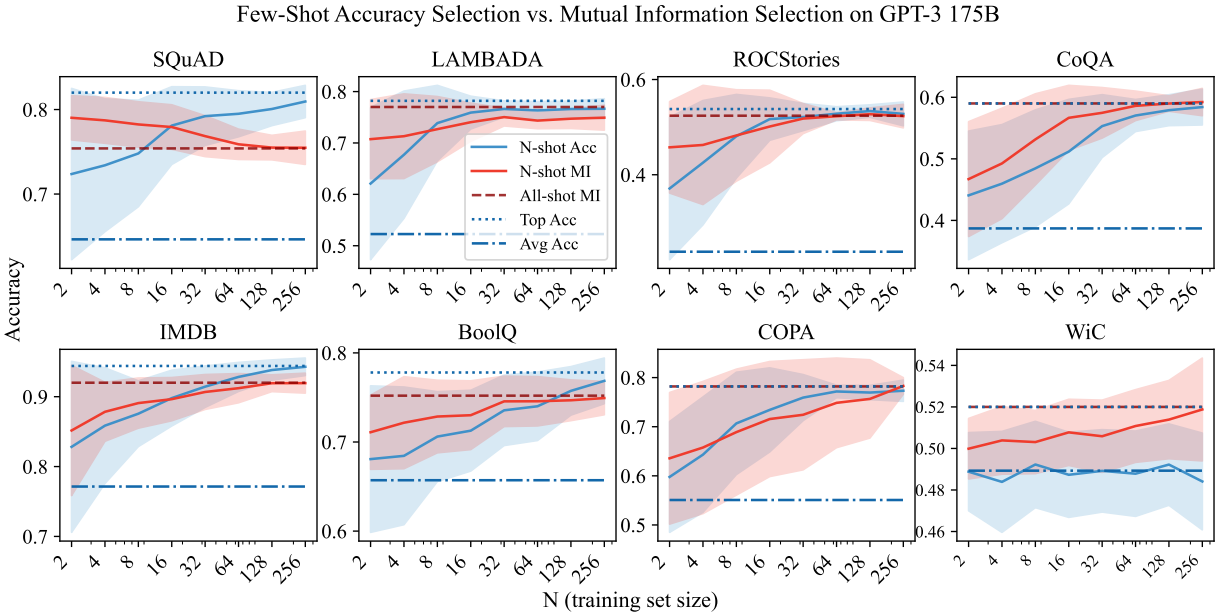

Figure 6: For $P = 100$ random train/test set partitions for each training size $N = 2, 4, 8, ..., 256$, we select a template based on accuracy (N-shot Acc) and based on mutual information based on just those N examples (N-shot MI). Then, we report accuracy of that template on the test set (size: $500 - N$). Error bars ($\pm\sigma$) are reported across the $P = 100$ partitions. For reference, the highest, average, and full-dataset MI template accuracy is also reported.

cussed, however, how generally MI and accuracy are correlated, except that the highest MI template tends to have anomalously high accuracy. Here, we establish that their correlation is high across all templates for the largest LMs. Each of the $K = 20$ templates has two corresponding measures: average accuracy and average MI. We can use these pairs to correlate MI and accuracy via Pearson's R.

We see in Figure 4 that the correlations are surprisingly high for the majority of models and datasets. For SQuAD, LAMBADA, ROCStories, and CoQA, this pattern holds across all model sizes; for the remainder, results are good on larger models and are much less reliable on smaller models. Overall, this is evidence that as mutual information increases, so does accuracy. In other words, mutual information can be used to make an educated guess about accuracy without having to use any ground truth labels, especially on larger models.

5.3 Compared to Few Labeled Examples

Next, we ask: How does our method compare to selecting a template based on the accuracy of a few-labeled examples? Also, how many unlabeled examples does MI need to be able to perform well?

Results with the largest model are reported in Figure 6. Note that with as few as $N = 2$ instances, MI selects a far better than average template, allowing performance gains even in the low-data, unlabeled regime. Additionally, for low N and across all eight datasets, MI even selects a better template on average than selecting based on labeled train set accuracy. This suggests that, even with labeled examples, selecting based off of MI may be preferable to test accuracy with few examples. Selecting by labeled train set accuracy often begins to perform better at higher N, but at the cost of requiring labeled data, while our method needs no labels.

5.4 Method Robustness and Ensembling

To explore our method's robustness we consider the question: what if we had included a different subset of templates, especially not including the top MI template? Figure 5 shows average MI/accuracy data for all $K = 20$ prompt templates on GPT-3 175B (similar plots for other models are found in Appendix B.1). For six of eight datasets, the results are robust; the top few prompt templates (by MI) are all high performers. The performance for COPA and WiC is more brittle; excluding the top-MI template would have resulted in a large drop in accuracy. This attests to the utility of generating a diverse slate of templates as recommended in Appendix A and also to the risk that outliers could compromise our method's effectiveness.

A comprehensive discussion of remedies for outliers is beyond the scope of this paper, but it is

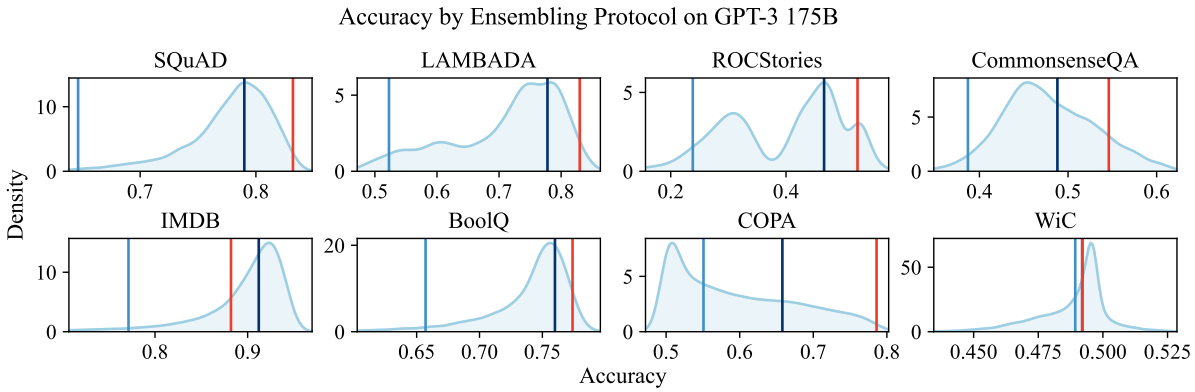

Figure 7: For each dataset the KDE plot represents accuracy over each of the $\binom{20}{5}$ ensembles of 5 templates from the 20 templates associated with the dataset. Each plot also includes lines representing the average accuracy of all single templates for the dataset, the accuracy of the ensemble of all 20 templates, and the accuracy of the ensemble of the top 5 templates chosen by MI. In only one case does all-20 beat top-5-MI, and it does so at 4× the cost.

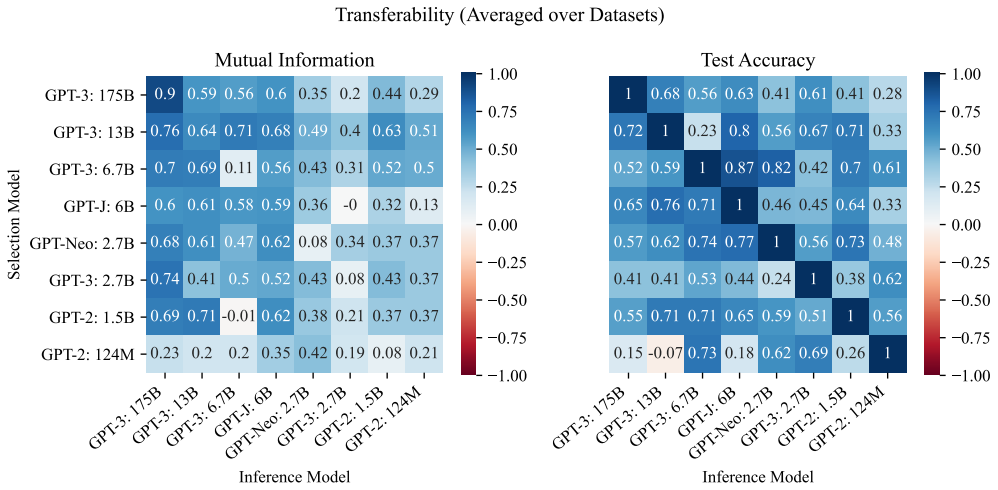

Figure 8: For each model/dataset pair, accuracies are normalized linearly so that 0 is the average prompt accuracy and 1 is the highest test accuracy. Using the prompt chosen by either MI or test accuracy on each selection model, average performance across datasets is reported for each inference model.

an important concern. Considering the strength of MI/accuracy correlations, one simple approach is to ensemble the top 5 MI templates.

To compare this principled top-5 ensemble to other possible ensembles of templates, we take all $\binom{20}{5}$ subsets of 5 templates from all 20 templates and calculate the accuracy of each ensemble. For each dataset, we plot this distribution's kernel density estimate, which models the p.d.e. of the random variable "accuracy of 5 random templates ensembled together". We then compare the top-5 MI ensemble to other possible ensembles. The results are shown in Figure 7.

We found that the top-5 MI ensemble does at least as well as the top-20 ensemble in all but one case. Two reasons to use MI are, then, that 1) the MI ensemble gets as good or better a result as ensembling all prompt templates and 2) at a fourth of the experimental cost. In short, ensembling by MI is a cheap and effective way to guard against anomalous high MI/low accuracy templates.

5.5 Transferability across Models

Finally, we explore how well-chosen templates generalize between models. Concretely, we choose templates by maximizing either test accuracy (oracle) or mutual information (our method) using a selection model ϕ_s, and then calculate test accuracy using a different inference model ϕ_i. We calculate absolute test accuracy and then normalize it such that 0 and 100 correspond to the average and maximum scores across templates for a model/dataset pair. We average our results across datasets and present the results in Figure 8. Prompt transfer for

each dataset can be found in Appendix B.2.

MI performance is best when the largest model (GPT-3 175B) is used as both the selection and inference model: on average, MI scores 90% on this normalized scale. Additionally, performance is most consistently high when the largest models are used either for selection or inference. But almost all transfer scores are well above 0 (only one negative average gain out of 64 transfer permutations), suggesting that transfer is often effective.

Overall, we have observed that prompt selection by mutual information is surprisingly effective across a variety of datasets and model sizes. This method works best on larger models and for tasks that the LM is capable of performing. Given the high diversity of tasks that we have explored, we expect this method to transfer well to many other NLP tasks, including regimes with little labeled data.

6 Conclusion

In this paper, we introduce a method for selecting prompts that effectively align language models to NLP tasks. Over a set of candidate prompts, our method selects the template that maximizes the mutual information between the input and the model output. We demonstrate that 1) mutual information is highly correlated with test accuracy and 2) selecting a prompt based on mutual information leads to significant accuracy gains over random choice, approaching oracle performance on GPT-3 175B, and it does so across model sizes and tasks.

Whereas other methods rely on ground truth labels and/or direct model access, ours requires neither. Many applications characterized by lack of computational resources, limited model access (e.g., inference only), and lack of ground truth data prohibiting testing of candidate prompts become feasible with our method.

7 Ethics

There are many ways to prompt a language model poorly, and there still seem to be NLP tasks which are beyond alignment regardless of model size or prompt quality. This method cannot align a LM to a task if the entire set of prompts is poor or, obviously, if the model cannot be aligned. High mutual information does not necessarily imply high accuracy despite the strong correlation we found. Thus, our method should only be employed on a task if there is some understanding of how high MI needs to be on a domain or set of templates to imply a sufficiently high accuracy for safe use.

Otherwise, we introduce no model, dataset, or other contribution that might warrant ethical concern.

Acknowledgements

We thank the anonymous reviewers for their helpful feedback. This material is based upon work supported by the National Science Foundation under Grant No. RI 2141680.

References

Asaf Amrami and Yoav Goldberg. 2018. Word Sense Induction with Neural biLM and Symmetric Patterns. pages 4860–4867.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET : Commonsense Transformers for Automatic Knowledge Graph Construction.

Zied Bouraoui, Jose Camacho-collados, and Steven Schockaert. Inducing Relational Knowledge from BERT.

Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. 1992. Class-based *n*-gram models of natural language. *Computational Linguistics*, 18(4):467–480.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *arXiv*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *CoRR*, abs/1905.10044.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm):4171–4186.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2021. PPT: Pre-trained Prompt Tuning for Few-shot Learning.

Ziwei Ji, Justin D. Li, and Matus Telgarsky. 2021. Early-stopped neural networks are consistent.

Lingpeng Kong, Cyprien de Masson d'Autume, Wang Ling, Lei Yu, Zihang Dai, and Dani Yogatama. 2019. A mutual information maximization perspective of language representation learning.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning.

Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. pages 4582–4597.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:1073–1094.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021a. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. pages 1–46.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. GPT Understands, Too.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James F. Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *CoRR*, abs/1604.01696.

Preetum Nakkiran and Yamini Bansal. 2020. Distributional generalization: A new kind of generalization. *CoRR*, abs/2009.08092.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and

Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. *CoRR*, abs/1606.06031.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True Few-Shot Learning with Language Models. (Cv):1–21.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. Language models as knowledge bases? *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 2463–2473.

Mohammad Taher Pilehvar and José Camacho-Collados. 2018. Wic: 10, 000 example pairs for evaluating context-sensitive representations. *CoRR*, abs/1808.09121.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. pages 1–53.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm.

Karl Stratos. 2019. Mutual information maximization for simple and accurate part-of-speech induction.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *CoRR*, abs/1811.00937.

Elena Voita and Ivan Titov. 2020. Information-Theoretic Probing with Minimum Description Length. pages 183–196.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Ningyu Zhang, Luoqiu Li, Xiang Chen, Shumin Deng, Zhen Bi, Chuanqi Tan, Fei Huang, and Huajun Chen. 2021. Differentiable Prompt Makes Pre-trained Language Models Better Few-shot Learners. pages 1–18.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-Shot Performance of Language Models.

Yukun Zuo, Quan Fang, Shengsheng Qian, Xiaorui Zhang, and Changsheng Xu. 2018. Representation Learning of Knowledge Graphs with Entity Attributes and Multimedia Descriptions. *2018 IEEE 4th International Conference on Multimedia Big Data, BigMM 2018*, pages 2659–2665.

A Prompt Engineering Process

In this section, we step through our method in detail. Again, note that this method uses no ground truth labels and does not require gradient updates or model parameter access. Given a task that can be represented in natural language with the OTR framework, the only requirements for our approach are a) several candidate prompt templates and b) some instances (X) on which to do inference.

1. **Generate a set of K prompt templatizing functions with corresponding collapsing functions.** Each prompt template function f_{θ_k} should take in an input from the dataset and output a prompt ready for processing by the language model. We chose to generate our template functions by hand, i.e., a human writes a sensible, custom natural language scaffolding that can be filled with input data (see examples in C).

Each template must also have a collapsing function c_{θ_k} that takes the language model output logprobs, exponentiates and sums "equivalent" logprobs, and normalizes the resulting probabilities to produce a distribution over targets. Equivalent logprobs are those that indicate the same answer. For example, a template might be designed for a question-answering task with possible answers "Yes" and "No". We consider all logits corresponding to possible lexical variants of each of these answers to be equivalent. For example, what logits should count toward the answer "Yes"? Not just the exact token "Yes", since "ye", " yes", and "YES" are all lexical variants of the same answer or the beginning of it, just with surrounding white space and alternative capitalization. The collapsing function lower-cases and strips white space from all logits, and if the lower-cased answer begins with a token, that token's probability (the exponentiated logprob) is added to the sum of probability for that answer. Finally, the sums of probabilities for all individual answers are normalized. Prompt template functions should be chosen to be as diverse as possible to increase the probability of finding high-quality prompts. For example, we use templates that frame input from datasets as test questions, back and forth dialogue between friends, Python code, test answer banks, etc. A sample of the prompt templates used in this work is provided in Appendix C. A good resource for coming up with prompt template function ideas is the OpenAI API examples collection[7].

[7] beta.openai.com/examples

While we aimed for as diverse a set of prompts as possible in this work, additional dimensions of variation in prompt templates could be explored in future work (e.g., ordering of few-shot examples).

2. **Playground.** For each chosen f_{θ_k}, calculate $g_\phi(f_{\theta_k}(x))$ for a few dataset samples. Do not look at associated ground truth labels for these samples. Simply check to ensure that g_ϕ puts high probability on the tokens one would expect given f_{θ_k} that could be reasonably collapsed by c_{θ_k} into $P(Y)$. For example, on the BoolQ reading comprehension task, the language model predicts the answer to a yes/no question with a corresponding passage. Given this task, we would expect the highest probability to be on tokens like "Yes" or "No". A poor prompt template, though, might put the highest probability on unrelated tokens like "I", "think", or "\n". Revise or replace any template that fails to put high probability mass on the tokens expected.

3. **Estimate mutual information for each template f_{θ_k}.** Choose how many data points N to use for estimating mutual information for each template function. A higher N will allow for estimation of mutual information based on a more representative sample of the dataset at the cost of more LM computation. Sample N samples from your dataset. Since we do not require any Y labels, one could even choose the X's on which you desire to do inference (as we do). Then, for each sample x and each template f_{θ_k}, calculate $P(Y|f_\theta(x))$ using Equation 1. Use the output to estimate MI for each prompt template with Equation 2.

For all of our experiments, c_θ takes in a distribution of tokens $g_\phi(f_{\theta_k}(x))$ and a mapping between the set of possible ground truth labels for $f_{\theta_k}(x)$ and model vocabulary T_ϕ. For a sentiment analysis task, that mapping would be from the ground truth labels "positive" and "negative" to the expected tokens "positive" and "negative" respectively. If a prompt template for the task was phrased as a yes/no question, the mapping for it would be from "positive" and "negative" to "yes" and "no" respectively. Our c function returns a probability over Y (target label space), and the highest probability label is treated as the prediction. To keep things simple, the values in our map are always single tokens. See examples in Appendix C.

4. **Choose prompt template(s) to use for inference based on mutual information.** For choosing

a single prompt template to use for inference, select the template with highest estimated mutual information. With an increased computational budget, one could also ensemble the top p prompt templates, as we describe in Section 5.4.

5. **Use chosen prompt template(s) to perform inference** Use chosen prompt template(s) $f_{\hat{\theta}}$ to calculate $c_{\hat{\theta}}(g_{\phi}(f_{\hat{\theta}}(x))$ for each dataset sample. Inference can be done with the language model used for estimating mutual information or a smaller model if cost is prohibitive (for information on performance statistics with this approach, see Figure 8).

B Additional Figures

B.1 Mutual Information vs. Accuracy

See Figure 9.

B.2 Per Dataset Transfer Heatmaps

See Figures 10-17.

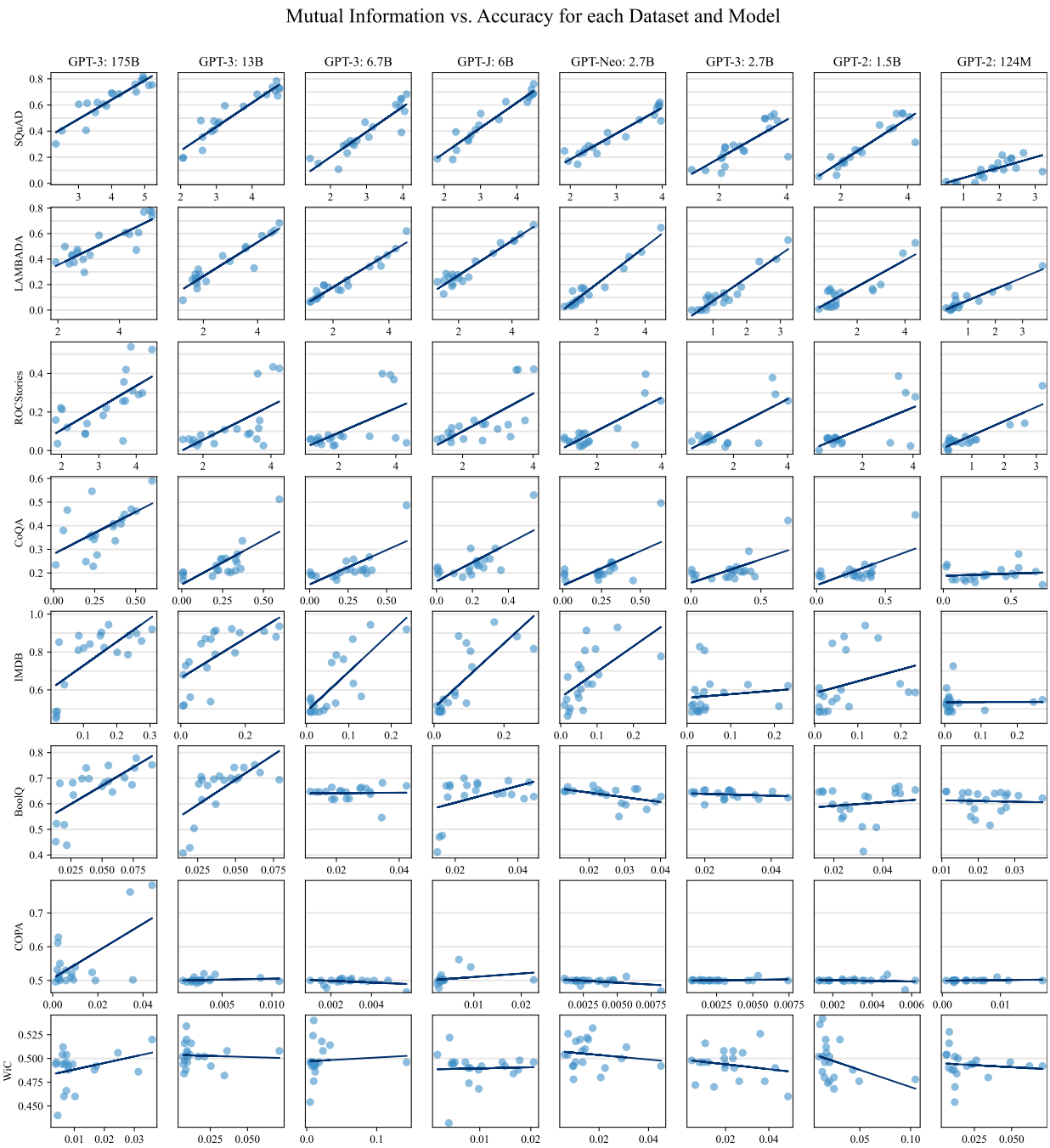

Figure 9: Mutual information plotted against accuracy per prompt for each dataset using GPT-3 175B with linear best fit (by MSE) lines to show overall trends

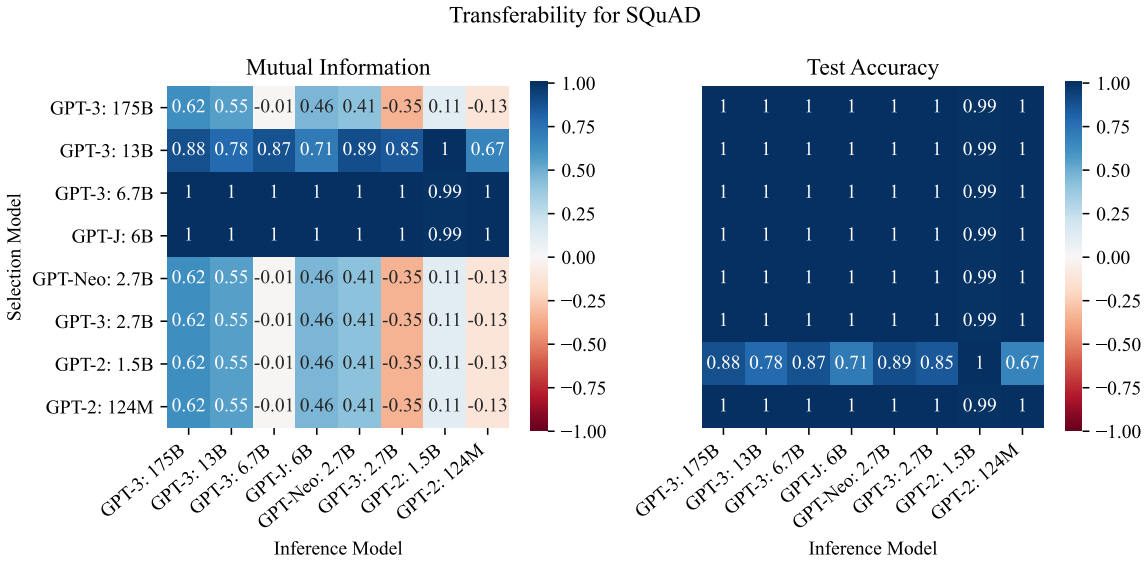

Figure 10: Prompt transfer performance for SQuAD

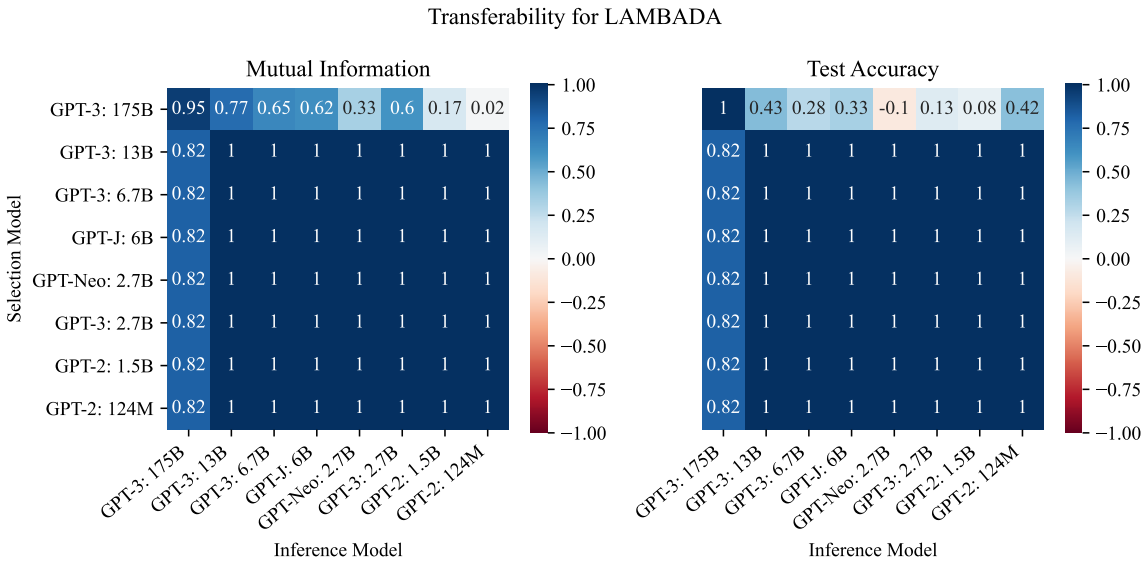

Figure 11: Prompt transfer performance for LAMBADA

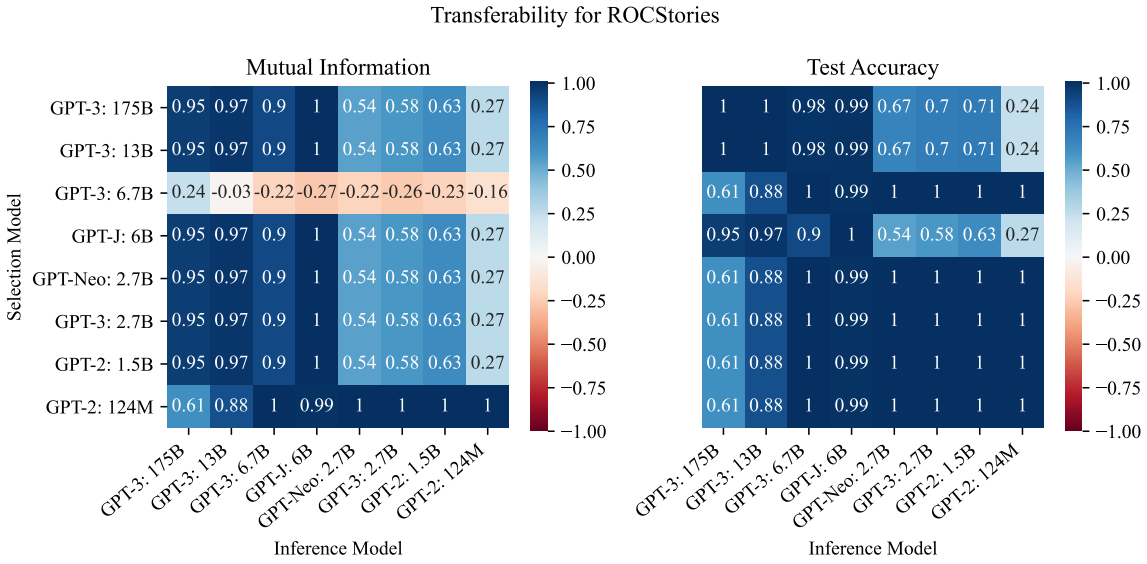

Figure 12: Prompt transfer performance for ROCStories

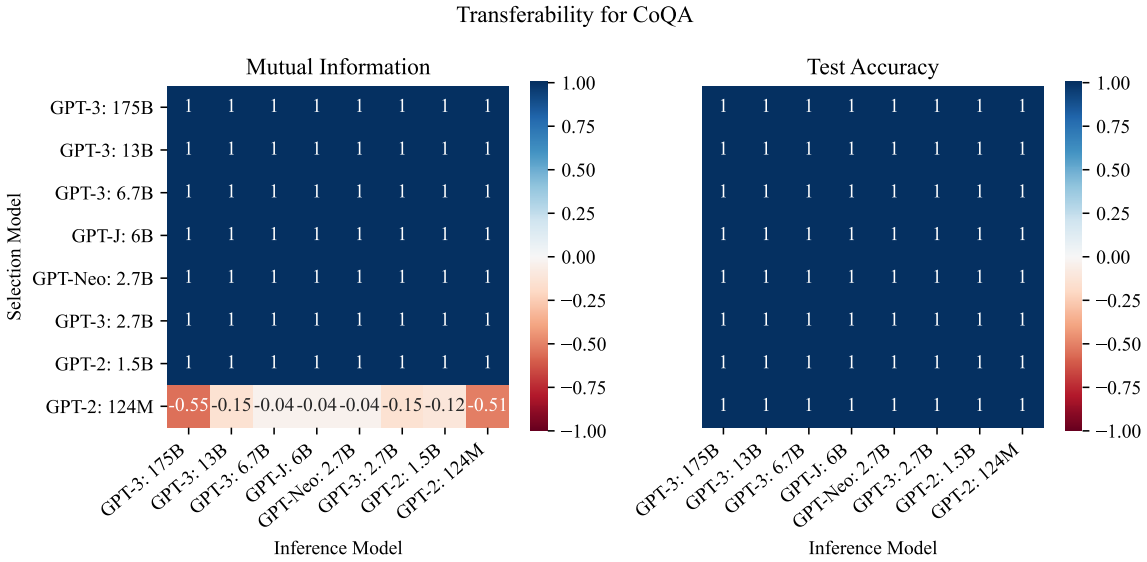

Figure 13: Prompt transfer performance for CoQA

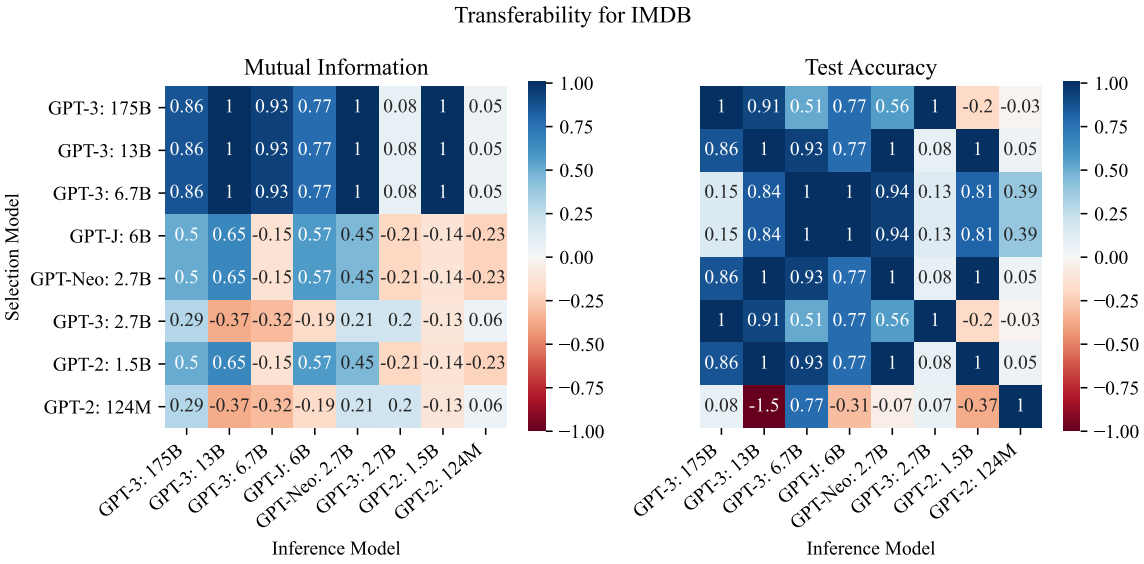

Figure 14: Prompt transfer performance for IMDB

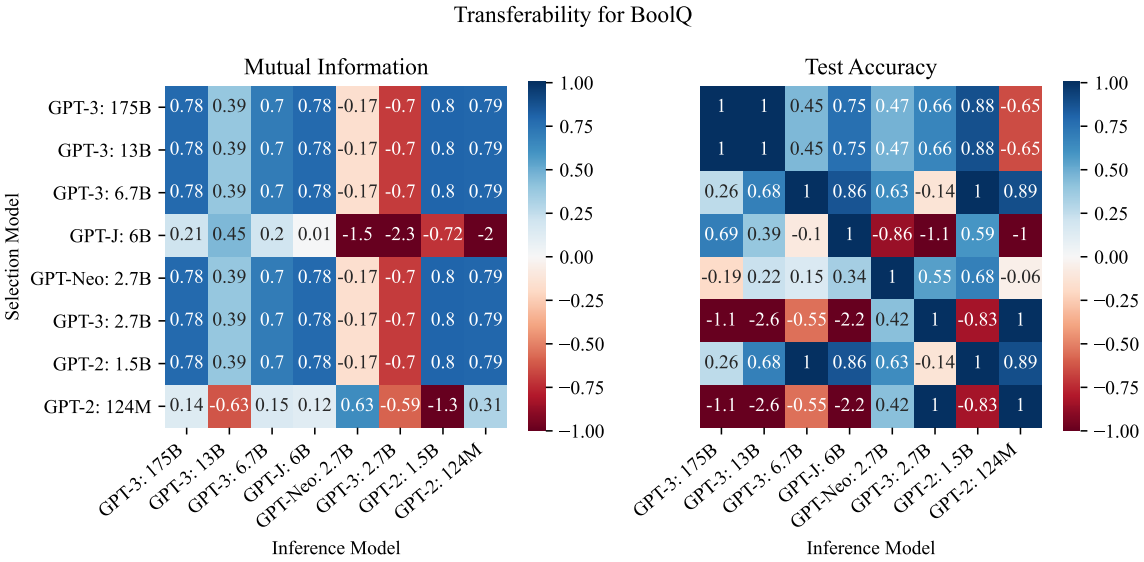

Figure 15: Prompt transfer performance for BoolQ

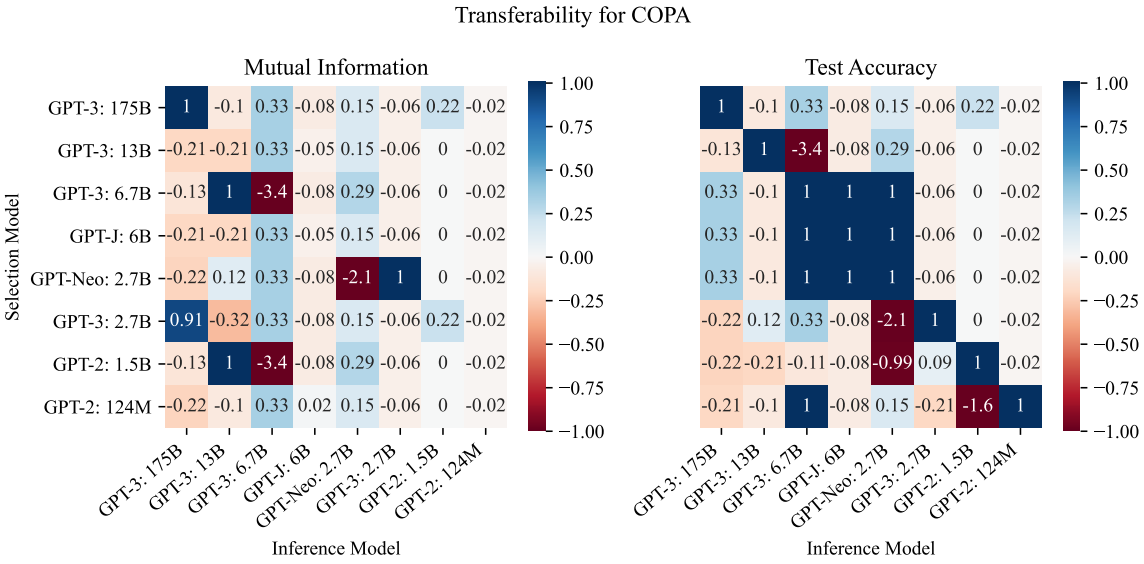

Figure 16: Prompt transfer performance for COPA

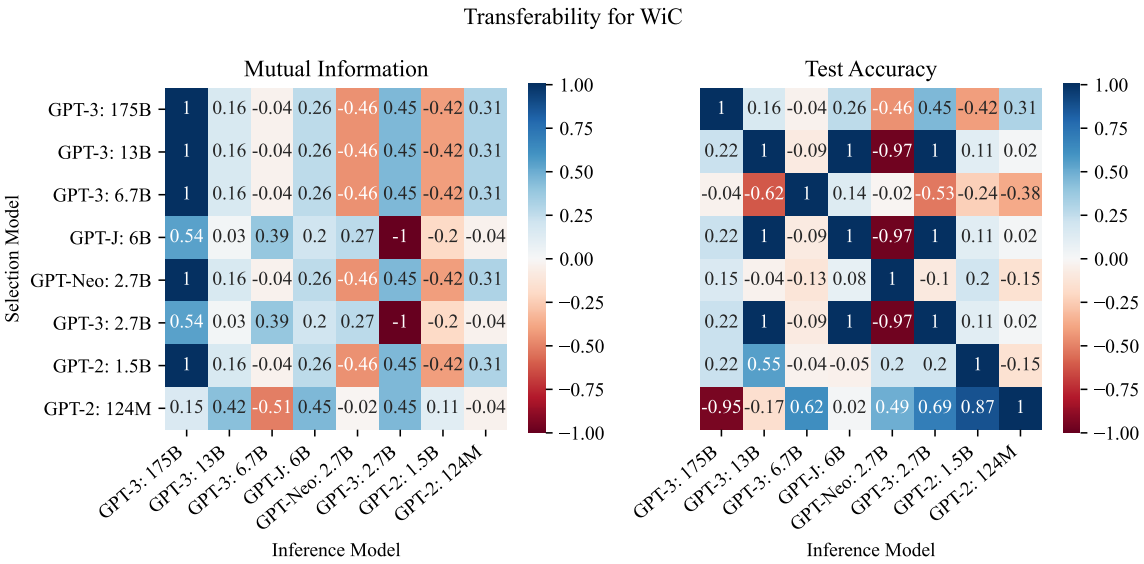

Figure 17: Prompt transfer performance for WiC

C Template Examples

The following are example template f_θs provided for each dataset. We include all used templates, ordered by accuracy. In blue, we highlight the data that is filled in from X; in red, we highlight the area where we ask the model to predict the next token; everything that is not highlighted is static from instance to instance. We also include the token sets used in the collapsing functions.

C.1 SQuAD

Prompt 1 (MI: 4.950, Acc: 0.820):

TASK: Answer the questions below using the phrasing from the context.

CONTEXT:
As of the census of 2000, there were 197,790 people, 84,549 households, and 43,627 families residing in the city. The population density was 3,292.6 people per square mile (1,271.3/km). There were 92,282 housing units at an average density of 1,536.2 per square mile (593.1/km). The racial makeup of the city was 38.3% White, 57.2% African American, 0.2% Native American, 1.3% Asian, 0.1% Pacific Islander, 1.5% from other races, and 1.5% from two or more races. Hispanic or Latino of any race were 2.6% of the population.

QUESTIONS:
1) In 2000, how many families lived in Richmond?
Answer: "43,627"

2) What percentage of the Richmond population of 2000 was Pacific Islander?
Answer: "

Collapsing token sets: None, all tokens are considered

Prompt 2 (MI: 4.965, Acc: 0.800):

Given the following passages and questions, provide a brief, correct answer from the text.

"BYU students arrive with superb preparation. The entering class has an average high school GPA of 3.71 (on a 4.0 scale) and an average ACT score that ranks in the 89th percentile nationally. The University consistently places in the top 20 for enrollment of National Merit Scholars.", "What high school GPA for BYU freshmen have on average?" -> "3.71""BYU students arrive with superb preparation. The entering class has an average high school GPA of 3.71 (on a 4.0 scale) and an average ACT score that ranks in the 89th percentile nationally. The University consistently places in the top 20 for enrollment of National Merit Scholars.", "What high school GPA for BYU freshmen have on average?" -> "3.71"
"In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleed, snow, graupel, and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called"showers".", "What causes precipitation to fall?" -> "gravity"
"As of the census of 2000, there were 197,790 people, 84,549 households, and 43,627 families residing in the city. The population density was 3,292.6 people per square mile (1,271.3/km). There were 92,282 housing units at an average density of 1,536.2 per square mile (593.1/km). The racial makeup of the city was 38.3% White, 57.2% African American, 0.2% Native American, 1.3% Asian, 0.1% Pacific Islander, 1.5% from other races, and 1.5% from two or more races. Hispanic or Latino of any race were 2.6% of the population.", "What percentage of the Richmond population of 2000 was Pacific Islander?" -> "

Collapsing token sets: None, all tokens are considered

Prompt 3 (MI: 4.965, Acc: 0.800):

Given the following passages and questions, provide a brief, correct answer from the text.

"BYU students arrive with superb preparation. The entering class has an average high school GPA of 3.71 (on a 4.0 scale) and an average ACT score that ranks in the 89th percentile nationally. The University consistently places in the top 20 for enrollment of National Merit Scholars.", "What high school GPA for BYU freshmen have on average?" -> "3.71""BYU students arrive with superb preparation. The entering class has an average high school GPA of 3.71 (on a 4.0 scale) and an average ACT score that ranks in the 89th percentile nationally. The University consistently places in the top 20 for enrollment of National Merit Scholars.", "What high school GPA for BYU freshmen have on average?" -> "3.71"
"In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleed, snow, graupel, and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called"showers".", "What causes precipitation to fall?" -> "gravity"
"As of the census of 2000, there were 197,790 people, 84,549 households, and 43,627 families residing in the city. The population density was 3,292.6 people per square mile (1,271.3/km). There were 92,282 housing units at an average density of 1,536.2 per square mile (593.1/km). The racial makeup of the city was 38.3% White, 57.2% African American, 0.2% Native American, 1.3% Asian, 0.1% Pacific Islander, 1.5% from other races, and 1.5% from two or more races. Hispanic or Latino of any race were 2.6% of the population.", "What percentage of the Richmond population of 2000 was Pacific Islander?" -> "

Collapsing token sets: None, all tokens are considered

Prompt 4 (MI: 4.901, Acc: 0.790):

TASK: Answer the questions below using the phrasing from the context.

CONTEXT:
BYU students arrive with superb preparation. The entering class has an average high school GPA of 3.71 (on a 4.0 scale) and an average ACT score that ranks in the 89th percentile nationally. The University consistently places in the top 20 for enrollment of National Merit Scholars.
QUESTIONS:
1) What high school GPA for BYU freshmen have on average?
Answer: "3.71"

CONTEXT:
As of the census of 2000, there were 197,790 people, 84,549 households, and 43,627 families residing in the city. The population density was 3,292.6 people per square mile (1,271.3/km). There were 92,282 housing units at an average density of 1,536.2 per square mile (593.1/km). The racial makeup of the city was 38.3% White, 57.2% African American, 0.2% Native American, 1.3% Asian, 0.1% Pacific Islander, 1.5% from other races, and 1.5% from two or more races. Hispanic or Latino of any race were 2.6% of the population.

QUESTIONS:
1) What percentage of the Richmond population of 2000 was Pacific Islander?
Answer: "

Collapsing token sets: None, all tokens are considered

Prompt 5 (MI: 4.711, Acc: 0.758):

P1: As of the census of 2000, there were 197,790 people, 84,549 households, and 43,627 families residing in the city. The population density was 3,292.6 people per square mile (1,271.3/km). There were 92,282 housing units at an average density of 1,536.2 per square mile (593.1/km). The racial makeup of the city was 38.3% White, 57.2% African American, 0.2% Native American, 1.3% Asian, 0.1% Pacific Islander, 1.5% from other races, and 1.5% from two or more races. Hispanic or Latino of any race were 2.6% of the population.
P2: In 2000, how many families lived in Richmond?
P1: 43,627
P2: What percentage of the Richmond population of 2000 was Pacific Islander?
P1:

Collapsing token sets: None, all tokens are considered

Prompt 6 (MI: 5.224, Acc: 0.754):

CHAPTER QUIZ

PASSAGE:
As of the census of 2000, there were 197,790 people, 84,549 households, and 43,627 families residing in the city. The population density was 3,292.6 people per square mile (1,271.3/km). There were 92,282 housing units at an average density of 1,536.2 per square mile (593.1/km). The racial makeup of the city was 38.3% White, 57.2% African American, 0.2% Native American, 1.3% Asian, 0.1% Pacific Islander, 1.5% from other races, and 1.5% from two or more races. Hispanic or Latino of any race were 2.6% of the population.

QUESTIONS:
1) In 2000, how many families lived in Richmond?
2) What percentage of the Richmond population of 2000 was Pacific Islander?

ANSWER KEY:
1) 43,627
2)

Collapsing token sets: None, all tokens are considered

Prompt 7 (MI: 5.126, Acc: 0.750):

CHAPTER QUIZ
PASSAGE: BYU students arrive with superb preparation. The entering class has an average high school GPA of 3.71 (on a 4.0 scale) and an average ACT score that ranks in the 89th percentile nationally. The University consistently places in the top 20 for enrollment of National Merit Scholars.
QUESTIONS:
1) What high school GPA for BYU freshmen have on average?

ANSWER KEY:
1) 3.71

CHAPTER QUIZ

PASSAGE:
As of the census of 2000, there were 197,790 people, 84,549 households, and 43,627 families residing in the city. The population density was 3,292.6 people per square mile (1,271.3/km). There were 92,282 housing units at an average density of 1,536.2 per square mile (593.1/km). The racial makeup of the city was 38.3% White, 57.2% African American, 0.2% Native American, 1.3% Asian, 0.1% Pacific Islander, 1.5% from other races, and 1.5% from two or more races. Hispanic or Latino of any race were 2.6% of the population.

QUESTIONS:
1) What percentage of the Richmond population of 2000 was Pacific Islander?

ANSWER KEY:
1)

Collapsing token sets: None, all tokens are considered

Prompt 8 (MI: 4.745, Acc: 0.700):

P1: BYU students arrive with superb preparation. The entering class has an average high school GPA of 3.71 (on a 4.0 scale) and an average ACT score that ranks in the 89th percentile nationally. The University consistently places in the top 20 for enrollment of National Merit Scholars.
P2: What high school GPA for BYU freshmen have on average?
P1: 3.71

P1: As of the census of 2000, there were 197,790 people, 84,549 households, and 43,627 families residing in the city. The population density was 3,292.6 people per square mile (1,271.3/km). There were 92,282 housing units at an average density of 1,536.2 per square mile (593.1/km). The racial makeup of the city was 38.3% White, 57.2% African American, 0.2% Native American, 1.3% Asian, 0.1% Pacific Islander, 1.5% from other races, and 1.5% from two or more races. Hispanic or Latino of any race were 2.6% of the population.
P2: What percentage of the Richmond population of 2000 was Pacific Islander?
P1:

Collapsing token sets: None, all tokens are considered

Prompt 9 (MI: 3.998, Acc: 0.692):

CHAPTER QUIZ

PASSAGE:
As of the census of 2000, there were 197,790 people, 84,549 households, and 43,627 families residing in the city. The population density was 3,292.6 people per square mile (1,271.3/km). There were 92,282 housing units at an average density of 1,536.2 per square mile (593.1/km). The racial makeup of the city was 38.3% White, 57.2% African American, 0.2% Native American, 1.3% Asian, 0.1% Pacific Islander, 1.5% from other races, and 1.5% from two or more races. Hispanic or Latino of any race were 2.6% of the population.

QUESTIONS:
1) What percentage of the Richmond population of 2000 was Pacific Islander?

ANSWER KEY:
1)

Collapsing token sets: None, all tokens are considered

Prompt 10 (MI: 4.037, Acc: 0.686):

TASK: Using words from the CONTEXT, answer the below QUESTIONS.

CONTEXT:
As of the census of 2000, there were 197,790 people, 84,549 households, and 43,627 families residing in the city. The population density was 3,292.6 people per square mile (1,271.3/km). There were 92,282 housing units at an average density of 1,536.2 per square mile (593.1/km). The racial makeup of the city was 38.3% White, 57.2% African American, 0.2% Native American, 1.3% Asian, 0.1% Pacific Islander, 1.5% from other races, and 1.5% from two or more races. Hispanic or Latino of any race were 2.6% of the population.

QUESTIONS:
1) What percentage of the Richmond population of 2000 was Pacific Islander?
Answer: "

Collapsing token sets: None, all tokens are considered

Prompt 11 (MI: 4.231, Acc: 0.684):

P1 tells P2 some information, P2 asks comprehension questions, and P1 answers.

P1: As of the census of 2000, there were 197,790 people, 84,549 households, and 43,627 families residing in the city. The population density was 3,292.6 people per square mile (1,271.3/km). There were 92,282 housing units at an average density of 1,536.2 per square mile (593.1/km). The racial makeup of the city was 38.3% White, 57.2% African American, 0.2% Native American, 1.3% Asian, 0.1% Pacific Islander, 1.5% from other races, and 1.5% from two or more races. Hispanic or Latino of any race were 2.6% of the population.
P2: What percentage of the Richmond population of 2000 was Pacific Islander?
P1: The answer is "

Collapsing token sets: None, all tokens are considered

Prompt 12 (MI: 3.568, Acc: 0.620):

P1: As of the census of 2000, there were 197,790 people, 84,549 households, and 43,627 families residing in the city. The population density was 3,292.6 people per square mile (1,271.3/km). There were 92,282 housing units at an average density of 1,536.2 per square mile (593.1/km). The racial makeup of the city was 38.3% White, 57.2% African American, 0.2% Native American, 1.3% Asian, 0.1% Pacific Islander, 1.5% from other races, and 1.5% from two or more races. Hispanic or Latino of any race were 2.6% of the population.
P2: What percentage of the Richmond population of 2000 was Pacific Islander?
P1: The answer is "

Collapsing token sets: None, all tokens are considered

Prompt 13 (MI: 3.261, Acc: 0.614):

Given the following passages and questions, provide a brief, correct answer from the text.
"As of the census of 2000, there were 197,790 people, 84,549 households, and 43,627 families residing in the city. The population density was 3,292.6 people per square mile (1,271.3/km). There were 92,282 housing units at an average density of 1,536.2 per square mile (593.1/km). The racial makeup of the city was 38.3% White, 57.2% African American, 0.2% Native American, 1.3% Asian, 0.1% Pacific Islander, 1.5% from other races, and 1.5% from two or more races. Hispanic or Latino of any race were 2.6% of the population.", "What percentage of the Richmond population of 2000 was Pacific Islander?" -> "

Collapsing token sets: None, all tokens are considered

Prompt 14 (MI: 3.760, Acc: 0.608):

As of the census of 2000, there were 197,790 people, 84,549 households, and 43,627 families residing in the city. The population density was 3,292.6 people per square mile (1,271.3/km). There were 92,282 housing units at an average density of 1,536.2 per square mile (593.1/km). The racial makeup of the city was 38.3% White, 57.2% African American, 0.2% Native American, 1.3% Asian, 0.1% Pacific Islander, 1.5% from other races, and 1.5% from two or more races. Hispanic or Latino of any race were 2.6% of the population.

What percentage of the Richmond population of 2000 was Pacific Islander?
The correct answer is:

Collapsing token sets: None, all tokens are considered

Prompt 15 (MI: 3.006, Acc: 0.606):

I read this in a book today:
As of the census of 2000, there were 197,790 people, 84,549 households, and 43,627 families residing in the city. The population density was 3,292.6 people per square mile (1,271.3/km). There were 92,282 housing units at an average density of 1,536.2 per square mile (593.1/km). The racial makeup of the city was 38.3% White, 57.2% African American, 0.2% Native American, 1.3% Asian, 0.1% Pacific Islander, 1.5% from other races, and 1.5% from two or more races. Hispanic or Latino of any race were 2.6% of the population.
From that context, did you catch What percentage of the Richmond population of 2000 was Pacific Islander?
Yes, the answer is

Collapsing token sets: None, all tokens are considered

Prompt 16 (MI: 3.843, Acc: 0.592):

TASK: Answer the questions below using the phrasing from the context.

CONTEXT:
As of the census of 2000, there were 197,790 people, 84,549 households, and 43,627 families residing in the city. The population density was 3,292.6 people per square mile (1,271.3/km). There were 92,282 housing units at an average density of 1,536.2 per square mile (593.1/km). The racial makeup of the city was 38.3% White, 57.2% African American, 0.2% Native American, 1.3% Asian, 0.1% Pacific Islander, 1.5% from other races, and 1.5% from two or more races. Hispanic or Latino of any race were 2.6% of the population.

QUESTIONS:
1) What percentage of the Richmond population of 2000 was Pacific Islander?
Answer: "

Collapsing token sets: None, all tokens are considered

Prompt 17 (MI: 3.508, Acc: 0.544):

I read this in a book today:
As of the census of 2000, there were 197,790 people, 84,549 households, and 43,627 families residing in the city. The population density was 3,292.6 people per square mile (1,271.3/km). There were 92,282 housing units at an average density of 1,536.2 per square mile (593.1/km). The racial makeup of the city was 38.3% White, 57.2% African American, 0.2% Native American, 1.3% Asian, 0.1% Pacific Islander, 1.5% from other races, and 1.5% from two or more races. Hispanic or Latino of any race were 2.6% of the population.
What percentage of the Richmond population of 2000 was Pacific Islander?
Answer:

Collapsing token sets: None, all tokens are considered

Prompt 18 (MI: 3.227, Acc: 0.406):

Context: As of the census of 2000, there were 197,790 people, 84,549 households, and 43,627 families residing in the city. The population density was 3,292.6 people per square mile (1,271.3/km). There were 92,282 housing units at an average density of 1,536.2 per square mile (593.1/km). The racial makeup of the city was 38.3% White, 57.2% African American, 0.2% Native American, 1.3% Asian, 0.1% Pacific Islander, 1.5% from other races, and 1.5% from two or more races. Hispanic or Latino of any race were 2.6% of the population.

Q: What percentage of the Richmond population of 2000 was Pacific Islander?

A:

Collapsing token sets: None, all tokens are considered

Prompt 19 (MI: 2.497, Acc: 0.402):

A friend of mine told me this:
As of the census of 2000, there were 197,790 people, 84,549 households, and 43,627 families residing in the city. The population density was 3,292.6 people per square mile (1,271.3/km). There were 92,282 housing units at an average density of 1,536.2 per square mile (593.1/km). The racial makeup of the city was 38.3% White, 57.2% African American, 0.2% Native American, 1.3% Asian, 0.1% Pacific Islander, 1.5% from other races, and 1.5% from two or more races. Hispanic or Latino of any race were 2.6% of the population.
My friend then asked: What percentage of the Richmond population of 2000 was Pacific Islander?
I answered:

Collapsing token sets: None, all tokens are considered

Prompt 20 (MI: 2.312, Acc: 0.302):

ANSWER KEY:

QUESTION1:
"As of the census of 2000, there were 197,790 people, 84,549 households, and 43,627 families residing in the city. The population density was 3,292.6 people per square mile (1,271.3/km). There were 92,282 housing units at an average density of 1,536.2 per square mile (593.1/km). The racial makeup of the city was 38.3% White, 57.2% African American, 0.2% Native American, 1.3% Asian, 0.1% Pacific Islander, 1.5% from other races, and 1.5% from two or more races. Hispanic or Latino of any race were 2.6% of the population." What percentage of the Richmond population of 2000 was Pacific Islander?
ANSWER1:

Collapsing token sets: None, all tokens are considered

C.2 LAMBADA

Prompt 1 (MI: 4.984, Acc: 0.782):

Fill in blank:

Alice was friends with Bob. Alice went to visit her friend ____. -> Bob
"I would speak to you privately," Bowen said, casting a glance around at the others milling about.

The worry in her eyes deepened, but she nodded hesitantly and awaited Bowen's directive.

He led her through the great hall, annoyance biting at him when he saw no place where people weren't congregated. He stepped outside the back of the keep, where, finally, he spied an area near the bathhouses, where it was quiet and ____. ->

Collapsing token sets: None, all tokens are considered

Prompt 2 (MI: 4.793, Acc: 0.770):

Fill in blank:

She held the torch in front of her.

She caught her breath.

"Chris? There's a step."

"What?"

"A step. Cut in the rock. About fifty feet ahead." She moved faster. They both moved faster. "In fact," she said, raising the torch higher, "there's more than a ____. -> step

"I would speak to you privately," Bowen said, casting a glance around at the others milling about.

The worry in her eyes deepened, but she nodded hesitantly and awaited Bowen's directive.

He led her through the great hall, annoyance biting at him when he saw no place where people weren't congregated. He stepped outside the back of the keep, where, finally, he spied an area near the bathhouses, where it was quiet and ____. ->

Collapsing token sets: None, all tokens are considered

Prompt 3 (MI: 5.062, Acc: 0.770):

Fill in blank:

Alice was friends with Bob. Alice went to visit her friend ____. -> Bob
George bought some baseball equipment, a ball, a glove, and a ____. -> bat
"I would speak to you privately," Bowen said, casting a glance around at the others milling about.

The worry in her eyes deepened, but she nodded hesitantly and awaited Bowen's directive.

He led her through the great hall, annoyance biting at him when he saw no place where people weren't congregated. He stepped outside the back of the keep, where, finally, he spied an area near the bathhouses, where it was quiet and ____. ->

Collapsing token sets: None, all tokens are considered

Prompt 4 (MI: 5.058, Acc: 0.736):

"I would speak to you privately," Bowen said, casting a glance around at the others milling about.

The worry in her eyes deepened, but she nodded hesitantly and awaited Bowen's directive.

He led her through the great hall, annoyance biting at him when he saw no place where people weren't congregated. He stepped outside the back of the keep, where, finally, he spied an area near the bathhouses, where it was quiet and

Collapsing token sets: None, all tokens are considered

Prompt 5 (MI: 4.194, Acc: 0.608):

P1: I'm going to tell you a story, but leave a word out. Once I'm done telling the story, pick the word that best fits in the blank.
I like to eat peanut butter and jelly ____.
P2: sandwiches
P1: I'm going to tell you a story, but leave a word out. Once I'm done telling the story, pick the word that best fits in the blank.
"I would speak to you privately," Bowen said, casting a glance around at the others milling about.

The worry in her eyes deepened, but she nodded hesitantly and awaited Bowen's directive.

He led her through the great hall, annoyance biting at him when he saw no place where people weren't congregated. He stepped outside the back of the keep, where, finally, he spied an area near the bathhouses, where it was quiet and ____.
P2:

Collapsing token sets: None, all tokens are considered

Prompt 6 (MI: 4.623, Acc: 0.608):

Fill in the blank for the following sentences.

"It was a cold night. The wind was whistling around the courtyard as I stepped out of the car and into the ____." -> "It was a cold night. The wind was whistling around the courtyard as I stepped out of the car and into the darkness."
""I would speak to you privately," Bowen said, casting a glance around at the others milling about.

The worry in her eyes deepened, but she nodded hesitantly and awaited Bowen's directive.

He led her through the great hall, annoyance biting at him when he saw no place where people weren't congregated. He stepped outside the back of the keep, where, finally, he spied an area near the bathhouses, where it was quiet and ____." -> ""I would speak to you privately," Bowen said, casting a glance around at the others milling about.

The worry in her eyes deepened, but she nodded hesitantly and awaited Bowen's directive.

He led her through the great hall, annoyance biting at him when he saw no place where people weren't congregated. He stepped outside the back of the keep, where, finally, he spied an area near the bathhouses, where it was quiet and

Collapsing token sets: None, all tokens are considered

Prompt 7 (MI: 4.328, Acc: 0.596):

P1: I'm going to tell you a story, but leave a word out. Once I'm done telling the story, pick the word that best fits in the blank.
It was a cold night. The wind was ____ around the courtyard as I stepped out of the car and into the darkness.
P2: whistling
P1: I'm going to tell you a story, but leave a word out. Once I'm done telling the story, pick the word that best fits in the blank.
"I would speak to you privately," Bowen said, casting a glance around at the others milling about.

The worry in her eyes deepened, but she nodded hesitantly and awaited Bowen's directive.

He led her through the great hall, annoyance biting at him when he saw no place where people weren't congregated. He stepped outside the back of the keep, where, finally, he spied an area near the bathhouses, where it was quiet and ____.
P2:

Collapsing token sets: None, all tokens are considered

Prompt 8 (MI: 3.338, Acc: 0.586):

Fill in the blank with the missing word to complete the sentence.

Passage: I like to eat peanut butter and jelly ____.
Missing Word: sandwiches

Passage: "I would speak to you privately," Bowen said, casting a glance around at the others milling about.

The worry in her eyes deepened, but she nodded hesitantly and awaited Bowen's directive.

He led her through the great hall, annoyance biting at him when he saw no place where people weren't congregated. He stepped outside the back of the keep, where, finally, he spied an area near the bathhouses, where it was quiet and ____.
Missing Word: "

Collapsing token sets: None, all tokens are considered

Prompt 9 (MI: 2.230, Acc: 0.498):

"I would speak to you privately," Bowen said, casting a glance around at the others milling about.

The worry in her eyes deepened, but she nodded hesitantly and awaited Bowen's directive.

He led her through the great hall, annoyance biting at him when he saw no place where people weren't congregated. He stepped outside the back of the keep, where, finally, he spied an area near the bathhouses, where it was quiet and ____.

The missing word in the story should be: "

Collapsing token sets: None, all tokens are considered

Prompt 10 (MI: 2.632, Acc: 0.474):

"I would speak to you privately," Bowen said, casting a glance around at the others milling about.

The worry in her eyes deepened, but she nodded hesitantly and awaited Bowen's directive.

He led her through the great hall, annoyance biting at him when he saw no place where people weren't congregated. He stepped outside the back of the keep, where, finally, he spied an area near the bathhouses, where it was quiet and ____.
Fill in the blank with the missing word or phrase.
What is the missing word? The missing word is "

Collapsing token sets: None, all tokens are considered

Prompt 11 (MI: 4.549, Acc: 0.470):

It was a cold night. The wind was ____ around the courtyard as I stepped out of the car and into the darkness.
Word: whistling

"I would speak to you privately," Bowen said, casting a glance around at the others milling about.

The worry in her eyes deepened, but she nodded hesitantly and awaited Bowen's directive.

He led her through the great hall, annoyance biting at him when he saw no place where people weren't congregated. He stepped outside the back of the keep, where, finally, he spied an area near the bathhouses, where it was quiet and ____.

Word:

Collapsing token sets: None, all tokens are considered

Prompt 12 (MI: 2.637, Acc: 0.454):

P1: I'm going to tell you a story, but leave a word out. Once I'm done telling the story, pick the word that best fits in the blank.
"I would speak to you privately," Bowen said, casting a glance around at the others milling about.

The worry in her eyes deepened, but she nodded hesitantly and awaited Bowen's directive.

He led her through the great hall, annoyance biting at him when he saw no place where people weren't congregated. He stepped outside the back of the keep, where, finally, he spied an area near the bathhouses, where it was quiet and ____.
P2: The word which fits best is "

Collapsing token sets: None, all tokens are considered

Prompt 13 (MI: 2.476, Acc: 0.434):

"I would speak to you privately," Bowen said, casting a glance around at the others milling about.

The worry in her eyes deepened, but she nodded hesitantly and awaited Bowen's directive.

He led her through the great hall, annoyance biting at him when he saw no place where people weren't congregated. He stepped outside the back of the keep, where, finally, he spied an area near the bathhouses, where it was quiet and ____.
Fill in the blank with the missing word or phrase to complete the sentence.
What is the missing word? The missing word is "

Collapsing token sets: None, all tokens are considered

Prompt 14 (MI: 3.043, Acc: 0.432):

Read the following sentences, and try to guess which word goes in the blank.
"I would speak to you privately," Bowen said, casting a glance around at the others milling about.

The worry in her eyes deepened, but she nodded hesitantly and awaited Bowen's directive.

He led her through the great hall, annoyance biting at him when he saw no place where people weren't congregated. He stepped outside the back of the keep, where, finally, he spied an area near the bathhouses, where it was quiet and ____.
Answer: "

Collapsing token sets: None, all tokens are considered

Prompt 15 (MI: 2.450, Acc: 0.428):

Fill in blank:

"I would speak to you privately," Bowen said, casting a glance around at the others milling about.

The worry in her eyes deepened, but she nodded hesitantly and awaited Bowen's directive.

He led her through the great hall, annoyance biting at him when he saw no place where people weren't congregated. He stepped outside the back of the keep, where, finally, he spied an area near the bathhouses, where it was quiet and ____. ->

Collapsing token sets: None, all tokens are considered

Prompt 16 (MI: 2.820, Acc: 0.398):

Fill in the blank with the missing word.
"I would speak to you privately," Bowen said, casting a glance around at the others milling about.

The worry in her eyes deepened, but she nodded hesitantly and awaited Bowen's directive.

He led her through the great hall, annoyance biting at him when he saw no place where people weren't congregated. He stepped outside the back of the keep, where, finally, he spied an area near the bathhouses, where it was quiet and ____.
Answer: "

Collapsing token sets: None, all tokens are considered

Prompt 17 (MI: 1.931, Acc: 0.376):

"I would speak to you privately," Bowen said, casting a glance around at the others milling about.

The worry in her eyes deepened, but she nodded hesitantly and awaited Bowen's directive.

He led her through the great hall, annoyance biting at him when he saw no place where people weren't congregated. He stepped outside the back of the keep, where, finally, he spied an area near the bathhouses, where it was quiet and ____.
Which word should we put in the blank to complete the story?
Let's use the word "

Collapsing token sets: None, all tokens are considered

Prompt 18 (MI: 2.530, Acc: 0.374):

P1: What word do you think fits best in the following story?
"I would speak to you privately," Bowen said, casting a glance around at the others milling about.

The worry in her eyes deepened, but she nodded hesitantly and awaited Bowen's directive.

He led her through the great hall, annoyance biting at him when he saw no place where people weren't congregated. He stepped outside the back of the keep, where, finally, he spied an area near the bathhouses, where it was quiet and ____.
P2: The word which fits best is "

Collapsing token sets: None, all tokens are considered

Prompt 19 (MI: 2.372, Acc: 0.364):

"I would speak to you privately," Bowen said, casting a glance around at the others milling about.

The worry in her eyes deepened, but she nodded hesitantly and awaited Bowen's directive.

He led her through the great hall, annoyance biting at him when he saw no place where people weren't congregated. He stepped outside the back of the keep, where, finally, he spied an area near the bathhouses, where it was quiet and ____.
Which word fills in the blank best?
The word that fills in the blank best is "

Collapsing token sets: None, all tokens are considered

Prompt 20 (MI: 2.860, Acc: 0.296):

Pick the best word to replace the blank.
Story: "I would speak to you privately," Bowen said, casting a glance around at the others milling about.

The worry in her eyes deepened, but she nodded hesitantly and awaited Bowen's directive.

He led her through the great hall, annoyance biting at him when he saw no place where people weren't congregated. He stepped outside the back of the keep, where, finally, he spied an area near the bathhouses, where it was quiet and ____.
Answer: "

Collapsing token sets: None, all tokens are considered

C.3 ROCStories

Prompt 1 (MI: 3.859, Acc: 0.538):

Fill in the blank for the following sentences.

"Marissa loved _____ pokemon go game. It is the biggest thing right now. She had done so much more walking since she started playing it. She walked all day and evening sometimes. She walked almost 10 miles in two days." -> "Marissa loved

Collapsing token sets: None, all tokens are considered

Prompt 2 (MI: 4.427, Acc: 0.524):

Fill in the blank for the following sentences.

"It was a cold night. The wind was _____ around the courtyard as I stepped out of the car and into the darkness." -> "It was a cold night. The wind was whistling around the courtyard as I stepped out of the car and into the darkness."
"Marissa loved _____ pokemon go game. It is the biggest thing right now. She had done so much more walking since she started playing it. She walked all day and evening sometimes. She walked almost 10 miles in two days." -> "Marissa loved

Collapsing token sets: None, all tokens are considered

Prompt 3 (MI: 3.728, Acc: 0.420):

Poke GO!

Marissa loved

Collapsing token sets: None, all tokens are considered

Prompt 4 (MI: 3.670, Acc: 0.356):

Fill in the blank with the missing word or phrase to complete the sentence.

I like to eat _____ and jelly sandwiches.
Answer: peanut butter

Marissa loved _____ pokemon go game. It is the biggest thing right now. She had done so much more walking since she started playing it. She walked all day and evening sometimes. She walked almost 10 miles in two days.
Answer:

Collapsing token sets: None, all tokens are considered

Prompt 5 (MI: 3.904, Acc: 0.310):

Fill in the blank with the missing word or phrase.

Sentence: I like to eat _____ and jelly sandwiches.
Missing Word/Phrase: peanut butter

Sentence: Marissa loved _____ pokemon go game. It is the biggest thing right now. She had done so much more walking since she started playing it. She walked all day and evening sometimes. She walked almost 10 miles in two days.
Missing Word/Phrase:

Collapsing token sets: None, all tokens are considered

Prompt 6 (MI: 4.167, Acc: 0.298):

P1: I'm going to tell you a story, but leave a word out. Once I'm done telling the story, pick the word that best fits in the blank.
It was a cold night. The wind was _____ around the courtyard as I stepped out of the car and into the darkness.
P2: whistling
P1: I'm going to tell you a story, but leave a word out. Once I'm done telling the story, pick the word that best fits in the blank.
Marissa loved _____ pokemon go game. It is the biggest thing right now. She had done so much more walking since she started playing it. She walked all day and evening sometimes. She walked almost 10 miles in two days.
P2:

Collapsing token sets: None, all tokens are considered

Prompt 7 (MI: 4.066, Acc: 0.290):

P1: I'm going to tell you a story, but leave a word out. Once I'm done telling the story, pick the word that best fits in the blank.
I like to eat _____ and jelly sandwiches.
P2: peanut butter
P1: I'm going to tell you a story, but leave a word out. Once I'm done telling the story, pick the word that best fits in the blank.
Marissa loved _____ pokemon go game. It is the biggest thing right now. She had done so much more walking since she started playing it. She walked all day and evening sometimes. She walked almost 10 miles in two days.
P2:

Collapsing token sets: None, all tokens are considered

Prompt 8 (MI: 3.707, Acc: 0.258):

Guess the word in the blank to complete the story.
Story: Marissa loved _____ pokemon go game. It is the biggest thing right now. She had done so much more walking since she started playing it. She walked all day and evening sometimes. She walked almost 10 miles in two days.
Answer:

Collapsing token sets: None, all tokens are considered

Prompt 9 (MI: 3.644, Acc: 0.256):

Pick the best word to replace the blank.
Story: Marissa loved _____ pokemon go game. It is the biggest thing right now. She had done so much more walking since she started playing it. She walked all day and evening sometimes. She walked almost 10 miles in two days.
Answer:

Collapsing token sets: None, all tokens are considered

Prompt 10 (MI: 1.979, Acc: 0.222):

Marissa loved _____ pokemon go game. It is the biggest thing right now. She had done so much more walking since she started playing it. She walked all day and evening sometimes. She walked almost 10 miles in two days.
Fill in the blank with the missing word or phrase.
What is the missing word? The missing word is "

Collapsing token sets: None, all tokens are considered

Prompt 11 (MI: 3.199, Acc: 0.220):

Fill in the blank with the missing word or phrase.
Marissa loved _____ pokemon go game. It is the biggest thing right now. She had done so much more walking since she started playing it. She walked all day and evening sometimes. She walked almost 10 miles in two days.
Answer:

Collapsing token sets: None, all tokens are considered

Prompt 12 (MI: 2.013, Acc: 0.214):

Marissa loved _____ pokemon go game. It is the biggest thing right now. She had done so much more walking since she started playing it. She walked all day and evening sometimes. She walked almost 10 miles in two days.
Fill in the blank with the missing word or phrase to complete the sentence.
What is the missing word? The missing word is "

Collapsing token sets: None, all tokens are considered

Prompt 13 (MI: 3.116, Acc: 0.182):

Read the following sentences, and try to guess which word goes in the blank.
Marissa loved _____ pokemon go game. It is the biggest thing right now. She had done so much more walking since she started playing it. She walked all day and evening sometimes. She walked almost 10 miles in two days.
Answer:

Collapsing token sets: None, all tokens are considered

Prompt 14 (MI: 1.843, Acc: 0.158):

Marissa loved _____ pokemon go game. It is the biggest thing right now. She had done so much more walking since she started playing it. She walked all day and evening sometimes. She walked almost 10 miles in two days.

The missing word in the story should be: "

Collapsing token sets: None, all tokens are considered

Prompt 15 (MI: 2.681, Acc: 0.140):

P1: I'm going to tell you a story, but leave a word out. Once I'm done telling the story, pick the word that best fits in the blank.
Marissa loved _____ pokemon go game. It is the biggest thing right now. She had done so much more walking since she started playing it. She walked all day and evening sometimes. She walked almost 10 miles in two days.
P2: The word which fits best is "

Collapsing token sets: None, all tokens are considered

Prompt 16 (MI: 2.150, Acc: 0.120):

Marissa loved _____ pokemon go game. It is the biggest thing right now. She had done so much more walking since she started playing it. She walked all day and evening sometimes. She walked almost 10 miles in two days.
Which word should we put in the blank to complete the story?
Let's use the word "

Collapsing token sets: None, all tokens are considered

Prompt 17 (MI: 2.634, Acc: 0.088):

Marissa loved _____ pokemon go game. It is the biggest thing right now. She had done so much more walking since she started playing it. She walked all day and evening sometimes. She walked almost 10 miles in two days.
Which word fills in the blank best?
The word that fills in the blank best is "

Collapsing token sets: None, all tokens are considered

Prompt 18 (MI: 2.637, Acc: 0.086):

P1: What word do you think fits best in the following story?
Marissa loved _____ pokemon go game. It is the biggest thing right now. She had done so much more walking since she started playing it. She walked all day and evening sometimes. She walked almost 10 miles in two days.
P2: The word which fits best is "

Collapsing token sets: None, all tokens are considered

Prompt 19 (MI: 3.648, Acc: 0.050):

It was a cold night. The wind was _____ around the courtyard as I stepped out of the car and into the darkness.
Word: whistling

Marissa loved _____ pokemon go game. It is the biggest thing right now. She had done so much more walking since she started playing it. She walked all day and evening sometimes. She walked almost 10 miles in two days.
Put the best word in the blank to complete the story.
Word:

Collapsing token sets: None, all tokens are considered

Prompt 20 (MI: 1.891, Acc: 0.036):

Marissa loved _____ pokemon go game. It is the biggest thing right now. She had done so much more walking since she started playing it. She walked all day and evening sometimes. She walked almost 10 miles in two days.
Choose a word to replace the blank.
Word: "

Collapsing token sets: None, all tokens are considered

C.4 CoQA

Prompt 1 (MI: 0.600, Acc: 0.590):

Instructions: For each question below, choose the answer from the answer bank corresponding to the question that best answers the question.

Question 1 Answer Bank: ladybug, bunny, goldfish, leopard, caterpillarQuestion: What animal would be most dangerous for a human to encounter in the wild?

Answer: leopard

Question 2 Answer Bank: wrong, pleasure, encouragement, depression, relief

Question: If you're still in love and end up stopping being married to your partner, what emotion are you likely to experience?

Answer:

Collapsing token sets: {'A': ['wrong'], 'B': ['pleasure'], 'C': ['encouragement'], 'D': ['depression'], 'E': ['relief']}

Prompt 2 (MI: 0.233, Acc: 0.546):

Common Sense Quiz Answer Key

Question 1: Where would people not typically go for fun?
A: theme park
B: movie theatre
C: carnival
D: waste management facility
E: beach
Correct Answer: D

Question 2: If you're still in love and end up stopping being married to your partner, what emotion are you likely to experience?
A: wrong
B: pleasure
C: encouragement
D: depression
E: relief
Correct Answer:

Collapsing token sets: ['A', 'B', 'C', 'D', 'E']

Prompt 3 (MI: 0.474, Acc: 0.470):

Given the following questions and choices, pick the choice that corresponds best to the question.

"I'm crossing the river, my feet are wet but my body is dry, where am I?", "bridge, waterfall, valley, pebble, mountain", -> "valley"
"In what Spanish speaking North American country can you get a great cup of coffee?", "mexico, mildred's coffee shop, diner, kitchen, canteen", -> "mexico"
"If you're still in love and end up stopping being married to your partner, what emotion are you likely to experience?", "wrong, pleasure, encouragement, depression, relief" -> "

Collapsing token sets: {'A': ['wrong'], 'B': ['pleasure'], 'C': ['encouragement'], 'D': ['depression'], 'E': ['relief']}

Prompt 4 (MI: 0.083, Acc: 0.466):

What would you use to put out a fire?
A: gasoline
B: poison
C: laundry detergent
D: water
E: pencil
Answer: D. water

If you're still in love and end up stopping being married to your partner, what emotion are you likely to experience?
A: wrong
B: pleasure
C: encouragement
D: depression
E: relief
Answer:

Collapsing token sets: ['A', 'B', 'C', 'D', 'E']

Prompt 5 (MI: 0.504, Acc: 0.462):

multiple choice quiz questions and answers

qa = ['q': 'What is France?', 'choices': ['state', 'city', 'country', 'continent', 'mountain range'], 'answer': 'country',], '[q': 'If you're still in love and end up stopping being married to your partner, what emotion are you likely to experience?', 'choices': [wrong, pleasure, encouragement, depression, relief], 'answer': '

Collapsing token sets: {'A': ['wrong'], 'B': ['pleasure'], 'C': ['encouragement'], 'D': ['depression'], 'E': ['relief']}

Prompt 6 (MI: 0.431, Acc: 0.448):

Given the following questions and choices, pick the choice that corresponds best to the question.

"I'm crossing the river, my feet are wet but my body is dry, where am I?", "bridge, waterfall, valley, pebble, mountain", -> "valley"
"If you're still in love and end up stopping being married to your partner, what emotion are you likely to experience?", "wrong, pleasure, encouragement, depression, relief" -> "

Collapsing token sets: {'A': ['wrong'], 'B': ['pleasure'], 'C': ['encouragement'], 'D': ['depression'], 'E': ['relief']}

Prompt 7 (MI: 0.417, Acc: 0.428):

Choose the best single answer to the question, and explain your answer.

Question: I'm crossing the river, my feet are wet but my body is dry, where am I?
Choices: bridge, waterfall, valley, pebble, mountain
Answer: "valley" is the best answer. While "bridge" also seems to make sense at first, your feet would not be wet if you crossed over a river on a bridge. Meanwhile, if you crossed the river at a valley, the river would be shallow, only getting your feet wet.

Question: In what Spanish speaking North American country can you get a great cup of coffee?
Choices: mildred's coffee shop, mexico, diner, kitchen, canteen
Answer: "mexico" is the best answer. It's true that you can get a cup of coffee in a coffee shop or a diner, but the question specifically asks for a Spanish speaking North American country. Mexico is the only country listed, so that must be the correct answer.

Question: If you're still in love and end up stopping being married to your partner, what emotion are you likely to experience?
Choices: wrong, pleasure, encouragement, depression, relief
Answer: "

Collapsing token sets: {'A': ['wrong'], 'B': ['pleasure'], 'C': ['encouragement'], 'D': ['depression'], 'E': ['relief']}

Prompt 8 (MI: 0.364, Acc: 0.408):

Q: What might a vegan eat for breakfast?

Choices: oats, bacon, sausage, omelet, ham

A: oats

Q: If you're still in love and end up stopping being married to your partner, what emotion are you likely to experience?

Choices: wrong, pleasure, encouragement, depression, relief

A:

Collapsing token sets: {'A': ['wrong'], 'B': ['pleasure'], 'C': ['encouragement'], 'D': ['depression'], 'E': ['relief']}

Prompt 9 (MI: 0.410, Acc: 0.408):

What would you use to put out a fire?
A: gasoline
B: poison
C: laundry detergent
D: water
E: pencil
Answer: water

If you're still in love and end up stopping being married to your partner, what emotion are you likely to experience?
A: wrong
B: pleasure
C: encouragement
D: depression
E: relief
Answer:

Collapsing token sets: ['A', 'B', 'C', 'D', 'E']

Prompt 10 (MI: 0.363, Acc: 0.396):

Choose the best single answer to the question, and explain your answer.

Question: I'm crossing the river, my feet are wet but my body is dry, where am I?
Choices: bridge, waterfall, valley, pebble, mountain
Answer: "valley" is the best answer. While "bridge" also seems to make sense at first, your feet would not be wet if you crossed over a river on a bridge. Meanwhile, if you crossed the river at a valley, the river would be shallow, only getting your feet wet.

Question: If you're still in love and end up stopping being married to your partner, what emotion are you likely to experience?
Choices: wrong, pleasure, encouragement, depression, relief
Answer: "

Collapsing token sets: {'A': ['wrong'], 'B': ['pleasure'], 'C': ['encouragement'], 'D': ['depression'], 'E': ['relief']}

Prompt 11 (MI: 0.059, Acc: 0.380):

Common Sense Quiz Answer Key

Question 1: If you're still in love and end up stopping being married to your partner, what emotion are you likely to experience?

A: wrong
B: pleasure
C: encouragement
D: depression
E: relief

Correct Answer:

Collapsing token sets: ['A', 'B', 'C', 'D', 'E']

Prompt 12 (MI: 0.233, Acc: 0.360):

Given the question, order the options from best answer to the question to worst answer to the question.

Question: I'm crossing the river, my feet are wet but my body is dry, where am I?
Choices: bridge, waterfall, valley, pebble, mountain
Answers (in order of best to worst): valley, bridge, waterfall, mountain, pebble

Question: If you're still in love and end up stopping being married to your partner, what emotion are you likely to experience?
Choices: wrong, pleasure, encouragement, depression, relief
Answers (in order of best to worst):

Collapsing token sets: {'A': ['wrong'], 'B': ['pleasure'], 'C': ['encouragement'], 'D': ['depression'], 'E': ['relief']}

Prompt 13 (MI: 0.255, Acc: 0.360):

Given the question, order the options from best answer to the question to worst answer to the question.

Question: I'm crossing the river, my feet are wet but my body is dry, where am I?
Choices: bridge, waterfall, valley, pebble, mountain
Answers (in order of best to worst): valley, bridge, waterfall, mountain, pebble

Question: In what Spanish speaking North American country can you get a great cup of coffee?
Choices: mildred's coffee shop, mexico, diner, kitchen, canteen
Answers (in order of best to worst): mexico, mildred's coffee shop, diner, kitchen, canteen

Question: If you're still in love and end up stopping being married to your partner, what emotion are you likely to experience?
Choices: wrong, pleasure, encouragement, depression, relief
Answers (in order of best to worst):

Collapsing token sets: {'A': ['wrong'], 'B': ['pleasure'], 'C': ['encouragement'], 'D': ['depression'], 'E': ['relief']}

Prompt 14 (MI: 0.222, Acc: 0.354):

Q: If you're still in love and end up stopping being married to your partner, what emotion are you likely to experience?

Choices: wrong, pleasure, encouragement, depression, relief

A:

Collapsing token sets: {'A': ['wrong'], 'B': ['pleasure'], 'C': ['encouragement'], 'D': ['depression'], 'E': ['relief']}

Prompt 15 (MI: 0.246, Acc: 0.342):

Teacher: I'm going to ask you a common sense question.

Student: Alright.

Teacher: If you're still in love and end up stopping being married to your partner, what emotion are you likely to experience?

Student: What are the possible answers?

Teacher: The answer is either "wrong," "pleasure," "encouragement," "depression," or "relief."

Student: I know the right answer - it's "

Collapsing token sets: {'A': ['wrong'], 'B': ['pleasure'], 'C': ['encouragement'], 'D': ['depression'], 'E': ['relief']}

Prompt 16 (MI: 0.376, Acc: 0.336):

questions,choices,answers
"What is France?","[state,city,country,continent,mountain range]",country
"If you're still in love and end up stopping being married to your partner, what emotion are you likely to experience?","[wrong,pleasure,encouragement,depression,relief]",

Collapsing token sets: {'A': ['wrong'], 'B': ['pleasure'], 'C': ['encouragement'], 'D': ['depression'], 'E': ['relief']}

Prompt 17 (MI: 0.265, Acc: 0.276):

Me: I watched the most recent episode of the "Is It Really Common Sense" game show yesterday night.
Friend: Oh, how was it?
Me: It was good. I remember one of the questions.
Friend: What was the question?
Me: If you're still in love and end up stopping being married to your partner, what emotion are you likely to experience?
Friend: What were the options?
Me: wrong, pleasure, encouragement, depression, or relief
Friend: Did the contestant get the answer right?
Me: Yep!
Friend: Which of the options was correct?
Me: The correct answer was

Collapsing token sets: {'A': ['wrong'], 'B': ['pleasure'], 'C': ['encouragement'], 'D': ['depression'], 'E': ['relief']}

Prompt 18 (MI: 0.197, Acc: 0.248):

Given the question, order the options from best answer to the question to worst answer to the question.

Question: If you're still in love and end up stopping being married to your partner, what emotion are you likely to experience?
Choices: wrong, pleasure, encouragement, depression, relief
Answers (in order of best to worst):

Collapsing token sets: {'A': ['wrong'], 'B': ['pleasure'], 'C': ['encouragement'], 'D': ['depression'], 'E': ['relief']}

Prompt 19 (MI: 0.013, Acc: 0.234):

If you're still in love and end up stopping being married to your partner, what emotion are you likely to experience?

A: wrong
B: pleasure
C: encouragement
D: depression
E: relief

Answer:

Collapsing token sets: ['A', 'B', 'C', 'D', 'E']

Prompt 20 (MI: 0.241, Acc: 0.228):

Teacher: I'm going to ask you a common sense question.

Student: Alright.

Teacher: What would you not expect to read about in a book on the founding of the United States?

Student: What are the possible answers?

Teacher: The answer is either "george washington," "declaration of independence," "boston tea party," "star spangled banner," or "vampire assassins."

Student: I know the right answer - it's "vampire assassins."

Teacher: That's right! Here's another common sense question for you. If you're still in love and end up stopping being married to your partner, what emotion are you likely to experience?

Student: What are the possible answers?

Teacher: The answer is either "wrong," "pleasure," "encouragement," "depression," or "relief."

Student: I know the right answer - it's "

Collapsing token sets: {'A': ['wrong'], 'B': ['pleasure'], 'C': ['encouragement'], 'D': ['depression'], 'E': ['relief']}

C.5 IMDB

Prompt 1 (MI: 0.175, Acc: 0.944):

P1: How was the movie?
P2: John Cassavetes is on the run from the law. He is at the bottom of the heap. He sees Negro Sidney Poitier as his equal and they quickly become friends, forming a sort of alliance against a bully of a foreman played by Jack Warden.

As someone who has worked in a warehouse myself when I was younger, I can tell you that the warehouse fights, complete with tumbling packing cases and flailing grappling hooks are as realistic as it gets. I've been in fights like these myself, although no one got killed.

The introduction of Sidney Poitier's widow is a variation on Shakespeare's Shylock "Do I not bleed?" This is an anti racist film, which, at the time, was much needed.

All the three principle characters - Warden, Cassavetes and Poitier - are superb, with Warden the most outstanding of the three.
P1: Would you say your review of the movie is negative or positive?
P2: I would say my review review of the movie is

Collapsing token sets: {'positive': ['positive'], 'negative': ['negative']}

Prompt 2 (MI: 0.306, Acc: 0.920):

P1: Could you give me a review of the movie you just saw?
P2: Sure, John Cassavetes is on the run from the law. He is at the bottom of the heap. He sees Negro Sidney Poitier as his equal and they quickly become friends, forming a sort of alliance against a bully of a foreman played by Jack Warden.

As someone who has worked in a warehouse myself when I was younger, I can tell you that the warehouse fights, complete with tumbling packing cases and flailing grappling hooks are as realistic as it gets. I've been in fights like these myself, although no one got killed.

The introduction of Sidney Poitier's widow is a variation on Shakespeare's Shylock "Do I not bleed?" This is an anti racist film, which, at the time, was much needed.

All the three principle characters - Warden, Cassavetes and Poitier - are superb, with Warden the most outstanding of the three.
P1: So, overall, would you give it a positive or negative review?
P2: I would give it a

Collapsing token sets: {'positive': ['positive'], 'negative': ['negative']}

Prompt 3 (MI: 0.154, Acc: 0.904):

Considering this movie review, determine its sentiment.

Review: John Cassavetes is on the run from the law. He is at the bottom of the heap. He sees Negro Sidney Poitier as his equal and they quickly become friends, forming a sort of alliance against a bully of a foreman played by Jack Warden.

As someone who has worked in a warehouse myself when I was younger, I can tell you that the warehouse fights, complete with tumbling packing cases and flailing grappling hooks are as realistic as it gets. I've been in fights like these myself, although no one got killed.

The introduction of Sidney Poitier's widow is a variation on Shakespeare's Shylock "Do I not bleed?" This is an anti racist film, which, at the time, was much needed.

All the three principle characters - Warden, Cassavetes and Poitier - are superb, with Warden the most outstanding of the three.

In general, was the sentiment positive or negative The sentiment was

Collapsing token sets: {'positive': ['positive'], 'negative': ['negative']}

Prompt 4 (MI: 0.260, Acc: 0.898):

P1: How was the movie?
P2: John Cassavetes is on the run from the law. He is at the bottom of the heap. He sees Negro Sidney Poitier as his equal and they quickly become friends, forming a sort of alliance against a bully of a foreman played by Jack Warden.

As someone who has worked in a warehouse myself when I was younger, I can tell you that the warehouse fights, complete with tumbling packing cases and flailing grappling hooks are as realistic as it gets. I've been in fights like these myself, although no one got killed.

The introduction of Sidney Poitier's widow is a variation on Shakespeare's Shylock "Do I not bleed?" This is an anti racist film, which, at the time, was much needed.

All the three principle characters - Warden, Cassavetes and Poitier - are superb, with Warden the most outstanding of the three.
P1: Would you say your review of the movie is positive or negative?
P2: I would say my review of the movie is

Collapsing token sets: {'positive': ['positive'], 'negative': ['negative']}

Prompt 5 (MI: 0.237, Acc: 0.888):

After reading the following review, classify it as negative or positive.

Review: John Cassavetes is on the run from the law. He is at the bottom of the heap. He sees Negro Sidney Poitier as his equal and they quickly become friends, forming a sort of alliance against a bully of a foreman played by Jack Warden.

As someone who has worked in a warehouse myself when I was younger, I can tell you that the warehouse fights, complete with tumbling packing cases and flailing grappling hooks are as realistic as it gets. I've been in fights like these myself, although no one got killed.

The introduction of Sidney Poitier's widow is a variation on Shakespeare's Shylock "Do I not bleed?" This is an anti racist film, which, at the time, was much needed.

All the three principle characters - Warden, Cassavetes and Poitier - are superb, with Warden the most outstanding of the three.

Classification:

Collapsing token sets: {'positive': ['positive'], 'negative': ['negative']}

Prompt 6 (MI: 0.151, Acc: 0.886):

Read the following movie review to determine the review's sentiment.

John Cassavetes is on the run from the law. He is at the bottom of the heap. He sees Negro Sidney Poitier as his equal and they quickly become friends, forming a sort of alliance against a bully of a foreman played by Jack Warden.

As someone who has worked in a warehouse myself when I was younger, I can tell you that the warehouse fights, complete with tumbling packing cases and flailing grappling hooks are as realistic as it gets. I've been in fights like these myself, although no one got killed.

The introduction of Sidney Poitier's widow is a variation on Shakespeare's Shylock "Do I not bleed?" This is an anti racist film, which, at the time, was much needed.

All the three principle characters - Warden, Cassavetes and Poitier - are superb, with Warden the most outstanding of the three.

In general, was the sentiment positive or negative? The sentiment was

Collapsing token sets: {'positive': ['positive'], 'negative': ['negative']}

Prompt 7 (MI: 0.086, Acc: 0.886):

Considering this movie review, determine its sentiment.

Review:
"""
John Cassavetes is on the run from the law. He is at the bottom of the heap. He sees Negro Sidney Poitier as his equal and they quickly become friends, forming a sort of alliance against a bully of a foreman played by Jack Warden.

As someone who has worked in a warehouse myself when I was younger, I can tell you that the warehouse fights, complete with tumbling packing cases and flailing grappling hooks are as realistic as it gets. I've been in fights like these myself, although no one got killed.

The introduction of Sidney Poitier's widow is a variation on Shakespeare's Shylock "Do I not bleed?" This is an anti racist film, which, at the time, was much needed.

All the three principle characters - Warden, Cassavetes and Poitier - are superb, with Warden the most outstanding of the three.
"""
In general, what was the sentiment of the review? The sentiment was

Collapsing token sets: {'positive': ['positive'], 'negative': ['negative']}

Prompt 8 (MI: 0.274, Acc: 0.858):

Yesterday I went to see a movie. John Cassavetes is on the run from the law. He is at the bottom of the heap. He sees Negro Sidney Poitier as his equal and they quickly become friends, forming a sort of alliance against a bully of a foreman played by Jack Warden.

As someone who has worked in a warehouse myself when I was younger, I can tell you that the warehouse fights, complete with tumbling packing cases and flailing grappling hooks are as realistic as it gets. I've been in fights like these myself, although no one got killed.

The introduction of Sidney Poitier's widow is a variation on Shakespeare's Shylock "Do I not bleed?" This is an anti racist film, which, at the time, was much needed.

All the three principle characters - Warden, Cassavetes and Poitier - are superb, with Warden the most outstanding of the three. Between positive and negative, I would say the movie was

Collapsing token sets: {'positive': ['positive'], 'negative': ['negative']}

Prompt 9 (MI: 0.026, Acc: 0.852):

Q: Is the sentiment of the following movie review negative or positive?
"""
John Cassavetes is on the run from the law. He is at the bottom of the heap. He sees Negro Sidney Poitier as his equal and they quickly become friends, forming a sort of alliance against a bully of a foreman played by Jack Warden.

As someone who has worked in a warehouse myself when I was younger, I can tell you that the warehouse fights, complete with tumbling packing cases and flailing grappling hooks are as realistic as it gets. I've been in fights like these myself, although no one got killed.

The introduction of Sidney Poitier's widow is a variation on Shakespeare's Shylock "Do I not bleed?" This is an anti racist film, which, at the time, was much needed.

All the three principle characters - Warden, Cassavetes and Poitier - are superb, with Warden the most outstanding of the three.
"""
A: The sentiment of the movie review was

Collapsing token sets: {'positive': ['positive'], 'negative': ['negative']}

Prompt 10 (MI: 0.119, Acc: 0.842):

Read the following movie review to determine the review's sentiment.

John Cassavetes is on the run from the law. He is at the bottom of the heap. He sees Negro Sidney Poitier as his equal and they quickly become friends, forming a sort of alliance against a bully of a foreman played by Jack Warden.

As someone who has worked in a warehouse myself when I was younger, I can tell you that the warehouse fights, complete with tumbling packing cases and flailing grappling hooks are as realistic as it gets. I've been in fights like these myself, although no one got killed.

The introduction of Sidney Poitier's widow is a variation on Shakespeare's Shylock "Do I not bleed?" This is an anti racist film, which, at the time, was much needed.

All the three principle characters - Warden, Cassavetes and Poitier - are superb, with Warden the most outstanding of the three.

In general, was the sentiment negative or positive? The sentiment was

Collapsing token sets: {'positive': ['positive'], 'negative': ['negative']}

Prompt 11 (MI: 0.162, Acc: 0.824):

Q: Is the sentiment of the following movie review positive or negative?
John Cassavetes is on the run from the law. He is at the bottom of the heap. He sees Negro Sidney Poitier as his equal and they quickly become friends, forming a sort of alliance against a bully of a foreman played by Jack Warden.

As someone who has worked in a warehouse myself when I was younger, I can tell you that the warehouse fights, complete with tumbling packing cases and flailing grappling hooks are as realistic as it gets. I've been in fights like these myself, although no one got killed.

The introduction of Sidney Poitier's widow is a variation on Shakespeare's Shylock "Do I not bleed?" This is an anti racist film, which, at the time, was much needed.

All the three principle characters - Warden, Cassavetes and Poitier - are superb, with Warden the most outstanding of the three.
A (positive or negative):

Collapsing token sets: {'positive': ['positive'], 'negative': ['negative']}

Prompt 12 (MI: 0.101, Acc: 0.822):

Q: Is the sentiment of the following movie review negative or positive?
John Cassavetes is on the run from the law. He is at the bottom of the heap. He sees Negro Sidney Poitier as his equal and they quickly become friends, forming a sort of alliance against a bully of a foreman played by Jack Warden.

As someone who has worked in a warehouse myself when I was younger, I can tell you that the warehouse fights, complete with tumbling packing cases and flailing grappling hooks are as realistic as it gets. I've been in fights like these myself, although no one got killed.

The introduction of Sidney Poitier's widow is a variation on Shakespeare's Shylock "Do I not bleed?" This is an anti racist film, which, at the time, was much needed.

All the three principle characters - Warden, Cassavetes and Poitier - are superb, with Warden the most outstanding of the three.
A (negative or positive):

Collapsing token sets: {'positive': ['positive'], 'negative': ['negative']}

Prompt 13 (MI: 0.084, Acc: 0.810):

Considering this movie review, determine its sentiment.

Review:
"""
John Cassavetes is on the run from the law. He is at the bottom of the heap. He sees Negro Sidney Poitier as his equal and they quickly become friends, forming a sort of alliance against a bully of a foreman played by Jack Warden.

As someone who has worked in a warehouse myself when I was younger, I can tell you that the warehouse fights, complete with tumbling packing cases and flailing grappling hooks are as realistic as it gets. I've been in fights like these myself, although no one got killed.

The introduction of Sidney Poitier's widow is a variation on Shakespeare's Shylock "Do I not bleed?" This is an anti racist film, which, at the time, was much needed.

All the three principle characters - Warden, Cassavetes and Poitier - are superb, with Warden the most outstanding of the three.
"""
In general, was the sentiment positive or negative? The sentiment was

Collapsing token sets: {'positive': ['positive'], 'negative': ['negative']}

Prompt 14 (MI: 0.201, Acc: 0.798):

P1: Could you give me a review of the movie you just saw?
P2: Sure, John Cassavetes is on the run from the law. He is at the bottom of the heap. He sees Negro Sidney Poitier as his equal and they quickly become friends, forming a sort of alliance against a bully of a foreman played by Jack Warden.

As someone who has worked in a warehouse myself when I was younger, I can tell you that the warehouse fights, complete with tumbling packing cases and flailing grappling hooks are as realistic as it gets. I've been in fights like these myself, although no one got killed.

The introduction of Sidney Poitier's widow is a variation on Shakespeare's Shylock "Do I not bleed?" This is an anti racist film, which, at the time, was much needed.

All the three principle characters - Warden, Cassavetes and Poitier - are superb, with Warden the most outstanding of the three.
P1: So overall was the sentiment of the movie negative or positive?
P2: I would give it a

Collapsing token sets: {'positive': ['positive'], 'negative': ['negative']}

Prompt 15 (MI: 0.234, Acc: 0.786):

After reading the following review, classify it as positive or negative.

Review: John Cassavetes is on the run from the law. He is at the bottom of the heap. He sees Negro Sidney Poitier as his equal and they quickly become friends, forming a sort of alliance against a bully of a foreman played by Jack Warden.

As someone who has worked in a warehouse myself when I was younger, I can tell you that the warehouse fights, complete with tumbling packing cases and flailing grappling hooks are as realistic as it gets. I've been in fights like these myself, although no one got killed.

The introduction of Sidney Poitier's widow is a variation on Shakespeare's Shylock "Do I not bleed?" This is an anti racist film, which, at the time, was much needed.

All the three principle characters - Warden, Cassavetes and Poitier - are superb, with Warden the most outstanding of the three.

Classification:

Collapsing token sets: {'positive': ['positive'], 'negative': ['negative']}

Prompt 16 (MI: 0.042, Acc: 0.628):

Q: Is the sentiment of the following movie review positive or negative?
"""
John Cassavetes is on the run from the law. He is at the bottom of the heap. He sees Negro Sidney Poitier as his equal and they quickly become friends, forming a sort of alliance against a bully of a foreman played by Jack Warden.

As someone who has worked in a warehouse myself when I was younger, I can tell you that the warehouse fights, complete with tumbling packing cases and flailing grappling hooks are as realistic as it gets. I've been in fights like these myself, although no one got killed.

The introduction of Sidney Poitier's widow is a variation on Shakespeare's Shylock "Do I not bleed?" This is an anti racist film, which, at the time, was much needed.

All the three principle characters - Warden, Cassavetes and Poitier - are superb, with Warden the most outstanding of the three.
"""
A: The sentiment of the movie review was

Collapsing token sets: {'positive': ['positive'], 'negative': ['negative']}

Prompt 17 (MI: 0.021, Acc: 0.486):

John Cassavetes is on the run from the law. He is at the bottom of the heap. He sees Negro Sidney Poitier as his equal and they quickly become friends, forming a sort of alliance against a bully of a foreman played by Jack Warden.

As someone who has worked in a warehouse myself when I was younger, I can tell you that the warehouse fights, complete with tumbling packing cases and flailing grappling hooks are as realistic as it gets. I've been in fights like these myself, although no one got killed.

The introduction of Sidney Poitier's widow is a variation on Shakespeare's Shylock "Do I not bleed?" This is an anti racist film, which, at the time, was much needed.

All the three principle characters - Warden, Cassavetes and Poitier - are superb, with Warden the most outstanding of the three.

Was the previous review negative or positive? The previous review was

Collapsing token sets: {'positive': ['positive'], 'negative': ['negative']}

Prompt 18 (MI: 0.016, Acc: 0.484):

John Cassavetes is on the run from the law. He is at the bottom of the heap. He sees Negro Sidney Poitier as his equal and they quickly become friends, forming a sort of alliance against a bully of a foreman played by Jack Warden.

As someone who has worked in a warehouse myself when I was younger, I can tell you that the warehouse fights, complete with tumbling packing cases and flailing grappling hooks are as realistic as it gets. I've been in fights like these myself, although no one got killed.

The introduction of Sidney Poitier's widow is a variation on Shakespeare's Shylock "Do I not bleed?" This is an anti racist film, which, at the time, was much needed.

All the three principle characters - Warden, Cassavetes and Poitier - are superb, with Warden the most outstanding of the three.

Was the previous review positive or negative? The previous review was

Collapsing token sets: {'positive': ['positive'], 'negative': ['negative']}

Prompt 19 (MI: 0.019, Acc: 0.462):

John Cassavetes is on the run from the law. He is at the bottom of the heap. He sees Negro Sidney Poitier as his equal and they quickly become friends, forming a sort of alliance against a bully of a foreman played by Jack Warden.

As someone who has worked in a warehouse myself when I was younger, I can tell you that the warehouse fights, complete with tumbling packing cases and flailing grappling hooks are as realistic as it gets. I've been in fights like these myself, although no one got killed.

The introduction of Sidney Poitier's widow is a variation on Shakespeare's Shylock "Do I not bleed?" This is an anti racist film, which, at the time, was much needed.

All the three principle characters - Warden, Cassavetes and Poitier - are superb, with Warden the most outstanding of the three.

Was the sentiment of previous review positive or negative? The previous review was

Collapsing token sets: {'positive': ['positive'], 'negative': ['negative']}

Prompt 20 (MI: 0.017, Acc: 0.450):

John Cassavetes is on the run from the law. He is at the bottom of the heap. He sees Negro Sidney Poitier as his equal and they quickly become friends, forming a sort of alliance against a bully of a foreman played by Jack Warden.

As someone who has worked in a warehouse myself when I was younger, I can tell you that the warehouse fights, complete with tumbling packing cases and flailing grappling hooks are as realistic as it gets. I've been in fights like these myself, although no one got killed.

The introduction of Sidney Poitier's widow is a variation on Shakespeare's Shylock "Do I not bleed?" This is an anti racist film, which, at the time, was much needed.

All the three principle characters - Warden, Cassavetes and Poitier - are superb, with Warden the most outstanding of the three.

Was the sentiment of previous review negative or positive? The previous review was

Collapsing token sets: {'positive': ['positive'], 'negative': ['negative']}

C.6 BoolQ

Prompt 1 (MI: 0.077, Acc: 0.778):

Given the passage and question, please answer the question with yes or no.

"'Turn on red – In Canada, left turn on red light from a one-way road into a one-way road is permitted except in some areas of Quebec, New Brunswick, and Prince Edward Island. Left turn on red light from a two-way road into a one-way road is permitted in British Columbia but only if the driver turns onto the closest lane and yields to pedestrians and cross traffic."', "'Can you turn left on red in canada?"' -> "'Yes"'

"'Pyruvic acid – Pyruvic acid (CHCOCOOH) is the simplest of the alpha-keto acids, with a carboxylic acid and a ketone functional group. Pyruvate (/paruvet/), the conjugate base, CHCOCOO, is a key intermediate in several metabolic pathways."', "'Is pyruvic acid and pyruvate the same thing?"' -> "'

Collapsing token sets: {'True': ['yes'], 'False': ['no']}

Prompt 2 (MI: 0.090, Acc: 0.752):

Passage: "Turn on red – In Canada, left turn on red light from a one-way road into a one-way road is permitted except in some areas of Quebec, New Brunswick, and Prince Edward Island. Left turn on red light from a two-way road into a one-way road is permitted in British Columbia but only if the driver turns onto the closest lane and yields to pedestrians and cross traffic."
Question: "Can you turn left on red in canada?"
Answer: "Yes"

Passage: "Pyruvic acid – Pyruvic acid (CHCOCOOH) is the simplest of the alpha-keto acids, with a carboxylic acid and a ketone functional group. Pyruvate (/paruvet/), the conjugate base, CHCOCOO, is a key intermediate in several metabolic pathways."
Question: "Is pyruvic acid and pyruvate the same thing?"
Answer: "

Collapsing token sets: {'True': ['yes'], 'False': ['no']}

Prompt 3 (MI: 0.055, Acc: 0.750):

Given the passage and question, please answer the question with yes or no.

"'Turn on red – In Canada, left turn on red light from a one-way road into a one-way road is permitted except in some areas of Quebec, New Brunswick, and Prince Edward Island. Left turn on red light from a two-way road into a one-way road is permitted in British Columbia but only if the driver turns onto the closest lane and yields to pedestrians and cross traffic."', "'Can you turn left on red in canada?"' -> "'Yes"'

"'Lord Voldemort – Lord Voldemort (known as Tom Marvolo Riddle) is a fictional character and the main antagonist in J.K. Rowling's series of Harry Potter novels. Voldemort first appeared in Harry Potter and the Philosopher's Stone, which was released in 1997. Voldemort appears either in person or in flashbacks in each book and its film adaptation in the series, except the third, Harry Potter and the Prisoner of Azkaban, where he is only mentioned."', "'Are tom riddle and lord voldemort the same person?"' -> "'Yes"'

"'Clerks – Clerks is a 1994 American independent black-and-white comedy film written, directed and co-produced by Kevin Smith. Starring Brian O'Halloran as Dante Hicks and Jeff Anderson as Randal Graves, it presents a day in the lives of two store clerks and their acquaintances."', "'Is the movie clerks in colors?"' -> "'No"'

"'Pyruvic acid – Pyruvic acid (CHCOCOOH) is the simplest of the alpha-keto acids, with a carboxylic acid and a ketone functional group. Pyruvate (/paruvet/), the conjugate base, CHCOCOO, is a key intermediate in several metabolic pathways."', "'Is pyruvic acid and pyruvate the same thing?"' -> "'

Collapsing token sets: {'True': ['yes'], 'False': ['no']}

Prompt 4 (MI: 0.076, Acc: 0.740):

Passage: "Turn on red – In Canada, left turn on red light from a one-way road into a one-way road is permitted except in some areas of Quebec, New Brunswick, and Prince Edward Island. Left turn on red light from a two-way road into a one-way road is permitted in British Columbia but only if the driver turns onto the closest lane and yields to pedestrians and cross traffic."
Question: "Can you turn left on red in canada?"
Answer: "Yes"

Passage: "Lord Voldemort – Lord Voldemort (known as Tom Marvolo Riddle) is a fictional character and the main antagonist in J.K. Rowling's series of Harry Potter novels. Voldemort first appeared in Harry Potter and the Philosopher's Stone, which was released in 1997. Voldemort appears either in person or in flashbacks in each book and its film adaptation in the series, except the third, Harry Potter and the Prisoner of Azkaban, where he is only mentioned."
Question: "Are tom riddle and lord voldemort the same person?"
Answer: "Yes"

Passage: "Clerks – Clerks is a 1994 American independent black-and-white comedy film written, directed and co-produced by Kevin Smith. Starring Brian O'Halloran as Dante Hicks and Jeff Anderson as Randal Graves, it presents a day in the lives of two store clerks and their acquaintances."
Question: "Is the movie clerks in colors?"
Answer: "No"

Passage: "Pyruvic acid – Pyruvic acid (CHCOCOOH) is the simplest of the alpha-keto acids, with a carboxylic acid and a ketone functional group. Pyruvate (/paruvet/), the conjugate base, CHCOCOO, is a key intermediate in several metabolic pathways."
Question: "Is pyruvic acid and pyruvate the same thing?"
Answer: "

Collapsing token sets: {'True': ['yes'], 'False': ['no']}

Prompt 5 (MI: 0.037, Acc: 0.740):

Given the passage and question, please answer the question with yes or no.

"'Turn on red – In Canada, left turn on red light from a one-way road into a one-way road is permitted except in some areas of Quebec, New Brunswick, and Prince Edward Island. Left turn on red light from a two-way road into a one-way road is permitted in British Columbia but only if the driver turns onto the closest lane and yields to pedestrians and cross traffic."', "'Can you turn left on red in canada?"' -> "'Yes"'

"'Lord Voldemort – Lord Voldemort (known as Tom Marvolo Riddle) is a fictional character and the main antagonist in J.K. Rowling's series of Harry Potter novels. Voldemort first appeared in Harry Potter and the Philosopher's Stone, which was released in 1997. Voldemort appears either in person or in flashbacks in each book and its film adaptation in the series, except the third, Harry Potter and the Prisoner of Azkaban, where he is only mentioned."', "'Are tom riddle and lord voldemort the same person?"' -> "'Yes"'

"'Pyruvic acid – Pyruvic acid (CHCOCOOH) is the simplest of the alpha-keto acids, with a carboxylic acid and a ketone functional group. Pyruvate (/paruvet/), the conjugate base, CHCOCOO, is a key intermediate in several metabolic pathways."', "'Is pyruvic acid and pyruvate the same thing?"' -> "'

Collapsing token sets: {'True': ['yes'], 'False': ['no']}

Prompt 6 (MI: 0.068, Acc: 0.702):

"Pyruvic acid – Pyruvic acid (CHCOCOOH) is the simplest of the alpha-keto acids, with a carboxylic acid and a ketone functional group. Pyruvate (/paruvet/), the conjugate base, CHCOCOO, is a key intermediate in several metabolic pathways."

For the question: "Is pyruvic acid and pyruvate the same thing?"
I would answer: "

Collapsing token sets: {'True': ['yes'], 'False': ['no']}

Prompt 7 (MI: 0.039, Acc: 0.698):

"Pyruvic acid – Pyruvic acid (CHCOCOOH) is the simplest of the alpha-keto acids, with a carboxylic acid and a ketone functional group. Pyruvate (/paruvet/), the conjugate base, CHCOCOO, is a key intermediate in several metabolic pathways."

When picking between yes or no For the question: "Is pyruvic acid and pyruvate the same thing?"
I would answer: "

Collapsing token sets: {'True': ['yes'], 'False': ['no']}

Prompt 8 (MI: 0.034, Acc: 0.698):

ANSWER KEY

Please read the following passage with the following question in mind: "Is pyruvic acid and pyruvate the same thing?"

Pyruvic acid – Pyruvic acid (CHCOCOOH) is the simplest of the alpha-keto acids, with a carboxylic acid and a ketone functional group. Pyruvate (/paruvet/), the conjugate base, CHCOCOO, is a key intermediate in several metabolic pathways.

Is pyruvic acid and pyruvate the same thing?
Answer key: "

Collapsing token sets: {'True': ['yes'], 'False': ['no']}

Prompt 9 (MI: 0.055, Acc: 0.688):

Passage: "Turn on red – In Canada, left turn on red light from a one-way road into a one-way road is permitted except in some areas of Quebec, New Brunswick, and Prince Edward Island. Left turn on red light from a two-way road into a one-way road is permitted in British Columbia but only if the driver turns onto the closest lane and yields to pedestrians and cross traffic."
Question: "Can you turn left on red in canada?"
Answer: "Yes"

Passage: "Lord Voldemort – Lord Voldemort (known as Tom Marvolo Riddle) is a fictional character and the main antagonist in J.K. Rowling's series of Harry Potter novels. Voldemort first appeared in Harry Potter and the Philosopher's Stone, which was released in 1997. Voldemort appears either in person or in flashbacks in each book and its film adaptation in the series, except the third, Harry Potter and the Prisoner of Azkaban, where he is only mentioned."
Question: "Are tom riddle and lord voldemort the same person?"
Answer: "Yes"

Passage: "Pyruvic acid – Pyruvic acid (CHCOCOOH) is the simplest of the alpha-keto acids, with a carboxylic acid and a ketone functional group. Pyruvate (/paruvet/), the conjugate base, CHCOCOO, is a key intermediate in several metabolic pathways."
Question: "Is pyruvic acid and pyruvate the same thing?"
Answer: "

Collapsing token sets: {'True': ['yes'], 'False': ['no']}

Prompt 10 (MI: 0.052, Acc: 0.682):

"Pyruvic acid – Pyruvic acid (CHCOCOOH) is the simplest of the alpha-keto acids, with a carboxylic acid and a ketone functional group. Pyruvate (/paruvet/), the conjugate base, CHCOCOO, is a key intermediate in several metabolic pathways."

For the question: "Is pyruvic acid and pyruvate the same thing?"
My answer would be: "

Collapsing token sets: {'True': ['yes'], 'False': ['no']}

Prompt 11 (MI: 0.026, Acc: 0.682):

Given the passage and question, please answer the question with yes or no.

"'Pyruvic acid – Pyruvic acid (CHCOCOOH) is the simplest of the alpha-keto acids, with a carboxylic acid and a ketone functional group. Pyruvate (/paruvet/), the conjugate base, CHCOCOO, is a key intermediate in several metabolic pathways."', "'Is pyruvic acid and pyruvate the same thing?"' -> "'

Collapsing token sets: {'True': ['yes'], 'False': ['no']}

Prompt 12 (MI: 0.016, Acc: 0.680):

"Pyruvic acid – Pyruvic acid (CHCOCOOH) is the simplest of the alpha-keto acids, with a carboxylic acid and a ketone functional group. Pyruvate (/paruvet/), the conjugate base, CHCOCOO, is a key intermediate in several metabolic pathways."

When picking between "true" or "false", For the question: "Is pyruvic acid and pyruvate the same thing?"
My answer would be: "

Collapsing token sets: {'True': ['true'], 'False': ['false']}

Prompt 13 (MI: 0.074, Acc: 0.674):

Please read the following passage with the following question in mind: "Is pyruvic acid and pyruvate the same thing?"

Pyruvic acid – Pyruvic acid (CHCOCOOH) is the simplest of the alpha-keto acids, with a carboxylic acid and a ketone functional group. Pyruvate (/paruvet/), the conjugate base, CHCOCOO, is a key intermediate in several metabolic pathways.

Is pyruvic acid and pyruvate the same thing?
Answer: "

Collapsing token sets: {'True': ['yes'], 'False': ['no']}

Prompt 14 (MI: 0.050, Acc: 0.668):

Read the following passage: "Pyruvic acid – Pyruvic acid (CHCOCOOH) is the simplest of the alpha-keto acids, with a carboxylic acid and a ketone functional group. Pyruvate (/paruvet/), the conjugate base, CHCOCOO, is a key intermediate in several metabolic pathways."

Given this question: "Is pyruvic acid and pyruvate the same thing?"
I would answer: "

Collapsing token sets: {'True': ['yes'], 'False': ['no']}

Prompt 15 (MI: 0.058, Acc: 0.646):

Read the following passage: "Pyruvic acid – Pyruvic acid (CHCOCOOH) is the simplest of the alpha-keto acids, with a carboxylic acid and a ketone functional group. Pyruvate (/paruvet/), the conjugate base, CHCOCOO, is a key intermediate in several metabolic pathways."

Given this question: "Is pyruvic acid and pyruvate the same thing?"
I would respond: "

Collapsing token sets: {'True': ['yes'], 'False': ['no']}

Prompt 16 (MI: 0.027, Acc: 0.634):

Based on the passage: "Pyruvic acid – Pyruvic acid (CHCOCOOH) is the simplest of the alpha-keto acids, with a carboxylic acid and a ketone functional group. Pyruvate (/paruvet/), the conjugate base, CHCOCOO, is a key intermediate in several metabolic pathways."

And answering the question: "Is pyruvic acid and pyruvate the same thing?"
By choosing yes or no
My answer would be: "

Collapsing token sets: {'True': ['yes'], 'False': ['no']}

Prompt 17 (MI: 0.013, Acc: 0.522):

Read the following passage: "Pyruvic acid – Pyruvic acid (CHCOCOOH) is the simplest of the alpha-keto acids, with a carboxylic acid and a ketone functional group. Pyruvate (/paruvet/), the conjugate base, CHCOCOO, is a key intermediate in several metabolic pathways.

Given this question: "Is pyruvic acid and pyruvate the same thing?"

If asked to choose "true" or "false", My answer would be: "

Collapsing token sets: {'True': ['true'], 'False': ['false']}

Prompt 18 (MI: 0.020, Acc: 0.518):

Read the following passage: "Pyruvic acid – Pyruvic acid (CHCOCOOH) is the simplest of the alpha-keto acids, with a carboxylic acid and a ketone functional group. Pyruvate (/paruvet/), the conjugate base, CHCOCOO, is a key intermediate in several metabolic pathways."

Given this question: "Is pyruvic acid and pyruvate the same thing?"
If asked to choose yes or no, My answer would be: "

Collapsing token sets: {'True': ['yes'], 'False': ['no']}

Prompt 19 (MI: 0.013, Acc: 0.452):

Read the following passage: "Pyruvic acid – Pyruvic acid (CHCOCOOH) is the simplest of the alpha-keto acids, with a carboxylic acid and a ketone functional group. Pyruvate (/paruvet/), the conjugate base, CHCOCOO, is a key intermediate in several metabolic pathways."

Given this question: "Is pyruvic acid and pyruvate the same thing?"
If asked to choose "true" or "false", I would answer: "

Collapsing token sets: {'True': ['true'], 'False': ['false']}

Prompt 20 (MI: 0.022, Acc: 0.438):

Read the following passage: "Pyruvic acid – Pyruvic acid (CHCOCOOH) is the simplest of the alpha-keto acids, with a carboxylic acid and a ketone functional group. Pyruvate (/paruvet/), the conjugate base, CHCOCOO, is a key intermediate in several metabolic pathways."

Given this question: "Is pyruvic acid and pyruvate the same thing?"
If asked to choose yes or no, I would answer: "

Collapsing token sets: {'True': ['yes'], 'False': ['no']}

C.7 COPA

Prompt 1 (MI: 0.044, Acc: 0.782):

For the following premises, choose the alternative that is either a cause or result of the premise, and justify your answer.

Premise: The man broke his toe. What was the CAUSE of this?
Alternative 1: He got a hole in his sock.
Alternative 2: He dropped a hammer on his foot.
Answer: Alternative 2. Getting a hole in your sock would not break your toe, unless there is additional information. Dropping a hammer (which is a heavy object), on the other hand, would almost certaintly break your toe. Thus, the best answer is Alternative 2.

Premise: I tipped the bottle. What happened as a RESULT?
Alternative 1: The liquid in the bottle froze.
Alternative 2: The liquid in the bottle poured out.
Answer: Alternative 2. Tipping a bottle causes liquid to fall out, not to freeze. Freezing is caused by being placed in a cold place. Pouring out (Alternative 2) is correct because it makes the most sense.

Premise: I knocked on my neighbor's door. What happened as a RESULT?
Alternative 1: My neighbor invited me in.
Alternative 2: My neighbor left his house.
Answer: Alternative 1. When you knock on a neighbor's door, it is likely that if they are home they will answer and invite you in. It does not make much sense, however, that a neighbor would leave their house without explanation. Therefore, Alternative 1 is the best result of the premise.

Premise: My foot went numb. What happened as a RESULT?
Alternative 1: I put my shoes on.
Alternative 2: I shook my foot.
Answer: Alternative

Collapsing token sets: {'1': ['1'], '2': ['2']}

Prompt 2 (MI: 0.034, Acc: 0.762):

The Choice Of Plausible Alternatives (COPA) evaluation provides researchers with a tool for assessing progress in open-domain commonsense causal reasoning. COPA consists of 1000 questions, split equally into development and test sets of 500 questions each. Each question is composed of a premise and two alternatives, where the task is to select the alternative that more plausibly has a causal relation with the premise. The correct alternative is randomized so that the expected performance of randomly guessing is 50%.

Examples

Premise: The man broke his toe. What was the CAUSE of this?
Alternative 1: He got a hole in his sock.
Alternative 2: He dropped a hammer on his foot.
Answer: Alternative 2

Premise: I tipped the bottle. What happened as a RESULT?
Alternative 1: The liquid in the bottle froze.
Alternative 2: The liquid in the bottle poured out.
Answer: Alternative 2

Premise: I knocked on my neighbor's door. What happened as a RESULT?
Alternative 1: My neighbor invited me in.
Alternative 2: My neighbor left his house.
Answer: Alternative 1

Premise: My foot went numb. What happened as a RESULT?
Alternative 1: I put my shoes on.
Alternative 2: I shook my foot.
Answer: Alternative

Collapsing token sets: {'1': ['1'], '2': ['2']}

Prompt 3 (MI: 0.003, Acc: 0.628):

What is the effect of the following premise: "My foot went numb."

Choice 1. I put my shoes on.
Choice 2. I shook my foot.
Answer: Choice

Collapsing token sets: {'1': ['1'], '2': ['2']}

Prompt 4 (MI: 0.002, Acc: 0.612):

Solve the following COPA task by choosing the sentence which makes the most sense after the premise.

Premise: My foot went numb.
Choice 1. I put my shoes on.
Choice 2. I shook my foot.
Answer: Choice

Collapsing token sets: {'1': ['1'], '2': ['2']}

Prompt 5 (MI: 0.003, Acc: 0.550):

If asked to pick between choice 1 ("I put my shoes on.") or choice 2 ("I shook my foot.") to see what the effect of this premise ("My foot went numb.") was, I would say: "choice

Collapsing token sets: {'1': ['1'], '2': ['2']}

Prompt 6 (MI: 0.010, Acc: 0.540):

Solve the following COPA tasks by choosing the sentence which makes the most sense after the premise.

Premise: The man broke his toe.
Choice 1. He got a hole in his sock.
Choice 2. He dropped a hammer on his foot.
Answer: Choice 2.

Premise: My foot went numb.
Choice 1. I put my shoes on.
Choice 2. I shook my foot.
Answer: Choice

Collapsing token sets: {'1': ['1'], '2': ['2']}

Prompt 7 (MI: 0.002, Acc: 0.532):

What is the effect of the following premise: "My foot went numb."

If asked to choose between Choice 1: "I put my shoes on." or Choice 2: "I shook my foot."
My answer would be: Choice

Collapsing token sets: {'1': ['1'], '2': ['2']}

Prompt 8 (MI: 0.006, Acc: 0.530):

I will give you a premise and you will choose either sentence 1) or 2) which is the better plausible alternative.
Premise: My foot went numb.
1) I put my shoes on.
2) I shook my foot.
The most plausible alternative is: Sentence

Collapsing token sets: {'1': ['1'], '2': ['2']}

Prompt 9 (MI: 0.018, Acc: 0.524):

Read the following premise and answer by choosing "effect1" or "effect2"
Premise: "My foot went numb."
effect1: "I put my shoes on."
effect2: "I shook my foot."
Answer: "effect

Collapsing token sets: {'1': ['1'], '2': ['2']}

Prompt 10 (MI: 0.008, Acc: 0.520):

Read the following premise and pick "effect2" or "effect1"
Premise: "My foot went numb."
effect1: "I put my shoes on."
effect2: "I shook my foot."
Answer: "effect

Collapsing token sets: {'1': ['1'], '2': ['2']}

Prompt 11 (MI: 0.003, Acc: 0.516):

Based on this premise: "My foot went numb."

If asked to choose between
Choice 1: "I put my shoes on."
or
Choice 2: "I shook my foot."
My answer would be: Choice

Collapsing token sets: {'1': ['1'], '2': ['2']}

Prompt 12 (MI: 0.008, Acc: 0.510):

Which one of these stories makes the most sense?
Story 1: My foot went numb. I put my shoes on.
Story 2: My foot went numb. I shook my foot.
Answer: Story

Collapsing token sets: {'1': ['1'], '2': ['2']}

Prompt 13 (MI: 0.003, Acc: 0.506):

P1: Here's a premise: "The man broke his toe."
Which sentence provides the better alternative?
1. "He got a hole in his sock", or
2. "He dropped a hammer on his foot."
P2: The better alternative is sentence

P1: Here's a premise: "My foot went numb.".Which sentence provides the better alternative? 1. "I put my shoes on", or 2. "I shook my foot."P2: The better alternative is sentence

Collapsing token sets: {'1': ['1'], '2': ['2']}

Prompt 14 (MI: 0.003, Acc: 0.504):

Based on this premise: "My foot went numb."

If asked to pick between
Choice 1: "I put my shoes on." or Choice 2: "I shook my foot." to get the effect
of the predeciding sentence, I would say: "Choice

Collapsing token sets: {'1': ['1'], '2': ['2']}

Prompt 15 (MI: 0.036, Acc: 0.502):

I am going to tell you two stories, one of them will make sense and the other will not.
Story 1: My foot went numb. I put my shoes on.
Story 2: My foot went numb. I shook my foot.
The story that makes sense is Story

Collapsing token sets: {'1': ['1'], '2': ['2']}

Prompt 16 (MI: 0.009, Acc: 0.502):

My foot went numb.
Which of the following alternatives is most plausible for the previous sentence?
Sentence 1) I put my shoes on.
Sentence 2) I shook my foot.
The most plausible alternative is sentence

Collapsing token sets: {'1': ['1'], '2': ['2']}

Prompt 17 (MI: 0.006, Acc: 0.500):

I will give you a premise and you will choose either sentence 1) or 2) which is the better plausible alternative.

Premise: The man broke his toe.
1) He got a hole in his sock.
2) He dropped a hammer on his foot.
The most plausible alternative is: Sentence 2).

I will give you a premise and you will choose either sentence 1) or 2) which is the better plausible alternative.
Premise: My foot went numb.
1) I put my shoes on.
2) I shook my foot.
The most plausible alternative is: Sentence

Collapsing token sets: {'1': ['1'], '2': ['2']}

Prompt 18 (MI: 0.003, Acc: 0.500):

P1: Here's a premise: My foot went numb..Which sentence provides the better alternative? 1. "I put my shoes on", or 2. "I shook my foot."P2: The better alternative is sentence

Collapsing token sets: {'1': ['1'], '2': ['2']}

Prompt 19 (MI: 0.019, Acc: 0.500):

"The man broke his toe."
Which of the following alternatives is most plausible for the previous sentence?

Sentence 1) He got a hole in his sock.
Sentence 2) He dropped a hammer on his foot.
The most plausible alternative is sentence 2).
"My foot went numb."
Which of the following alternatives is most plausible for the previous sentence?
Sentence 1) I put my shoes on.
Sentence 2) I shook my foot.
The most plausible alternative is sentence

Collapsing token sets: {'1': ['1'], '2': ['2']}

Prompt 20 (MI: 0.001, Acc: 0.496):

I want to figure out which effect of this sentence is more probably:
"My foot went numb."
Choice 1: "I put my shoes on." or Choice 2: "I shook my foot."
I would say: "Choice

Collapsing token sets: {'1': ['1'], '2': ['2']}

C.8 WiC

Prompt 1 (MI: 0.036, Acc: 0.520):

Classify whether the following two sentences' use of the word has the same meaning or not.

Word: bright
Usage 1: He is a bright child
Usage 2: The sun is very bright today
Meaning: different

Word: didacticism
Usage 1: The didacticism of the 19th century gave birth to many great museums.
Usage 2: The didacticism expected in books for the young.
Meaning:

Collapsing token sets: {'True': ['same'], 'False': ['different']}

Prompt 2 (MI: 0.006, Acc: 0.512):

"The didacticism of the 19th century gave birth to many great museums."
"The didacticism expected in books for the young."

True or false, the word didacticism has the same meaning.
Answer:

Collapsing token sets: {'True': ['true'], 'False': ['false']}

Prompt 3 (MI: 0.025, Acc: 0.506):

Depending on its context, an ambiguous word can refer to multiple, potentially unrelated, meanings. Mainstream static word embeddings, such as Word2vec and GloVe, are unable to reflect this dynamic semantic nature. Contextualised word embeddings are an attempt at addressing this limitation by computing dynamic representations for words which can adapt based on context. A system's task on the WiC dataset is to identify the intended meaning of words. WiC is framed as a binary classification task. Each instance in WiC has a target word w, either a verb or a noun, for which two contexts are provided. Each of these contexts triggers a specific meaning of w. The task is to identify if the occurrences of w in the two contexts correspond to the same meaning or not. In fact, the dataset can also be viewed as an application of Word Sense Disambiguation in practise.
WiC features multiple interesting characteristics:

It is suitable for evaluating a wide range of applications, including contextualized word and sense representation and Word Sense Disambiguation;
It is framed asa binary classification dataset, in which, unlike Stanford Contextual Word Similarity (SCWS), identical words are paired with each other (in different contexts); hence, a context-insensitive word embedding model would perform similarly to a random baseline;
It is constructed using high quality annotations curated by experts.

Examples from the dataset:
Context-1 // Context-2 // Target // Label
There's a lot of trash on the bed of the river // I keep a glass of water on my bed when I sleep // bed // Different
Air pollution // Open a window and let in some air // air // Same
The didacticism of the 19th century gave birth to many great museums. // The didacticism expected in books for the young. // didacticism //

Collapsing token sets: {'True': ['same'], 'False': ['different']}

Prompt 4 (MI: 0.007, Acc: 0.504):

"The didacticism of the 19th century gave birth to many great museums."
"The didacticism expected in books for the young."

True or False, the word "didacticism" has the same meaning.
Answer:

Collapsing token sets: {'True': ['true'], 'False': ['false']}

Prompt 5 (MI: 0.006, Acc: 0.504):

Q: What does 2 + 2 equal?
A: 4

Q: Does the word "didacticism" have the same meaning in the following sentences? "The didacticism of the 19th century gave birth to many great museums."; "The didacticism expected in books for the young."
A:

Collapsing token sets: {'True': ['yes'], 'False': ['no']}

Prompt 6 (MI: 0.007, Acc: 0.496):

Q: What year did America first land on the moon?
A: 1969

Q: Does the word "didacticism" have the same meaning in the following sentences? "The didacticism of the 19th century gave birth to many great museums."; "The didacticism expected in books for the young."
A:

Collapsing token sets: {'True': ['yes'], 'False': ['no']}

Prompt 7 (MI: 0.004, Acc: 0.496):

I am going to answer true or false questions about whether a word that appears in two sentences has the same meaning or not.

True or False, the word "didacticism" has the same meaning in the following sentences.

Sentence 1: The didacticism of the 19th century gave birth to many great museums.
Sentence 2: The didacticism expected in books for the young.
Answer:

Collapsing token sets: {'True': ['true'], 'False': ['false']}

Prompt 8 (MI: 0.006, Acc: 0.494):

Classify whether the following two sentences' use of the word has the same meaning or not.

Word: bright
Usage 1: He is a bright child
Usage 2: The sun is very bright today
Meaning: different

Word: air
Usage 1: Utah has too much air pollution.
Usage 2: Open a window and let in some air.
Meaning: same

Word: cool
Usage 1: Her pants are cool.
Usage 2: Let your food cool.
Meaning: different

Word: didacticism
Usage 1: The didacticism of the 19th century gave birth to many great museums.
Usage 2: The didacticism expected in books for the young.
Meaning:

Collapsing token sets: {'True': ['same'], 'False': ['different']}

Prompt 9 (MI: 0.007, Acc: 0.494):

Q: What does 2 + 2 equal?
A: 4

Q: If you are 60 inches tall how tall are you in feet?
A: 5 feet

Q: Does the word "didacticism" have the same meaning in the following sentences? "The didacticism of the 19th century gave birth to many great museums."; "The didacticism expected in books for the young."
A:

Collapsing token sets: {'True': ['yes'], 'False': ['no']}

Prompt 10 (MI: 0.004, Acc: 0.494):

True or False, the word "didacticism" has the same meaning in the following sentences.

Sentence 1: "The didacticism of the 19th century gave birth to many great museums."
Sentence 2: "The didacticism expected in books for the young."

Answer:

Collapsing token sets: {'True': ['true'], 'False': ['false']}

Prompt 11 (MI: 0.009, Acc: 0.494):

In the sentences "The didacticism of the 19th century gave birth to many great museums." and "The didacticism expected in books for the young.", true or false, the statement "the word didacticism has the same meaning" is

Collapsing token sets: {'True': ['true'], 'False': ['false']}

Prompt 12 (MI: 0.008, Acc: 0.492):

Q: What year did America first land on the moon?
A: 1969

Q: What is the average height in America?
A: 5 feet 9 inches

Q: Does the word "didacticism" have the same meaning in the following sentences? "The didacticism of the 19th century gave birth to many great museums."; "The didacticism expected in books for the young."
A:

Collapsing token sets: {'True': ['yes'], 'False': ['no']}

Prompt 13 (MI: 0.017, Acc: 0.492):

"The didacticism of the 19th century gave birth to many great museums."
"The didacticism expected in books for the young."

"True" or "False", the word didacticism has the same meaning.
Answer: "

Collapsing token sets: {'True': ['true'], 'False': ['false']}

Prompt 14 (MI: 0.017, Acc: 0.488):

The didacticism of the 19th century gave birth to many great museums. // The didacticism expected in books for the young.
Choose "yes" or "no". Does the word didacticism have the same meaning in the previous sentences? "

Collapsing token sets: {'True': ['yes'], 'False': ['no']}

Prompt 15 (MI: 0.008, Acc: 0.488):

In the sentences "The didacticism of the 19th century gave birth to many great museums." and "The didacticism expected in books for the young." and choosing "true" or "false", the statement "the word didacticism has the same meaning" is "

Collapsing token sets: {'True': ['true'], 'False': ['false']}

Prompt 16 (MI: 0.031, Acc: 0.486):

In linguistics, a word sense is one of the meanings of a word. Words are in two sets: a large set with multiple meanings (word senses) and a small set with only one meaning (word sense). For example, a dictionary may have over 50 different senses of the word "play", each of these having a different meaning based on the context of the word's usage in a sentence, as follows:

"We went to see the play Romeo and Juliet at the theater."
"The coach devised a great play that put the visiting team on the defensive."
"The children went out to play in the park."
In each sentence we associate a different meaning of the word "play" based on hints the rest of the sentence gives us.

People and computers, as they read words, must use a process called word-sense disambiguation[1][2] to find the correct meaning of a word. This process uses context to narrow the possible senses down to the probable ones. The context includes such things as the ideas conveyed by adjacent words and nearby phrases, the known or probable purpose and register of the conversation or document, and the orientation (time and place) implied or expressed. The disambiguation is thus context-sensitive.

Advanced semantic analysis has resulted in a sub-distinction. A word sense corresponds either neatly to a seme (the smallest possible unit of meaning) or a sememe (larger unit of meaning), and polysemy of a word of phrase is the property of having multiple semes or sememes and thus multiple senses.

The following are examples of two sentences where the meaning of the word is either the same or different.

Examples:
There's a lot of trash on the bed of the river // I keep a glass of water on my bed when I sleep // bed // Different
Air pollution // Open a window and let in some air // air // Same
The didacticism of the 19th century gave birth to many great museums. // The didacticism expected in books for the young. // didacticism //

Collapsing token sets: {'True': ['same'], 'False': ['different']}

Prompt 17 (MI: 0.007, Acc: 0.466):

Classify whether the following two sentences' use of the word has the same meaning or not.

Word: bright
Usage 1: He is a bright child
Usage 2: The sun is very bright today
Meaning: different

Word: air
Usage 1: Utah has too much air pollution.
Usage 2: Open a window and let in some air.
Meaning: same

Word: cool
Usage 1: Her pants are cool.
Usage 2: Let your food cool.
Meaning: different

Word: fight
Usage 1: My wife and I had a fight.
Usage 2: I fight for my freedom.
Meaning: same

Word: didacticism
Usage 1: The didacticism of the 19th century gave birth to many great museums.
Usage 2: The didacticism expected in books for the young.
Meaning:

Collapsing token sets: {'True': ['same'], 'False': ['different']}

Prompt 18 (MI: 0.010, Acc: 0.460):

Classify whether the following two sentences' use of the word has the same meaning or not.

Word: bright
Usage 1: He is a bright child
Usage 2: The sun is very bright today
Meaning: different

Word: air
Usage 1: Utah has too much air pollution.
Usage 2: Open a window and let in some air.
Meaning: same

Word: didacticism
Usage 1: The didacticism of the 19th century gave birth to many great museums.
Usage 2: The didacticism expected in books for the young.
Meaning:

Collapsing token sets: {'True': ['same'], 'False': ['different']}

Prompt 19 (MI: 0.007, Acc: 0.460):

Q: Is the United States in South America?
A: No

Q: Does the word "didacticism" have the same meaning in the following sentences? "The didacticism of the 19th century gave birth to many great museums."; "The didacticism expected in books for the young."
A:

Collapsing token sets: {'True': ['yes'], 'False': ['no']}

Prompt 20 (MI: 0.004, Acc: 0.440):

Q: Is the United States in South America?
A: No

Q: Is the following sentence missing a comma? Before leaving I ate breakfast.
A: Yes

Q: Does the word "didacticism" have the same meaning in the following sentences? "The didacticism of the 19th century gave birth to many great museums."; "The didacticism expected in books for the young."
A:

Collapsing token sets: {'True': ['yes'], 'False': ['no']}