BEHAVIORBOX: Automated Behavioral Comparison of Language Models

Anonymous ACL submission

Abstract

Language model evaluation is a daunting task: prompts are brittle, corpus-level perplexities are vague, and the choice of benchmarks are endless. Choosing examples that show meaningful, generalizable differences between two LMs is crucial to understanding where one model succeeds and another fails. Can this process be done automatically? In this work, we propose methodology for automated behavioral comparison of language models that uses performance-aware contextual embeddings to find fine-grained features of text where one LM outperforms another. Our method, which we name BEHAVIORBOX, is able to extract coherent features that also demonstrate statistically significant differences with respect to the ease of generation between two LMs. We apply BEHAVIORBOX to compare models that vary in size, model family, and post-training, and enumerate insights into specific contexts that illustrate meaningful differences in performance.[1]

1 Introduction

Where does one language model perform better than another? This deceptively simple question holds a near-endless number of complications. Practitioners must select from a dizzying array of evaluation methods, datasets, benchmarks, and metrics. Seemingly innocuous changes to evaluation pipelines, like the formatting of prompts, have been shown to drastically impact accuracy on a wide range of tasks (Sclar et al., 2023). Even evaluating language models based on their original training objective—next token prediction—is not so straightforward. While metrics like perplexity (Jelinek et al., 1977) on a held-out corpus are commonly used and are generally correlated with downstream performance (e.g. Adiwardana et al. 2020; Isik et al. 2024), the use of corpus-level perplexity on extremely large, diverse data often masks finer-grained differences on particular subgroups and domains (Magnusson et al., 2023).

What could an alternative to collections of benchmarks and corpus-level perplexity look like? One solution would be to partition the data into slices and report performance across these sub-corpora, as was done in Paloma (Magnusson et al., 2023). However, such an approach depends on both knowing the relevant partitions ahead of time and having sufficient metadata such that these partitions can be made. But what if we could instead discover what the relevant features of these partitions are, and automatically generate a report telling practitioners specific and coherent groups of text where one model outperforms another?

We attempt to tackle this problem using our new evaluation method—BEHAVIORBOX—which discovers fine-grained, human-interpretable features of data where one LM performs better than another. Unlike evaluations that depend on predetermined domains of data, BEHAVIORBOX is a bottom-up approach that finds semantic and/or structural features of text where one model outperforms another, and does so independently of the domain or corpus the text originates from. As a consequence, BEHAVIORBOX is capable of finding specific features and relationships in text that span across documents and domains, without the need to partition these domains ahead of time.

To find these features, BEHAVIORBOX not only considers the *context* of a text sample (via a contextual embedding), but also factors in the evaluated LMs' *performance* on that sample (via the probabilities the models assign to the text), forming a performance-aware contextual representation of each text sample. After generating a large dataset of these representations, we then train a sparse autoencoder (SAE), which learns simple linear decompositions of the dense representations, with each component of the sparse representation acting as a discovered feature. Finally, using the groups of

[1] Code/data for this work will be released open-source.

data determined by the SAE features, we generate natural language descriptions of each group.

We demonstrate the efficacy of BEHAVIORBOX in discovering fine-grained differences between models in the language modeling task by comparing models that differ in size, family, and in types of post-training; specifically, we look at base and post-trained models of two sizes (7B and 13B parameters) across two model families, Llama 2 (Touvron et al., 2023) and OLMo 2 (OLMo et al., 2024). Using BEHAVIORBOX, we are able to find extremely fine-grained features in data that point to larger models' ability to better predict long-tailed text (e.g. uncommon or archaic phrases and terms), as well as show particular features related to dialogue and conversation where chat/RLHF-ed models excel. We are also able to discover differences between models that otherwise show near-identical performance with respect to perplexity, such as differences in predicting particular structure or formatting in text or different parts of speech in specific contexts. The insights provided by BEHAVIORBOX provide a more holistic and detailed perspective on LM performance, and can be used to augment existing methods for evaluation and interpretability.

2 Background

BEHAVIORBOX draws both conceptually and methodologically from two well-established areas of research: the problem of *slice finding* and the *behavioral evaluation* of black-box NLP systems.

2.1 Slice Finding

A key component in debugging and building better machine learning and NLP systems is identifying where and when a system underperforms. When we evaluate these systems, we may use overall metrics, such as accuracy on a benchmark or perplexity on a large corpus. However, overall performance may obfuscate stark differences in performance across subgroups; thus, if we are interested in the performance on groups within the larger dataset, we may partition the data into predetermined categories, and compare performance within these groups. Nevertheless, it is often difficult to know *a priori* what the relevant groups of data are with respect to model performance. The task of automatically identifying salient groups of data where a model underperforms is known as *slice finding* (Chung et al., 2018), and is applicable across all sorts of tasks and modalities, from image classification to question answering.

Early works in slice finding often relied on metadata to find relevant slices (Chung et al., 2018), but such an approach depends on the appropriate metadata categories to be specified and present in the data, which may not necessarily be the case. To solve this problem, slice finding methods such as George (Sohoni et al., 2020), Spotlight (d'Eon et al., 2022), and Domino (Eyuboglu et al., 2022) utilize learned representations of the data to find semantically similar clusters of underperforming samples. These methods have primarily focused on image classification tasks, and a few constrained natural language tasks, such as sentiment analysis.

BEHAVIORBOX takes a similar approach as these works by utilizing contextual representations, but differs in two major ways. First, we focus on the language modeling task, which involves a significantly more complex output space compared to the tasks explored in prior work. Second, we focus on model *comparison* as opposed to where a single model is "incorrect", as such a distinction is much less clear in the context of text generation.

2.2 Behavioral Evaluation in NLP

As NLP systems have grown ever more complex, efforts to better understand these largely black-box systems have become increasingly important. One approach to better understanding such a system is to generate explanations for a system's decisions, which can be viewed as *behaviors* (Ribeiro et al., 2020). Explanations usually take the form of a relationship between a particular feature in the data and the resulting prediction, e.g. the impact of the use of negation on the predictions of a sentiment analysis model. These explanations not only need to faithfully capture model behaviors, but should also be human interpretable (Ribeiro et al., 2016; Lundberg and Lee, 2017).

In the context of explaining errors of NLP systems, works like Errudite (Wu et al., 2019) and CheckList (Ribeiro et al., 2020) provide frameworks for practitioners to stress-test models on precise hypotheses regarding the impact of specific features. Nevertheless, these hypotheses still need to be specified ahead of time. BEHAVIORBOX can be seen as a complementary approach by serving as a form of hypothesis discovery, where such hypotheses can then be further explored in various other behavior evaluation frameworks.

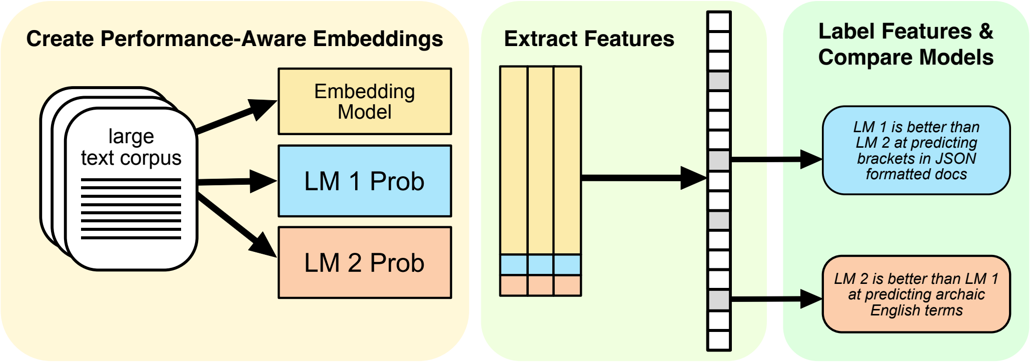
Create Performance-Aware Embeddings
large text corpus
Embedding Model
LM 1 Prob
LM 2 Prob
Extract Features
Label Features & Compare Models
LM 1 is better than LM 2 at predicting brackets in JSON formatted docs
LM 2 is better than LM 1 at predicting archaic English terms

Figure 1: BEHAVIORBOX is a three-part automatic behavior comparison pipeline that discovers fine-grained features where one LM differs from another.

3 Method Overview: BEHAVIORBOX

As shown in Figure 1, BEHAVIORBOX is a automatic behavior comparison pipeline for language modeling, comprised of three parts:

1. **Data generation**, which consists of calculating contextual embeddings and aligning these embeddings with probabilities under LMs for the same text (§4),

2. **Extracting features** that are coherent and capture similarities and differences regarding performance between models (§5), and

3. **Labeling and synthesizing the performance differences** between LMs within the discovered data slices (§6).

The unit of data used in this method can, in theory, be as small as a token or as large as a document. However, in our experiments, we focus on characterizing performance at (roughly) the word level.

We decided on this level of abstraction to balance both granularity and salience, as well as for engineering convenience. While tokens serve as the atomic unit of generation and are closest to the training objective, they may be less human-interpretable and are harder to work with when aligning the different tokenizers of the embedding model and various LMs. On the other hand, larger structures like phrases and sentences may be easier to categorize in terms of salient groups, but may be difficult to parse for certain types of documents commonly included in pretraining and evaluation (e.g. code, mathematical expressions, or other non-linguistic textual data), and could furthermore mask more granular trends that may be of interest.

4 Data Generation

Prior work has shown that incorporating learned representations of the input data along with a model's predictions and gold labels helps with identifying unlabeled classes of data where said model underperforms (Eyuboglu et al., 2022; Sohoni et al., 2020). Drawing from these works, BEHAVIORBOX uses contextual embeddings to provide semantic information about each word, along with probabilities generated by the evaluated LMs, which serve as a measure of the LMs' performance. For contextual embeddings, we use the last hidden layer of Longformer (Beltagy et al., 2020).

As previously mentioned, we use BEHAVIORBOX to slice our data (some arbitrary text dataset) into groups of words. Collecting and aligning contextual embeddings and probabilities per word across models that utilize different tokenization processes requires a number of engineering decisions, such as determining the boundaries of words, subsequently combining or splitting token log probabilities when necessary, and handling strings longer than the context window of different models.

We use Longformer's pre-tokenizer, which largely splits on whitespace, as our method of determining word boundaries. To aggregate token representations within a single word, we average the embeddings of the constituent tokens. For probabilities, we multiply the probabilities of constituent tokens. For instance, given a word that spans tokens n to m in a sequence $w = \{t_n, t_{n+1}, \dots, t_m\}$

$$\mathbf{e}_w = \frac{1}{(m-n+1)} \sum_{j=n}^{m} \mathbf{e}_{t_j}, \quad \mathbf{e}_{t_j} \in \mathbb{R}^{768} \qquad (1)$$

$$p_w = \prod_{j=n}^{m} P(t_j | t_1, \dots, t_{j-1}) \qquad (2)$$

Each datapoint in the resulting dataset is a vector of dimension 770, where the first 768 dimensions are from the Longformer embedding and the last two are the probabilities of the language models being compared.

$$\mathbf{x}_w = \begin{bmatrix} \mathbf{e}_w \\ p_{w,1} \\ p_{w,2} \end{bmatrix} \in \mathbb{R}^{770} \tag{3}$$

As the source of text we use to create the dataset of performance-aware representations, we use portions of the Dolma Dataset (Soldaini et al., 2024), an open dataset for language modeling containing a diverse mix of web content, academic publications, code, books, and encyclopedic materials. We sample 1000 documents across six of the data sources included in Dolma (Common Crawl, The Stack, C4, PeS2o, Project Gutenberg, and Wikipedia), totaling in approximately 80M words of data.

5 Extracting Features

Once we have a dataset of aligned words and probabilities for the LMs we wish to compare, we now have to find a way to extract and label fine-grained slices of data. This needs to be done in such a way that the slices are composed of *coherent* sets of words in context and the labels adequately *explain* the slice in a human-interpretable manner.

Previous works in automatic slice finding that incorporate learned representations have used various clustering algorithms such as k-means clustering (Sohoni et al., 2020; d'Eon et al., 2022) and Gaussian mixture models (Eyuboglu et al., 2022). However, as opposed to finding (hard) partitions in the data, we want to find specific *features* associated with text where one model performs better of worse than another. These features need not form a true mathematical partition of the entire corpus, but can instead be treated as linear decompositions of each text sample, where each word in context is comprised of some number of these features.

Finding simple, linear decompositions of otherwise complex representations is a problem in a wide variety of settings in NLP, such as creating more interpretable word embeddings (Faruqui et al., 2015) and—more recently—interpreting the internal states of transformer models (Cunningham et al., 2023; Lieberum et al., 2024; Gao et al., 2024, *inter alia*). We take a similar methodological approach to these works by using sparse autoencoders to extract features relevant to performance differences between two LMs. Using the SAE, we can then extract slices corresponding to each feature by finding the words whose representations that lead to the highest activation value of that feature.

5.1 Sparse Autoencoder Training

Recall that the features we are looking for ideally have the following characteristics: they should be coherent, fine-grained, and capture performance differences between models. Balancing each of these criteria inform our use of various hyperparameters and regularization choices.

The sparse autoencoder consists of an encoder and decoder: the encoder takes as input a vector $\mathbf{x}$, which is a concatenation of the contextual word embedding and LM probabilities, and creates a sparse representation $\mathbf{f}(\mathbf{x})$. The decoder then reconstructs the input (denoted as $\hat{\mathbf{x}}$) from this sparse representation. $\sigma(\cdot)$ denotes the activation function.

$$\mathbf{f}(\mathbf{x}) = \sigma(\mathbf{W}_{enc}\mathbf{x} + \mathbf{b}_{enc}) \tag{4}$$

$$\hat{\mathbf{x}} = \mathbf{f}(\mathbf{x})\mathbf{W}_{dec} + \mathbf{b}_{dec} \tag{5}$$

For $\sigma(\cdot)$, we use RELU (Agarap, 2018) to ensure non-negative values, as we conceptually want our features to be additive. The weights of $\mathbf{W}_{enc}$, $\mathbf{b}_{enc}$, $\mathbf{W}_{dec}$, and $\mathbf{b}_{dec}$ are learned by minimizing the L_2 distance between the reconstruction $\hat{\mathbf{x}}$ and the original input $\mathbf{x}$, using AdamW (Loshchilov and Hutter, 2017) as our optimizer.

5.2 Enforcing Sparsity

While allowing us to create a faithful representation of the original input, the above setup does not constrain the autoencoder to be sparse. As a way to enforce sparsity, we apply a batch-wise top-k operation to the pre-RELU SAE hidden state (Makhzani and Frey, 2013; Gao et al., 2024; Bussmann et al., 2024): for some value k, we flatten the batch (of size N), and zero out all activations that are not in the top $N \times k$ activations. This allows us to directly enforce $\mathbb{E}[\mathrm{L}_0]$ at the batch level, as opposed to using a proxy such as adding an L_1 penalty to the loss (Bricken et al., 2023).

$$\mathbf{f}_{sparse}(\mathbf{x}) = \text{BatchTopK}\Big(\sigma(\mathbf{W}_{enc}\mathbf{x} + \mathbf{b}_{enc}),\ k\Big) \tag{6}$$

5.3 Balancing Context and Performance Awareness

Including the probabilities in the input to the SAE on its own does not guarantee that the SAE will utilize that information. This may arise due to a number of reasons, which we address with two modifications in training. One reason why the SAE may not utilize probabilities is simply because these two features are overwhelmed by the large number of embedding features' contribution to the L_2 loss. Thus, we up-weigh the probability features so that the magnitude of the probability components make up 20% of the total magnitude of the input.

Another issue we address is the potential for the SAE to learn representations that do not depend on the probability features, i.e. the decoder weights for these features are very close to zero. To account for this, we introduce a penalty term in the loss that is high when the decoder weights associated with the probability features are small. This decoder penalty, denoted as $\mathcal{L}_{dec}$, is defined as

$$\mathcal{L}_{dec} = \lambda \sum_i \frac{1}{2} \exp(-d_i) \tag{7}$$

where λ is a hyperparameter to control the weighting of the penalty and d_i is a term in the decoder weights associated with the probability feature. Thus, the total loss we minimize is

$$\mathcal{L} = ||\mathbf{x} - \hat{\mathbf{x}}||_2 + \mathcal{L}_{dec} \tag{8}$$

Hyperparameters The dimension of the sparse representation we learn is 3000 with $k = 50$ and a decoder penalty coefficient $\lambda = 1e - 4$. We include a table of all SAE training hyperparameters and additional training details in Appendix A.1.

6 Labeling and Synthesizing Features

6.1 Extracting and Labeling Slices

After training, we now need to extract the slice of words associated with each feature of the SAE. We do this by taking the same dataset of words used to train the SAE and find the top 50 words that lead to the highest activation value for each feature in $\mathbf{f}_{sparse}(\mathbf{x})$. For each feature, we filter out words that have a zero activation value, as well as those that have an activation value that is both in the bottom quartile (of the top 50) and ≤ 0.75 times the max activation value. Then, for each feature and associated words, we get the context of that word from the document it originated and concatenate the preceding and following 10 words.

However, not every feature is indicative of a significant performance difference between models. To exclude those that are not, we use a two-sided t-test and filter out features that have a non-significant ($p > 0.05$) difference in mean probability between the two models.

As manually inspecting every slice across multiple SAE runs would take a prohibitive amount of time, we partially automate this process by using a strong LLM (Claude 3.5 Sonnet, Anthropic 2024) as an annotator. For a given feature, we prompt Claude to first determine if a group of words and their contexts form a coherent group, and if so to provide a label describing this group.[2]

6.2 Synthesizing Meta Features

After the labeling step, we now have a list of features indicative of performance differences between the two models. While this on its own is interesting, we would ideally like to synthesize broader categories of features to make understanding these differences easier.

As we want to focus on the set of features that lead to the largest gaps in performance, we do an additional filtering step to select features that show a difference in mean probability greater than some cutoff (in our experiments, we set this cutoff to $\Delta = 0.02$). Out of these features, we validate the labels by feeding the same label with examples to the LLM annotator, asking it to either keep the original label if it is appropriate, provide a new one if the current label does not accurately describe the examples, or invalidate the feature if the group is not coherent.[3] From these, we then perform a qualitative analysis to find larger "meta features" that contrast the two groups of labels.

7 Differentiating Model Performance with BEHAVIORBOX

Interpretability methods are notoriously hard to evaluate effectively (Lipton, 2018; Arora et al., 2022), and thus in this work we follow previous work on slice finding (Chung et al., 2018) and largely rely on qualitative inspection of the trends discovered by our method to demonstrate its utility. Specifically, we use BEHAVIORBOX to perform comparisons on language models across three axes of variation:

- **Model family:** Llama 2 (Touvron et al., 2023) and OLMo 2 (OLMo et al., 2024) (henceforth simply Llama and OLMo, respectively)
- **Model size:** 7B and 13B

[2] We include the prompt used in Appendix A.2.

[3] Additional details on this validation step are included in Appendix A.2.

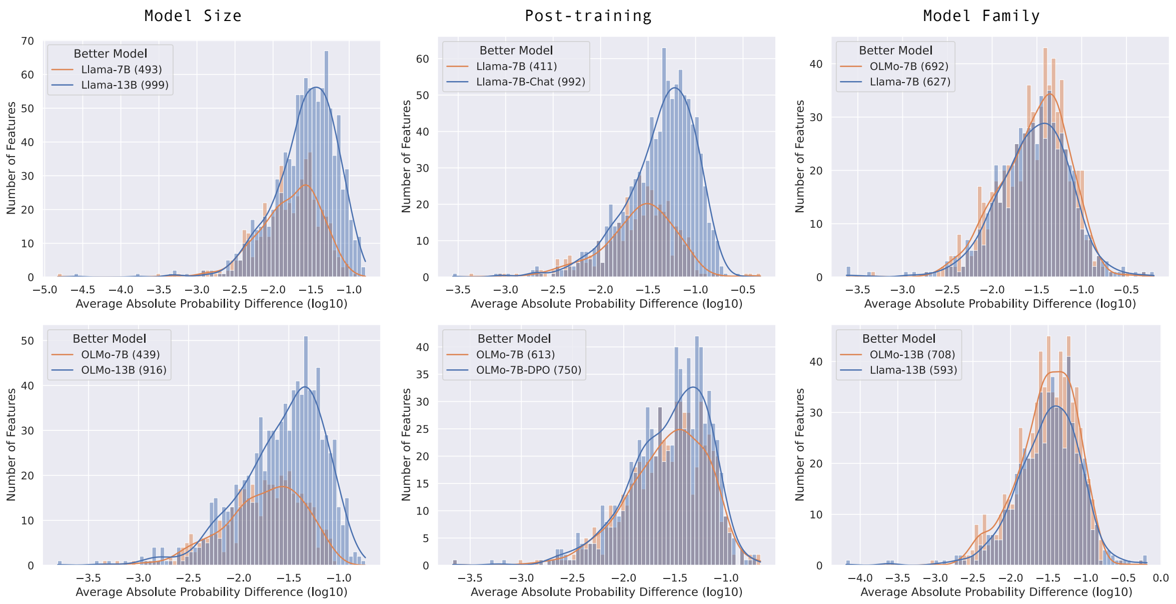

Figure 2: Histograms illustrating the distribution of average probability differences between two models across features; each curve represents the better performing model for that subset of features.

Model	Perplexity	Base Δ
Llama-7B	9.856	-
Llama-7B-Chat	13.911	+4.055
Llama-13B	8.773	-
Llama-13B-Chat	11.386	+2.613
OLMo-7B	9.803	-
OLMo-7B-DPO	12.762	+2.959
OLMo-13B	8.756	-
OLMo-13B-DPO	9.567	+0.811

Table 1: Perplexity per word for each of the models studied (lower is better). Base Δ indicates the change in perplexity from the base model of the same size.

- **Post-training:** we compare base models along with their chat variants (for Llama) and DPO variants (for OLMo)

Table 1 reports each model's perplexity per word on the subset of Dolma we use in our experiments. For each axis of comparison, we first report (1) overall trends in the number and magnitude of discovered features, then (2) take a closer look at the specific features output by BEHAVIORBOX for a particular model pair. The latter analysis is performed on a subset of features that meet the criteria as described in §6.2.

In the below sections, we indicate the number of validated features where the model is better in parentheses. For each comparison, we include a table of meta features and example instances with references to original labels, which can be found in Appendix A.3. Shared meta features are indicated in blue , contrasting in pink .

7.1 Model Size

To compare trends across model sizes, we focus our analysis on the difference across base models (i.e. Llama-7B vs. Llama-13B and OLMo-7B vs. OLMo-13B). First, we compare the magnitude of probability differences discovered for each model in Figure 2. We can see that among both model families, BEHAVIORBOX finds over twice as many features where the 13B models (blue) outperform the 7B ones (orange), reflective of the overall trend of larger models being more performant. Next, we derive qualitative insight from the extracted features, which are shown in Table 2

Llama-7B (83) vs. 13B (219) The primary differences between the 7B and 13B Llama models fall into two categories. First, Llama 13B excels at less common *genres/domains*, where 7B shows some advantages in dialogue and narrative text. Further, the clusters from Llama 13B reference *uncommon terms*, generally those that are more formal or archaic. Both of these trends indicate increased capacity of the larger model to better memorize long-tail phenomena. Additionally, both hmodels ave features related to *punctuation and formatting*, again with 13B excelling on more unusual punctuation/formatting phenomena.

Table 2: Comparison of features across models varying in number of parameters.

Meta Feature	Llama-7B	Llama-13B
Punctuation / Formatting	question marks and commas (2083, 272), text containing quotation marks (2361), closing angle brackets (1877)	parentheses (1273), comment markers (1192), section breaks (1790), markup special characters (1444), punctuation in bibliographic entries (537)
Text Diversity	**Primarily dialogue and narrative text from literary and historical sources** (2083, 272, 251, 174, 2129, etc.)	**Greater diversity in genre/style:** dramatic and literary dialogue (1273, 2811, 2843, 997, 1648, etc.), formal/archaic writing (810, 2962, 2288, 409, 719, etc.), religious/philosophical contexts (1402, 1928).
Uncommon Terms	**Primarily commonly used words or phrases**	**Includes relatively uncommon/archaic phrases:** "fellow" as a prefix (1625), had/hath/hast (409)
Meta Feature	**OLMo-7B**	**OLMo-13B**
Temporal Relations	"and" to emphasize continuity or repetition (2033), phrases describing effects, sequences, and passage of time (1091, 2796, 1534), cooking time (215), time adverbs (174)	references to periods/durations in casual conversation (830), sequential/temporal markers (343, 961), phrases indicating periodic/intermittent occurrences (1943), "until" to indicate cooking time (2142)
Connective Phrases	**Includes a single reference to the use of phrases used as temporal transitions** (1534)	**Variety of references to connecting or transitional phrases** (639, 716, 1230, 1019)
Adjectives (of degree)	**No reference to adjectives**	**Text involving adjectives used to describe (comparative) degree, intensity, or extent** (2093, 2096, 738, 2570, 2766)

OLMo-7B (62) vs. 13B (223) Unlike the Llama models, OLMo-7B and 13B appear to have a closer overlap with respect to domain. Both sets of features contain references to *temporal relations*; of these, both include features related to the phrase "no sooner" as well as features related to cooking instructions, but these instances come from different contexts. OLMo-13B has more features related to *connective phrases* as well as the use of *adjectives*.

7.2 Model Families

From both the distribution of features in Figure 2 as well as perplexity, it appears that both size pairs of Llama and OLMo models show very similar performance. Nevertheless, we can still find features that distinguish them.

OLMo-7B (101) vs. Llama-7B (39) Compared to Llama-7B, OLMo-7B appears to perform better on a wider range of *structured formatting*, such as whitespace and curly braces in theatrical scripts, as well as the use of *pronouns* in various contexts; Llama-7B has features associated with particular *numerical* values. Both have a substantial number of features involving *question marks*.

OLMo-13B (172) vs. Llama-13B (131) Unsurprisingly, the trends between the 13B models are very similar to those of the 7B models, such as both models having many features related to *question marks*. As before, OLMo outperforms Llama on *whitespace formatting*, though Llama performs better on formatting in *code/configuration files*.

7.3 Post-training

Finally, we look at model pairs that differ only in whether or not they have undergone post-training, specifically comparing base and Chat/DPO variants. Unlike the comparisons across size and model family, we see an unexpected trend where models with *higher* perplexity (the post-trained models) have a greater amount of features where they perform better, especially for Llama.

Why might this be the case? These results can partially be explained by the fact that models that have undergone RLHF restrict their generations to templates or restricted blueprints (Li et al., 2024). Thus, these differences may be more semantically coherent and tend to occur in very localized contexts. From this, we hypothesize that the regions where the post-trained model significantly outperforms the base model are more easily learned by the SAE, as these local performance differences are also consistent with local embedding similarity.

Llama-7B (89) vs. 7B-Chat (256) Touvron et al. 2023 report that Llama-7B-Chat, in contrast to the base model, was optimized for dialogue; we can see this reflected in a greater emphasis on *dialogue and conversations* in features. Interestingly, many of these phrases are used to hedge or qualify statements. We also find that while the base model has two features related to *code/numerical formatting*, these are not present in the 7B-Chat features. Both models have features related to *punctuation*, but with a greater focus on document formatting/sectioning for 7B and citations for 7B-Chat.

OLMo-7B (145) vs. 7B-DPO (163) Compared to the Llama models, the OLMo-7B base and DPO models display a less stark difference in performance when comparing features (as well as a lower comparative gain in perplexity). Like Llama, we find that the post-trained model has more features

Table 3: Comparison of features across model families.

Meta Feature	OLMo-7B	Llama-7B
Question Marks	after expressing uncertainty (1831), in dialogue (2233), QA pairs (2339)	in literary/scholarly texts (2328), in dialogue (127, 2088), followed by quotation marks (310)
Formatting	**Variety of structured formatting:** whitespace and line breaks between sentences/dialogue (1395, 635, 80, 71, 2196), brackets and curly braces (1877, 2480, 488)	**Primarily delimiters or equal signs** (2672, 2813, 222)
Pronouns	**Pronouns in varying contexts, grammatical person and gender** (1699, 2111, 1229, 910, 2797, etc.)	**Primarily possessive pronouns** (92, 475)
Numerical Values	**Various numbers appearing in particular contexts:** organizational elements such as footnotes and references (2826, 1046, 2480), in lists (1911), at beginning of lines (220)	**Particular numerical values:** the number 1 in various contexts (2756), 0 or 2 in technical contexts (678)

Meta Feature	OLMo-13B	Llama-13B
Question Marks	in dialogue or quotations (1627, 2233, 2154), after expressing uncertainty and in rhetorical questions (1831)	in dialogue or quotations (310, 2581, 2088, 127, 445)
Formatting	**Whitespace and line breaks in literary text:** indicating a line, sentence, or dialogue break (1359, 635, 80, 2196), before capitalized personal pronouns (1699), between a question and answer (2339)	**Code and configuration files:** delimiters and equals signs in database/config mappings (2672, 2813, 222), comment symbols (1192), property/attribute labels (302), forward slashes in filepaths (2919)

Table 4: Comparison of features across base vs. post-trained models.

Meta Feature	Llama-7B	Llama-7B-Chat
Punctuation / Formatting	section separators/breaks (1790, 2127), punctuation at the end of quoted text (1627, 2083)	punctuation after quoted text (585), commas (537, 1645, 1108), delimiters in text (1190), question marks (2088)
Dialogue / Conversations	**Only has features related to punctuation in dialogue** (1627, 2083)	**Various dialogue/conversational phrases:** "so" or "sometimes" as affirmative responses or qualifiers (2511), "kind" or "sort" used to deny or qualify (2811), forms of "think" in contemplative dialogue (1038), "what's the matter" to express concern (1088)
Code / Numerical Formatting	**Includes numerical formatting and code syntax features:** filepaths, decimals, and versions (1855), color codes and syntax (2714)	**No reference to code or numerical formatting**

Meta Feature	OLMo-7B	OLMo-7B-DPO
Punctuation / Formatting	before line breaks (2546), at end of statements or questions (548, 966), commas to indicate pauses (1250), after bibliographic entries (175)	punctuation at the end of quotations (585, 310), XML closing markers and curly braces (1877, 2171), opening parentheses (1273), equal signs in code (2813)
Dialogue / Conversations	**Primarily informational, literary, or narrative text** (2942, 1359, 2588, 215, 271, etc.)	**Various dialogue/conversational phrases:** variations of "as I said" (1912), "tell" used in dialogue, "of the sort" to indicate refusal or denial (1145)
Formality	**More formal text**	**Casual terms/contexts:** informal phrases using "sort" and "course" (633), informal/casual dialogue (830, 2793)

involving *conversational phrases and dialogue*; furthermore, we can see a distinction in *formality*, with the post-trained model containing more features related to casual speech. Like the Llama models, both have *punctuation and formatting* related features, with 7B-DPO having a greater focus on non-punctuation formatting.

8 Discussion and Conclusion

In this work, we introduced BEHAVIORBOX, an automated pipeline for the behavioral comparison of language models that bridges the gap between aggregated metrics and fine-grained performance analysis. By integrating contextual embeddings with model probabilities into a unified, performance-aware representation and leveraging a sparse autoencoder to extract human-interpretable features, our approach enables the discovery of coherent data slices where one model outperforms another. Our experiments—spanning variations in model family, size, and post-training regimes—demonstrate that BEHAVIORBOX can uncover nuanced performance differences, such as distinctions in formatting, domain-specific language, and syntactic patterns, that are often masked by conventional evaluation metrics like perplexity.

Beyond its utility for detailed performance diagnostics, BEHAVIORBOX serves as a hypothesis generation tool for further behavioral analysis. The automatic labeling and synthesis of meta features facilitate a deeper understanding of language model behavior, thereby supporting more informed decisions in model development and deployment. Overall, our method represents a step toward more transparent and actionable insights into the inner workings of large-scale language models.

9 Limitations

While BEHAVIORBOX shows promise as an interpretability and diagnostic tool, several limitations warrant discussion. First, the approach is dependent on the quality and compatibility of the underlying contextual embeddings and probability estimates. Any misalignment between the embedding space and the performance signals can obscure meaningful differences. Second, aggregating token-level probabilities into word-level metrics may introduce noise, particularly when tokenization strategies differ across models.

Additionally, the sparse autoencoder, despite its design for interpretability, may not capture all relevant behavioral nuances, and its performance is sensitive to hyperparameter choices such as the sparsity level and the weighting of probability features. The automated labeling process—while efficient—relies on a strong LLM annotator, which can sometimes generate inconsistent or suboptimal descriptions. Finally, our experiments have been conducted on a subset of language modeling tasks and datasets; thus, the generalizability of BEHAVIORBOX to other tasks, domains, or non-textual modalities remains to be fully explored. Future work may address these limitations by refining the representation alignment, exploring alternative aggregation strategies, and broadening the scope of evaluation.

Ethical Considerations

BEHAVIORBOX provides new tools for practicioners to better understand the behavior of language models, and particular the differences between multiple language models. On the whole, this has the potential for easing the ethical deployment of language models by identifying potential issues in advance of deployment and rectifying them before their deployment. Overall, we foresee few ethical risks in the existence of such a framework, although as with all automatic tools, users must be cautious in jumping to conclusions based solely on the tool output without careful thought.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *ArXiv*, abs/1803.08375.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Siddhant Arora, Danish Pruthi, Norman Sadeh, William W Cohen, Zachary C Lipton, and Graham Neubig. 2022. Explain, edit, and understand: Rethinking user study design for evaluating model explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5277–5285.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2.

Bart Bussmann, Patrick Leask, and Neel Nanda. 2024. Batchtopk sparse autoencoders. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*.

Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, and Steven Euijong Whang. 2018. Slice finder: Automated data slicing for model validation. *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1550–1553.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *ArXiv*, abs/2309.08600.

Greg d'Eon, Jason d'Eon, James R. Wright, and Kevin Leyton-Brown. 2022. The spotlight: A general method for discovering systematic errors in deep learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1962–1981, New York, NY, USA. Association for Computing Machinery.

Sabri Eyuboglu, Maya Varma, Khaled Kamal Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Re. 2022. Domino: Discovering systematic errors with cross-modal embeddings. In *International Conference on Learning Representations*.

Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah A. Smith. 2015. Sparse overcomplete word vector representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1491–1500, Beijing, China. Association for Computational Linguistics.

Leo Gao, Tom Dupr'e la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. Scaling and evaluating sparse autoencoders. *ArXiv*, abs/2406.04093.

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

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Margaret Li, Weijia Shi, Artidoro Pagnoni, Peter West, and Ari Holtzman. 2024. Predicting vs. acting: A trade-off between world modeling & agent modeling. *ArXiv*, abs/2407.02446.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, J'anos Kram'ar, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *ArXiv*, abs/2408.05147.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Ian Magnusson, Akshita Bhagia, Valentin Hofmann, Luca Soldaini, A. Jha, Oyvind Tafjord, Dustin Schwenk, Pete Walsh, Yanai Elazar, Kyle Lo, Dirk Groeneveld, Iz Beltagy, Hanna Hajishirzi, Noah A. Smith, Kyle Richardson, and Jesse Dodge. 2023. Paloma: A benchmark for evaluating language model fit. *ArXiv*, abs/2312.10523.

Alireza Makhzani and Brendan J. Frey. 2013. k-sparse autoencoders. *CoRR*, abs/1312.5663.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2024. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. 2020. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *arXiv preprint*.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melissa Hall Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2019. Errudite: Scalable, reproducible, and testable error analysis. In *Proceedings of the*

Hyperparameter	Value
Batch size	128
Learning rate	10^{-4}
AdamW β_1	0.9
AdamW β_2	0.99
Dict size	3000
k	50
Probability feature weight	0.2
Decoder penalty λ	10^{-4}

Table 5: Hyperparameters used to train SAEs.

57th Annual Meeting of the Association for Computational Linguistics, pages 747–763, Florence, Italy. Association for Computational Linguistics.

A Appendix

A.1 SAE Hyperparameters and Training

Hyperparameters used to train the SAE are included in Table 5. Hyperparameters were chosen based on a number of heuristics, including the number of dead latents, distribution of values of the encoder vs. probability feature decoder weights, and the number of resulting coherent features as determined by the LLM annotation process.

Additionally, to mitigate the presence of dead latents during training, we follow the methodology in (Bricken et al., 2023) and periodically re-initialize encoder and decoder weights for features that have no non-zero activations on a hold-out eval set during training.

A.2 LLM Annotation

The following prefix was used to generate a first-pass of annotation labels. The top 10 words and their contexts (samples with highest activation) for a feature were provided in a list following the prefix.

Your job is to determine if a group of words (surrounded by asterisks, e.g. *word*) in specific contexts form a coherent group that can be described concisely. I will provide you with a list of words surrounded by asterisks and the context in which they appear, usually within a sentence or a block of text. Each word and how it appears in context will be its own item in a list.

Here are some examples:
- conservation: Efforts in *conservation* are essential for protecting endangered species.
- habitat: The loss of *habitat* is a significant threat to biodiversity.
- ecosystem: An *ecosystem* needs a balance of various species to thrive.

Your job is to determine if the words form a coherent group that can be described concisely. If the words do form a coherent group, please provide a concise description of the group. Provide your answer in the following format:

<BEGIN ANSWER>
Coherent: <YES or NO>
Description: <if YES above, your description here; otherwise NONE>
<END ANSWER>

Do not provide any additional text after <END ANSWER>. Only respond with YES or NO for the "Coherent" field. If you respond with YES, you must provide a description in the "Description" field. Descriptions should be concise, ideally a single sentence. For the above example, descriptions may be something like "Nouns describing environmental conservation" or "Words related to biodiversity". Note that groups and descriptions may also pertain to formatting, such as "Punctuation before whitespaces in documents discussing logic" or "Series of whitespaces in documents discussing visual art". The description should NOT refer to the asterisks, those are only there to help you identify the words.

Please categorize the following list of words and their contexts as coherent or not coherent, and provide a description if needed:

To filter and validate the features, we have an additional round of LLM annotation, which takes as input the original label from the annotator LLM along with the top 20 words and their contexts. In our qualitative analyses, we only consider the labels output from this labeling stage that were scored $\geq$ 1 (as labels scored 1 or 2 were re-labeled).

Your job is to determine if a group of words (surrounded by asterisks, e.g. *word*) in specific contexts form a coherent group that is accurately described by a given label. I will provide you with a list of words surrounded by asterisks and the context in which they appear, usually within a sentence or a block of text. Each word and how it appears in context will be its own item in a list. Determine if the words form a group that is accurately described by the label by providing a numerical score (0 to 3, and -1). Scores are defined as follows:

- 0: The label is not accurate and the words do not form any coherent groups.
- 1: The label is not accurate, but the words form a coherent group.
- 2: The label is accurate, but fails to capture a more specific trend.
- 3: The label is accurate and captures a specific trend.
- -1: There are two coherent groups.

Additionally, if you give a score of 1 or 2, provide an alternative label that you believe would be more accurate. If you give a label of -1, provide a label for each group. Each label should be separated with <SEP>. This label should be precise, concise, and accurate, ideally a single sentence, Otherwise, leave the alternative label field blank.

Provide your answer in the following format, be sure to include both "Score" and "Label" fields:

<BEGIN ANSWER>
Score: <a number between 1-3 or -1>
Label: <label(s) if original score is 1, 2, or -1, empty otherwise>
<END ANSWER>

Do not provide any additional text after <END ANSWER>. Only respond a number between 0 and 3 or -1 in the Score field. The description should NOT refer to the asterisks, those are only there to help you identify the words. If there are double asterisks in the text, assume the word of interest is the whitespace between them.

Please score the following list of words and their label, and provide a new label if necessary:

A.3 Feature Labels

We include all labels used in our qualitative analysis in the tables below.

Table 6: Features for Llama-7B and Llama-13B comparison

Feature	Llama-7B	Mean Δ Prob
2083	Question marks followed by quotation marks at the end of questions in dialogue	0.0933
272	Comma-quote punctuation sequences at the end of dialogue in narrative text	0.0918
1091	Uses of "sooner" in phrases indicating immediate sequence of events, typically in the pattern "no sooner... than"	0.0739
1518	The phrase "at any rate" used as a transitional expression to qualify or shift between thoughts in various contexts	0.0717
251	Honorific title "Mrs." used before women's surnames in various literary contexts	0.0695
2965	Variations of the word "born" (including "Born") at the start of biographical entries, typically followed by location and date information	0.0668
1378	Verbs and nouns related to mental processes, reasoning, and discovery	0.0668
174	Words used as degree modifiers or intensifiers in literary narrative contexts	0.0592
2129	Common English words ("I", "to") appearing in dialogue or narrative text from likely historical or literary sources	0.0588
2361	Text fragments containing punctuation marks and quotation marks around dialogue or quoted text	0.0586
1534	Parenthetical or transitional phrases using common words ('time', 'way', 'while', 'same', 'other') to connect or qualify statements	0.0575
1254	Uses of "at last" as a temporal phrase indicating the end or culmination of a waiting period	0.0552
2302	Verbs and related nouns describing acts of focus, observation, or compliance	0.0549
1619	Uses of the word "way" describing manner or behavior in narrative contexts	0.0545
903	Double asterisks marking dialogue or section breaks in literary text	0.0528
1877	XML-style closing angle brackets followed by quotation marks in software configuration files	0.0508
1931	Honorifics and articles appearing before periods or whitespace in formal writing	0.0490
897	Instances of the pronoun "it" in various sentence contexts showing direct object or subject usage	0.0489
2290	Instances of "*thought*" followed by "of" expressing concern or distress in narrative contexts	0.0488
1960	Present tense verbs used to describe actions or states in various narrative contexts	0.0481
774	Words commonly used as function words (prepositions, conjunctions, auxiliaries) in various literary and instructional texts	0.0459
1779	Occurrences of time phrases using "two" or temporal words in literary passages, primarily in the pattern "minute or two"	0.0458
1287	Personal and impersonal pronouns appearing at the start of clauses in formal documents	0.0445
1244	Phrase "with a view" used to express purpose or intention in formal writing	0.0443
2111	Religious pronouns and articles referring to divine entities in spiritual texts	0.0440
2233	Question marks at the end of interrogative sentences in dialogue or narrative text	0.0431
2988	Terms referring to different dimensions or elements of a subject matter in analytical or explanatory contexts	0.0420
1399	Possessive apostrophe-s endings in various literary texts	0.0409
1054	Instances of "that" used as a conjunction in old-fashioned descriptive phrases of the form "[noun], [adjective/noun] that [pronoun] was"	0.0407
2134	Uses of "for" preceding time duration specifications in recipe and instruction texts	0.0398
1903	Variations of the phrase "take/taking into account/consideration" used to express consideration or inclusion of factors	0.0388
2033	Repetitive use of "and" in narrative text to emphasize scale, distance, or quantity	0.0381
1667	Expressions using "other" or "another" to indicate alternation or movement between two options or positions	0.0366
2485	Variations of the word "other(s)" used to indicate alternative items or choices in narrative texts	0.0365
1077	Pronouns used as subjects in narrative text, often appearing after "no" or describing individuals	0.0363
2573	Forms of the phrase "take/took a liking to" expressing fondness or preference in narrative contexts	0.0361
921	Uses of the word "way" in prepositional phrases indicating means, manner, or direction	0.0358
2256	Unit of measurement (pounds) used in dyeing recipe instructions	0.0352
972	Time-related adverbs appearing in narrative texts describing sequence or immediacy of events	0.0344
1197	Uses of "or" in expressions indicating an unspecified alternative, typically following "somehow," "one," or "some"	0.0334
1211	Numbers appearing at the end of sentences or items in numbered lists across different texts	0.0326
802	The word "to" used as a preposition indicating direction, destination, or relationship between quantities in various contexts	0.0322

Feature	Llama-7B	Mean Δ Prob
615	Common qualifying phrases in dialogue or narrative text using "sort", "same", "least", and related terms to express disagreement or qualification	0.0319
135	Uses of "other" in contexts describing comparisons or conflicts between two opposing parties	0.0318
977	Articles and time-related words appearing in narrative or descriptive text passages	0.0317
1080	Instances of "matter" in dialogue expressing concern about someone's condition or well-being, typically in the phrase "what's the matter"	0.0311
912	Relative adverbs expressing quantity or degree used in comparative or qualifying statements	0.0309
930	Common English grammatical elements including auxiliary verbs, contractions, articles, punctuation, and connectors in various contexts	0.0309
2316	Uses of common transition words and conjunctions in narrative text	0.0304
725	Uses of "or" in phrases expressing indefinite quantity or choice, typically following "one" or referring to unspecified options	0.0304
2262	Italicized titles or proper nouns in bibliographic or literary contexts	0.0302
2635	Uses of the word "part" in formal or diplomatic contexts indicating involvement, responsibility, or agency	0.0302
968	Question marks appearing at the end of sentences in philosophical or analytical texts	0.0301
868	Comma followed by quotation mark at the end of spoken dialogue in narrative text	0.0292
1338	Conjunction words used to establish logical sequence or parallel relationships between clauses in formal prose	0.0292
28	Forms of the word "being" referring to humans or living creatures, often preceded by "human"	0.0290
2622	Personal pronouns and nouns appearing at the start of sentences or clauses	0.0285
1644	Third-person pronouns appearing at the start of sentences or stage directions in narrative texts	0.0282
2731	Prepositions (mostly "in") at the start of clauses or sentences in various literary contexts	0.0276
899	Right curly brace characters appearing at the end of dialogue in theatrical scripts	0.0275
1778	Variations of "sort" and "spite" in phrases expressing contrast or negation, particularly in constructions like "in spite of" and "nothing of the sort"	0.0272
733	Instances of common conjunctions and auxiliary verbs ('if', 'is', 'such') used in conditional or comparative contexts	0.0267
129	Uses of "the whole of" as a phrase indicating entirety or completeness in various contexts	0.0267
2118	Words describing formal activities, interactions, and observations in narrative prose	0.0264
1185	Forms of "say" or "we" appearing in dialogue or narrative contexts, typically followed by quotation marks or statements	0.0251
407	Commas followed by whitespace in a listing of historical names with page numbers	0.0249
500	Past tense and modal verbs expressing possibility or occurrence in narrative contexts	0.0248
1604	Sentence-initial words and punctuation marks appearing at the end of dialogue or section titles	0.0245
398	Comma usage as a separator in various contexts including addresses, lists, and numbers	0.0244
548	Punctuation marks (question marks and commas) appearing at the end of phrases or clauses in various texts	0.0244
2473	Contractions and auxiliary verbs in questions or statements indicating uncertainty or seeking confirmation	0.0243
848	Words related to identity, naming, and personal history appearing in biographical or historical contexts	0.0243
20	Personal pronouns serving as sentence subjects in narrative texts	0.0241
2437	Forms of pronouns and auxiliary verbs used as function words in narrative texts	0.0239
1220	Exclamatory expressions followed by double asterisks in dialogue or narrative text	0.0239
1845	Instances of "or" in phrases expressing inevitability or uncertainty, typically in patterns like "somehow or other" and "sooner or later"	0.0236
675	Preposition 'from' used to indicate the starting point of a geographic or spatial range	0.0235
883	Common English function words and phrases appearing in formal or literary dialogue and narration	0.0228
372	Special characters followed by punctuation marks at the end of text segments, primarily closing parentheses and curly braces with periods	0.0217
155	Numerical format "8" appearing in book edition details, specifically in "Cr. 8 vo." format specifications	0.0213
497	Words denoting type, category, or nonspecific reference (kind, sort, thing) used in contexts of classification or description	0.0213
742	Page number references in an index or bibliography separated by commas and hyphens	0.0210
2196	Asterisk markers appearing after text segments ending with punctuation marks in dialogue or narrative prose	0.0208

Feature	Llama-13B	Mean Δ Prob
1463	Uses of "sorts" to indicate various types or varieties within lists or descriptions of items or behaviors	-0.1504
1273	Opening parenthesis followed by italicized character names or references in dramatic or literary dialogue formatting	-0.1427
1625	Uses of "fellow" as a prefix meaning "other people" or "fellow humans", often with a hyphenated form "fellow-men"	-0.1386
2811	Phrases using "kind" or "sort" to express negation or categorization in dialogue and narrative text	-0.1376
514	Contextual references to previously mentioned items or comparisons, often paired with words like "latter" or appearing in discussions of prior elements	-0.1346
810	Instances of "in the course of" or similar phrases indicating temporal progression in formal writing	-0.1313
2808	Past or conditional uses of the verb "have" in narrative contexts	-0.1243
1398	Use of "in order" as a subordinating conjunction to express purpose or intention in complex sentences	-0.1232
830	Words used in phrases expressing duration or temporal experiences	-0.1196
119	Uses of "if" and similar words in similes and hypothetical comparisons within descriptive narratives	-0.1172
2962	Possessive pronoun "your" used in formal or narrative contexts addressing a second person	-0.1156
2843	First-person personal pronouns and possessive adjectives in narrative dialogue or internal monologue	-0.1144
755	Verbs indicating the transfer of knowledge or instruction in educational contexts	-0.1092
1402	Instances of possessive pronoun "My" in religious or spiritual texts, with one outlier ("I") and one unrelated term ("Sergeant")	-0.1057
1192	Comment markers ('#') and descriptive text in configuration/code files	-0.1056
2096	The word "slightest" used in contexts emphasizing minimal or negligible degree or extent	-0.1051
1790	Section breaks or subtitle separators in biographical text about musicians and artists	-0.1035
2078	Usage of the word "in" following instructions to place or put items, primarily in recipe/cooking contexts	-0.1016
2288	Words and punctuation marks used as text separators or connectors in formal or literary prose	-0.0991
905	Usage of "parts" to describe geographic or spatial divisions of territories, regions, or physical locations	-0.0986
1048	Usage of "the most" in phrases about maximizing or taking advantage of opportunities	-0.0965
1753	The word "and" used in numeric expressions to connect whole numbers with additional quantities	-0.0962
409	Past tense auxiliary verbs (had/hath/hast) in formal or archaic English texts	-0.0961
343	Usage of "no sooner" followed by "than" in narrative text to express immediate sequence of events	-0.0955
719	Instances of the phrase "at the same time" used as a temporal or logical conjunction in formal prose	-0.0955
738	Instances of "and" in phrases expressing increasing or progressive change using comparative adjectives or adverbs	-0.0931
2953	Past tense verbs used in biographical or narrative contexts	-0.0921
1444	Special characters and symbols appearing at the end of text segments, often in technical or markup contexts	-0.0877
901	Personal pronouns and possessive forms used in narrative fiction	-0.0848
537	Punctuation marks (commas and similar delimiters) appearing in bibliographic or reference entries	-0.0838
2395	Words indicating temporal or simultaneous relationships within narrative texts	-0.0831
1062	Words referring to types of work, jobs, or career categories in discussions about employment and social roles	-0.0821
1095	Words used in narrative contexts to indicate timing, causation, or consequence	-0.0815
997	Instances of the word "tell" used in dialogue or direct speech requesting information from someone	-0.0802
201	Uses of "same" in phrases indicating simultaneous actions, consistently appearing in the pattern "and at the same time"	-0.0795
1648	Instances of "can" and "his" used in personal dialogue and narrative descriptions, primarily in question-asking and possessive contexts	-0.0793
1036	Instances of the word "ever" appearing in various literary contexts, often expressing continuity or permanence	-0.0773
1370	Instances of words in text where something moves or passes "through" a collective group or physical space	-0.0769
2088	Question marks at the end of dialogue or questions in literary text	-0.0764
1767	Possessive pronoun "my" and "your" in emotional or dramatic exclamations and declarations	-0.0764

Feature	Llama-13B	Mean Δ Prob
2748	Instances of the word "same" used for expressing equality, identity, or similarity in various contexts	-0.0764
633	Uses of "sort" and "course" as part of common phrases indicating type, manner, or progression ("that sort of thing", "in the course of")	-0.0750
874	Uses of function words "When" and "One" as sentence starters or list items in various written contexts	-0.0728
1555	Phrases using "so much as" to emphasize minimal or threshold actions that are prohibited or never occurred	-0.0722
1230	Common words used as qualifiers or modifiers in narrative text to indicate progression, extent, or manner	-0.0716
320	Variations of the word "wouldn't" appearing in dialogue or questions across different texts	-0.0711
1928	Words related to inquiry, correctness, and judgment appearing in philosophical or argumentative contexts	-0.0707
716	Common English conjunctions and auxiliary verbs used in connecting clauses and forming conditional statements	-0.0707
332	First-person and second-person pronouns in direct dialogue exchanges	-0.0704
275	Instances of the word "beside" describing physical proximity or positioning of people relative to others	-0.0696
256	Common conversational interjections and expressions of surprise/emphasis used in dialogue	-0.0695
845	Punctuation marks (, and :) followed by whitespace in various document contexts including bibliographic entries and translations	-0.0693
2099	Common English function words (articles, auxiliaries, pronouns) appearing in narrative prose	-0.0691
2418	Words used as referential or linking terms in various contexts, typically serving as anaphoric references to previously mentioned concepts	-0.0685
1575	Instances of "matter" used as a question to inquire about someone's well-being or problem	-0.0680
1471	Conditional conjunctions 'if' and 'or' used in narrative text to express uncertainty or alternatives	-0.0676
2987	Words expressing types, varieties, or categories used to refer to multiple similar items or instances	-0.0663
2844	The word "or" used in threats or ultimatums giving two alternatives, where the second is a negative consequence	-0.0662
85	Uses of "or" in phrases expressing indefinite or unspecified alternatives, particularly in constructions like "something or other" and "somehow or other"	-0.0659
639	Instances of "of" where many appear in the phrase "all of a sudden" in narrative contexts	-0.0657
2857	Instances of the phrase "all sorts of" used to describe various unspecified items or activities	-0.0649
2817	Words appearing after "never" or preceded by comparative terms ("more", "most") in narrative dialogue and descriptions	-0.0648
1282	Instances of the words "ever" and "forever" appearing in emotional dialogue expressing permanence	-0.0647
2592	Verbs and adjectives expressing problems, restrictions, or negative outcomes in narrative contexts	-0.0640
1210	Articles and pronouns in literary or poetic contexts exhibiting informal or archaic language use	-0.0624
700	Common English prepositions and basic function words appearing in various narrative and descriptive contexts	-0.0620
1349	Possessive uses of "of" in dialogue or narrative text where "of" connects to a personal reference or ownership	-0.0614
689	Words expressing personal possession or individual autonomy in discussions of self-interest and family relationships	-0.0611
2856	Punctuation marks, connecting words, and phrases used in bibliographic or reference formatting from historical texts	-0.0603
2700	Demonstrative pronouns and nouns used in comparative contexts to reference previously mentioned subjects	-0.0589
1569	Instances of words used as temporal references or comparisons to previous states in narrative contexts	-0.0587
2293	Hedging or qualifying words/phrases used to express possibility, occurrence, or factual statements in narrative contexts	-0.0586
1798	Words appearing in different contexts with similar formatting issues or encoding problems, particularly with special characters or unusual quotation marks	-0.0584
2759	Words expressing consideration, uncertainty, or inquiry ('regard', 'whether', 'whom', 'require') in formal or literary prose contexts	-0.0583
2056	Instances of "or" in phrases indicating indefinite or non-specific circumstances, typically following "some" and preceding "other" or "another"	-0.0581
1276	Past tense and gerund forms of the verb "do" in various narrative contexts	-0.0569

Feature	Llama-13B	Mean △ Prob
818	Function words and punctuation marks appearing in literary or formal texts, typically serving connecting or subordinating roles	-0.0567
1463	Uses of "sorts" to indicate various types or varieties within lists or descriptions of items or behaviors	-0.1504
1273	Opening parenthesis followed by italicized character names or references in dramatic or literary dialogue formatting	-0.1427
1625	Uses of "fellow" as a prefix meaning "other people" or "fellow humans", often with a hyphenated form "fellow-men"	-0.1386
2811	Phrases using "kind" or "sort" to express negation or categorization in dialogue and narrative text	-0.1376
514	Contextual references to previously mentioned items or comparisons, often paired with words like "latter" or appearing in discussions of prior elements	-0.1346
810	Instances of "in the course of" or similar phrases indicating temporal progression in formal writing	-0.1313
2808	Past or conditional uses of the verb "have" in narrative contexts	-0.1243
1398	Use of "in order" as a subordinating conjunction to express purpose or intention in complex sentences	-0.1232
830	Words used in phrases expressing duration or temporal experiences	-0.1196
119	Uses of "if" and similar words in similes and hypothetical comparisons within descriptive narratives	-0.1172
2962	Possessive pronoun "your" used in formal or narrative contexts addressing a second person	-0.1156
2843	First-person personal pronouns and possessive adjectives in narrative dialogue or internal monologue	-0.1144
755	Verbs indicating the transfer of knowledge or instruction in educational contexts	-0.1092
1402	Instances of possessive pronoun "My" in religious or spiritual texts, with one outlier ("I") and one unrelated term ("Sergeant")	-0.1057
1192	Comment markers ('#') and descriptive text in configuration/code files	-0.1056
2096	The word "slightest" used in contexts emphasizing minimal or negligible degree or extent	-0.1051
1790	Section breaks or subtitle separators in biographical text about musicians and artists	-0.1035
2078	Usage of the word "in" following instructions to place or put items, primarily in recipe/cooking contexts	-0.1016
2288	Words and punctuation marks used as text separators or connectors in formal or literary prose	-0.0991
905	Usage of "parts" to describe geographic or spatial divisions of territories, regions, or physical locations	-0.0986
1048	Usage of "the most" in phrases about maximizing or taking advantage of opportunities	-0.0965
1753	The word "and" used in numeric expressions to connect whole numbers with additional quantities	-0.0962
409	Past tense auxiliary verbs (had/hath/hast) in formal or archaic English texts	-0.0961
343	Usage of "no sooner" followed by "than" in narrative text to express immediate sequence of events	-0.0955
719	Instances of the phrase "at the same time" used as a temporal or logical conjunction in formal prose	-0.0955
738	Instances of "and" in phrases expressing increasing or progressive change using comparative adjectives or adverbs	-0.0931
2953	Past tense verbs used in biographical or narrative contexts	-0.0921
1444	Special characters and symbols appearing at the end of text segments, often in technical or markup contexts	-0.0877
901	Personal pronouns and possessive forms used in narrative fiction	-0.0848
537	Punctuation marks (commas and similar delimiters) appearing in bibliographic or reference entries	-0.0838
2395	Words indicating temporal or simultaneous relationships within narrative texts	-0.0831
1062	Words referring to types of work, jobs, or career categories in discussions about employment and social roles	-0.0821
1095	Words used in narrative contexts to indicate timing, causation, or consequence	-0.0815
997	Instances of the word "tell" used in dialogue or direct speech requesting information from someone	-0.0802
201	Uses of "same" in phrases indicating simultaneous actions, consistently appearing in the pattern "and at the same time"	-0.0795
1648	Instances of "can" and "his" used in personal dialogue and narrative descriptions, primarily in question-asking and possessive contexts	-0.0793
1036	Instances of the word "ever" appearing in various literary contexts, often expressing continuity or permanence	-0.0773
1370	Instances of words in text where something moves or passes "through" a collective group or physical space	-0.0769
2088	Question marks at the end of dialogue or questions in literary text	-0.0764

Feature	Llama-13B	Mean Δ Prob
1767	Possessive pronoun "my" and "your" in emotional or dramatic exclamations and declarations	-0.0764
546	Article "The" and one "#" appearing at the start of titles, quotes, or references in text	-0.0563
148	Forms of possessive pronouns and determiners used in sequential interpersonal interactions	-0.0559
2669	Third-person masculine possessive pronoun "his" primarily used in narrative contexts describing physical actions or movements	-0.0545
2895	Uses of the word "or" in phrases indicating an indefinite time or manner (e.g., "somehow or other", "day or two")	-0.0540
2312	Words expressing degree, extent, or manner appearing in narrative prose contexts	-0.0539
2451	Uses of "former" and "respective" as referential adjectives in formal or academic writing contexts	-0.0538
2456	Forms of the verb "be" used in various grammatical constructions expressing states of being or existence	-0.0534
847	Past conditional uses of "have" in narrative fiction dialogues and prose	-0.0533
488	Stage direction markers ending with a period and closing brace in a theatrical script	-0.0520
1912	Instances of "said" appearing after phrases like "as I" or "like I" in dialogue	-0.0518
956	Instances of "one another" used to describe reciprocal relationships or interactions between people	-0.0518
2785	Usage of "or" and similar words in contexts expressing uncertainty or indefinite choices	-0.0517
1784	Verbs and nouns related to observation, probability, or determination of facts	-0.0516
2305	Instances of "one" and "case" used in expressions indicating uncertainty, selection, or generalization	-0.0501
1179	Instances of the definite article "the" in religious and philosophical texts discussing divine or moral concepts	-0.0495
2865	Instances of 'were' used in hypothetical or conditional statements expressing uncertainty or possibility	-0.0487
122	Connecting words and punctuation used as conjunctions or transitions in various texts	-0.0484
57	Function words ('were', 'so', 'as', 'least') used as comparative or conditional modifiers in literary prose	-0.0483
1631	Numbers appearing in sequential lists or page references in bibliographic or indexing contexts	-0.0481
2165	Words serving as conjunctions or relative pronouns in text showing varying formatting and spacing patterns	-0.0477
2408	Uses of the relative pronoun "which" in various literary contexts, primarily introducing dependent clauses	-0.0471
1017	Uses of the word "latter" to refer to the second of two previously mentioned options	-0.0458
2813	Equal signs used as assignment or comparison operators in software configuration and logging contexts	-0.0458
103	Instances of "The" or "the" at the beginning of sentences or clauses, often following punctuation marks	-0.0456
1408	Possessive pronouns used within narrative texts describing personal actions or relationships	-0.0451
2435	Instances of the word "means" used to express methods, ways, or resources to accomplish something	-0.0451
1592	Phrases using "in the least" to indicate minimal or no degree of something	-0.0451
21	Conditional conjunction words ('if' and 'than') appearing at sentence transitions or clause boundaries in narrative text	-0.0450
2717	Verbs and function words expressing uncertainty or concern about future outcomes, often in the form "what will become of" or similar questioning patterns	-0.0441
1940	Words discussing fundamental concepts of being and deity (one, existence, name) in philosophical or religious contexts	-0.0433
661	Instances of "then" following the phrase "every now and" in narrative contexts	-0.0430
1733	The word "and" used as a coordinating conjunction linking related elements in various contexts	-0.0427
1152	Instances of the word "and" used in repetitive sequences to emphasize continuous or ongoing actions	-0.0426
408	Words commonly used as comparative or contrastive conjunctions in various texts	-0.0426
2544	Words indicating requirement or necessity appearing in various contexts	-0.0420
660	Common English function words (articles, prepositions, auxiliary verbs) appearing in various text contexts	-0.0419
2472	Occurrences of "if only" expressing wishes or regrets in narrative contexts	-0.0417
563	Instances of "as if" used in similes or hypothetical comparisons within text	-0.0408
1354	Articles and commas appearing after various words and before whitespace in historical texts	-0.0391
2428	The word "Born" appearing in biographical headers showing birth years of historical figures	-0.0390
1902	Common English function words (articles, pronouns, conjunctions) and punctuation marks appearing in narrative or descriptive text	-0.0388

Feature	Llama-13B	Mean ∆ Prob
2660	Forms of the verb "depend" used to express relationships of causation or reliance in academic or philosophical texts	-0.0381
1294	Adverbial expressions of time or sequence used in narrative passages	-0.0380
637	Uses of "or" in phrases expressing alternative or uncertainty, typically following "some/one form/way" or similar constructions	-0.0378
1731	Words commonly used in temporal expressions or time-related phrases in narrative text	-0.0374
1539	Personal pronouns in texts discussing relationships, religion, and social dynamics	-0.0373
2585	Possessive and determinative pronouns/articles used in narrative prose contexts	-0.0371
1787	Instances of "case" in phrases indicating alternative scenarios or conditions, often preceded by "either" or "which"	-0.0369
2747	Words meaning perpetuity or continuity ('always', 'ever', 'soon') appearing in literary texts with varied formatting and punctuation	-0.0367
2693	Usage of "might" as a modal verb expressing possibility or hypothetical scenarios in narrative contexts	-0.0365
365	Common English words used as conjunctions or relative pronouns to connect clauses in narrative text	-0.0365
2430	Words indicating sequence or contrast ('followed', 'other') appearing in narrative text with consistent formatting and similar usage patterns	-0.0362
288	Words used as comparative or preferential conjunctions in narrative text	-0.0362
1181	Phrases containing "out" or similar uncertain qualifiers, often in constructions like "X out of Y" or "X cases out of ten" expressing probability or measurement	-0.0357
1797	Commas used as separators in lists or sequences of items, names, or ingredients	-0.0355
92	Pronouns and pronoun phrases used to refer to previously mentioned people or things in narrative contexts	-0.0355
2026	Uses of "if" in conditional statements expressing hypothetical situations or wishes	-0.0354
290	Phrases expressing futility, typically in the form "no/of no use" in dialogue	-0.0347
970	Instances of common English function words ('the', 'be', 'other') appearing in narrative text contexts	-0.0344
2401	Usage of function words and prepositions to indicate comparison, extent, or relative position in narratives.	-0.0343
2535	References to groups of people or parties in historical and political contexts	-0.0341
1535	The word "purpose" used in contexts describing intentions, goals, or objectives in historical or narrative texts	-0.0341
1170	Forms of the verb "have" used in various tenses and contexts across different documents	-0.0335
2916	Modal auxiliary verbs 'had' and 'may' in hypothetical or wishful contexts, with one unrelated outlier ('precisely')	-0.0333
603	Commas and "and" used as separators in various numerical, temporal, and list contexts	-0.0333
2789	Possessive pronoun 'his' used in formal or literary prose contexts	-0.0331
34	Stage directions in a theatrical script ending with a period and closing brace, often followed by character dialogue or actions	-0.0328
933	Instances of the conjunction "and" and similar connecting words appearing in descriptive prose text, often in measurements or attempts	-0.0327
1717	Instances of the word "known" preceded by "never be" in narrative contexts	-0.0325
358	Uses of "and" in contexts describing ranges or intervals between two points, values, or entities	-0.0320
1108	Commas used as list or clause separators in formal or technical writing	-0.0320
981	The word "If" or "or" appearing at the start of conditional clauses in various literary contexts	-0.0319
1227	Instances of the word "shared" in contexts discussing China's international relations and diplomatic concepts	-0.0319
495	Commas and other words serving as sentence connectors in literary and historical texts	-0.0318
2208	Phrase "nothing of the sort" and variations appearing at the end of dialogue or statements	-0.0316
2475	Articles and possessive pronouns ('the', 'his', 'their') used as grammatical determiners in narrative texts	-0.0309
2520	Past tense verbs expressing indifference or lack of concern, primarily "cared" and "mattered"	-0.0309
2035	Usage of "No" and "contrary" in dialogue and argumentative contexts as negative responses or contrasting statements	-0.0309
2492	Words used as conjunctions or adverbs to express contrast, simultaneity, or degree in various contexts	-0.0309
86	Prepositions and adverbs used to express direction, accompaniment, or absence in narrative contexts	-0.0308
13	Instances of words or symbols indicating sequence or order in text, such as "latter" and chapter numbers	-0.0307
2223	Common English articles and prepositions ('the' and 'in') appearing in narrative or descriptive text	-0.0305

Feature	Llama-13B	Mean △ Prob
2915	Uses of the word "means" referring to methods, ways, or resources to accomplish something	-0.0305
764	Capitalized nouns and pronouns appearing as characters or entities in narrative text	-0.0300
2466	Demonstrative and possessive pronouns ('His' and 'Those/those') appearing at the start of clauses in various texts	-0.0299
1906	Article "the" appearing in various texts, with one instance capitalized as "The"	-0.0299
2701	Instances of relative pronouns and conjunctions used as connecting words in various texts	-0.0299
2778	Instances of the article "a" appearing after prepositions, conjunctions, or punctuation, followed by temporal or quantitative phrases	-0.0290
1081	Prepositions used to express rates and relationships between two entities	-0.0289
801	Words and phrases related to sustaining life, including references to making a living and being alive	-0.0288
1630	Words serving as sentence connectors or basic pronouns in fragmented text containing unusual whitespace or formatting	-0.0287
2562	Words appearing in phrases expressing negation or qualification ("of the kind", "else", "thought of")	-0.0285
235	Word "some" used in parallel structures to indicate contrast or variety in descriptive lists	-0.0285
411	Common English intensifiers and comparators used in dialogue and narrative prose	-0.0283
262	Instances of 'how' in exclamatory or descriptive passages expressing intensity or degree	-0.0281
467	Instances of the demonstrative pronoun "those" used as a reference to previously mentioned items or groups in academic or formal text	-0.0278
2662	Instances of "Very" appearing as part of the character name "Very Young Man" in a narrative text	-0.0277
1298	Uses of "for" in recipe instructions specifying cooking durations	-0.0275
2661	Instances of the preposition "upon" used in formal or academic contexts, often following "based" to indicate foundations or dependencies	-0.0273
2189	Personal pronouns and words referring to individual human beings in various contexts	-0.0269
2219	Closing curly brace and dot appearing after character names in play dialogue formatting	-0.0265
793	Uses of "other" in comparative phrases structured as "one... other" or similar parallel constructions	-0.0264
2936	Common plural nouns expressing portions, opportunities, or varieties in narrative contexts	-0.0264
1285	Different types of punctuation and common words appearing in varied document formats and contexts, often at line or section boundaries	-0.0259
2236	Punctuation marks used as delimiters in lists and parenthetical expressions	-0.0257
1568	Forms of "never" and "of course" used as emphatic expressions in dialogue or emotional contexts	-0.0256
2028	The pronoun "it" used as a subject in various narrative contexts	-0.0252
104	Numbers appearing in square brackets as reference citations or footnote markers in academic or literary texts	-0.0252
1010	Interrogative words ('?' and 'Or') at the start of questions or alternative propositions in literary texts	-0.0252
1970	Commas separating country names in lists	-0.0251
797	Instances of common words ("one", "of") and basic terms appearing in narrative or dialogue contexts	-0.0250
294	Uses of "so" and "was" in phrases beginning with "As it was" or containing "not so with", indicating comparison or contrast in narrative contexts	-0.0245
1318	Multiple instances of the word "case" used in conditional phrases indicating alternative scenarios or circumstances	-0.0244
1886	Common English articles and auxiliary verbs used in various prose contexts	-0.0243
2930	Instances of the word "way" used to describe obstacles, hindrances, or means of accomplishing something	-0.0242
564	Common English function words appearing after whitespace in narrative text	-0.0241
14	References to past time periods in narrative or historical contexts	-0.0240
1169	Article "the" appearing in various literary and academic contexts	-0.0238
730	Punctuation sequences ending with a closing parenthesis and separator character in bibliographic or reference contexts	-0.0236
1206	Common English phrases "not a bit" and "as a matter of fact" used as fixed expressions in various contexts	-0.0235
2442	Words introducing hypothetical or counterfactual situations, primarily using "but for" constructions	-0.0234
2100	Personal pronouns and words used in self-introductions or identifications in dialogue	-0.0230
1031	Uses of the word "partly" and variations of "total" in different narrative contexts indicating partial or complete amounts	-0.0227

Feature	Llama-13B	Mean Δ Prob
1460	Usage of "for" in cooking instructions indicating duration of time	-0.0222
2090	Pronouns and general nouns used in narrative contexts describing actions or events	-0.0218
1068	Prepositions and spatial terms used to describe movement or relative position in narrative texts	-0.0217
325	Punctuation marks followed by single quotes in dialogue or quoted text	-0.0215
1785	Nouns referring to participants in text-based interactions (reader, critic, visitor) appearing in formal or literary contexts	-0.0214
2243	Second-person pronoun "you" and third-person pronouns "he/him" appearing in dialogue with non-standard English or dialectal speech	-0.0214
2655	Uses of the word "of" in phrases describing geographical locations or parts of places	-0.0212
938	Instances of "again" (or "Again") appearing in repetitive phrases like "again and again" in narrative contexts	-0.0211
2123	Instances of the word "and" used as a conjunction connecting two related clauses or phrases in various texts	-0.0211
1137	Verbs expressing persuasion or influence over others' actions	-0.0210
70	Commas used after dates or numbers in various types of documents	-0.0210
2291	Instances of the word "Of" at the start of independent clauses or sentences, typically following punctuation	-0.0210
2634	First-person singular pronoun 'I' used as a subject at the beginning of dialogue responses	-0.0208
760	Numbers appearing in sequences of comma-separated lists in various document indices or references	-0.0208
2816	Temporal words (mostly "first" and "before") marking sequence or timing in narrative text	-0.0207
1330	Commas used as separators in lists of geographic locations or institutions	-0.0206
1822	Uses of "or" and similar verbs as informal connectors in casual or uncertain statements	-0.0205
1484	Numbers or identifiers (often '2') used for section or list enumeration in various texts	-0.0203
2685	Possessive pronouns in various narrative and descriptive contexts	-0.0201

Table 7: Features for OLMo-7B and OLMo-13B comparison

Feature	OLMo-7B	Mean Δ Prob
2033	Repetitive use of 'and' as a conjunction to emphasize continuity, distance, or repetition	0.0843
2597	Past and present tense forms of the verb "to go" and "to be" in narrative contexts	0.0763
2111	Religious or spiritual references to divine entities, specifically "The Lord" and "His" in religious texts	0.0751
1091	Variations of "no sooner... than" and "depended" used in narrative sequences describing cause and immediate effect	0.0735
725	Uses of "or" in phrases indicating an unspecified member of a set or an approximate quantity	0.0720
1378	Verbs and pronouns related to human cognition, perception, and emotional states	0.0709
1790	Punctuation marks indicating section breaks in biographical text about musicians	0.0687
1667	Variations of "other" used in phrases indicating alternation or reciprocity, often in patterns like "one...or other" or "each other"	0.0666
1877	Closing angle brackets followed by quotation marks in XML/markup files	0.0658
1534	Phrases indicating duration or passage of time, often used as temporal transitions in narrative text	0.0636
972	Temporal adverbs appearing at the start or middle of sentences in narrative text	0.0594
241	Instances of 'to' describing spatial positions, directions, or physical movements, particularly involving body postures and orientations	0.0559
215	Instances of cooking and baking time specifications in recipe instructions	0.0553
174	Adverbs indicating timing, degree, or extent in narrative contexts	0.0532
2976	Words or titles from formal or professional contexts indicating roles, positions, or organizational entities	0.0522
1972	Equals signs followed by text strings in software configuration or property files	0.0514
42	Words referring to unspecified groups or quantities in text discussing human activities and choices	0.0513
2501	Possessive pronouns and forms indicating ownership or personal connection in literary dialogue and narrative text	0.0511
274	Variations of the word "ground" used to express reasoning or basis for actions/claims in legal and argumentative contexts	0.0510
2796	Instances of 'no sooner' used as a temporal phrase in narrative sequences	0.0507
1571	The word "former" used in comparative contexts to reference a previously mentioned item or subject	0.0491
1295	Uses of "which" as a relative pronoun in complex sentence structures	0.0458
183	Page numbers used as references in bibliographic or index entries	0.0449
2290	The word "thought" followed by "of" in contexts expressing concern, worry, or contemplation	0.0448
236	Uses of the phrase "in order" as a subordinating conjunction to express purpose or reasoning	0.0443
2246	Common English verbs and function words used in narrative prose	0.0440
2503	Words indicating spatial or temporal boundaries and extents in narrative texts	0.0438
1314	Uses of 'than' in comparative phrases indicating temporal or quantitative measurements	0.0413
2886	Past tense verbs describing completed actions in narrative contexts	0.0394
246	File path separators in video game ability file paths	0.0381
1607	Text sequences where punctuation marks and common words appear immediately before whitespace characters, typically in narrative contexts	0.0374
2635	Words used in formal or bureaucratic writing to indicate roles, actions, or positions of entities	0.0362
715	Words appearing after hyphens in compound expressions using numerical quantities	0.0357
1752	Abbreviated measurements of weight (lb.) appearing in texts about dyeing and chemical processes	0.0346
802	Preposition 'to' used to indicate direction, purpose, or relationship between elements in various contexts	0.0341
1845	Variations of the phrase "somehow/sometime/sooner or other/later" used as informal expressions of uncertainty or inevitability	0.0339
272	Commas followed by quotation marks in dialogue punctuation	0.0336
1931	Honorific titles and article "The" appearing at the start of sentences or proper nouns in formal writing	0.0334
225	Instances of common English pronouns and prepositions used in narrative prose	0.0323
251	Honorific title for married women appearing in narrative prose	0.0319
1664	Forms of pronouns and determiners used in conversational and narrative contexts	0.0316
1615	Past tense verbs related to knowledge, understanding, or recognition	0.0313
1177	References to the number "15" appearing in various document contexts, often as section numbers, footnotes, or page numbers	0.0306

Feature	OLMo-7B	Mean Δ Prob
1553	Past tense verbs and pronouns used in narrative storytelling contexts	0.0293
2926	Prepositions used in comparative or descriptive phrases indicating difference or relation	0.0280
1808	Words commonly used in formal or legal writing to reference previous statements or establish context	0.0273
129	Uses of "whole" referring to complete durations or entireties of time periods	0.0272
2630	Formatting marks and special characters in various document contexts, including italics markers and punctuation	0.0272
2647	Instances of the word "refers" used as a verb to indicate citation or reference to other sources in academic or literary contexts	0.0269
500	Past tense verbs expressing possibility, occurrence, or reflection	0.0255
2828	Nouns used in 19th century prose describing social and property relations	0.0254
2373	Reference numbers in square brackets appearing in academic or historical texts	0.0254
2485	Variations of "other" and "others" used as pronouns to reference alternative or additional items in a sequence	0.0241
1903	Phrases using variations of "take into account/consideration" meaning to consider or factor in something	0.0235
1017	Uses of "latter" as a reference to the second of two previously mentioned options or items	0.0232
1287	Personal and impersonal pronouns used as sentence subjects in English text	0.0231
675	Uses of the word ’from’ in geographical or spatial descriptions indicating starting points of routes, paths, or boundaries	0.0227
416	Instances of "a" and "per" used as function words in formal or regulatory contexts, often following "as"	0.0227
19	Ampersands and other punctuation marks used as abbreviations in bibliographic or reference entries	0.0220
1650	Page or reference number 203 appearing in academic citations and footnotes	0.0215
869	Third-person singular pronoun "it" used as a subject in complex sentences	0.0214
2339	Double asterisks followed by dialog or exclamatory text in literary works	0.0210

Feature	OLMo-13B	Mean Δ Prob
830	Context-specific references to periods or durations in casual conversation	-0.1599
905	References to geographical, physical, or organizational divisions of spaces, regions, or entities	-0.1456
585	Punctuation marks followed by quotation marks in bibliographic and literary texts	-0.1368
639	Instances of "of" and "too" in narrative prose, typically appearing in transitional or descriptive phrases	-0.1342
716	Common English conjunctions and auxiliary verbs used in connecting clauses and forming questions	-0.1242
39	Instances of common English articles and function words in various literary contexts	-0.1229
2987	Words indicating various categories or classifications used to group or describe things	-0.1217
485	Common English verbs (and one noun) used in various everyday contexts	-0.1183
21	Subordinating conjunctions used to introduce conditional or comparative clauses in narrative prose	-0.1164
2093	Superlative adjectives expressing degree or intensity in various contexts	-0.1147
2011	Words expressing mental states or cognitive processes in dialogue or narrative contexts	-0.1120
343	Usage of ’no sooner’ in narrative text to indicate immediate sequence of events	-0.1078
2475	Definite articles and possessive pronouns functioning as grammatical determiners in various narrative contexts	-0.1071
1753	The word "and" used in numeric expressions and measurements	-0.1050
1471	Common conjunctions used to express uncertainty or alternatives in narrative prose	-0.1030
1463	Multiple instances of the word "sorts" used to indicate variety or different types, along with some other general categorical terms	-0.1018
2096	The word "slightest" used as an adjective to emphasize minimal or negligible degree or extent	-0.1011
1036	Instances of the word "ever" used in various literary contexts with different meanings and connotations	-0.1003
1840	References to a character called "the Very Young Man" in a narrative text	-0.1001
2915	Uses of the word "means" referring to methods, resources, or ways of achieving something	-0.0988
1349	Uses of "of" in possessive constructions following demonstrative pronouns (this/that)	-0.0981
1170	Forms of the verb "to have" in various sentence contexts	-0.0978
1501	Instances of words used for equivalence or identity across different contexts	-0.0944
2142	Word "until" used in cooking instructions to indicate duration endpoint	-0.0943
1625	Usage of "fellow" as an adjective describing other humans or citizens in formal moral/ethical discourse	-0.0939
2740	Adverbs of quantity/degree used in narrative prose contexts	-0.0937

Feature	OLMo-13B	Mean △ Prob
901	Personal pronouns and possessive markers used in narrative texts	-0.0932
792	Personal possessive pronouns in religious or spiritual texts	-0.0925
119	Uses of "if" in similes expressing hypothetical comparisons	-0.0920
961	Words used as sequential or temporal markers in narrative texts	-0.0919
392	Religious or biblical references to a male divine figure or moral actor, typically in formal or archaic English	-0.0910
257	Question words and forms ("what", "is", "?") in interrogative or philosophical contexts	-0.0910
738	Repetitive use of "and" between comparative adjectives or adverbs to emphasize gradual change or progression	-0.0840
1679	The phrase "now and then" used to indicate occasional or intermittent occurrences in narrative texts	-0.0834
2570	Pronouns and adjectives used in comparative or referential contexts within narrative or argumentative text	-0.0808
2857	The phrase "all sorts" used to indicate various or multiple types of something	-0.0804
1230	Common English words used as connective or transitional phrases in narrative text	-0.0797
1861	Words related to referring to or bringing up something in conversation or text	-0.0787
69	Instances of the auxiliary verb "have" used to express desire, intention, or requirement in historical texts	-0.0771
25	Instances of 'than' in contexts describing immediate sequential actions, often following 'no sooner'	-0.0770
1767	First-person possessive pronouns expressing personal ownership or relation in emotional or dramatic contexts	-0.0769
659	Conditional or introductory words used in narrative contexts	-0.0761
1626	Uses of the word "possible" expressing feasibility or potential in various contexts	-0.0758
1943	Temporal phrases using "now" and similar words to indicate periodic or intermittent occurrences in narrative text	-0.0749
2766	Adverbs or adjectives expressing certainty or extent in narrative contexts	-0.0740
2811	Phrases indicating an unspecified type, variety, or category, often used in denials or general references	-0.0726
2037	Words commonly used in constructions indicating extent, manner, or degree in narrative texts	-0.0722
1019	Common English words used to connect logical statements or express factual relationships in formal writing	-0.0718
1912	Variations of speech reporting verbs ("said", "told", "don't") preceded by "as I" or "I" in dialogue	-0.0717
2430	Forms of "followed" and "other hand" used in narrative transitions and sequential actions	-0.0712
2374	Words related to disagreement, opposition, or deviation from a norm in formal discourse	-0.0711
2207	Uses of the word "same" indicating similarity or identical nature across various contexts	-0.0710
2165	Conjunctions and relative pronouns used as connecting words in various texts	-0.0701
2810	References to time markers in narrative texts, primarily using "o'clock" notation	-0.0691
1352	Words indicating comparison, contrast, or consideration between multiple viewpoints or alternatives	-0.0686
537	Punctuation marks (commas, brackets) appearing at the end of text segments in bibliographic or reference-style entries	-0.0680
1179	The definite article 'the' appearing in various religious and philosophical texts	-0.0672
320	Contractions of "would not" used in questions, typically appearing at the end of dialogue	-0.0669
2383	Auxiliary verbs and related words expressing conditional or hypothetical situations in narrative contexts	-0.0646
1401	Phrases indicating short time periods or durations	-0.0636
1648	Common English auxiliary verbs and pronouns used in dialogue and narrative context	-0.0631
1559	File references and structural elements in PHP-related configuration or documentation	-0.0630
2808	Modal verb "have" used in conditional or hypothetical statements	-0.0624
2505	Instances of "face to face" encounters or direct confrontations between people or animals	-0.0622
1152	Repeated use of "and" in sequences describing continuous or repetitive actions	-0.0616
197	Equal signs used as section or line markers in academic or annotated texts	-0.0610
318	Personal and impersonal pronouns used as sentence subjects in narrative text	-0.0606
2236	Punctuation marks following list items in educational or reference texts	-0.0597
23	The word "only" appears predominantly after "if" in phrases expressing wishful thinking or regret	-0.0592
2661	The word "upon" used in contexts of dependency, basis, or relationship between concepts	-0.0586
1048	Variations of the phrase "make/making the most of" used to describe taking advantage of opportunities	-0.0585
1910	Apostrophes in contractions and informal speech representing dropped letters	-0.0580
2484	Indefinite and possessive articles/pronouns used in narrative or formal text	-0.0577

Feature	OLMo-13B	Mean △ Prob
563	Instances of "as if" used to make hypothetical comparisons or analogies in text	-0.0573
2756	Numbers or fractions containing "1" in various technical and instructional contexts	-0.0568
2104	Common English function words and determiners used in narrative text	-0.0564
2379	Uses of the word "information" in privacy policy and data collection contexts	-0.0555
793	Uses of the word "other" in comparative or alternative contexts, often paired with "one" or describing relationships between entities	-0.0554
201	Uses of the phrase "at the same time" to indicate simultaneous actions or concurrent conditions	-0.0543
1569	Instances of the word "before" (and similar temporal words) used as an adverb to reference a previous state or action	-0.0543
2088	Question marks at the end of dialogue or interrogative statements	-0.0543
2705	Instances of common pronouns ("what" and "it") at the start of quoted speech or sentences	-0.0543
2511	Instances of "so" used as an affirmative response or agreement in dialogue	-0.0530
481	Uses of the word "latter" in comparison/reference contexts indicating the second of two previously mentioned items	-0.0530
1451	People in authority or leadership roles mentioned in narrative contexts	-0.0530
1088	Instances of "what is the matter" used as a questioning phrase to express concern or inquiry	-0.0529
2009	Text segments showing dialogue breaks or transitions in aboriginal or pidgin English narratives	-0.0529
2429	Uses of "all kinds" to indicate comprehensive variety or completeness in lists or descriptions	-0.0527
2156	Phrases using "all parts" or "many parts" to describe geographic distribution or widespread locations	-0.0521
2189	Personal pronouns and words referring to individual human beings in texts discussing human nature and relationships	-0.0509
1276	Past tense and gerund forms of the verb "do" in various narrative contexts	-0.0506
730	Right parentheses followed by commas in bibliographic or reference citations	-0.0501
2544	Modal or auxiliary words expressing requirement, necessity, or possibility in various contexts	-0.0499
2662	References to "The Very Young Man" as a character in a narrative text	-0.0494
2560	Words and characters appearing in formal or antiquated transitional phrases like "be that as it may" or "somehow or other"	-0.0492
78	Instances of "part" in phrases indicating actions, behaviors, or responsibilities of specific parties	-0.0492
2368	Right curly braces appearing at the start of lines in dramatic or poetic text, followed by various dialogue or narrative content	-0.0492
2487	Uses of the preposition 'of' in various grammatical constructions and contexts	-0.0491
1269	Usage of "so" and variations of "kinds" as function words in descriptive prose	-0.0487
2100	Personal pronouns and given names in dialogue from narrative texts	-0.0482
1206	Common fixed phrases "not a bit" and "as a matter of fact" used as conversational expressions	-0.0478
1354	Articles and commas appearing in various historical or literary texts with inconsistent spacing around them	-0.0471
1281	Articles preceding descriptions of sudden sounds or events in narrative texts	-0.0471
2145	Forms of "result" and its synonym "consequence" used to describe causation or outcomes	-0.0469
2789	Instances of the possessive pronoun "his" in literary or formal texts	-0.0465
513	Numbers used as reference markers or page numbers in academic or bibliographic contexts	-0.0462
57	Function words in literary/narrative text indicating comparison, degree, or hypothetical states	-0.0461
2769	References to footnote number 75 in various documents	-0.0458
1890	Uses of the word "necessary" indicating requirement or essential need across various contexts	-0.0456
2086	Articles and common conjunctions appearing in various literary and technical contexts	-0.0449
2281	Articles and conjunctions used as connecting words in formal or historical texts	-0.0446
14	Words referring to temporal or identifying labels in historical or nostalgic contexts	-0.0444
828	Words expressing uncertainty, probability, or relative states in various contexts	-0.0444
645	Uses of the word "time" and its contextual appearances in narrative text	-0.0438
1714	Dialogue markers showing character reactions or transitions, followed by quoted speech	-0.0435
631	Single quotes appearing after semicolons in dialogue sequences	-0.0430
2141	Page numbers, punctuation marks and other separators used in document indices or bibliographic entries	-0.0429
2043	Common English pronouns and auxiliary verbs appearing in dialogue or narrative text	-0.0417
2919	Forward slashes appearing in file paths and API endpoints in technical documentation	-0.0413

Feature	OLMo-13B	Mean Δ Prob
1716	Commas and the word "other" used as separators or references in various narrative contexts	-0.0411
2535	References to collective groups of citizens or inhabitants in historical and political texts	-0.0411
1254	The phrase "at last" used as a temporal marker indicating the end of a period of waiting or anticipation	-0.0408
2905	Words and punctuation marks appearing at the end of clauses or phrases in academic texts	-0.0407
2552	Past and present tense auxiliary and action verbs in narrative contexts	-0.0404
943	Words appearing in questions or statements about decision-making and obligations	-0.0403
1061	Words appearing in phrases expressing uncertainty or hypothetical situations, primarily using "as if" constructions	-0.0402
719	Instances of the phrase "at the same time" used as a transition or conjunction in formal writing	-0.0399
2300	Assignment or equality operators in configuration and code settings	-0.0398
262	Instances of "how" used as an intensifier in literary or formal prose	-0.0397
2860	Common English prepositions and conjunctions used in various literary contexts	-0.0396
2930	Uses of "way" meaning obstacle or impediment in formal prose	-0.0395
836	Modal verbs ("should" and "ought") expressing obligation or recommendation in instructional contexts	-0.0393
192	Instances of the word "manner" (and similar terms) used to describe ways or methods of doing things	-0.0392
1689	Forms of "to be" verbs and "who" used in narrative or documentary contexts	-0.0390
1645	Right parenthesis and period punctuation pair appearing after author names or references in a bibliography or catalog	-0.0389
2711	Function words used in dialogue and narrative prose to express conditional, temporal, or modal meanings	-0.0386
1190	Punctuation marks followed by whitespace in various document contexts	-0.0383
801	Words related to life, survival, and existence used in various contexts	-0.0383
1741	Words used in descriptive narrative contexts to discuss states, situations, or characteristics	-0.0376
2956	Commas serving as text separators in various bibliographic and literary contexts	-0.0373
2114	References to footnote or figure number 167 in various academic texts	-0.0371
2028	Pronoun "it" used as subject or object in narrative prose	-0.0369
2305	Words used as pronouns or expressions indicating individual instances or hypothetical situations in formal text	-0.0368
2451	Words indicating previously mentioned items or relative positioning in comparative contexts	-0.0367
2489	Uses of "and" following numbers between one hundred and three hundred in numeric expressions	-0.0366
2101	Uses of the word "kinds" to indicate variety or multiple types of items in a list or collection	-0.0366
661	The phrase "every now and then" used to indicate periodic or occasional occurrences in various contexts	-0.0359
2685	Possessive pronouns in literary or narrative contexts	-0.0356
205	Usage of "were" in constructions involving hypothetical or figurative comparisons, particularly in the phrase "as it were"	-0.0355
865	Adjectives and adverbs describing increasing volume or intensity in narrative contexts	-0.0353
1612	Variations of the phrase "did/do the same" indicating copied or repeated actions	-0.0353
748	Personal pronouns and related variations in religious and philosophical texts discussing existence and self	-0.0351
1568	Instances of common expressions using "of course" and emphatic repetitions of "never" in dialogue	-0.0350
495	Commas and other punctuation marks used as separators in lists of proper names	-0.0348
1408	Personal and possessive pronouns used in narrative contexts	-0.0347
2449	Past-tense usage of "done" indicating completion or conclusion of actions or events	-0.0347
2442	Instances of the words ’but’ and ’only’ used as conjunctions or qualifiers in hypothetical or conditional statements	-0.0346
637	The word "or" used in phrases expressing alternatives or uncertainty, often following "some," "one," or similar patterns	-0.0334
2971	Past tense forms of "to be" and references to time appearing in narrative contexts	-0.0333
2186	Words used in textual or numerical enumerations and listings, often appearing in formal or dated documents	-0.0333
132	Opening quotation marks at the start of dialogue in narrative text	-0.0332
2863	Uses of "as soon as" in narrative texts to indicate immediate temporal sequence	-0.0332
2288	Common conjunctions and punctuation marks in religious or formal texts	-0.0326
1143	Forms of the basic words "same" and "knew" used as verbs or adjectives in narrative contexts	-0.0326
1285	Words and symbols marking the end of data entries or records in various document formats	-0.0318

Feature	OLMo-13B	Mean ∆ Prob
325	Punctuation marks followed by single quotes in dialogue or quoted text	-0.0316
2171	Closing curly braces at the end of numerical data arrays in a programming or configuration file with Asian character annotations	-0.0312
2747	Words indicating perpetuity or temporal continuation appearing in literary dialogue and narrative	-0.0309
2579	Commas used as separators in various textual contexts, appearing between words or phrases	-0.0308
2737	Verbs and nouns describing human actions, behaviors, and personal stakes in various contexts	-0.0301
70	Commas appearing after dates, numbers, or geographic locations in various documents	-0.0300
1732	Commas used as separators in lists or sequences	-0.0297
1403	Uses of "and" in numerical expressions between one hundred and three hundred	-0.0296
2672	Special characters used as delimiters in database or code mappings	-0.0295
408	Comparative and connective words used in evaluative contexts	-0.0291
1130	Phrases describing physical appearance or coloring in a person's face	-0.0289
1605	Past tense verbs expressing mental states or sensory experiences in narrative contexts	-0.0286
633	Instances of "sort" and "course" used as references to unspecified actions or situations	-0.0286
147	The word "worse" appearing in variations of the phrase "for better or worse" in different contexts	-0.0283
832	Words used as pronouns referring to previously mentioned entities or actions in formal writing	-0.0282
840	Punctuation and dialogue markers in dramatic or theatrical text formatting	-0.0282
1961	Articles and pronouns used as grammatical function words in 19th century English prose	-0.0281
1969	Underscores appearing before text ending in periods or commas, typically in formatting or reference contexts	-0.0281
2865	Past subjunctive form of "to be" used in hypothetical or conditional statements	-0.0279
569	Common transitional or qualifying phrases in formal writing	-0.0277
358	Uses of 'and' in phrases expressing ranges or intervals between two values or endpoints	-0.0275
712	Articles and pronouns used as common grammatical elements in narrative text	-0.0275
2526	Common English prepositions and function words used in various narrative contexts	-0.0274
1108	Punctuation marks used as delimiters in lists and clauses across various academic and technical documents	-0.0272
581	Uses of "and" in contexts involving numbers between 150-250	-0.0271
409	Archaic or biblical forms of auxiliary verbs meaning "have" or "had" used in literary or religious texts	-0.0267
391	Words indicating temporal or spatial position within a sequence or area	-0.0266
2456	Forms of the verb "be" used as auxiliaries or main verbs in various contexts	-0.0264
1380	Possessive pronouns referring to male subjects in narrative contexts	-0.0264
1235	Words occurring in informal dialogue or narrative text with colloquial language patterns	-0.0264
1039	Third-person pronouns used as sentence subjects in narrative text	-0.0260
2353	Double asterisks marking dialogue breaks or speaker changes in literary text	-0.0257
726	Instances of the words "few" and "same" used as determiners or adjectives in various contexts	-0.0255
1950	Command action labels in a software interface, typically describing file and application operations	-0.0252
2844	The word "or" used as a conjunction to introduce threatening alternatives or consequences	-0.0250
2533	Forms of the preposition "in" appearing in recipes and texts, primarily used to indicate incorporation or fitting within a context	-0.0241
1809	Verbs expressing possession, requirement, or maintaining/losing control	-0.0240
2052	Words functioning as common English auxiliary verbs or prepositions in various narrative contexts	-0.0239
1138	Words related to knowledge and understanding appearing in narrative contexts	-0.0237
1449	Instances of "same" and "hand" used in temporal or sequential phrases, primarily in the construction "at the same time"	-0.0235
1407	Instances of common quantifiers or numerals used in various contexts	-0.0234
1060	Commas used as thousand separators in numerical values within various texts	-0.0234
112	Temporal expressions at the start of sentences indicating the beginning of narrative events	-0.0233
222	Equal signs before text labels in software configuration or properties files	-0.0233
908	Auxiliary verb "have" used to form future, conditional, or perfect tenses in different narrative contexts	-0.0229
2058	References and citations using numerical or punctuation markers in bibliographic or footnote contexts	-0.0225
1372	Uses of the word "time" indicating a specific moment or concurrent events, with one preposition "from" and one conjunction "and" as outliers	-0.0225

Feature	OLMo-13B	Mean △ Prob
120	Uses of "other" as an adjective to refer to a second person or entity previously mentioned in the context	-0.0224
535	Common prepositions and possessive markers appearing after conditional phrases or similar qualifying statements	-0.0223
2057	Usage of "exception" in phrases describing exclusions from a larger group, typically following "with the"	-0.0217
122	Common English conjunctions and connective words used to link phrases or clauses in formal writing	-0.0213
2554	Articles appearing in various academic or technical contexts	-0.0211
1676	Verbs related to showing, proving, or explaining something in academic or narrative contexts	-0.0210
127	Question marks at the end of various queries or interrogative statements	-0.0210
1348	Words indicating type, manner, or category used as nouns in formal or literary contexts	-0.0209
1602	The word "and" appearing in the phrase "now and then" indicating periodic occurrence in narrative texts	-0.0206
1642	Uses of the relative pronoun "which" in literary or formal texts	-0.0204
2587	Vertical bar symbols used as delimiters in tabular or structured text data	-0.0201

Table 8: Features for Llama-7B and Llama-7B-chat comparison

Feature	Llama-7B	Mean Δ Prob
1790	Double hyphens used as section separators in biographical texts about musicians and artists	0.4829
2127	Words and punctuation used as section breaks or transitional elements in formal texts from a similar time period	0.3644
2780	Uses of "afterwards" indicating a subsequent event or time period in narrative text	0.3299
1627	Punctuation sequence ".’ " appearing at the end of quoted dialogue in literary text	0.0954
209	Uses of the word "respect" with the preceding phrases "in this" or "in that" to indicate specific reference to a previously mentioned aspect	0.0952
1232	Forms of the verbs "to be" and "from" used in questions and statements about past experiences or movement	0.0877
50	Uses of the phrase "the sort of" to describe types or categories of people, places, or things	0.0868
1139	Words commonly used as grammatical function words or determiners in English sentences	0.0830
1855	Periods appearing in various numerical or technical contexts, including file paths, decimal numbers, and version numbers	0.0785
957	Words appearing in common English phrases or expressions like "out of", "by the way", "in the aftermath of", and "at the same time"	0.0771
744	Starting words of sentences or clauses in literary or academic text	0.0736
1534	Common transitional or temporal phrases used in narrative text	0.0713
394	Uses of "share" referring to a rightful portion or contribution in a distribution or responsibility	0.0703
174	Words used as intensifiers or modifiers in narrative descriptions, often indicating degree or extent	0.0701
743	Instances where words express uncertainty, lack of knowledge, or unresolved situations in narrative contexts	0.0699
1628	Function words showing comparison or relation, specifically ’to’ and ’which’ used in formal or literary contexts	0.0632
1194	Words referring to concepts of truth, proportion, and components in philosophical or religious texts	0.0629
1779	Uses of words indicating brief time periods in narrative passages	0.0614
2083	Question marks at the end of quoted dialogue in literary texts	0.0600
2015	Words commonly used as general references or placeholders in discussions of actions, events, or topics	0.0584
2033	Instances of the word "and" in contexts describing distances or repetitive sequences	0.0538
2290	Uses of the word "thought" expressing mental distress or anxiety about future events	0.0537
2134	Preposition used in cooking/preparation instructions to indicate duration of time	0.0534
1971	Noun used to introduce or emphasize established or claimed truths in formal writing	0.0529
2640	Second-person pronoun "thee" and similar archaic words used in formal or religious texts	0.0525
2714	Programming color codes and syntax formatting patterns in code documentation	0.0519
1063	Words expressing variety, timing, or extent (’kinds’, ’some’, ’soon’) in different contexts	0.0512
402	Words that function as connective or transitional terms in formal literary prose	0.0502
2629	Single period characters followed by whitespace in document page references and citations	0.0487
1667	Expressions using ’other’ or variations to indicate alternation between two options or sides	0.0483
2121	Second-person pronouns and related words appearing in various narrative contexts	0.0472
2784	Verbs expressing personal preferences or opinions	0.0460
28	References to humans or living creatures in literary text discussing personhood or humanity	0.0457
2621	Words describing methods or approaches, specifically "manner" and "way" used in explanatory contexts	0.0456
1287	Personal and impersonal pronouns used as subjects in formal and narrative texts	0.0448
604	Prepositions used in measuring or describing physical distances	0.0447
971	Article "a" appearing at the start of noun phrases in various contexts	0.0446
968	Question marks at the end of sentences in philosophical or contemplative discourse	0.0430
34	Stage directions indicating character names followed by periods in a play script	0.0405
1987	Uses of the word "way" and similar terms expressing manner or method in narrative contexts	0.0400
2449	Instances of the past participle "done" occurring at the end of clauses, often in phrases like "said and done"	0.0393
2520	Past tense verbs expressing indifference or lack of concern	0.0387
2796	Instances of "No sooner" followed by "than" in narrative sequences	0.0386
2802	Use of "been" in contexts describing first-time experiences or prior experiences	0.0378
2111	Religious or spiritual references to God/Lord using reverential capitalization	0.0374

Feature	Llama-7B	Mean △ Prob
629	Terms related to human society, work, and social organization, with numbers appearing in sequential lists	0.0367
2025	Character names followed by a period in stage directions or dialogue markers from theatrical scripts	0.0360
1046	Reference numbers appearing in footnotes, citations, or page numbers in academic or historical texts	0.0348
1819	Function words ('is' and 'and') used as connectors in narrative texts	0.0346
903	Question marks or exclamation marks followed by quotation marks in dialogue or emotional text	0.0340
1348	Generic nouns used as references to previously mentioned items or categories	0.0335
55	Verbs expressing desire or preference, primarily using "like" in polite or formal contexts	0.0333
374	Pronouns referring to God or divine entities in religious or spiritual texts	0.0325
1866	Conjunction "Or" used at the start of sentences or clauses to present alternatives or questions	0.0324
2708	Words beginning sentences that establish time sequences in narratives	0.0322
2264	Different uses of the word "way" in contexts meaning "in terms of" or "regarding"	0.0319
1845	Instances of "or" used in phrases expressing indefinite time or manner, such as "sooner or later" and "somehow or other"	0.0315
725	Uses of the word "or" as a conjunction connecting alternatives, typically following the number "one"	0.0314
319	Words and symbols used as units of measurement or time in recipes and technical instructions	0.0310
2679	Instances of the article "a" in various literary and historical texts	0.0308
1708	Second-person pronoun "you" used as direct address in various narrative and dialogue contexts	0.0305
637	The word "or" appearing in phrases expressing alternatives or variations, typically in patterns like "some way or other" and "one form or another"	0.0300
2713	Past participle form of "to be" in narrative contexts describing past experiences or states	0.0297
533	Words appearing in phrases indicating sudden or unexpected events, particularly "all of a sudden" and similar constructions	0.0279
1664	Words functioning as pronouns or determiners referring to unspecified people, things, or instances	0.0275
2220	Examples of "followed" used to describe physical pursuit or movement in narrative sequences	0.0271
1674	Prepositions used in sequences or dependencies indicating repetition or accumulation	0.0268
1386	Instances of the word "soon" appearing after "as" in narrative sequences describing immediate actions	0.0266
812	Character dialogue markers in dramatic text using a period and underscore notation	0.0266
1211	Numbers appearing in educational or instructional contexts, often as question or section markers	0.0265
147	Instances of "worse" appearing in variations of the marriage vow phrase "for better or worse"	0.0265
1833	Words and phrases used as conjunctions or transitions to indicate alternatives or temporal sequence, particularly in patterns like "sooner or later"	0.0259
1769	A word used to refer to the second of two previously mentioned items in comparative contexts	0.0257
2035	Words used as negative or contradictory responses in dialogue	0.0253
599	Prepositions and adverbs indicating temporal or spatial relationships in narrative text	0.0252
2582	Punctuation marks serving as delimiters or separators in various document contexts including indexes, citations, and numerical values	0.0251
250	Repeated instances of "each other" in phrases describing mutual relationships or interactions	0.0249
77	Words followed by punctuation marks in file paths, configuration settings, and chapter headings	0.0245
2616	Numbers and text appearing in various index, reference, or listing contexts	0.0243
912	Adverbs used in comparative or relative clauses expressing degree or extent	0.0242
1822	Uses of "or" in phrases expressing non-specific alternatives, typically following words like "some" or "something"	0.0239
2577	Nouns describing types, categories, or domains of things	0.0229
2208	Instances of "nothing/anything of the sort" and similar phrases used as emphatic denials in dialogue	0.0225
1224	Phrases starting with "For some time" or "After a time" used as temporal transitions in narrative text	0.0220
2879	Common words used as measure words or quantifiers in formal writing	0.0217
515	The word "of" appearing in sudden or unexpected situations, often in the phrase "all of a sudden"	0.0211

Feature	Llama-7B	Mean Δ Prob
26	Single periods appearing in text layouts with consistent spacing patterns and line formatting	0.0211
2765	Variations of "there/There" used as existential pronouns at the start of statements	0.0210
802	The word "to" used as a preposition connecting two elements in various instructional or descriptive contexts	0.0207

Feature	Llama-7B-Chat	Mean Δ Prob
1893	Archaic words used to indicate movement or direction from a previously mentioned place	-0.2896
2222	Critical apparatus entries in scholarly editions showing textual variants, marked by parallel bars and containing manuscript sigla	-0.1834
585	Punctuation marks following quoted text in various literary contexts	-0.1690
2915	Uses of the word "means" referring to methods, ways, or instruments for achieving a purpose	-0.1683
537	Commas appearing in bibliographic and reference entries	-0.1636
1466	Instances of the coordinating conjunction "and" in academic citations and references	-0.1606
2865	Past subjunctive form of "to be" used in hypothetical or conditional statements	-0.1557
1281	Articles preceding descriptions of sudden sounds or disturbances in narrative texts	-0.1510
2101	Uses of the word "kinds" to indicate variety or multiple types of items in a list or collection	-0.1496
1190	Punctuation marks used as delimiters in various textual contexts	-0.1484
997	Forms of "tell" used in questions or requests for information	-0.1468
1767	Personal pronoun "my" used in emotional or dramatic contexts expressing personal loss, suffering, or deep feeling	-0.1389
52	Third-person male pronoun "he/He" used as the subject of various narrative sentences	-0.1386
2511	Variations of ’so’ and ’sometimes’ used as affirmative responses or qualifiers in dialogue and descriptive text	-0.1350
1048	Phrases using "make/making the most" to express maximizing or taking full advantage of opportunities or situations	-0.1348
1626	Uses of the word "possible" in contexts expressing feasibility or capability	-0.1322
205	Uses of ’were’ in hypothetical or figurative expressions, often appearing with ’as it’ or conditional phrases	-0.1317
2295	Possessive pronouns used in formal or historical writing	-0.1253
741	Phrases indicating customary or habitual behavior, often using "as was/is" followed by words like "wont," "custom," or "fashion"	-0.1216
1038	Forms of "think" and auxiliary verbs in questioning or contemplative dialogue	-0.1207
2147	Articles, pronouns, prepositions, and currency symbols appearing at the start of document lines or after punctuation	-0.1180
2088	Question marks at the end of interrogative sentences in dialogue	-0.1176
164	Personal pronouns and function words used as dialogue in narrative text	-0.1114
1645	Author or publication citations ending with a closing parenthesis and period in bibliographic entries and references	-0.1111
409	Past tense auxiliary verbs (had, hath, hast) used in archaic or formal religious texts	-0.1107
495	Grammatical elements (punctuation and conjunctions) used in narrative texts to connect clauses and phrases	-0.1100
1403	Usage of "and" connecting numerical values in measurements and counts	-0.1097
313	Conditional statements starting with "if" followed by the word "only"	-0.1073
235	Uses of "some" as part of contrasting pairs or lists describing varying qualities or actions	-0.1069
563	Usage of ’if’ in similes or hypothetical comparisons indicated by ’as if’ constructions	-0.1066
2463	Past tense forms of "to be" followed by "to" in various narrative contexts	-0.1066
2967	Forms of the verb "to be" used in various contexts showing existence or state	-0.1059
361	Words expressing reference to additional or different people or things	-0.1056
1170	Forms of the verb "have" used in various grammatical contexts	-0.1052
2811	Phrases using "kind" or "sort" to deny or qualify statements, often in dialogue or formal writing	-0.1043
905	Nouns referring to geographic or spatial divisions of regions or territories	-0.1028
1108	Commas used as separators in lists or clauses across various technical and academic contexts	-0.1020
2971	Past tense forms of the verb "to be" and temporal words in narrative contexts	-0.1019
2475	Articles and possessive pronouns used as grammatical determiners in narrative texts	-0.0996
159	Words occurring in similar book titles following the pattern "[TOPIC] EVERY CHILD SHOULD KNOW"	-0.0996
1088	Interrogative usage of "what is/what’s the matter" in dialogue expressing concern or inquiry about a problem	-0.0994
1169	Instances of common English articles and function words in various literary contexts	-0.0986
1471	Conjunctions used to express uncertainty, possibility, or alternatives in narrative prose	-0.0985
2489	The word "and" appearing in numerical expressions between hundreds and smaller numbers	-0.0984
2456	Forms of the verb "be" used as auxiliary or linking verbs in various contexts	-0.0978

Feature	Llama-7B-Chat	Mean ∆ Prob
2972	Common English transition words and prepositions used in various narrative contexts	-0.0702
1325	Equal signs used as delimiters between professional titles/occupations and other information in directory-style listings	-0.0681
165	Words referring to people or human attributes in literary or academic texts	-0.0675
2300	Assignment or equality operators in configuration and metadata files	-0.0674
970	Words commonly used as articles or determiners in English appearing in narrative text	-0.0673
2444	Synonyms used as general references or placeholders in casual speech and writing	-0.0672
417	Common English function words (pronouns, conjunctions, prepositions) appearing at the start of clauses in formal or archaic text	-0.0667
1648	Common English function words (pronouns and modal verbs) used in narrative prose contexts	-0.0666
1243	References to medical patients and diseases in healthcare contexts	-0.0657
859	Action or state words (verbs and adjectives) indicating permission, potential, or capability in various contexts	-0.0656
78	Uses of "part" in phrases indicating actions, behaviors, or responsibilities of specific parties	-0.0656
156	Double asterisks used as separators for dialogue or quotations in literary text	-0.0646
2893	Articles and common connecting words appearing in historical and biographical texts	-0.0644
2930	Instances of the word "way" used to describe obstacles, interference, or opposition in formal text	-0.0639
645	Form of the phrase "at the same time" and other temporal expressions using the word "time"	-0.0638
1920	The word "and" used to connect numerical values in measurements, quantities, or counts	-0.0630
86	Prepositions and adverbs used in motion or accompaniment contexts	-0.0630
1635	Lines beginning with colons followed by dialogue or conversation snippets in dramatic or theatrical texts	-0.0627
1597	Relative and interrogative pronouns and adverbs used in narrative contexts	-0.0626
546	Articles (mainly "The") appearing at the start of titles or publication names in various texts	-0.0622
1568	Common expressions using "of course" and "never" for emphasis in dialogue or emotional statements	-0.0620
245	Common prepositions and articles appearing at line breaks in formatted text	-0.0618
2857	Phrases using "all sorts" to indicate variety or multiple instances of something	-0.0609
425	Numbers appearing as reference markers, footnotes, or section numbers in academic or literary texts	-0.0607
1913	Forms of auxiliary verbs 'is' and 'has' appearing in literary or formal prose contexts	-0.0598
343	The word "sooner" appearing in temporal phrases with "no" and "than" to indicate immediate sequence of events	-0.0597
2236	Punctuation marks used as separators in lists and parenthetical expressions	-0.0596
760	Numbers appearing in index or reference lists with surrounding commas and page numbers	-0.0596
1759	The word "and" used as a conjunction to connect time ranges or intervals	-0.0596
916	Common verbs and adjectives used in dialogue and narrative contexts, primarily 'said' and 'possible'	-0.0589
1714	Double asterisks appearing in dialogue or quoted speech indicating pauses or breaks in conversation	-0.0586
2664	Words indicating relative ranking or comparison within a group	-0.0584
1039	Personal pronouns used as sentence subjects in narrative texts	-0.0579
1552	Words denoting sequence or enumeration in various forms (numerals, ordinals, and related terms)	-0.0578
2759	Words expressing consideration or uncertainty in formal discourse	-0.0570
635	Double asterisks followed by quotation marks marking dialogue breaks in literary text	-0.0568
2305	Phrases indicating singularity or individual instances in various contexts	-0.0565
320	Variations of "would not" appearing in dialogue or questions	-0.0562
787	Usage of "and" in numerical expressions between one hundred and two thousand	-0.0562
2211	Articles "the" and "a" used as determiners in various academic and professional texts	-0.0550
1095	Words indicating causation, outcome, or reasoning in formal texts	-0.0547
1066	Past tense and modal verbs expressing states of mind, belief, or experience	-0.0541
462	Words expressing finality, intent, or extent of commitment in narrative contexts	-0.0536
2572	Common English conversational phrases and transitional expressions used to qualify or modify statements	-0.0536
461	Uses of the word "pages" in book and publication metadata describing length	-0.0535
1662	Occurrences of "ever" in religious or spiritual contexts discussing eternity or perpetuity	-0.0533
2737	Common verbs and nouns describing human actions, emotions, and experiences in formal prose	-0.0521
222	Equal signs followed by labels or status messages in software configuration files	-0.0520
719	Phrase "at the same time" used as a transition or conjunction in literary text	-0.0519
1395	Double asterisks appearing after the end of narrative segments in prose text	-0.0518

Feature	Llama-7B-Chat	Mean Δ Prob
1848	Common English articles and determiners used in various written contexts	-0.0518
1016	Prepositions used to establish relationships between subjects in various written contexts	-0.0505
1555	Phrases using "so much as" to emphasize minimal or threshold actions that are prohibited or notable	-0.0505
2435	Uses of the word "means" referring to methods, tools, or ways of accomplishing something	-0.0505
746	Words appearing in phrases expressing duration or time relationships, mainly in constructions with "as long as" and similar temporal expressions	-0.0500
1950	Assignment operators in configuration or properties files for a software system	-0.0494
868	Commas followed by quotation marks in dialogue endings across literary texts	-0.0490
357	Relative pronouns and determiners used in narrative prose	-0.0487
436	Common English prepositions and conjunctions used in narrative text	-0.0487
2608	Commas separating clauses in historical or narrative text	-0.0487
1749	Words used as conjunctions or connecting words in narrative text	-0.0486
2038	Configuration comments denoted by '#' in a vehicle settings file	-0.0486
442	Common English words used to indicate relationships or comparisons between elements in diverse contexts	-0.0486
358	Uses of "and" in phrases describing ranges or intervals between two values or endpoints	-0.0481
2581	Question marks and common prepositions at the end of text fragments in literary dialogue	-0.0477
2471	Uses of the article "a" in contexts of searching, finding, or creating paths and solutions	-0.0476
1487	Phrases containing common words ("tell", "with", etc.) in formal or narrative dialogue from literature	-0.0473
1244	Instances of "view" used in the phrase "with a/the view of" to indicate purpose or intention	-0.0473
2291	Uses of the word "Of" at the start of clauses in formal or historical texts	-0.0471
1732	Commas used as list separators in various enumerations	-0.0469
1080	Instances of "matter" used in dialogue to inquire about problems or concerns	-0.0467
275	Uses of the preposition "beside" indicating physical proximity or adjacency between people	-0.0463
2189	Personal pronouns and words referring to individuals or people in various contexts	-0.0453
294	Instances of "as it was" and "it is not so" used as contrastive phrases in narrative writing	-0.0447
1745	Uses of the indefinite article 'a' in various contexts, often appearing in phrases indicating time or condition	-0.0447
2529	Instances of "for" in cooking instructions specifying duration of cooking steps	-0.0445
85	Uses of "or" in phrases expressing indefinite alternatives, typically in constructions like "somehow or other" and "some form or other"	-0.0438
70	Commas used after dates in various contexts	-0.0436
1413	Personal and demonstrative pronouns in narrative contexts	-0.0435
1276	Past tense forms of "do" used as auxiliary or main verbs in narrative contexts	-0.0421
1525	Words indicating imminent or upcoming events in narrative passages	-0.0419
262	Adverb 'how' used for emphasis in exclamatory or emotional expressions in literary text	-0.0415
846	Words indicating location, possession, or measurement in formal or administrative contexts	-0.0415
2685	Possessive pronouns used in various narrative and descriptive contexts	-0.0415
22	Forms of the verb "to be" (is/are) in various literary contexts	-0.0414
2844	The word "or" used as a conjunction to introduce threats or ultimatums in dialogue	-0.0412
867	Multiple instances of "out of the question" used as a phrase indicating impossibility or refusal	-0.0409
568	Articles and possessive pronouns used as determiners in literary prose passages	-0.0406
938	Repetitive usage of words to emphasize continuous or repeated actions	-0.0405
2654	Conjunctions used as sentence connectors in narrative text	-0.0405
1858	Temporal expressions beginning sentences that establish new narrative events	-0.0401
1372	Instances of "time" and "from" used as temporal and spatial prepositions in narrative contexts	-0.0396
1318	Instances of the word "case" used to indicate conditional or alternative scenarios	-0.0395
1844	Third-person masculine pronouns and their possessive forms in narrative contexts	-0.0395
1068	Directional and spatial movement words in narrative contexts describing physical transitions or distances	-0.0394
1474	Second-person and first-person plural pronouns in dialogue or direct speech	-0.0393
1254	Phrases using "at last" to indicate the final occurrence or conclusion of a waiting period	-0.0392
1787	Phrases using "case" in conditional or alternative scenarios, typically following "in either" or "in which"	-0.0390
1036	Instances of the word "ever" in literary passages expressing permanence or continuity	-0.0382
1380	Possessive pronouns in action-oriented narrative passages	-0.0382
1670	Phrases indicating short, indefinite time durations, often using "or" to connect two numbers	-0.0372
2695	Common prepositions used to express temporal or spatial relationships in English text	-0.0365
2937	Word "in" used as part of phrasal verbs or prepositions in narrative contexts	-0.0360

Feature	Llama-7B-Chat	Mean Δ Prob
1408	Personal and possessive pronouns used to refer to people in narrative contexts	-0.0354
1620	Indefinite article ’a’ used to introduce new people, objects, or situations in narrative contexts	-0.0349
1530	Personal pronouns and verbs related to personal interaction or relationships	-0.0339
1503	Words describing someone’s behavior, demeanor, or way of conducting themselves in social situations	-0.0339
196	Forms of common function words (articles, pronouns, auxiliary verbs) appearing in written prose	-0.0338
497	Words indicating type, nature, or category (including "being", "kind", "sort", "thing") used in descriptive contexts	-0.0337
421	Numbers preceded by "No." in various catalog or inventory listings	-0.0335
2376	Words appearing in comparative constructions using "the... the..." pattern in English sentences	-0.0335
120	Uses of the word "other" as an adjective to refer to a second person, entity, or thing previously mentioned	-0.0334
2026	Instances of "if" in contexts expressing hypothetical or comparative situations, often following "as"	-0.0332
2588	Conjunctions and connective words appearing at the start of poetic or literary lines	-0.0329
2325	Determiners and pronouns used at the start of sentences in formal or academic text	-0.0327
201	Uses of the phrase "at the same time" indicating simultaneous actions or states	-0.0324
2682	Page numbers appearing in citations and footnotes in scholarly texts	-0.0321
1294	Common English function words appearing after "all at" or "the" in narrative text	-0.0317
2954	Words used to reference or distinguish between previously mentioned items in comparative contexts	-0.0315
1716	Separators or connectors between items in lists and paired concepts in text	-0.0306
407	Commas followed by names in a bibliographic or directory listing	-0.0304
391	Words indicating temporal or spatial position within a sequence or area	-0.0301
1856	Adjectives and pronouns denoting additional or alternative items in a sequence or group	-0.0299
273	Words referring to types, categories, or instances of things	-0.0298
1747	Commas used for clause separation or list delimiting in formal written text	-0.0292
1901	The word "the" appearing as a definite article in various narrative contexts	-0.0292
2672	Special characters used as delimiters in database or programming variable assignments and transformations	-0.0290
2355	First and second person pronouns used in informal or dialectal dialogue	-0.0283
2207	Instances of words ’same’ and ’place’ used as references to indicate identical conditions or position replacement	-0.0277
1528	Chapter number ’V’ appearing in various table of contents and chapter headings	-0.0275
2560	Words appearing in idiomatic expressions expressing uncertainty or alternatives, such as "be that as it may" and "somehow or other"	-0.0273
2655	Instances of "of" following words referring to geographic subdivisions or regions	-0.0268
322	Conjunctions expressing alternatives or conditions in narrative contexts	-0.0265
2646	Uses of the indefinite article ’a’ in formal prose discussing society, governance, and morality	-0.0265
2505	Phrases indicating direct encounters or confrontations between two parties	-0.0262
2114	Reference number 167 appearing in various scholarly citation and footnote contexts	-0.0257
793	Instances of "other" appearing in phrases expressing alternatives, relationships, or duality between two entities	-0.0254
631	Single quotation marks following dialogue in literary text, typically appearing after semicolons	-0.0253
597	Connective or transitional words used in formal or literary text to reference previously mentioned items or temporal sequences	-0.0251
849	Numbers and symbols used for enumeration or section marking in various texts	-0.0244
185	Common English articles and prepositions used in various contexts	-0.0243
657	The word "latter" used to refer to the second of two previously mentioned items in contrasting contexts	-0.0243
1399	Possessive apostrophe-s used across different literary contexts	-0.0240
615	Common English phrases using intensifiers or qualifiers (e.g., "nothing of the sort", "all the same")	-0.0239
20	Personal pronouns used as sentence subjects in narrative texts	-0.0239
1010	Question marks and the word ’Or’ appearing as single-word lines or section titles in literary works	-0.0237
827	Relative pronouns used in formal or literary prose	-0.0230
2407	Common function words (’if’, ’a’, ’is’, ’than’, ’I’) appearing in various narrative contexts	-0.0228

Feature	Llama-7B-Chat	Mean △ Prob
2837	Instances of "part" (with one exception) used in phrases about participation or involvement in activities	-0.0228
329	Nouns describing social interaction or accompaniment in various contexts	-0.0228
582	Function words used in connecting or modifying clauses in formal or literary prose	-0.0226
1970	Commas separating country names in lists	-0.0225
2421	Numbers appearing in sequential listings or references within academic or bibliographic contexts	-0.0221
1484	Sequential numbers used as reference markers or section numbers in academic or literary texts	-0.0218
661	Phrases using "now and then" to indicate periodic or intermittent occurrences	-0.0216
2039	Prepositions used as function words in formal or academic texts to express relationships between elements	-0.0213
220	Numbers appearing at the beginning of questions or sections in educational or literary texts	-0.0208
2447	Forms of comparison and evaluation words "worse" and "worst" plus contextually similar verbs of communication expressing negative situations	-0.0204
2082	Sequential or transitional words and punctuation marks used in various literary and document contexts	-0.0202
82	Personal pronouns used in narrative dialogue and prose	-0.0201

Table 9: Features for OLMo-7B and OLMo-7B-DPO comparison

Feature	**OLMo-7B**	**Mean Δ Prob**
1790	Punctuation mark sequences separating section headings in biographical texts about musicians	0.2166
2942	Words and their contexts describing formal or ceremonial robes and clothing in various settings	0.2046
2780	Uses of "afterwards" to indicate subsequent events in narrative text	0.1492
28	Forms of the word "being" used to refer to human or living entities	0.1437
1667	Variations of the phrase "one ... or other" expressing alternatives or opposites	0.1431
717	Commas used as separators in various texts spanning different languages and contexts	0.1096
1359	Abstract nouns referring to existence and entities in philosophical or metaphysical discussions	0.1035
1324	Commas separating numbers in numerical lists or sequences	0.1021
2588	The word "And" used as a coordinating conjunction at the beginning of lines in poetic or literary text	0.0990
2127	Words and punctuation marks used as transitional elements in narrative or sequential text	0.0986
215	The word "for" used in cooking instructions to specify duration of cooking time	0.0973
271	Words serving as subjects or objects in sentences about governance and public behavior	0.0947
1585	Words expressing general categories, observations, or casual inspection ("kind(s)", "look") with variations	0.0940
1649	Instances of "up" used in phrases about maintaining pace or keeping pace with others while moving	0.0920
2551	Common nouns used in academic or philosophical discourse to refer to topics of discussion	0.0892
2026	Instances of "if" used in similes or hypothetical comparisons, often following "as" or "than"	0.0887
2677	Personal pronouns and possessive determiners used in narrative contexts	0.0872
251	Honorific title "Mrs." used to address or refer to married women in narrative text	0.0851
805	Common English words and punctuation marks appearing at the end of index entries or in narrative text	0.0838
2177	Common English words used as quantifiers or modifiers in narrative text	0.0829
2683	Forms of "to be" verbs (is/was) used in descriptive passages	0.0822
1460	Time duration indicators in cooking recipe instructions	0.0807
2501	Personal possessive pronoun "my" used in emotional or dramatic expressions in literary texts	0.0805
236	Uses of the phrase "in order" followed by "to" or "that" to express purpose or intention	0.0758
2546	Punctuation marks and common words appearing after text segments and before line breaks or continuations	0.0742
1298	Usage of ’for’ in cooking instructions indicating duration of cooking or processing time	0.0731
1022	Common operators and symbols used in programming assignments and configuration files	0.0725
548	Punctuation marks at the end of questions or statements in various texts	0.0711
1069	Uses of the phrase "of any kind" and similar variations in different contexts	0.0699
1080	Uses of "matter" in questions expressing concern about someone’s wellbeing or condition	0.0676
690	Instances of common function words used in formal or literary prose passages	0.0674
2333	Various uses of the word "belongs" expressing ownership, categorization, or membership	0.0668
2134	The word "for" used as a preposition to indicate duration of time in recipes and instructions	0.0667
533	Uses of the phrase "all of a sudden" in narrative text describing unexpected events	0.0667
2831	Common phrases using determiners or quantifiers (kind/kinds, most, way) to express measure, type, or extent	0.0648
2121	Second-person pronouns in direct speech or narrative addressing the reader/listener	0.0643
2225	Words indicating initial or immediate instances in narrative contexts	0.0642
129	Uses of the word "whole" to indicate complete time periods or entirety in historical and narrative texts	0.0639
957	Instances of "way" appearing in the phrase "by the way" used as a conversational transition	0.0636
1250	Commas used for narrative pauses in literary prose passages	0.0630
1229	Personal pronouns and function words used in narrative text passages	0.0629
2770	References to legal and intellectual constructs in formal or published texts	0.0629
2529	Instances of "for" in recipe instructions indicating cooking duration or waiting time	0.0615
802	Preposition "to" used as a connector between elements in various contexts	0.0614
966	Question marks at the end of dialogue in narrative text	0.0609
2603	Present and past tense verbs related to knowledge, presence, and existence	0.0598
1389	Words expressing sympathy or compassion in narrative contexts	0.0575
175	Punctuation marks at the end of numerical or bibliographic entries	0.0573
2111	References to divine or religious pronouns and articles in spiritual or religious texts	0.0567
306	Usage of "way" in phrases indicating manner or means of assistance/support	0.0565

Feature	OLMo-7B	Mean Δ Prob
993	Commas used as separators in various types of bibliographic or index entries	0.0564
640	Instances of "keep up" meaning to maintain pace or match speed with someone/something	0.0561
683	References to specific time periods measured in years within historical or biographical texts	0.0561
1314	Comparative conjunction ’than’ used after expressions indicating measurement or quantity	0.0559
1704	Scientific measurements and quantitative metrics used in technical or research contexts	0.0553
2049	Uses of "of" in phrases expressing emphasis or ranking, particularly following variations of "most"	0.0552
191	Words and numbers used in comparative or sequential contexts with corresponding numerical data or ordinal relationships	0.0548
117	Placeholder "0" used as a parameter in error messages and system logs	0.0531
1644	Third-person pronouns and simple past-tense verbs in narrative contexts	0.0521
795	Instances of "there" or "There" used as an existential or locative marker in narrative text	0.0504
2646	Instances of the indefinite article ’a/an’ in literary or formal prose passages	0.0499
593	The word "there" used as an existential or locative term in various sentence contexts	0.0498
2033	Words ’and’ and ’some’ used in repetitive or emphatic sequences to convey magnitude or intensity	0.0478
968	Question marks appearing at the end of sentences in literary or philosophical texts	0.0476
1464	Comparative and relative terms used to express degrees of similarity or difference	0.0470
1197	Uses of "or" in phrases expressing indefinite alternatives, typically in the pattern "some [time/way/one] or [other/another]"	0.0461
693	Interjections expressing emotions or reactions in dialogue	0.0458
1176	Words marking transitions or progression in text, including punctuation and comparative terms	0.0453
1226	Question marks at the end of interrogative sentences in literary or academic texts	0.0450
1447	Double asterisks followed by text demonstrating dialogue, thoughts or narrative breaks	0.0441
1998	Words expressing time permanence or continuity in narrative texts	0.0441
873	Punctuation marks and comparison words used as sentence separators or connectors in various texts	0.0433
2724	Exclamation marks at the end of emotional or emphatic statements in literary texts	0.0432
2599	Uses of the word "kind" indicating type, variety, or classification in various contexts	0.0430
1287	Personal and impersonal pronouns used as sentence subjects in various texts	0.0428
40	Forms of the word "thought" used as a past-tense verb indicating mental activity or consideration	0.0421
1674	Words commonly used as temporal or sequential prepositions in narrative contexts	0.0419
2290	Instances of "thought" expressing mental distress or worry about future events	0.0404
2432	Single dots appearing in rows of dots used for text alignment or spacing in document layouts	0.0402
2811	Phrases using "kind" or "sort" to indicate a generic type or category, often in dismissive or speculative contexts	0.0398
2035	Negations and contrasts used in dialogue and argumentative text	0.0394
20	Personal pronouns used as sentence subjects in narrative text	0.0393
725	Uses of "or" in phrases expressing indefinite quantities or choices, often following "one" or "something"	0.0380
971	Singular indefinite article ’a’ used in narrative time descriptions and measurements	0.0379
2634	First-person singular pronoun "I" appearing in direct speech or dialogue	0.0377
2490	Common English articles and conjunctions used as grammatical connectors in various texts	0.0376
2797	Articles and pronouns appearing at the start of sentences or clauses in various texts	0.0373
1518	"At any rate" used as a transitional phrase to qualify or modify previous statements	0.0371
462	Modal verbs and nouns expressing intention, duration, or degree in formal or archaic writing	0.0366
969	Words expressing degrees of certainty or factuality in various contexts	0.0364
1053	Uses of ’as’ in constructions forming comparisons or indicating extent, typically followed by ’to’	0.0363
1279	Forms of common linking verbs and conjunctions in literary or philosophical text passages	0.0362
1479	Commas used as punctuation marks in narrative text separating clauses or items in a series	0.0357
1957	Prepositions indicating physical location or spatial relationships in narrative descriptions	0.0357
1403	The word "and" appearing in numeric expressions between hundreds and smaller numbers	0.0356
246	Backslashes appearing as file path separators in video game ability file paths	0.0354
529	Demonstrative pronouns used in literary or philosophical texts to refer to previously mentioned concepts	0.0353
2655	Prepositional phrases using "of" to describe geographical divisions or sections of locations	0.0346
531	Function words used in comparative or referential contexts within sentences	0.0343
2503	Distance or time measurements using comparative phrases with ’more’ and temporal endpoints	0.0339

Feature	OLMo-7B	Mean ∆ Prob
2123	Instances of ’and’ used as a coordinating conjunction connecting two related clauses or phrases in literary or philosophical texts	0.0337
2057	Phrases using "with the exception of" to indicate exclusion from a larger group	0.0336
1294	Common English function words used as temporal markers or connectors in narrative texts	0.0332
1991	Words indicating instances, situations, or occurrences used in narrative contexts	0.0331
812	Full stops followed by speaker changes in dramatic dialogue	0.0323
2025	Period symbols following character names in dramatic play dialogues and stage directions	0.0318
135	Usage of "other" in contexts describing conflict or competition between two parties	0.0315
55	Verbs and adjectives expressing desire, preference, or likelihood in formal or literary dialogue	0.0313
483	Personal pronouns used as subjects in narrative dialogue and prose	0.0312
1845	Phrases using "or" as part of expressions indicating eventual or inevitable occurrence, typically in the form "sooner or later" or "some day or other"	0.0311
753	Instances of the definite article "the" in various literary and historical texts	0.0310
1372	Words used as temporal and spatial connectors in narrative text	0.0305
2069	Articles and prepositions used in various academic or historical texts	0.0304
50	Uses of "the sort of" followed by a noun describing a person, place, or thing in comparative statements	0.0299
1206	Common phrases "not a bit of it" and "as a matter of fact" used as speech or narrative transitions	0.0288
2879	Nouns used in historical or biographical contexts to describe involvement, time periods, and types	0.0285
1500	Comparative word "less" used to indicate reduced quantity or value	0.0281
134	Uses of "than" in comparative phrases indicating temporal or spatial measurements	0.0276
921	Uses of the word "way" in phrases indicating manner, method, or direction	0.0270
912	Adverbs used as part of comparative phrases beginning with "so" or "as"	0.0268
1681	Commas used as separators in various bibliographic and catalog-style entries	0.0265
1137	Verbs expressing persuasion or facilitation of actions by others	0.0264
1992	Instances of the word "regard" (plus one "content") used in formal writing to express consideration or attention to something	0.0264
1195	Commas separating authors’ first and last names in a bibliographic index	0.0263
2207	Words ’same’ and ’place’ used in contexts of replacement or equivalence	0.0263
1325	Equal signs used as field separators in structured records containing professional or occupational information	0.0251
2291	Instances of "Of" at the start of clauses introducing numerical or quantitative information	0.0250
2763	Uses of "there" as an expletive or existential pronoun to introduce statements about existence or presence	0.0248
1770	Past tense verbs indicating the start or progression of actions in narrative contexts	0.0247
1911	Commas followed by whitespace in lists of proper nouns or numbers	0.0246
2081	Commas used as separators in various textual contexts including titles, dialogue, and lists	0.0242
1749	Uses of "as" and similar comparative words in various sentence constructions	0.0238
1822	Connecting words or conjunctions used in phrases expressing uncertainty or alternatives	0.0236
2577	Words indicating types, varieties, or categories in different contexts	0.0227
26	Single periods appearing after multiple dots in document section headers and references	0.0226
2339	Exclamatory or emphatic interjections followed by dialogue or narrative text	0.0224
1592	Forms of the phrase "not in the least" or similar expressions using "least" to indicate minimal or no degree of something	0.0223
1054	Uses of "that" and related phrases ("that is") serving as explanatory or clarifying conjunctions in various contexts	0.0222
2739	Possessive forms using apostrophe-s in various text excerpts	0.0221
2141	Document formatting elements including periods, page numbers, and references appearing at line endings or in indices	0.0220
2956	Punctuation marks and words acting as separators in various bibliographic and poetic contexts	0.0220
1409	Personal pronouns and words related to obligation or timing in narrative prose	0.0213
2679	Uses of the indefinite article ’a’ in various measurements, comparisons, and descriptions	0.0212
329	Uses of the word "company" and similar terms expressing accompaniment or association with others	0.0208
2738	Phrases indicating the ending or duration of a time period	0.0204

Feature	OLMo-7B-DPO	Mean ∆ Prob
585	Punctuation marks at the end of quoted text followed by additional punctuation	-0.1902
2142	Forms of "until" used in cooking instructions to indicate cooking duration or completion state	-0.1855
1877	XML-style closing markers in technical documentation	-0.1470

Feature	OLMo-7B-DPO	Mean △ Prob
2798	Uses of filler phrases indicating variety or necessity ("manner of", "need of", "sort of") in literary contexts	-0.1392
1700	Prepositions and adverbs used idiomatically in expressions indicating manner, direction, or degree	-0.1205
2905	Instances of 'others' used in lists or comparisons to contrast with 'some' or similar terms	-0.1168
1273	Opening parenthesis followed by underscore in literary dialogue or stage directions	-0.1086
1569	Instances of the word "before" used as an adverb to indicate a previous time or state	-0.1043
1625	Uses of the word "fellow" in contexts discussing relationships and obligations between humans in society	-0.1018
2645	Words indicating subsequent or additional items in sequences or listings	-0.1002
310	Question marks appearing at the end of quoted dialogue or rhetorical questions in literary text	-0.0997
905	Instances of "part(s)" and "area" used to describe geographic or spatial divisions	-0.0876
2718	Phrases using "such a" followed by a word indicating method or type	-0.0865
324	Words expressing degree, extent, or threshold in narrative contexts	-0.0833
2756	The number "1" appearing in measurements, quantities, or data values across different contexts	-0.0817
2472	The word "only" appearing in conditional "if only" expressions indicating wishes or regrets	-0.0814
2778	The indefinite article "a" used as a determiner before singular nouns in various narrative contexts	-0.0797
2444	Phrases using "sort" or "sorts" as a qualifier to describe or reference a type or category of something	-0.0786
633	Informal phrases using 'sort' and 'course' to reference previously mentioned concepts or situations	-0.0776
2451	Words describing temporal or relational order, particularly in reference to previously mentioned items	-0.0746
2171	Closing curly braces in game configuration data with numeric values and foreign text annotations	-0.0735
1070	Transitional words and phrases used to indicate timing, sequence, or temporal relationships in narrative contexts	-0.0733
2511	Words functioning as affirmative responses or qualifiers in dialogue and narrative contexts	-0.0716
830	Temporal expressions used in informal dialogue or personal reflections	-0.0714
1281	Indefinite article 'a' preceding descriptions of sudden sounds or events in action sequences	-0.0712
2100	Personal pronouns and names in dialogue indicating character speech or identification	-0.0711
1784	Verbs and phrases related to observation, verification, or occurrence in formal or narrative contexts	-0.0710
2813	Equals signs appearing in configuration or property assignment statements across software logs and settings	-0.0687
5	Special characters used as formatting or list markers at the start of lines in bibliographic or reference texts	-0.0683
1912	Variations of the phrase "as I said" used in dialogue or narrative speech	-0.0680
792	Possessive pronouns 'my' and 'his' used in religious or spiritual contexts referring to a divine Lord	-0.0675
1858	Sentence-initial instances of "One" followed by time-of-day or time-period words to begin narrative sequences	-0.0673
997	Forms of the verb "tell" used in dialogue or reported speech requesting or sharing information	-0.0662
313	Instances of conditional phrases or statements using "if only" and similar constructions expressing wishes or conditions	-0.0659
2088	Question marks at the end of dialogue or questions in literary text	-0.0659
2864	Words and phrases used as transitional or temporal connectors in narrative text	-0.0656
901	Personal pronouns and possessive markers used in narrative contexts	-0.0655
1048	Usage of "make/making the most" phrases indicating maximizing opportunities or benefits	-0.0645
1626	Variations of the word "possible" used to express feasibility or extent in different contexts	-0.0642
1533	Common English function words used in narrative or descriptive contexts	-0.0640
2915	Uses of the word "means" referring to methods, resources, or ways of achieving something	-0.0640
278	Words commonly used to express uncertainty, dependency, or approximation in narrative contexts	-0.0637
2793	Common linking or transitional words used in casual dialogue and narrative prose	-0.0632
21	Subordinating conjunctions used to introduce conditional or comparative clauses in narrative text	-0.0621
1145	Phrases using "of the sort" or similar expressions to indicate refusal or denial	-0.0617
2022	Punctuation marks used as delimiters or separators in various textual contexts	-0.0613
1187	Verbs requesting information or asking someone to share knowledge in dialogue	-0.0611
1181	Forms of the phrase "X out of Y" where X and Y are numbers or measures in various contexts	-0.0593

Feature	OLMo-7B-DPO	Mean △ Prob
2256	Unit of measurement (pounds) in dyeing instructions	-0.0590
361	Words expressing various forms of "other" or plurality in texts discussing social relationships and interactions	-0.0580
481	Uses of "latter" to refer to the second of two previously mentioned items	-0.0576
1276	Past tense forms of the verb "do" in narrative contexts	-0.0568
1451	Characters or people mentioned in narrative text who have distinct roles or identities	-0.0556
151	Indefinite pronouns or words expressing uncertainty in narrative contexts	-0.0555
2810	Time-related indicators in text, primarily occurrences of "o'clock" and temporal references	-0.0551
668	Common English articles and conjunctions appearing in various narrative and religious texts	-0.0548
2475	Common English articles and possessive pronouns used in narrative prose	-0.0547
438	Pronouns used as sentence subjects to refer to previously mentioned concepts	-0.0541
439	Common English function words appearing at the start of sentences or clauses	-0.0540
2351	Time expressions indicating subsequent events, typically following a numeric duration	-0.0534
164	Common English pronouns and prepositions appearing at the start of sentences or clauses in narrative text	-0.0533
1950	Text labels for file and application management actions in a software interface	-0.0527
2641	Adverb indicating occasional or intermittent occurrence used in parallel sentence structures	-0.0523
281	Uses of "or" in phrases expressing uncertainty or vagueness, typically in the pattern "something or other"	-0.0520
1271	Common English function words ('if', 'of', 'better') used in various grammatical constructions	-0.0519
1306	Forms of verbs and the noun "men" appearing in contexts about human behavior and social interactions	-0.0518
832	Words used as referential pronouns to indicate simultaneity or previous mention in formal text	-0.0517
2844	The word "or" used in threats or ultimatums presenting negative consequences	-0.0513
1730	Dialogue punctuation marks and common conversational words in literary text	-0.0512
474	Qualifiers or comparative words used to express similarity, degree, or extent	-0.0507
661	Uses of the phrase "every now and then" indicating periodic or occasional occurrences	-0.0502
320	Contracted form of "would not" appearing in informal dialogue or questions	-0.0501
1230	Common English phrases indicating manner, extent, or progression ("in the way of", "in the course of", "make the most of")	-0.0501
2493	Uses of the definite article "the" in various written contexts discussing authority, governance, and education	-0.0498
332	First-person pronouns and contractions of "do not" in casual dialogue	-0.0497
1532	Uses of the word "ever" in phrases indicating perpetuity or eternal duration	-0.0486
2808	Forms of the verb "have" used in various narrative contexts, plus one instance of "her" as an outlier	-0.0484
2456	Various uses of the verb "be" in modal or future constructions expressing possibility, necessity, or prediction	-0.0480
1759	Uses of "and" in phrases describing time intervals or periods between two points	-0.0474
867	Words or phrases in quotation-style contexts discussing impossibility or negative responses	-0.0471
2099	Common English auxiliary verbs and pronouns used in narrative contexts	-0.0461
2395	Time-related adverbs and phrases indicating simultaneous or concurrent events	-0.0460
2189	Personal pronouns and words referring to people or individuals in narrative contexts	-0.0460
2293	Common phrases expressing uncertainty, factuality, or qualification in formal writing	-0.0454
2300	Equals signs used as assignment or definition operators in configuration or code files	-0.0450
1474	Second-person pronouns used in direct speech showing confrontational or aggressive addressing	-0.0448
201	Uses of "same" in phrases indicating simultaneous actions or conditions, typically following "at the"	-0.0448
808	Variations of the word "poor" used as expressions of sympathy in narrative contexts	-0.0444
615	Common English phrases containing terms like "sort", "same", "least" that express negation or comparison, often following "of the" or "all the"	-0.0440
2212	Book or publication titles marked with underscores in bibliographic or literary contexts	-0.0436
126	Common English conjunctions and auxiliary verbs ("but" and "was") used in narrative prose	-0.0435
1085	Instances of the word "terms" being used to describe relationships or standing between people or groups	-0.0431
358	The word "and" used in phrases describing ranges or intervals between two points	-0.0429
260	Words indicating comparison, alternation, or reference to different members of a group in descriptive texts	-0.0428

Feature	OLMo-7B-DPO	Mean ∆ Prob
275	Preposition indicating physical proximity or adjacency between people in narrative contexts	-0.0428
1809	Verbs expressing necessity, loss, or transition in narrative texts	-0.0420
1221	Personal pronouns and indefinite quantifiers used in narrative text	-0.0418
2660	Forms of the verb "depend" used to express reliance or conditional relationships	-0.0417
868	Comma punctuation marks followed by quotation marks in literary dialogue	-0.0414
78	Uses of "part" in phrases indicating actions, behaviors, or responsibilities of specific parties	-0.0412
2096	Uses of the word "slightest" as an adjective indicating minimal or negligible degree	-0.0411
383	Common phrases involving modal verbs and function words in narrative contexts	-0.0406
1605	Past tense verbs and temporal words expressing prior experiences or attempts	-0.0405
660	Common English function words (that, of, the, be, a) used in various grammatical contexts	-0.0399
1617	Variations of the phrase "one way or another/other" in different contexts	-0.0390
2368	Right curly braces appearing at the start of lines in dramatic or poetic text	-0.0373
2429	Uses of "kinds" in phrases describing various types or varieties of items in lists or inventories	-0.0370
52	Masculine singular pronoun used in various narrative contexts	-0.0370
2789	Possessive pronoun "his" used to indicate male ownership or association in narrative contexts	-0.0367
1556	Possessive pronouns in various narrative contexts	-0.0365
874	Instances of temporal or quantitative words ("When" and "One") at the start of clauses or recipe measurements	-0.0364
1558	Pronouns used in formal or literary contexts to refer to people	-0.0364
2941	Words commonly used in formal or literary English to express relationships, sequence, or qualification within sentences	-0.0360
631	Single quotes followed by semicolons in dialogue punctuation	-0.0358
1448	Uses of the definite article ’the’ in formal writing discussing governance, society, and ethics	-0.0353
1152	Repeated instances of the conjunction "and" used to create rhythm or emphasis in narrative text	-0.0353
222	Equals signs used as assignment operators in configuration or properties files	-0.0352
33	Words and numbers serving as section or chapter markers in document structure, along with transitional words in narrative text	-0.0352
1733	Coordinating conjunctions and prepositions used to connect related elements in various contexts	-0.0348
1038	Question-initiating words in interrogative sentences expressing doubt, contemplation, or inquiry	-0.0347
2723	References to page numbers, footnotes, or bibliographic citations in academic or historical texts	-0.0344
1890	Instances of the word "necessary" used to indicate requirement or essential need	-0.0336
569	Common transition phrases and expressions used in formal writing to connect or contrast ideas	-0.0335
1393	The definite article "the" used in formal or historical documents	-0.0327
2580	Punctuation marks and words appearing in dialogue or quoted speech	-0.0326
192	Uses of the word "manner" in formal or administrative contexts referring to methods, ways, or conduct	-0.0316
409	Archaic English auxiliary verbs (hath/had) used in formal or religious texts	-0.0316
879	Uses of "than" in comparative statements across various literary contexts	-0.0316
2034	Instances of "others" used in parallel constructions with "some" to indicate contrasting groups or alternatives	-0.0313
245	Prepositions and nouns related to temporal sequences or completion in narrative text	-0.0309
1179	Article "the" used in religious or spiritual texts discussing faith, morality, and divine guidance	-0.0302
85	Phrases using "or" and similar words to express indefinite alternatives or possibilities	-0.0297
2156	Phrases using "all parts" to describe geographic distribution or locations	-0.0291
637	The word "or" appearing in phrases expressing uncertainty or alternatives, often in constructions like "some way or other" and "one form or another"	-0.0291
392	Biblical or religious pronouns used in scripture quotes or religious text discussing divine figures	-0.0290
2471	Instances of the article "a" preceding the phrase "way" in contexts discussing paths, solutions, or escape routes	-0.0287
2065	Phrases using "in the way of" to indicate limited extent or capacity of something	-0.0285

Feature	OLMo-7B-DPO	Mean Δ Prob
721	Negative constructions using "not to" expressing prohibition, impossibility, or unsuitability	-0.0279
1444	Numbers and punctuation symbols appearing at the beginning of lines in literary references and annotations	-0.0277
597	Words used to reference previously mentioned items or sequences in formal text	-0.0275
235	Uses of the word "some" in parallel list structures or contrasting pairs	-0.0274
2624	Conjunctions connecting alternative or additional options in various contexts	-0.0272
2960	Phrases used as clarifying expressions or verbal fillers, typically following "that/nothing of the" and serving to elaborate or negate a point	-0.0268
391	Words indicating temporal or spatial position within a larger context	-0.0263
915	Conditional word "if" used to start hypothetical scenarios in various narrative contexts	-0.0262
1589	Variations of the words "need" and "love" used as emotional or necessity expressions in narrative text	-0.0262
1670	Expressions of approximate time durations using measurements like minutes, hours, years	-0.0252
57	Words used as qualifying or comparative terms in narrative prose	-0.0251
1297	Numbers or numerals used for sequential list or section numbering in texts	-0.0247
2505	Words used in phrases describing direct encounters or confrontations between people or animals	-0.0247
1019	Common nouns used to reference evidence, methods, or components in formal or academic writing	-0.0245
1092	Common English conjunctions and auxiliary verbs appearing at the start of clauses in literary texts	-0.0234
535	Prepositions and possessive markers in hypothetical or conditional statements	-0.0225
1849	Preposition 'to' used in various grammatical constructions linking actions, destinations, or relationships	-0.0225
719	Phrase "at the same time" used as a transitional expression in formal writing	-0.0221
1575	Interrogative phrases expressing concern using "what" and variations of "matter" in dialogue	-0.0217
825	Personal pronouns used in first-person narratives and formal documents	-0.0214
1471	Common conjunctions ('if' and 'or') used in literary prose to express uncertainty or alternatives	-0.0213
2919	Forward slashes appearing in file paths and API endpoint specifications	-0.0212
1505	Uses of "the other" in contexts describing the second item in a pair or comparison	-0.0211
1418	Phrase "not in the least" used as a negative response or denial in dialogue	-0.0210
1017	A word used as a reference to the second of two previously mentioned items or options	-0.0202
1526	Quotation marks followed by underscores used to mark titles or quoted phrases in text	-0.0201

Table 10: Features for OLMo-7B and Llama-7B comparison

Feature	OLMo-7B	Mean Δ Prob
1395	Double asterisks appearing between sentences in narrative text	0.4717
1627	Single quotes followed by periods appearing at the end of dialogue segments	0.3235
2337	Punctuation marks (commas and colons) following exclamatory phrases or statements	0.2974
1699	Personal pronouns appearing after exclamation marks or strong statements in dialogue	0.1458
635	Double asterisks marking dialogue breaks or speaker changes in literary text	0.1416
1831	Question marks appearing at the end of text segments expressing uncertainty or interrogation	0.1285
2033	Usage of the word "and" (or "And") as a conjunction in various contexts, predominantly in repetitive patterns or measurements	0.1252
2233	Question marks appearing at the end of questions in dialogue or narrative text	0.1157
1667	Variations of the phrase "one way or other" and similar expressions indicating alternation or choice	0.1116
2128	Special characters acting as delimiters or operators in structured data formats	0.1050
1378	Verbs related to mental processes, cognition, and decision-making	0.1025
1485	Equal signs used as assignment operators in code or configuration files	0.1013
2111	Religious or sacred pronouns and articles referring to God or divine entities in spiritual texts	0.0981
2290	Expressions using "thought of" to convey mental contemplation of an undesirable or concerning scenario	0.0976
903	Question-ending text snippets in dialogue or narrative contexts	0.0973
80	Double asterisks followed by text at sentence or paragraph boundaries	0.0973
1877	XML closing angle brackets followed by special characters in technical documentation	0.0921
251	The title "Mrs" used as a formal address for married women in narrative text	0.0915
1599	Words denoting groups, varieties, or collective human characteristics in literary texts	0.0904
1091	Words describing temporal sequence or dependency in narrative contexts	0.0897
648	Words describing circumstances, fate, or physical location in historical or narrative contexts	0.0887
634	Prepositions used in sentences about carrying or accompanying items or people	0.0885
1229	Third-person pronouns used in narrative prose	0.0881
71	Double asterisks appearing at line breaks in literary text passages	0.0870
910	Common English function words (articles, pronouns, and auxiliary verbs) used in narrative prose	0.0865
2309	Configuration and parameter placeholders in camera-related software settings	0.0855
2797	Common English pronouns and articles appearing at the start of sentences	0.0852
2826	Numbers appearing in document organizational elements like footnotes, lists, and chapter markers	0.0841
20	Personal pronouns at the start of sentences in narrative texts	0.0836
95	Common verbs of perception and interaction used in dialogue and narrative prose	0.0832
398	Commas used as separators in lists, addresses, and numerical values	0.0828
215	Instructions for cooking duration in recipe steps	0.0824
1790	Double hyphens separating sections in biographical text about musicians	0.0800
2677	Personal pronouns used as subject or possessive determiners in narrative contexts	0.0780
2339	Question-answer pairs in dialogue where the response begins with punctuation marks	0.0744
2796	Instances of "sooner" in the phrase "no sooner... than" in narrative contexts	0.0733
733	Words used as conjunctions or prepositions to express conditional or comparative relationships in narrative text	0.0731
1991	Words indicating multiple instances or occurrences in explanatory contexts	0.0726
1845	The word "or" used in phrases indicating an unspecified future time or manner, typically following "somehow," "sooner," or "day"	0.0722
2599	Uses of "kind" as a noun meaning type, sort, or category in various contexts	0.0698
2679	Uses of the article "a" in various narrative and descriptive contexts	0.0696
1534	Phrases used to express time, sequence, or transitions in narrative flow	0.0694
2879	Common words indicating reference or relation (century, kind, part, place) used in contexts describing historical events, participation, or order	0.0689
510	Articles and pronouns used as grammatical function words in various contexts	0.0684
2546	Punctuation marks and words appearing at sentence or phrase boundaries in various text fragments	0.0681
1911	Names of people, places, or numbers followed by commas in a list format	0.0672
2922	Question marks and punctuation appearing at the ends of questions or interrogative statements	0.0655
394	Instances of the word "share" referring to a portion, contribution, or fair allocation of resources or responsibilities	0.0649
1600	Common transition or qualification phrases in English writing	0.0647
897	Forms of the pronoun "it" used in various contexts as an object or subject	0.0639

Feature	OLMo-7B	Mean Δ Prob
1973	Uses of the word "Some" at the start of sentences or clauses to describe different groups of people and their actions	0.0631
827	Relative pronouns introducing dependent clauses in literary or formal text	0.0620
82	Personal pronouns used at the start of sentences in narrative text	0.0609
770	Common appearances of the indefinite article 'a' in various texts, often in phrases like "as a rule"	0.0604
971	The article "a" appearing in various narrative and descriptive contexts	0.0597
2143	Uses of "latter" referring to a previously mentioned second element, with one instance of "matter" appearing to be an OCR error	0.0589
483	Third-person and second-person pronouns in dialogue and narrative contexts	0.0585
2689	Phrases using prepositions (into, over, to) in contexts of consideration or inclusion	0.0583
1197	Uses of 'or' in phrases expressing uncertainty or alternatives, often paired with words like 'another', 'other', or 'another'	0.0582
1649	Phrases using "keep up" or similar variations to describe maintaining pace or following along with others	0.0581
2127	Words or symbols indicating sequence or transition in written text, including temporal terms and punctuation marks	0.0581
2377	Words expressing uncertainty or indefinite reference in narrative contexts	0.0580
2635	Uses of words related to participation, attribution, or involvement in formal documents	0.0573
28	Words referring to living entities or persons, particularly in discussions of human and divine existence	0.0564
1295	Examples of "which" used as a relative pronoun to connect dependent clauses to main clauses	0.0559
24	Determiners used as function words in various narrative and instructional contexts	0.0552
2597	Verbs indicating movement, continuation, or past presence	0.0550
2333	Variations of the verb "belongs" expressing ownership, classification, or attribution in academic or philosophical contexts	0.0549
2196	Double asterisks followed by text indicating dialogue or narrative transitions	0.0549
1083	Instances of words expressing relationships or comparisons between entities, primarily using "between" along with other relational terms	0.0547
568	Definite and possessive articles appearing in quoted or literary text passages	0.0542
42	Words functioning as indefinite pronouns referring to unspecified members of a group	0.0541
1500	Uses of "less" indicating reduced quantity, value, or degree in comparative contexts	0.0539
2238	Commas used in lists of character names within narrative text	0.0529
2669	Third-person masculine possessive pronouns referring to male characters or animals in narrative contexts	0.0527
220	Sequential question numbers appearing at the beginning or end of lines in educational or literary texts	0.0525
717	Commas appearing in various textual excerpts containing citations, names, and lists	0.0516
1431	Punctuation marks appearing in text followed by whitespace or line breaks	0.0511
2573	Words expressing personal relationships, preferences, or responses in social interactions	0.0509
725	Uses of 'or' in phrases expressing indefinite choice between alternatives, typically following 'one' or with 'other'	0.0507
2435	Instances of the word "means" used to indicate methods, tools, or ways of accomplishing something	0.0501
2363	References to variant readings and textual notes in a critical apparatus of classical or medieval texts	0.0499
1017	"Latter" used as a reference word to compare or contrast with a previously mentioned option	0.0496
977	Common English function words appearing in prose text surrounded by spaces	0.0495
1836	Programming-related instances of the word "instruction" describing control flow in code documentation	0.0495
1556	Possessive pronouns in various literary and narrative contexts	0.0494
34	Stage directions and dialogue markers in a theatrical script showing character names followed by periods	0.0486
1314	Usage of 'than' in comparative phrases indicating time duration or measurement	0.0484
604	Prepositions used in spatial or directional descriptions of physical proximity	0.0481
2837	Instances of 'took/take part in' describing participation in activities or events	0.0479
2782	Articles and conjunctions appearing after peculiar whitespace formatting in various texts	0.0478
1769	Usage of "latter" in phrases contrasting with "former" to reference the second of two previously mentioned items	0.0473
2056	Uses of "or" in phrases expressing uncertainty or vagueness, typically in constructions like "something or other" and "some way or other"	0.0460
1046	Reference numbers in footnotes, citations, and bibliographic entries in academic or historical texts	0.0458

Feature	OLMo-7B	Mean △ Prob
1236	Phrases using "other" and similar words to contrast or present alternative viewpoints in formal writing	0.0457
322	Conjunction words expressing alternatives or possibilities in narrative text	0.0457
319	Units of measurement and time durations in cooking and recipe instructions	0.0457
336	Words indicating temporal or logical sequence in argumentative or procedural text	0.0444
2480	Reference numbers appearing at the end of sentences or paragraphs, typically in brackets or parentheses	0.0443
1903	Phrasal uses of "into" in combination with forms of "take" and "account" to express consideration or inclusion	0.0439
488	Period-curly brace combinations marking speaker changes in a theatrical script or play dialogue	0.0430

Feature	Llama-7B	Mean △ Prob
132	Opening quotation marks following colons in dialogue passages	-0.6420
585	Punctuation marks followed by quotation marks in various literary contexts	-0.3941
1557	Reference markers or bracketed annotations in academic or scholarly texts	-0.2688
2672	Special characters used as delimiters in variable naming and data transformation contexts	-0.2259
2813	Equals signs used as delimiters in configuration or property files	-0.1847
310	Question marks immediately followed by quotation marks at the end of dialogue or questions in text	-0.1837
2008	Underscore characters used as word separators in database or code documentation	-0.1836
1466	Literary conjunctions 'and' with surrounding textual references and punctuation, plus a few outlier words	-0.1566
2142	Forms of the word "until" used in cooking instructions to indicate duration or completion of a step	-0.1414
2088	Question marks at the end of dialogue or interrogative statements	-0.1229
2756	The number "1" appearing in various numerical contexts including measurements, recipes, and specifications	-0.1227
1840	References to a character called "the Very Young Man" in a narrative text	-0.1224
202	Variations of the phrase "thing to do" used to describe actions or decisions in narrative contexts	-0.1206
537	Punctuation marks appearing at the end of bibliographic or reference entries in scholarly texts	-0.1015
1798	Words related to cognition, parenting, and behavioral patterns in documents discussing family and personal matters	-0.0997
845	Punctuation marks appearing at the end of parenthetical or list elements in bibliographic or reference entries	-0.0995
222	Equal signs used as assignment operators in configuration or properties files	-0.0936
792	Possessive pronouns used to indicate relationships between people in religious or narrative contexts	-0.0931
2328	Question marks at the end of sentences in literary or scholarly texts	-0.0896
2300	Assignment or equality operators in software configuration and metadata files	-0.0882
678	Numerical values (0 or 2) appearing in technical or statistical contexts	-0.0854
1463	Uses of the word "sorts" to indicate various types or varieties within different contexts	-0.0844
127	Question marks appearing at the end of dialogue or interrogative sentences	-0.0828
887	Words describing abstract personal qualities or attributes in various contexts	-0.0827
39	Common English articles and auxiliary verbs in various sentence contexts	-0.0821
716	Common function words (if, is, that, or) used in connecting clauses and expressing relationships between ideas in written text	-0.0813
1198	Past tense verbs in literary dialogues and narratives discussing past events or memories	-0.0797
1575	Interrogative phrases using "matter" to inquire about problems or concerns	-0.0785
2037	Words serving as connectors or transitions in narrative text, including punctuation and common linking words	-0.0781
2475	Common English articles and possessive pronouns used in narrative prose	-0.0775
2804	Words indicating contrast, limitation, or qualification in narrative text	-0.0768
2156	References to "all parts" indicating geographic or spatial distribution in various contexts	-0.0765
85	Variations of the phrase "or other/another" used to express indefiniteness or uncertainty	-0.0726
2208	Usage of "nothing of the sort" as a phrase of strong disagreement or denial in dialogue	-0.0719
1218	Verbs describing temporal existence, occurrence, or dependency	-0.0705
1858	Words introducing temporal transitions, specifically indicating the start of narrative events	-0.0704
1471	Conjunctions ('if' and 'or') used in expressions of uncertainty or hypothetical situations in narrative text	-0.0704
738	Repeated word "and" in phrases expressing increasing or decreasing intensity using "more," "further," or similar comparative terms	-0.0701
1372	Instances of the word "time" and other temporal connectors in narrative contexts	-0.0700

Table 11: Features for OLMo-13B and Llama-13B comparison

Feature	OLMo-13B	Mean Δ Prob
1395	Double asterisks followed by sentence fragments indicating subsequent narrative actions or observations in literary text	0.5308
1627	Quotation marks followed by a period in literary dialogue	0.2947
635	Double asterisks marking dialogue breaks or speaker changes in literary text	0.1636
1182	Instances of first-person pronoun "I" followed by "suppose" in dialogue within narrative text	0.1580
2035	Words expressing negation or contradiction, typically at the start of responses in dialogue	0.1333
1197	Variations of the phrase "one or other/another" used to express alternatives or uncertainty	0.1257
1831	Question marks appearing at the end of sentences expressing uncertainty or rhetorical questions	0.1251
1666	Punctuation marks in dates across various historical documents and letters	0.1196
1699	Words or phrases followed by two asterisks and a capitalized personal pronoun in English text	0.1188
129	Uses of the word "whole" referring to complete time periods or entirety of something	0.1173
1091	Temporal words used in narrative sequences indicating immediate succession or timing of events	0.1166
1440	Various text delimiters and punctuation marks used to separate or end sections in literary or transcribed text	0.1158
231	Personal pronouns used as subject or object in historical narrative text	0.1150
2033	Usage of words as repetitive connectors emphasizing distance, quantity, or continuation in narrative text	0.1086
693	Interjections expressing emotions or reactions in dialogue	0.1046
1566	Roman numeral XIII appearing in chapter or section numbering contexts	0.1044
80	Double asterisks followed by text at line beginnings in structured document contexts	0.1019
2797	Common English pronouns and articles appearing at the start of sentences	0.1004
2233	Question marks (with or without quotes) appearing at the end of questions in dialogue	0.0960
2622	Third-person pronouns used at the start of sentences in narrative text	0.0893
1911	Commas followed by whitespace in text listings of names, numbers, or locations	0.0885
756	Contractions used in question tags following statements	0.0860
174	Adverbs and prepositions used as modifiers in narrative descriptions indicating degree, extent, or minimal amount	0.0857
1518	The phrase "at any rate" used as a transitional expression in various contexts	0.0854
28	Forms of the word "being" (and one "fellow") referring to living entities or persons in literary texts	0.0850
2613	Personal pronoun "he" used as a subject referring to male individuals in various narrative contexts	0.0831
1022	Special characters (=, ?, 0) used as assignment or delimiter operators in configuration and code files	0.0818
2134	Uses of "for" to indicate time duration in recipe and food preparation instructions	0.0812
2111	Articles and possessive pronouns referring to God or the Lord in religious texts	0.0810
2826	Numbers and letters used as section or list item markers in document organization	0.0804
1553	Past tense verbs and personal pronouns used in narrative contexts	0.0801
1534	Words and phrases indicating temporal continuity or transitional expressions in narrative text	0.0792
1003	Words expressing degree, manner, or type in various contexts (kind, possible, between)	0.0787
271	Instances of the word "people" used to refer to the general public or citizenry in political and social contexts	0.0781
1277	Personal pronouns used in religious or spiritual texts referring to divine entities and followers	0.0767
2551	Instances of words referring to topics or matters under discussion in academic or intellectual contexts	0.0765
1325	Role or occupation designators following names and addresses in directory entries	0.0761
241	Preposition ’to’ used in various directional and spatial contexts	0.0740
1973	Instances of the word "some" at the start of clauses describing different groups or individuals within a larger population	0.0732
912	Words indicating degree or extent used in comparative or correlative phrases	0.0729
1972	Equals signs used in system configuration or property assignments within software code	0.0729
2121	Personal pronouns or possessive adjectives used in direct address or instruction contexts	0.0716
762	References to "the people" as a collective body in political and social contexts	0.0708
737	First-person singular pronoun "I" used in dialogue and personal statements	0.0706
1845	Variations of the phrase "sooner or later" or "somehow or other" used to express eventual or uncertain timing	0.0700
1835	Instances of words used in formal or archaic transitional phrases in historical or literary texts	0.0691

Feature	OLMo-13B	Mean ∆ Prob
391	Words indicating temporal or sequential position within text or historical contexts	0.0688
1190	Punctuation marks used for separating or ending textual elements in various document types	0.0684
2599	Uses of the word "kind" to express type, variety, or nature of something	0.0665
1254	Phrase "at last" used as a temporal marker indicating the end or culmination of waiting or events	0.0661
634	Prepositions appearing in narrative or instructional text, primarily "with" and "between"	0.0653
1009	References to geographic regions or locations within a larger area or territory	0.0648
2763	Instances of the word "there" used as an existential marker to indicate presence or existence	0.0642
2837	Active participation in events or activities indicated by 'take/took part in'	0.0639
1910	Contractions using apostrophes in informal or dialectal speech	0.0633
753	Instances of the article "the" in various literary and historical texts	0.0622
1284	Common variations of the phrase "What's the matter?" in dialogue	0.0618
2262	Italicized titles of published works or references in bibliographic and literary contexts	0.0618
374	Pronouns used in religious or spiritual texts referring to a divine entity or deity	0.0617
30	Common English function words appearing at the start of clauses in literary texts	0.0609
2501	Possessive pronouns used in emotional or dramatic narrative contexts	0.0599
2154	Conditional words and question marks in dialogue or interrogative contexts	0.0592
1759	The word "and" used in phrases describing time ranges or intervals between two points	0.0592
220	Numbers appearing as section or question markers in educational or literary texts	0.0584
55	Instances of "like" and related words expressing desire or preference in dialogue	0.0580
1943	Instances of the phrases "every now and then" or temporal adverbs signifying periodic occurrence in narrative texts	0.0576
1348	Words used as generic qualifiers or references to types/categories in various contexts	0.0572
1123	Literary conjunctions or connective words used in narrative prose	0.0567
2988	Uses of the word "aspects" and "occupations" in contexts discussing different parts or categories of systems, life, or activities	0.0564
673	Functional words appearing in contexts expressing uncertainty, likelihood, or supposition	0.0564
1903	Phrase pattern "take/taking into account" used to describe consideration of factors or conditions	0.0563
2679	The word "a" used as an indefinite article before nouns describing measurements, movements, or complete entities	0.0557
1360	Personal pronouns and prepositions used in narrative text describing people and relationships	0.0542
717	Commas used as separators in various types of document excerpts	0.0537
1649	Instances of "up" used in phrases about maintaining pace or speed with others while moving	0.0535
2541	Third-person masculine pronouns (he/him) used as subject or object in narrative contexts	0.0525
236	Uses of the phrase "in order" followed by "to" or "that" to express purpose or intention in formal writing	0.0506
1032	Instances of "one" and "the" in religious or philosophical texts emphasizing absolute durations or divine authority	0.0504
1931	Honorific title "Mr." followed by periods in various written contexts	0.0503
2196	Double asterisks appearing after complete phrases or at sentence boundaries	0.0502
1674	Prepositions used in temporal or spatial sequences within narrative texts	0.0499
675	Uses of 'from' to indicate the starting point of a geographical or physical span or distance	0.0498
2765	Usage of impersonal pronouns 'it' and 'there' at the start of sentences in narrative texts	0.0498
1369	Exclamation marks followed by dialogue or quoted text expressing strong emotions	0.0492
1287	Personal and impersonal pronouns used at the start of sentences or clauses	0.0491
2936	Words indicating portions, varieties, or possibilities within descriptive contexts	0.0479
380	Instances of common English words used in narrative prose from a similar time period	0.0476
1080	Instances of "matter" in dialogue asking about someone's wellbeing or condition	0.0475
1822	Instances of the words "or" and related helping verbs in casual or uncertain expressions	0.0471
2577	Words describing categories or types followed by "of" in academic or analytical contexts	0.0470
2012	Words commonly used as honorifics or qualifiers in formal dialogue and text	0.0469
1856	Various forms of the word "other" used as an indefinite pronoun or adjective to refer to additional or alternative items in a list or group	0.0469
416	Uses of "a" and related articles in phrases containing "as a whole" or "as a rule"	0.0452
2117	Common English articles and punctuation marks used in various written contexts	0.0446
1399	Possessive apostrophe-s appearing in various literary contexts	0.0446
802	Preposition 'to' used in various contexts to indicate movement, direction, or proportion	0.0440
2998	Punctuation and formatting elements used as section or chapter separators in various texts	0.0435
192	Words used as function words to describe how actions or processes are carried out	0.0430

Feature	OLMo-13B	Mean ∆ Prob
868	Commas followed by quotation marks in dialogue ending with exclamations or strong statements	0.0425
2049	Uses of the word "of" in contexts expressing ranking or ordering (e.g., "most of", "last of", "best of")	0.0420
2708	Temporal words ("One") appearing at the start of narrative sequences describing past events	0.0419
2333	Forms of the verb "belongs" used in discussions of ownership, categorization, or attribution	0.0412
1237	The word "and" used as a conjunction to connect contrasting or complementary concepts in philosophical and abstract discussions	0.0407
2145	Forms of words expressing causation or outcome in academic or formal writing	0.0401
1460	Instances of "for" followed by cooking time durations in recipe instructions	0.0399
2480	Reference numbers in brackets or parentheses used as citations or footnotes in academic or formal texts	0.0393
1250	Commas used as separators in narrative prose passages	0.0383
1380	Possessive pronouns used to indicate personal ownership or relationship in narrative contexts	0.0380
2339	Double asterisks followed by dialogue or text showing responses to questions or statements	0.0374
70	Commas used after dates, numbers, or location names in various texts	0.0365
36	Phrases containing "one" followed by common function words or generic nouns in various contexts	0.0361
135	Usage of "other" in contexts describing competition, conflict, or comparison between two parties	0.0360
2771	Words appearing in contexts asking about degree, amount, or duration	0.0359
2661	Preposition 'upon' used in formal or structured contexts, often following 'based' to indicate classification or dependency	0.0356
1769	Instances of 'latter' used in former/latter comparisons in text	0.0354
1999	Polite expressions of gratitude or emphasis using "much" and "things" in dialogue or narrative contexts	0.0351
969	Words expressing varying levels of certainty or possibility in formal writing	0.0349
691	Forms of "to be" and other auxiliary verbs in conditional or speculative statements	0.0346
1046	Reference numbers in footnotes, section markers, or bibliographic citations in academic texts	0.0346
2621	Words describing methods or approaches used to accomplish actions	0.0346
2956	Commas used as separators in various textual contexts including poetry, lists, and bibliographic entries	0.0340
250	Reciprocal pronoun "each other" used in contexts describing mutual relationships or interactions	0.0337
2225	Words expressing temporal sequence or immediacy in narrative contexts	0.0336
533	Instances of "all of a sudden" used to indicate an abrupt or unexpected occurrence in narrative text	0.0336
848	Words appearing in sentences discussing names, titles, or designations of people, places, or things	0.0335
1620	Instances of the article "a" preceding descriptions of people or objects in narrative prose	0.0334
483	Personal pronouns used in dialogue or narrative text showing character interactions	0.0325
2028	Instances of the pronoun "it" used at the beginning of clauses or after prepositions	0.0323
1371	Words functioning as subordinating conjunctions or pronouns at the start of dependent clauses in narrative text	0.0323
531	Instances of the word "same" used for expressing equality or identity between things	0.0322
1107	Connector words used for transition and joining in text, including possessives and prepositions	0.0322
1468	Conditional or hypothetical expressions using common function words in narrative text	0.0320
2069	Common English articles and prepositions used in connecting phrases across various academic texts	0.0319
60	Words expressing requirements, objectives, or desires in various contexts	0.0318
2311	Instances of "depends/depended" used to express dependency or reliance relationships in various contexts	0.0311
1176	Words indicating proximity, sequence, or punctuation in narrative texts	0.0304
1376	Pronouns used in dialogue or narrative text, often appearing after verbs like "had" or before verbs like "occurred"	0.0303
1244	Uses of "view" (and one "viewer") in contexts indicating purpose or intention, typically following "with a view"	0.0302
563	Uses of "as if" in comparative expressions showing hypothetical similarity	0.0300
2911	Words functioning as referential terms in narrative contexts indicating relationships between entities or events	0.0293
725	Phrase "one or" followed by words indicating non-specific selection from alternatives	0.0291

Feature	OLMo-13B	Mean Δ Prob
1224	Phrases starting with "For some" or "After a" followed by "time" in narrative contexts describing duration	0.0282
2220	Verbs describing sequences of movement or consequences, primarily used to show one person or event following or resulting from another	0.0281
420	Uses of "from the very" in phrases indicating time or beginning points	0.0270
155	Book format specification "8vo" (octavo) in bibliographic entries	0.0270
2066	Forms of the word "mental" used in context of metaphysical planes or bodies in spiritual/esoteric texts	0.0270
1708	Second-person pronoun 'you' used in direct address or dialogue across various texts	0.0267
1755	Repetitive uses of "closer" to indicate increasing proximity or decreasing distance over time	0.0266
407	Commas appearing between names and title/suffix in a catalog or directory listing	0.0263
2658	Interrogative phrases using "matter" to express concern or inquire about a problem or situation	0.0260
2646	Instances of the indefinite article 'a/an' in literary or philosophical texts	0.0260
836	Modal verbs (should/ought) expressing obligation or recommendation in various contexts	0.0255
661	Instances of "then" appearing in the phrase "every now and then" in narrative contexts	0.0255
215	Instructions specifying cooking duration in recipe steps	0.0254
604	Prepositions used in spatial or distance descriptions	0.0254
723	Words expressing quantity or completion used in evaluative contexts	0.0250
2446	The word "morning" used in temporal sequences describing events occurring on subsequent days	0.0245
2353	Double asterisks used as dialogue separators in literary text	0.0244
827	Relative pronouns used to connect clauses in literary or formal text	0.0241
1072	Words used in negative or restrictive contexts indicating finality or limitation	0.0239
1877	XML closing tags and markers in technical configuration files	0.0239
2585	Third-person pronouns and articles used in narrative prose passages	0.0238
1403	Uses of "and" in numeric expressions between one hundred and three hundred	0.0233
607	Punctuation and words appearing before descriptive phrases or clauses in narrative text	0.0231
2153	Common function words (articles, pronouns, prepositions, conjunctions) used in narrative text	0.0226
1748	Instances of "other" and similar words used to refer to one of two alternatives or members of a pair	0.0223
2650	Reference numbers appearing in footnotes and citations in academic or literary works	0.0217
2884	Present participle "going" (often in future tense constructions) and "dealing" used in narrative contexts	0.0214
1164	Words appearing in "if [word]" conditional phrases expressing wishes or regrets	0.0212
370	Uses of the word "or" in uncertain or alternative situations, along with similar connecting words	0.0210
2906	Uses of the pronoun "it" at the start or middle of narrative sentences	0.0208
491	The word "or" used in contexts of alternatives, choices, or numerical ranges	0.0201

Feature	Llama-13B	Mean Δ Prob
132	Opening quotation marks at the beginning of dialogue in literary text	-0.6293
1444	Special characters appearing after closing brackets in document markup or metadata notation	-0.2930
2672	Special characters used as delimiters in database or configuration file field mappings	-0.2605
1557	Reference numbers or markers in brackets appearing at the beginning of lines in scholarly or annotated texts	-0.2271
2813	Equal signs used as assignment or comparison operators in configuration or log files	-0.1865
310	Question marks immediately followed by quotation marks at the ends of dialogue or questions	-0.1700
2581	Question marks followed by quotation marks at the end of dialogue or questions in literary text	-0.1386
1997	Personal pronoun "he/He" used as a subject in narrative contexts	-0.1240
2088	Question marks at the end of dialogue or interrogative statements	-0.1234
1932	Common function words and operators appearing in various textual contexts	-0.1164
1790	Double hyphens used as section breaks in biographical text about musicians	-0.1063
2843	First-person singular pronouns and references to human experience in narrative contexts	-0.1047
2793	Words functioning as temporal or logical connectors in narrative text	-0.1034
1223	Nouns referring to existence or ways of being, including physical life, spiritual essence, and forms of existence	-0.1017
537	Punctuation marks at the end of bibliographic or reference entries	-0.0941
2808	Past tense auxiliary verb 'have' used in speculative or conditional statements	-0.0935

Feature	Llama-13B	Mean △ Prob
2576	Uses of "less" in contexts specifying minimum measurements, quantities, or thresholds	-0.0931
1273	Opening parenthesis followed by italicized character names or text references in dialogue or theatrical scripts	-0.0927
2690	Verbs describing actions or states in narrative contexts	-0.0870
1575	Forms of the phrase "What's the matter?" in dialogue expressing concern	-0.0866
2379	References to personal data or user information in privacy policy documents	-0.0865
2328	Question marks at the end of interrogative sentences	-0.0852
896	Punctuation marks appearing between chapter titles or section headings in a table of contents	-0.0838
901	Personal pronouns and possessive markers in narrative fiction texts	-0.0815
2868	Verbs related to knowledge, learning, or responding used in expressions of uncertainty or questioning	-0.0808
2096	Uses of "slightest" to emphasize complete absence or minimal degree of something	-0.0764
633	Phrases using "sort" or "course" as part of expressions indicating type, manner, or category	-0.0754
1964	Double hyphens used as section breaks or dramatic pauses in narrative text	-0.0747
85	Usage of "or other" as a phrase indicating an unspecified alternative in narrative text	-0.0745
738	Repetitive phrases using "and" to indicate increasing or decreasing progression over time	-0.0741
1540	Words commonly used in narrative prose discussing personal conditions or states	-0.0739
2493	Uses of the definite article 'the' in formal or literary contexts	-0.0721
222	Equal signs followed by text labels or configuration settings in software interface files	-0.0711
658	Punctuation marks used for standard English syntax in formal writing	-0.0687
1324	Commas separating numbers in sequences of reference citations or numerical lists	-0.0669
343	The phrase "no sooner" used to indicate immediate sequential actions, typically followed by "than"	-0.0665
2811	Expressions using "kind" or "sort" to indicate denial or classification in dialogue and narrative	-0.0652
678	Instances of "0" appearing in technical or encoded data strings	-0.0650
2987	Words expressing types, categories, or variations in different contexts	-0.0649
104	Numbers or words appearing in square brackets as reference markers in text	-0.0648
256	Common expressions of emphasis or intensification in dialogue and narrative text	-0.0643
812	Period-underscore punctuation pairs appearing in dialogue formatting in dramatic texts	-0.0636
1654	Past tense verbs and common expressions involving movement or desire in narrative contexts	-0.0636
320	Contractions or variations of "would not" used in rhetorical questions	-0.0618
1471	Conjunctions used to express alternatives or hypotheticals in narrative text	-0.0616
1192	Configuration comments indicated by '#' symbol in server configuration files	-0.0608
1858	Temporal phrases beginning with "One" that introduce a new narrative event or scene	-0.0608
2833	Forms of the verbs "to be" and occurrences of "same" in various literary and technical contexts	-0.0598
660	Common English function words appearing in various literary and administrative texts	-0.0595
2463	Past tense form of "be" used in first-person narratives describing past experiences	-0.0576
201	Uses of "same" in phrases indicating simultaneous actions or conditions, typically following "at the"	-0.0572
127	Question marks at the end of dialogue or interrogative sentences	-0.0559
2295	Possessive pronouns referring to authority figures in formal or historical texts	-0.0552
1757	Instances of "afraid" expressing personal fears and anxieties in first-person narratives	-0.0548
1584	Phrases expressing uncertainty or alternatives, often using variations of "or" and "other"	-0.0538
1010	Question marks and the word "Or" appearing as punctuation or conjunctions at the start of sentences in literary texts	-0.0532
689	Words related to individual identity and self-interest in discussions of personal and social dynamics	-0.0524
1170	Forms of the verb "have" used in various tenses and contexts	-0.0502
908	Modal verb 'have' used in conditional or hypothetical expressions indicating preference, likelihood, or obligation	-0.0490
1784	Verbs indicating acts of determining, understanding, or convincing in formal prose	-0.0486
2267	Stage directions in theatrical scripts indicating character actions or emotions	-0.0479
379	Phrases expressing variability or alternatives, often using formulaic expressions like "one reason or another"	-0.0476
2778	Indefinite article 'a' followed by time-related phrases or quantities	-0.0474
1767	Personal pronouns in exclamatory or emotional literary passages	-0.0472
409	Archaic forms of the auxiliary verb "have" used in formal or religious texts	-0.0471
547	Instances of the verb "find" and pronouns "They/Those" used in analytical or discovery contexts	-0.0460

Feature	Llama-13B	Mean Δ Prob
69	Commonly used English words appearing at the start of dependent clauses or phrases in historical texts	-0.0456
646	Instances of "less" used in phrases indicating quantity or measurement, often in the form "no less than" or similar constructions	-0.0452
1592	Negative constructions using "in the least" or "the worst" to express minimal or negative degree	-0.0444
1913	Forms of the auxiliary verbs "has" and "is" appearing in statements expressing temporal or comparative conditions	-0.0438
1749	Words and punctuation used as connective or comparative elements in complex sentence structures	-0.0431
302	Property or attribute labels in software configuration files	-0.0428
2411	The word "a" used as an indefinite article preceding time-related phrases in narrative text	-0.0425
2766	Variations of phrases expressing certainty or verification, primarily using "make sure" and "in the least"	-0.0419
2011	Words expressing mental states or perceptions used in dialogue or internal monologue	-0.0413
2919	Forward slashes appearing in configuration and system file paths	-0.0413
1152	Words indicating repetition or continuation in narrative text, primarily using variations of "and"	-0.0412
1349	Possessive constructions using "of" following demonstrative pronouns	-0.0411
905	Uses of the word "parts" or "area" referring to geographic or spatial divisions	-0.0405
1732	Commas used as list separators in various types of enumerations	-0.0400
1532	The word "ever" used in expressions of eternal or infinite time, often in religious or emotional contexts	-0.0385
1019	Words indicating factual or methodological concepts used to support arguments or observations	-0.0383
1085	Uses of the word "terms" to describe relationships or social connections between people	-0.0375
2776	Articles and intensifiers used in narrative descriptions of historical or dramatic events	-0.0369
956	Uses of "one another" expressing mutual or reciprocal actions between people or groups	-0.0362
2207	Word pairs where 'same' or 'place' indicates replacement, similarity, or position in various contexts	-0.0360
1048	Instances of the phrase "make/making the most of" used to express maximizing opportunities or benefits	-0.0360
2368	Closing curly braces appearing at the start of lines in what appears to be a dramatic or poetic text	-0.0356
1501	Different usages of the word "same" referring to identical or equivalent things, with one outlier each for "product", "sensor", and "employee"	-0.0355
2533	Uses of "in" and "less" as part of phrasal verbs or expressions indicating inclusion or incorporation	-0.0353
2634	First-person singular pronoun 'I' used as the subject of sentences in dialogue	-0.0351
2376	Words appearing in comparative constructions using variations of "the more/less... the more/less" pattern	-0.0350
2505	Phrases indicating direct physical confrontation or close proximity between individuals	-0.0344
1206	Common phrases "not a bit of it" and "as a matter of fact" used as discourse markers	-0.0341
332	First-person and negated "do" contractions in informal dialogue exchanges	-0.0338
275	Preposition used to indicate physical proximity or adjacency between people or objects	-0.0327
78	Uses of "part" in phrases expressing actions, behaviors, or responsibilities of individuals or groups	-0.0323
1733	The word "and" used as a coordinating conjunction to connect related elements in various contexts	-0.0321
1716	Commas and the word "other" used as list separators or connectors in various texts	-0.0318
1950	GUI labels and error messages for file operations and system management in a software application	-0.0313
1408	Personal and possessive pronouns used in narrative contexts	-0.0311
535	Common prepositions or contractions used in conditional or qualifying statements	-0.0308
1928	Words expressing concepts related to judgment, reasoning, and decision-making	-0.0302
2099	Common English auxiliary and function words used in narrative contexts	-0.0299
719	Uses of "at the same time" as a transitional phrase indicating simultaneity or contrast	-0.0298
597	Words referring to sequential or temporal ordering in various contexts	-0.0285
2652	Time measurements and references in cooking instructions and recipes	-0.0282
2056	Instances of "or" used in vague expressions following "some" to indicate uncertainty or indefiniteness	-0.0281
1133	Forms of the verbs "occupy" and "is" used to describe physical or conceptual space	-0.0269
2201	References to accompaniment or presence of others in narrative contexts	-0.0265
1591	Personal pronouns used in dialogue and narrative text	-0.0264

Feature	Llama-13B	Mean △ Prob
2487	Instances of the word "of" in various texts showing its usage as a preposition	-0.0262
2003	Common English pronouns and determiners used in formal or archaic writing	-0.0256
847	Past conditional uses of the word "have" in narrative contexts	-0.0256
2442	Words indicating hypothetical or alternate scenarios in historical or narrative texts	-0.0255
1391	Phrases containing modifiers like "all the" or "its" followed by words indicating sameness or possession	-0.0255
1281	Articles preceding descriptions of sounds or events in narrative prose	-0.0250
2466	Demonstrative and possessive pronouns appearing at the start of clauses in formal or archaic texts	-0.0250
61	Metadata property assignments and configurations in a software system, particularly related to virtual appliance operations	-0.0249
2475	Articles and possessive pronouns used as grammatical determiners in narrative prose	-0.0244
2456	Forms of the verb "be" used as auxiliary or linking verbs in complex sentences	-0.0238
1857	Possessive pronouns used to indicate ownership or belonging in various contexts	-0.0233
2305	Function words used to indicate singularity or possibility in formal written text	-0.0230
86	Prepositions and adverbs used in phrases describing movement, accompaniment, or lack thereof	-0.0226
2946	Verbs expressing desire, possession, or self-reference in narrative dialogue	-0.0221
546	Articles ("The" and "A") and symbols used at the beginning of sentences or titles in various texts	-0.0217
2817	Words expressing prior knowledge, comparison, or reflection in conversational dialogue	-0.0216
100	Punctuation marks used to end dialogue or statements in dramatic or theatrical text	-0.0214
445	Question marks at the end of sentences in literary dialogue or narrative text	-0.0204
1451	Nouns referring to human roles or social positions in narrative contexts	-0.0204
636	Closing parenthesis followed by comma in various academic and technical texts	-0.0204