DSAI: Unbiased and Interpretable Latent Feature Extraction for Data-Centric AI

Hyowon Cho
KAIST AI
hyyoka@kaist.ac.kr

Soonwon Ka
NAVER AI Platform
soonwon.ka@navercorp.com

Daechul Park
NAVER AI Platform
daechul.park@navercorp.com

Jaewook Kang
NAVER AI Platform
jaewook.kang@navercorp.com

Bokyung Son
NAVER AI Platform
bo.son@navercorp.com

Abstract

Large language models (LLMs) often struggle to objectively identify latent characteristics in large datasets due to their reliance on pre-trained knowledge rather than actual data patterns. To address this data grounding issue, we propose Data Scientist AI (DSAI), a framework that enables unbiased and interpretable feature extraction through a multi-stage pipeline with quantifiable prominence metrics for evaluating extracted features. On synthetic datasets with known ground-truth features, DSAI demonstrates high recall in identifying expert-defined features while faithfully reflecting the underlying data. Applications on real-world datasets illustrate the framework's practical utility in uncovering meaningful patterns with minimal expert oversight, supporting use cases such as interpretable classification [1].

Keywords

Latent Feature Extraction, Open Information Extraction, Bias, Interpretability

ACM Reference Format:
Hyowon Cho, Soonwon Ka, Daechul Park, Jaewook Kang, and Bokyung Son. 2025. DSAI: Unbiased and Interpretable Latent Feature Extraction for Data-Centric AI. In . ACM, New York, NY, USA, 31 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The ability to analyze large-scale datasets is a cornerstone of deriving actionable business insights. Traditionally, this task has been managed by human *data scientists*, but it faces several key challenges: (1) the large volume of data makes it difficult to review all information comprehensively, (2) human analysis can often include subjective bias, and (3) collaboration with domain experts is often required, leading to high operational costs.

[1]The title of our paper is chosen from multiple candidates based on DSAI-generated criteria.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

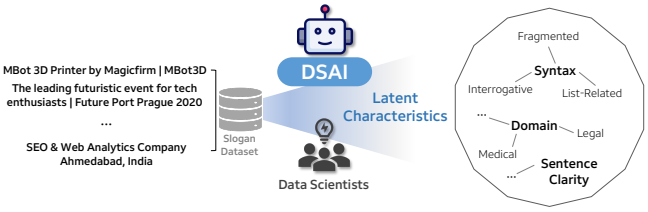


Figure 1: DSAI automates the extraction of latent features like syntax, domain, and clarity

Large language models (LLMs) have emerged as powerful tools for identifying patterns within massive datasets, leveraging their ability to process and generate language in context [1, 2, 9, 15, 21, 24]. However, their application to data analysis is limited by critical shortcomings. First, LLMs often struggle to identify latent characteristic patterns in large datasets due to inherent *data grounding issues*, where outputs rely on pre-trained knowledge rather than the specific nuances of the input data [4, 7, 28]. Second, the difficulty in verifying LLM-generated responses and the lack of quantitative evaluation methods require expert oversight, which can be prohibitively expensive at scale.

To address these limitations, we propose **Data Scientist AI (DSAI)**, a framework that systematically applies LLMs to extract and refine latent features from data. Unlike direct feature extraction approaches, DSAI adopts a bottom-up approach, starting with detailed analysis of individual data points, aggregating their characteristics, and deriving actionable features. The process is guided by defined *perspectives* which provide LLMs with a consistent framework for interpreting data points while minimizing subjective bias.

The DSAI pipeline operates in five stages: **#1 Perspective Generation** identifies data-driven perspectives from a small subset of data. **#2 Perspective-Value Matching** assigns values to individual data points by evaluating them against these perspectives. **#3 Clustering** groups values with shared characteristics to reduce redundancy. **#4 Verbalization** converts extracted features into a compact criterion form. **#5 Prominence-based Selection** determines which features to use based on a prominence intensity metric that quantifies the discriminative power of each extracted feature.

Throughout these stages, the LLM remains task-agnostic – we do not reveal the specific domain or the "correct answer" context during feature generation. This design minimizes bias and ensures that the identified features are grounded in the data rather than the model's background knowledge.

We validate our framework on datasets curated for our experiments, including a research titles dataset [25] and an advertising slogans dataset [3], each with expert-defined criteria that serve as ground truth for evaluation. We then demonstrate DSAI's practical value on three real-world datasets: news headlines with click-through rate (CTR) labels [27], a spam detection dataset [5], and Reddit comments with community engagement metrics [13].

Our main contributions are as follows:

- **Minimizing Bias**: We ensure that LLMs focus on latent characteristics present in the data, thus mitigating the tendency to rely on domain-specific prior knowledge (addressing the data grounding issue).
- **Prominence Metric**: We introduce a quantitative metric for feature prominence, which serves as a proxy for the discriminative power of each extracted feature.
- **Interpretability**: We improve interpretability through feature-to-source traceability, allowing users to trace each extracted feature back to the data points that support it.
- **Efficiency**: We enable thorough examination of large datasets with minimal human labor by systematically guiding LLMs through the analysis process.

2 Related Works

2.1 Latent Feature Extraction with LLM

Recent advancements in LLMs have demonstrated their effectiveness in extracting latent features, particularly in identifying perspectives and matching values in data [17]. Studies using LLM-based clustering techniques have shown promising results in extracting high-level concepts[9, 18, 23, 24], demonstrating their utility for analyzing large datasets [8, 26]. Research has also shown that decomposing complex tasks into multiple stages or aspects enhances performance [11, 19]. Our framework combines perspective generation, multi-stage feature construction, and clustering using LLMs to conduct comprehensive latent feature extraction.

While prior work largely focuses on grouping inputs by topic or inducing abstract semantic concepts, they typically do not aim to extract explicit linguistic features that explain fine-grained distinctions such as text quality. The outputs of these methods—such as clusters or latent topic labels—are often evaluated using coherence or grouping accuracy, rather than how well they reflect interpretable, discriminative patterns grounded in the data. As such, these approaches differ in both objective and output format from methods that seek to generate human-readable criteria tied to specific outcome variables.

2.2 Bias of LLMs

LLMs face challenges in adapting to new patterns due to pre-existing knowledge biases. [7] show that in-context learning (ICL) struggles to overcome these biases even with explicit prompts or many-shot examples. Using external knowledge bases also has limited effectiveness in reducing hallucinations and reliance on internal biases [4, 10]. Advanced LLMs fail to align with provided context in about 40% of predictions when the given context conflicts with their prior knowledge [28]. These findings emphasize the need for robust techniques to mitigate bias and enhance grounding, especially for applications where data-driven conclusions are crucial.

Reference Criteria	Comparison Criteria	Recall (%)	
		Slogan	Title
FlippedPosData	PosData	89.5%	94.1%
FlippedMixedData	MixedData	81.8%	89.5%
NoData	PosData	100.0%	95.6%
	MixedData	100.0%	92.3%
	FlippedPosData	90.0%	94.1%
	FlippedMixedData	96.4%	85.2%
Expert	NoData	88.9%	83.3%
	PosData	66.7%	54.2%
	MixedData	50.0%	75.0%
	FlippedPosData	94.4%	45.8%
	FlippedMixedData	77.8%	83.3%
	DSAI (Thres: [0])	100.0%	83.3%
	DSAI (Thres: [0.348])	88.9%	83.3%
	DSAI (Thres: [0.692])	77.8%	75.0%

Table 1: Recall of features derived from different approaches across slogans and titles.

3 Challenges in LLM-Driven Data Analysis

In this section, we examine the behavior of a state-of-the-art LLM when tasked with directly generating features from data. These exploratory experiments reveal the data-grounding challenges that motivate our DSAI approach.

3.1 Setting

Dataset Annotation and Sampling. We use two expert-annotated text datasets where the "ground truth" latent features are known (defined by domain experts). The first dataset consists of research paper titles and the second contains advertising slogans. In each dataset, experts [6, 14, 16, 22] have outlined a set of specific criteria or features that good examples should exhibit. We annotated each sample for the presence or absence of each criterion. This yields a detailed profile of which features are present in each data point. We then designated *positive* examples as those that satisfy many of the expert criteria (top-scoring "high-quality" samples) and *negative* examples as those that violate or lack many of the criteria (bottom-scoring "low-quality" samples). This dataset construction emphasizes concrete feature differences rather than a subjective quality judgment.

Model. All experiments in this section use GPT-4o [15] as the LLM, which is shown to have strong capabilities in understanding nuanced text and performing annotation-like tasks [20]. The model was prompted in a zero-shot manner to generate dataset features under various conditions.

3.2 Data Grounding Issues

Using the above datasets, we explored straightforward ways of prompting the LLM to extract latent features, and evaluated whether the LLM's outputs were truly grounded in the input data. We tried two input configurations for the prompt:

- **PosData**: Provide the LLM with only positive examples and ask it to identify characteristics common to these examples.

- **MixedData**: Provide the LLM with both positive and negative examples, and ask it to find distinguishing features.

In both cases, we formatted the prompt to list a set of representative samples and requested the model to output a list of key features or criteria describing the positive group. We found that both PosData and MixedData prompts yielded feature lists that substantially overlapped with the expert-defined criteria (Table 1). On the surface, this suggests the model can produce reasonable-sounding features. However, because our datasets come from well-known domains, we must question whether the LLM actually derived these features from the given data, or if it simply regurgitated its own prior knowledge about the domain. To investigate this, we designed four experiments around the following questions:

(a) Does the LLM Adapt to Input Data? We tested whether the LLM truly uses the input examples to adapt its feature generation. If the model is grounding its output in the provided data, then changing the data labels should lead to corresponding changes in the generated features. To check this, we flipped the class labels in the input and observed the effect on the output features:

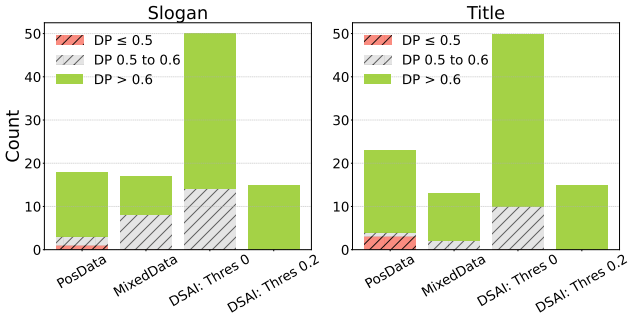
- **FlippedPosData**: We took the negative examples but misled the LLM by labeling them as if they were positive ("high-quality") examples.
- **FlippedMixedData**: We presented the LLM with the same set of positive and negative examples as MixedData, but swapped their labels (positives labeled as "low-quality" and negatives labeled as "high-quality").

If the model adapted to the data, FlippedPosData should produce features for low-quality texts, and FlippedMixedData should invert the original MixedData features. Instead, we observed the opposite. As shown in Table 1, FlippedPosData produced features nearly identical to PosData, and FlippedMixedData closely matched MixedData. This suggests that flipping labels had minimal impact, indicating the model relied on prior notions of "high-quality" text rather than adapting to the input data.

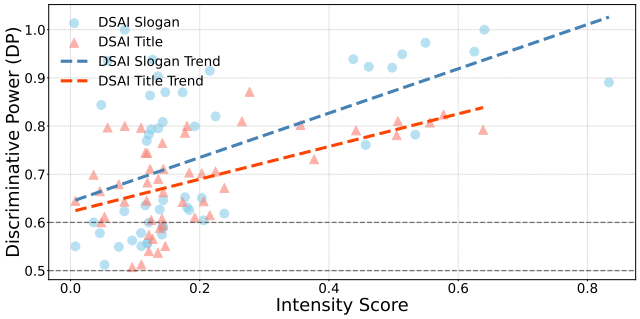
(b) Is Pre-existing Knowledge the Primary Source of Generated Features? The above result raises the question: would the LLM have generated a similar list of features even if we gave it no data at all? To test this, we prompted the LLM to list features of high-quality text without providing any example data. Here, the model must rely solely on its internal knowledge.

- **NoData**: The LLM is asked to imagine or recall what characteristics define good content, without seeing any specific dataset samples.

Remarkably, NoData features showed high recall of expert-defined criteria, strongly overlapping with PosData and MixedData (Table 1). Even without input data, the model reproduced nearly all expert criteria—fully for slogans and almost entirely for titles. This suggests the LLM relied on prior knowledge rather than capturing from the provided data, highlighting both its strong knowledge base and limited data-driven adaptation. While this demonstrates the model's impressive knowledge base, it also underscores the lack of true data grounding in these direct generation approaches.



(a) DSAI shows superior grounding over direct feature generation methods, with no DP score below 0.5. When prominence exceeds 0.2, all scores remain above 0.6.



(b) The graph shows that higher prominence leads to higher DP scores for DSAI features.

Figure 2: DP scores for direct feature generation and DSAI methods.

(c) Are the Generated Features Truly Reflective of Positive Data's Latent Characteristics? We next examined the quality of the features generated by the LLM in terms of how well they actually characterize the positive class in the data. Just because a feature sounds like a good guideline doesn't guarantee that it differentiates our positive and negative examples. For instance, the model might output "Use simple language" as a feature of good text, but if both our positive and negative examples equally exhibit (or fail to exhibit) this trait, then that feature isn't really capturing what makes the positive group unique in our dataset.

To assess this, we quantify the discriminative power (DP) of each feature. We define a feature's *DP score* as a measure of how well that feature separates positive examples from negative ones in the dataset. In practice, we calculate DP score as the fraction of examples that exhibit the feature which belong to the positive class: $P(\text{positive}|\text{feature-present})$. One can think of this like a precision of the feature for identifying positive samples. A DP of 0.5 means the feature appears just as often in negatives as in positives – effectively no discriminative value. A DP closer to 1 means the feature is mostly present in positives (strong positive indicator), whereas a DP below 0.5 means the feature is actually more common in negatives, which would indicate a misidentified or inversely correlated feature.

Using this metric, we evaluated the features generated under the PosData and MixedData prompts. We found that several of

those features had low DP scores, some even below 0.5 (Figure 2a). This means the model sometimes proposed features that were more prevalent in the "bad" examples than the "good" ones. These results highlight a limitation of directly using an LLM for feature generation: many of the features it outputs, even if they sound plausible, do not truly reflect the distinguishing characteristics of the positive data.

(d) Does Removing Task Context Improve Data Grounding? One hypothesis to improve grounding was that the LLM's knowledge might be overly triggered by the context of the task in the prompt. To test if removing task-specific context reduces bias and improves data grounding, we used:

- **NoContext**: The LLM is given two unlabeled sets of texts but is not told which set is "high-quality" or "low-quality", nor even that the goal is to identify high-quality text features. Essentially, we ask the model to compare two groups of texts without naming the task.

In this setting, the LLM defaulted to generating vague and generic descriptors. This suggests that task context removal alone does not improve data grounding and highlights the need for structured guidance during feature extraction.

Summary of Findings. The experiments above reveal that a naive application of LLMs – feeding them data and asking for patterns – is largely dominated by the models' prior knowledge and can fail to capture dataset-specific characteristics. The LLM-generated features overlapped with known human criteria but were not genuinely learned from the input samples (as evidenced by label-flip and no-data conditions), and many lacked true discriminative value for the task at hand. These challenges motivate our proposed solution. In the next section, we introduce the DSAI framework, which is specifically designed to provide the LLM with the necessary structure and constraints to ensure that the extracted features are grounded in the data and relevant to the task.

4 Data Scientist AI

DSAI is a five-stage framework for automated latent feature generation that leverages LLM capabilities within a structured process. The pipeline, illustrated in Figure 3, is designed to overcome the shortcomings identified in §3.2 by guiding the LLM through controlled steps.

#1 Perspective Generation. In this first stage, we prompt the LLM to generate a diverse set of perspectives – these are different angles or aspects under which the data can be described. The key idea is to break down the analysis into multiple facets. To achieve this, we feed the LLM a small subset of the dataset, including a few positive and negative examples. By keeping the task context hidden, we minimize domain bias while still providing concrete data for the model to analyze. We then ask the LLM to propose distinct perspectives that might explain differences in the data.

Each perspective generated by the LLM comes with a structured description: we instruct the model to give each perspective a name, a brief evaluation criterion, a suggested process for analyzing a text from that perspective, and an example from the provided data illustrating the perspective in action (Figure 3 (1)). By having the LLM produce this structured output, we ensure consistency in how it conceptualizes each aspect. After generation, we also apply a deduplication step to remove redundant or overlapping perspectives. The result of Stage 1 is a list of candidate perspectives, which serve as the conceptual foundation for the subsequent analysis.

#2 Perspective-Value Matching. In Stage 2, we systematically evaluate each data point against the perspectives identified in Stage 1. For every *(perspective, data point)* pair, the LLM is prompted to assign a value that describes the data point with respect to that perspective. After this stage, each data point is associated with a set of *(perspective, value)* pairs – one for each perspective – capturing how the LLM perceives that point on each aspect.

#3 Value Clustering. To make the features more interpretable and reduce redundancy, we cluster similar values within each perspective. We employ the LLM for this clustering task, drawing on its semantic understanding. The process works in two sub-steps for each perspective: (1) Cluster Label Generation: The LLM examines all the values it assigned under a given perspective and proposes a smaller set of representative labels (cluster names). (2) Value Assignment: Next, the LLM assigns each raw value to one of the generated cluster labels.

The outcome is that for each perspective, we now have a handful of feature categories (the cluster labels), each representing a group of similar values. Each data point can thus be described in terms of these cluster labels.

#4 Verbalization. This stage transforms *(perspective, label)* pairs into verbal *criteria*. For each pair, we compute $P(\text{positive}|D_{p,l})$, the proportion of positive examples in dataset $D_{p,l}$ corresponding to the *(perspective, label)* pair (p, l). Using this, we calculate a *directional score* $2 \times P(\text{positive}|D_{p,l}) - 1$, which determines how the pair should be verbalized. Pairs with positive directional scores (>0) are directly verbalized as features describing positive data. Pairs with negative directional scores (<0) are transformed into "avoid" statements, which indirectly characterize positive data by specifying features to avoid.[2] This dual transformation approach ensures coverage of both desired and undesired traits.

#5 Prominence-based Selection. Finally, DSAI employs *prominence intensity* as the feature selection metric, defined as the absolute value of directional score $\|2 * P(\text{positive}|D_{p,l}) - 1\|$. Using prominence as a metric, we can now select the most impactful features by setting a prominence threshold (which can be adjusted by the user) and retain only features above that threshold. The key is that DSAI is not just outputting an unstructured list of features – it provides a way to prioritize them. This addresses the issue with the baseline LLM approach (§3.2), which gave no indication of which features were more important or reliable. By looking at the prominence scores, users can decide how many features to consider or where to draw the line between major and minor features.

After Stage 5, the final output of DSAI is a curated set of features, each in a clear natural-language form, typically accompanied by their prominence scores. These features are intended to be data-grounded and interpretable, providing insight into the data. In the

[2]For instance, if *(clarity, low)* receives a negative directional score, it is verbalized as "Avoid sentences with low clarity."

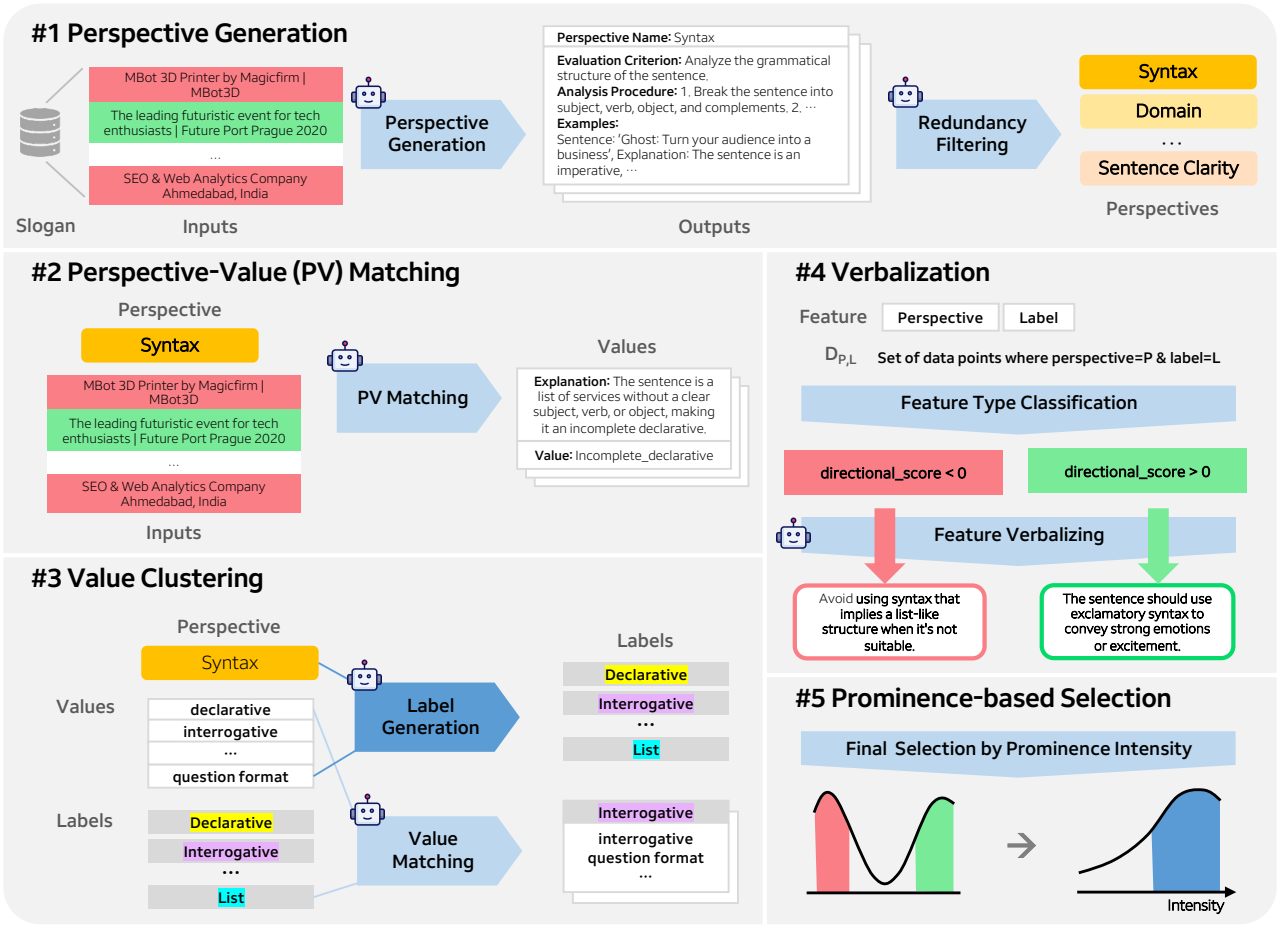


Figure 3: Overview of the DSAI pipeline: Perspectives are first generated to guide analysis (#1), then used to match values to data points (#2). These values are clustered to reduce redundancy (#3), verbalized into concise criteria (#4), and prioritized based on their prominence (#5).

next section, we evaluate how well this pipeline works in practice, especially in comparison to the direct LLM approach.

Traceability and Transparency. The feature generation process maintains the link of each feature to its perspective-label combination, the original data samples that contributed to its creation, and the quantitative prominence score that justified its selection. Such traceability enables users to verify the legitimacy of extracted features by examining their source data and the complete derivation process.

5 Validation of Methodology Using Expert-Driven Annotation Dataset

This section validates our methodology through experiments on various datasets, focusing on three key aspects: recall of expert-defined criteria (§5.1), discriminative power of generated criteria (§5.2), and reliability of pipeline modules (§5.3).

5.1 Recall of Expert-Defined Criteria

One way to gauge DSAI's effectiveness is to see if it can rediscover the ground-truth criteria that domain experts have defined for these datasets. We applied the full DSAI pipeline to the slogans and research titles data annotated as described in §3.1.

Criteria Generation. We generated criteria through our pipeline and retained those with $|D_{p,l}| > 6$ and positive prominence intensity scores. This yielded 235 criteria for slogans and 198 criteria for research titles. Examples of the generated criteria are shown in Table 2.

Human Feature Matching. For recall evaluation, one annotator initially performed loose matching of generated criteria, which was then validated through majority voting among three annotators.

Recall. Our methodology showed strong performance in reproducing expert-defined criteria, even at high prominence intensity thresholds. All 9 expert criteria for slogans were captured at a threshold of 0.348, and 83% recall (10/12) was achieved for research titles at a threshold as high as 0.692 (Table 1). While PosData and

Requirement	Prominence	Frequency
Top Requirements		
The advertising tone should convey a focus on quality, using optimistic and aspirational language.	0.8571	14
The sentence should convey an optimistic advertising tone that encourages engagement.	0.8333	108
The sentence should incorporate cultural references that align with consumptive themes.	0.8333	12
The sentence should employ indirect methods to engage the audience effectively.	0.7857	28
Ensure the sentence contains a component that emotionally appeals to the reader.	0.7831	83
Bottom Requirements		
The sentence should fully and effectively communicate its intended message.	0.0267	1,159
The sentence should avoid merely providing information without an intended action or emotion.	0.0250	1,122
Ensure the sentence includes references to cultural significance.	0.0248	1,009
Avoid using imperative or overly complex grammatical structures in titles.	0.0244	41
Ensure the use of inclusive language in the sentence.	0.0225	1,109
Avoid sentences that lack necessary cultural references.	0.0224	1,115

Table 2: Top and Bottom Requirements of Slogan Dataset based on Prominence.

MixedData also showed decent coverage, their results relied on LLM's pre-existing knowledge as discussed in §3.2. In contrast, our approach, by design, minimizes such potential bias by withholding task-specific context, while still achieving comparable or better recall rates.

Recall Dynamics across Various Thresholds. The adjustable prominence intensity thresholds allow users to tailor their analyses by balancing discriminative power and coverage. We provide a detailed analysis of recall dynamics across different threshold values in Appendix D. In summary, we categorized expert-defined criteria, ranked their importance, and examined at which thresholds they were filtered out. More important criteria persisted at higher thresholds, while less critical ones were eliminated at lower thresholds.

5.2 Discriminative Power (DP)

Having shown that DSAI can reproduce expert-defined criteria, we next evaluate whether the features it identifies truly differentiate high-quality examples from low-quality ones. To do so, we assess each feature's *discriminative power* (DP)—a measure of how well a feature separates positive samples (e.g., high-quality data) from negative ones. Formally, we define the DP score as follows:

$$DP\ Score = \begin{cases} P(\text{positive}|\text{feature-present}) & \text{if} \\ \qquad\qquad \text{directional\_score} > 0, & \\ P(\text{negative}|\text{feature-absent}) & \text{elif} \\ \qquad\qquad \text{directional\_score} < 0. & \end{cases}$$

A DP score close to 1.0 implies that the feature is a strong indicator for the intended class, while a score around 0.5 implies the feature is no more informative than random chance. For example, in an advertising slogan dataset, consider the feature "uses emotional, uplifting language." If this feature is present in 50 slogans and 40 of them are labeled positive, then $DP = 0.8$, indicating that the feature is fairly predictive. Similarly, in a spam classification setting, a feature like "contains phrases such as 'Act Now' or 'Limited Time Offer'" might appear in 100 messages, of which 95 are spam—yielding $DP = 0.95$, a highly discriminative indicator. In contrast, features with $DP \approx 0.5$—such as "contains a phone number" in the same context—offer little predictive value.

To evaluate how DSAI's internal scoring aligns with actual discriminative utility, we took all the DSAI-generated criteria and binned them into five buckets based on their prominence intensity scores. From each bucket, we sampled 10 criteria for manual examination. For each sampled feature, we annotated a set of examples to determine whether the feature was present or absent, excluding non-applicable[3] cases. We then computed the DP score of each feature as described above (§3.2(c)), and compared the resulting scores across the different prominence buckets.

The results confirmed our expectations: criteria with higher prominence scores generally exhibited higher discriminative power. As illustrated in Figure 2b, all of the DSAI-generated features achieved $DP > 0.5$, indicating at least some positive correlation with class membership. Notably, features in the top prominence buckets frequently reached $DP > 0.8$, while lower-prominence features hovered closer to the decision boundary of 0.5. These findings reinforce that the prominence intensity metric is not only internally consistent, but also externally predictive.

5.3 Reliability of Pipeline Operations

DSAI's multi-stage pipeline relies on the LLM's output at several steps. While the final outcomes are validated to be effective, we also place strong emphasis on ensuring the reliability of each intermediate stage. To this end, we designed a dedicated LLM evaluator to verify the outputs of key stages. However, given the risk of self-confirmation bias in LLM-based verification, we adopted a rigorous process to ensure that the evaluator itself is reliable and aligned with human judgment.

Human-Guided Evaluator Optimization. For each evaluation stage, we curated a representative subset of samples and annotated gold-standard labels via expert review. We then iteratively refined the LLM evaluator prompts and instructions until they reproduced these labels with 100% agreement.

[3] *e.g.*, feature about "use of sales terms" in examples unrelated to business models.

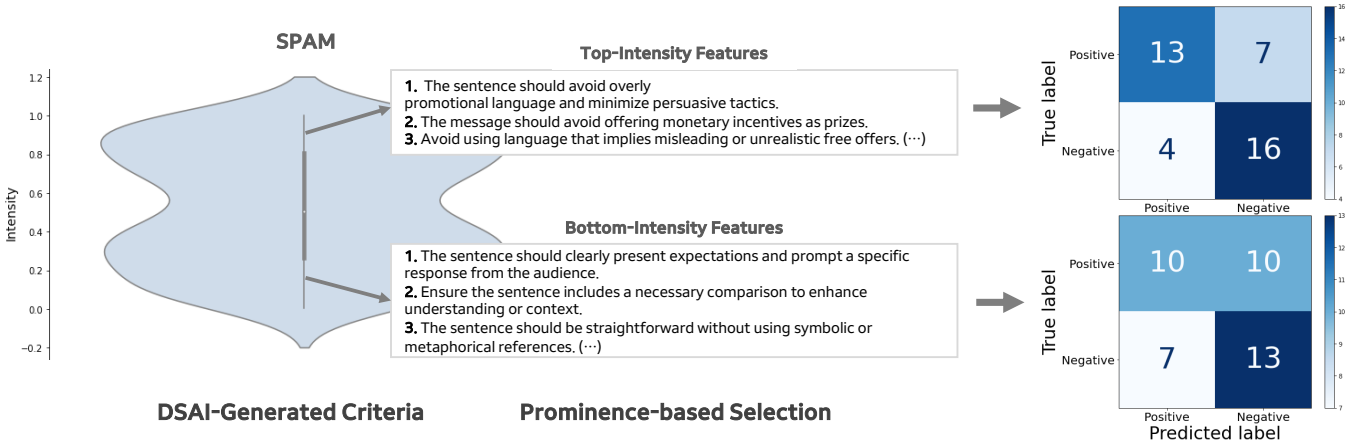


Figure 4: Example of interpretable spam classification: The figure shows how feature prominence guides criteria selection, with high-prominence criteria improving spam classification performance.

Triple Cross-Check Validation. After the LLM evaluator produced verification results on the broader dataset, we conducted a triple cross-check procedure. Three independent human reviewers each examined the original input and the LLM's assigned labels. The evaluator's decisions showed high agreement with human reviewers across all stages, further confirming that the model was not merely reinforcing its own bias.

High Evaluator Reliability. Following this optimization and validation process, we applied the LLM evaluator to verify key pipeline stages:

- Stage 2 (Value Matching): For each perspective and data point, after the LLM assigned a value, we asked the LLM to confirm whether that assignment was correct given the data point's content.
- Stage 3 (Clustering): After the LLM clustered values, we gave it each value along with the cluster label and asked if that assignment was appropriate.
- Stage 4 (Verbalization): We asked the LLM to verify that each verbalized feature correctly described the intended cluster and perspective, and that the phrasing (direct or "avoid") corresponded to the sign of the directional score.

The LLM's self-audit verification process showed high consistency rates of >98%, 94%, and 98% for stages #2, #3, and #4 respectively. These results indicate that the pipeline's internal operations are reliable; the LLM is largely consistent and does not contradict itself when asked to re-check its work.

6 Real-World Application with Quantitative Datasets

Having validated DSAI on datasets with known ground truth criteria, we now apply it to several real-world datasets to demonstrate its practicality and versatility. We selected three real-world user feedback datasets critical for business insights [12]: (1) MIND [27], analyzing engagement features in news headlines with high CTR as the positive group; (2) spam detection [5], identifying patterns in spam messages as the positive group; and (3) Reddit [13], exploring interaction-promoting linguistic features in highly upvoted comments as the positive group. These datasets differ significantly in content and domain, which allows us to see how DSAI adapts to different domains without any domain-specific tuning.

Prominence Distribution and Sample Insights. Running DSAI on each dataset, we observed that the distribution of feature prominence scores differs by domain. This is expected: each domain has its own characteristics and noise levels, so the threshold for what constitutes a strongly discriminative feature will vary. For instance, in the spam dataset, it might be crucial to set a higher prominence threshold to avoid any features that could lead to false positives. In the news headline dataset, one might choose a threshold that balances identifying strong engagement drivers while not missing out on subtler but interesting patterns. The Reddit data might show a different spread, capturing nuances of informal language or humor that drive upvotes.

Importantly, DSAI captured not only general traits (like "uses urgent language" might be a spam trait common across many messages, or "mentions specific names/events" for news headlines) but also fine-grained nuances specific to subsets of each dataset (like "sarcastic undertone" or "emotionally intense"). These are the kinds of details often overlooked by simpler LLM analyses or manual inspection. For example, DSAI might find that in the news dataset, a certain style of phrasing has a subtle impact on CTR, which wouldn't be obvious without this kind of analysis. These observations highlight the benefit of a data-centric approach: DSAI can adapt to the particular domain and context of each dataset, rather than relying on one-size-fits-all features. We provide detailed breakdowns and examples from each dataset in Appendix G.

Potential for Downstream Tasks. Because DSAI produces human-readable criteria, the extracted features have the potential to readily support various downstream applications such as style transfer (rewriting content to meet certain criteria), generating annotation guidelines (for human labelers to follow), or directly for classification tasks.

To illustrate this, we conducted a toy spam classification experiment using 20 spam and 20 ham samples. Using the five criteria with the highest prominence intensity scores led to effective classification performance, while using the five criteria with the lowest prominence resulted in poor performance (See Figure 4). This small experiment demonstrates that DSAI's prominence scoring correlates with real utility: the features deemed important by DSAI indeed helped in a classification task, while those deemed unimportant were not useful.

7 Conclusion

In this paper, we proposed DSAI, a faithful data-driven feature extraction framework that ensures LLMs identify latent characteristics from data without relying on their domain-related biases. DSAI automates thorough examination of large datasets while minimizing human labor and enhancing interpretability through source-to-feature traceability. Through empirical validation, we confirmed its capability to extract meaningful features, suggesting its potential for applications requiring interpretable and efficient feature extraction.

8 Limitation

While our results confirm that DSAI can effectively guide LLMs to produce data-grounded features, several limitations remain. First, the framework's performance depends heavily on the quality of the underlying LLM. If the model struggles to assign values, cluster them appropriately, or generate coherent perspectives, the outputs can be error-prone or require significant manual intervention. In brief tests with smaller open-source models such as Mistral-7B, we observed notably lower performance—particularly in the perspective generation step—suggesting that these models lacked the capacity to reliably execute the full pipeline. For this reason, we focused our experiments on a strong model like GPT-4o, which is better equipped to handle the multi-stage reasoning and generation required by DSAI. Our aim is to demonstrate that even such high-performing models can fall short in data-grounded interpretation—and that a structured framework like DSAI is necessary to mitigate those limitations.

Second, although GPT-4o was used for annotation and evaluation in most of our experiments due to its strong generation performance, relying on model-based annotations introduces potential risks of bias or label noise. We mitigated this by referencing expert-defined criteria and conducting manual reviews, but model-generated supervision remains an inherent source of variability.

Third, DSAI has thus far been demonstrated primarily on textual data. Extending it to other modalities such as images, audio, or structured logs may require architecture-specific adaptations or integration with multimodal LLMs. Additionally, the multi-stage structure of the pipeline—especially when evaluating numerous perspectives per data point—can become computationally expensive on large-scale datasets. Optimization or sampling strategies may be necessary to improve scalability in such cases.

Finally, although DSAI reduces reliance on domain experts by automatically discovering interpretable features, some degree of human oversight remains important, particularly in high-stakes domains like law or medicine where correctness and accountability are critical.

References

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

[2] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[3] Yiping Jin, Akshay Bhatia, Dittaya Wanvarie, and Phu TV Le. 2023. Towards improving coherence and diversity of slogan generation. *Natural Language Engineering* 29, 2 (2023), 254–286.

[4] Krishnaram Kenthapadi, Mehrnoosh Sameki, and Ankur Taly. 2024. Grounding and evaluation for large language models: Practical challenges and lessons learned (survey). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 6523–6533.

[5] Esther Kim. 2016. SMS Spam Collection Dataset. *https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset* (2016).

[6] Chiranjeev Kohli, Lance Leuthesser, and Rajneesh Suri. 2007. Got slogan? Guidelines for creating effective slogans. *Business horizons* 50, 5 (2007), 415–422.

[7] Jannik Kossen, Yarin Gal, and Tom Rainforth. 2024. In-context learning learns label relationships but is not conventional learning. In *The Twelfth International Conference on Learning Representations*.

[8] Sehyun Kwon, Jaeseung Park, Minkyu Kim, Jaewoong Cho, Ernest K Ryu, and Kangwook Lee. 2023. Image clustering conditioned on text criteria. *arXiv preprint arXiv:2310.18297* (2023).

[9] Michelle S Lam, Janice Teoh, James A Landay, Jeffrey Heer, and Michael S Bernstein. 2024. Concept Induction: Analyzing Unstructured Text with High-Level Concepts Using LLooM. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–28.

[10] Hyunji Lee, Sejune Joo, Chaeeun Kim, Joel Jang, Doyoung Kim, Kyoung-Woon On, and Minjoon Seo. 2023. How Well Do Large Language Models Truly Ground? *arXiv preprint arXiv:2311.09069* (2023).

[11] Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024. HD-Eval: Aligning Large Language Model Evaluators Through Hierarchical Criteria Decomposition. *arXiv preprint arXiv:2402.15754* (2024).

[12] Hanyang Luo, Wanhua Zhou, Wugang Song, and Xiaofu He. 2022. An Empirical Study on the Differences between Online Picture Reviews and Text Reviews. *Information* 13, 7 (2022), 344.

[13] Samuel Magnan. 2019. 1 million Reddit comments from 40 subreddits. *https://www.kaggle.com/datasets/smagnan/1-million-reddit-comments-from-40-subreddits* (2019).

[14] Lakshmi Balachandran Nair and Michael Gibbert. 2016. What makes a 'good'title and (how) does it matter for citations? A review and general model of article title attributes in management science. *Scientometrics* 107 (2016), 1331–1359.

[15] OpenAI. 2024. Hello gpt-4o. *https://openai.com/index/hello-gpt-4o/* (2024).

[16] A Padrakali and K Chitra Chellam. 2017. Advertising Slogan-It's Emphasis and Significance in Marketing. *International Research Journal of Management and Commerce* 4, 11 (2017), 37–46.

[17] Letian Peng, Yuwei Zhang, and Jingbo Shang. 2023. Controllable Data Augmentation for Few-Shot Text Mining with Chain-of-Thought Attribute Manipulation. *arXiv preprint arXiv:2307.07099* (2023).

[18] Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A Prompt-based Topic Modeling Framework. arXiv:2311.01449 [cs.CL] https://arxiv.org/abs/2311.01449

[19] Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2023. Branch-solve-merge improves large language model evaluation and generation. *arXiv preprint arXiv:2310.15123* (2023).

[20] Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large Language Models for Data Annotation: A Survey. arXiv:2402.13446 [cs.CL] https://arxiv.org/abs/2402.13446

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

[22] Milind S Tullu. 2019. Writing the title and abstract for a research paper: Being concise, precise, and meticulous is the key. *Saudi journal of anaesthesia* 13, Suppl 1 (2019), S12–S17.

[23] Vijay Viswanathan, Kiril Gashteovski, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2024. Large Language Models Enable Few-Shot Clustering. *Transactions of the Association for Computational Linguistics* 12 (2024), 321–333. doi:10.1162/tacl_a_00648

[24] Mengting Wan, Tara Safavi, Sujay Kumar Jauhar, Yujin Kim, Scott Counts, Jennifer Neville, Siddharth Suri, Chirag Shah, Ryen W White, Longqi Yang, et al. 2024. Tnt-llm: Text mining at scale with large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5836–5847.

[25] Qingyun Wang, Zhihao Zhou, Lifu Huang, Spencer Whitehead, Boliang Zhang, Heng Ji, and Kevin Knight. 2018. Paper Abstract Writing through Editing Mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 260–265. doi:10.18653/v1/P18-2042

[26] Zihan Wang, Jingbo Shang, and Ruiqi Zhong. 2023. Goal-driven explainable clustering via language descriptions. *arXiv preprint arXiv:2305.13749* (2023).

[27] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. 2020. Mind: A large-scale dataset for news recommendation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 3597–3606.

[28] Kevin Wu, Eric Wu, and James Zou. 2024. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A LLM Annotation with Expert-Defined Criteria

This section describes our methodology employed for data annotation using expert-defined criteria and evaluates the effectiveness of each criterion through distribution analysis.

A.1 Method

We established a gold standard by manually annotating 10–20 samples per dataset and using expert-defined criteria. After optimizing prompts to maximize alignment with this manual annotation, we used the most aligned prompt to annotate 3,000 data points with GPT-4o, classifying each as "Feature-Present" or "Feature-Absent". This score for each data point was calculated as the total number of conforming criteria. Based on these scores, we selected the top and bottom 600 samples as *positive* (high-quality text) and *negative* (low-quality text) samples respectively. This annotation process proved cost-effective compared to human annotation, with an average cost of $3.5 USD per 3,000 annotations.

A.2 Ensuring Distinction Between Positive and Negative Groups

Given that some expert criteria may lack sufficient discriminative power despite their theoretical importance, we conducted an analysis comparing the ratios of "Feature-Present" for each criterion between top- and bottom-ranked samples. Note that a truly discriminative criterion should show high Feature-Present ratio in the top group and low Feature-Present ratio in the bottom group.

Slogan (Top vs. Bottom) Table 3 demonstrate significant Feature-Present ratio disparities between the top and bottom groups for most criteria. Several criteria demonstrate maximal distinction (*e.g.*, "Concise but not too simple" (100% vs. 0%)), and some show moderate but still meaningful differentiation (*e.g.*, "Include the brand name" (99% vs. 30%)). "No exaggeration" (99% vs. 91%) and "Direct/Straight-forward" (99% vs. 61%) have smaller gaps.

Title (Top vs. Bottom) Table 4 shows even stronger distinction between the top and bottom groups. Multiple criteria (*e.g.*, "Simple format" "Direct", "Concise and precise") achieve 100% Feature-Present ratio in the top group compared to as low as 0–3% Feature-Present ratio in the bottom group, indicating their high predictive value for title quality. The large differences across all criteria confirm their strong discriminative power.

B Implementation Details and Cost Analysis

Details of Each Inference. In Stage #1 Perspective Generation, we generated 50 perspectives at per forwarding step across three steps, iteratively concatenating each step's output with the few-shot example in the prompt for the next step. For Stage #2 Perspective-Value Matching, each forwarding step processed one sentence and three perspectives as input. Stage #3 Perspective-Oriented Value Clustering used all generated values per perspective in order to effectively cluster shared characteristics. In Stage #4 Verbalization, we transformed each perspective-label pair into a compact sentence (*criterion*). The prompts used in this process will be released through github repository.

Criterion	Top Feature-Present (%)	Bottom Feature-Present (%)
Direct/Straight-forward	99	61
Consise but not too simple	100	0
Pleasant to hear	84	1
Includes a sales idea	98	4
No exaggeration	99	91
Future-oriented	81	1
Clear positioning	93	2
Highlight the brand's unique traits	99	0
Include the brand name	99	30

Table 3: Ratio of Feature-Present in Top and Bottom for Slogan expert-defined criteria.

Checklist Item	Top Feature-Present (%)	Bottom Feature-Present (%)
Simple format	100	3
Direct	100	1
Informative and specific	100	0
Functional (with scientific keywords)	100	3
Concise and precise	100	0
Include the main theme	100	0
Not too long or too short	100	3
Avoid whimsical words	100	60
Avoid jargon	100	44
Mention place/sample size if valuable	100	72
Important terms at the beginning	100	11
Descriptive titles preferred	100	9

Table 4: Ratio of Feature-Present in Top and Bottom for Title expert-defined criteria.

Cost Analysis. The total cost of processing 10 perspectives and 100 sentences is $2.43746, broken down as follows:

In **Step 1: Perspective Generation**, generating 10 perspectives costs $0.0304. Each perspective costs $0.00304. With a total of 2,100 input tokens and an average of 82.90 output tokens per perspective, the cost is calculated as:

$$10 \times 0.00304 = 0.0304$$

In **Step 2: Perspective-Value Matching**, each sentence requires $\lceil 10/3 \rceil = 4$ inferences, as each inference processes 3 perspectives. For 100 sentences, this results in:

$$100 \times 4 = 400 \text{ inferences.}$$

With a cost per inference of $0.00496, the total cost for this step is:

$$400 \times 0.00496 = 1.984$$

On average, each inference involves 2,206 input tokens and 440 output tokens.

In **Step 3: Perspective Value Clustering**, each perspective incurs a clustering cost of $0.0087125. This includes two components:

- **Label Generation**, costing $0.00579625 for 597 input tokens and approximately 1,010 output tokens.
- **Value Matching**, costing $0.00291625 for 269 input tokens and 516 output tokens.

The total clustering cost per perspective is:

$$0.00579625 + 0.00291625 = 0.0087125$$

For 10 perspectives, the total clustering cost is:

$$10 \times 0.0087125 = 0.087125$$

In **Step 4: Verbalization**, it is assumed that the 10 perspectives cluster into 5 groups each, resulting in 50 perspective-label pairs. Verbalizing each pair costs $0.0018, so the total cost for this step is:

$$50 \times 0.0018 = 0.09$$

Summing all the costs, the total processing cost is:

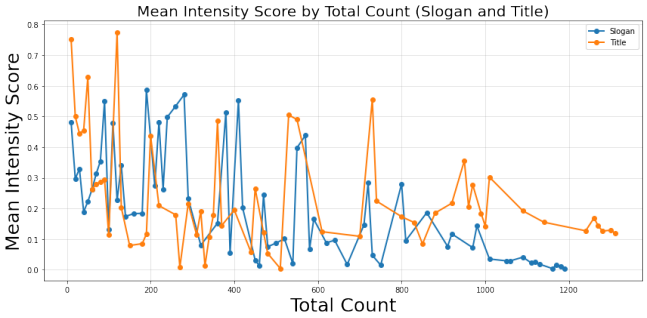
$$0.0304 + 1.984 + 0.087125 + 0.09 = 2.43746$$

C Prominence Intensity and Occurrence Analysis

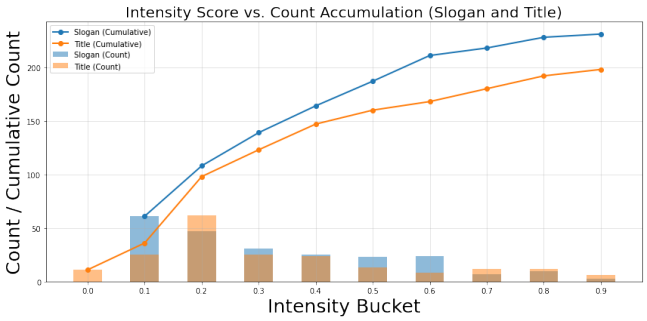
This section explores the relationship between feature frequency, prominence, and their impact on discriminative power. By analyzing feature occurrence and prominence distributions, we provide insights to help readers determine appropriate thresholds for these factors, aiding in the selection of features that optimize performance in distinguishing positive and negative traits within datasets.

C.1 Influence of Feature Occurrence on Discriminative Power

We analyzed the influence of feature occurrence frequency on discriminative power by grouping features into frequency buckets (e.g., <=10, 20-100, and >100) and evaluating mean/maximum prominence metrics (Figure 5a)



(a) Average of Prominence Score per Frequency Bucket



(b) Data Coverage per Prominence Bucket

Figure 5: Comparison of Prominence scores and data coverage across frequency and prominence buckets.

Low-Frequency Features. Features with low frequencies (<= 10) demonstrated higher discriminative power, with mean prominence as high as 0.482 (some reaching even 1.0) in the slogan dataset. This suggests that low-frequency features effectively represent specific traits of positive data, though their rarity may limit generalization.

Medium-Frequency Features. Features with medium frequencies 20–100) displayed a balance between specificity and generality. They demonstrated consistent performance in discriminative tasks through moderate mean prominence values.

High-Frequency Features. Features with high frequency (> 100) were found to be more generic, showing low mean prominence values (0.087 in the slogan dataset for frequencies > 500). Though less discriminative, they provide valuable insights into the dataset's baseline characteristics. Their high frequency makes them suitable for applications where specificity is less critical.

C.2 Influence of Feature Prominence on Discriminative Power

We also analyzed the distribution of features across different prominence score levels of by examining frequency accumulation within prominence buckets (Figure 5b). This analysis revealed how features were concentrated across low, moderate, and high prominence scores.

Slogan. The features were concentrated in the lower prominence range, with 61 features below 0.1 and 108 features below 0.2. This distribution suggests that while the majority of features exhibit limited discriminative power, a select subset of features with higher intensities plays a disproportionate role in capturing latent positive traits.

Title. An even more pronounced skew toward lower intensities was observed, with 11 features showing zero prominence and 62 features below 0.2. This distribution highlights that while low-prominence features comprise the majority of the dataset, they have limited effectiveness in distinguishing positive traits.

D Threshold Analysis on Slogans

In this section, we track how different prominence and frequency thresholds affect the recall of expert criteria, with particular attention to which criteria types are most resistant to threshold increases. In general, we observe progressive filtering of features as thresholds increase, which aligns with our criteria importance categorization. However, we also observed some exceptions diverging from their importance categorization. These findings illustrate how DSAI can identify discrepancies between theoretical feature importance and actual implementation patterns in the data.

D.1 Analysis of Expert-defined Criteria

We categorize expert criteria based on their importance to slogan effectiveness, ranging from critical to supplementary features. This categorization provides a framework for analyzing which features persist across different thresholds.

(a) Core Message Delivery features (Critical Importance). These features are essential for effective communication of the brand's core sales proposition and unique characteristics. They are considered critical because they directly influence a slogan's ability to capture and deliver the brand's main message to consumers.

- 4. Convey the sales idea clearly and concisely
- 8. Emphasize the brand's unique traits or the benefits it provides

(b) Structure and Expression features (High Importance). These features focus on clarity and readability, making a slogan easy to understand and leaving a lasting impression. They significantly impact audience engagement and retention, though they may not directly affect the message's content.

- 1. It should be direct and straightforward
- 2. Keep it simple, but not overly simple
- 5. Avoid misleading or exaggerated words

(c) Tone and Atmosphere features (Moderate Importance). These features create an emotional connection with the audience through engaging and positive tone. While they enhance a slogan's appeal, they are not as pivotal as core message delivery or structural clarity.

- 3. It should have a pleasant tone
- 6. It should be future-oriented

(d) Supplementary features (Lower Importance). These features can enhance the slogan's overall quality or effectiveness, but are not essential for a slogan's primary function.

- 9. Include the brand name in the slogan
- 7. Clear positioning through comparison or closeness

D.2 Prominence Intensity Threshold Analysis

Table 5 displays the effect of different prominence thresholds on recall and data size.

Recall	Threshold (Prominence)	Data Size
1	0.003	235
0.889	0.348	83
0.778	0.549	44
0.667	0.714	15
0.556	0.750	12
0.444	0.833	6
0.333	0.857	5
0	1	0

Table 5: Recall and Data Size Across Prominence Thresholds

Recall remains stable until reaching a relatively high threshold (0.348), indicating robust baseline coverage of our pipeline. The relationship between importance levels and threshold resilience shows an overall aligned pattern, though with some complexity: while lower-importance criteria are consistently filtered out early, criteria of moderate importance and above show varied retention patterns, with some excluded at mid-level thresholds and others persisting until the highest thresholds.

(1) **Early Exclusions (<0.6)**: *7. Clear positioning through comparison or closeness* (lower importance) and *9. Include the brand name in the slogan‘ (lower importance)* are excluded first, consistent with their supplementary nature.
(2) **Mid-level Exclusions (<0.8)**: General criteria such as *2. Keep it simple but not overly simple* (high importance) and *4. Convey the sales idea clearly and concisely* (critical importance) are excluded only at higher thresholds (0.721 and above), underscoring their broad applicability.
(3) **Late (No) Exclusions**: Multiple features with moderate (*6. Future-oriented*) or above (*e.g., 1. Directness* (high importance), *8. Emphasize the unique traits and benefits* (critical importance)) persist until the highest thresholds.

D.3 Frequency Threshold Analysis

Table 6 illustrates how recall changes as we filter features based on their frequency in the dataset.

Recall remains stable up to a frequency threshold of 93, indicating strong coverage of expert criteria even when considering only frequently observed patterns. The relationship between theoretical importance and threshold resilience reveals both expected alignments and interesting disparities. Critical features persist at high thresholds, and certain lower-importance features being filtered early. Some theoretically important features drop out earlier than expected, while certain lower-importance features demonstrate surprisingly high frequency in practice.

(1) **Early Exclusions (<500)**: Notably, *5. Avoid misleading words* (high importance) and *6. Future-oriented tone* (moderate importance) drop out first. According to our theoretical categorization, these are not supplementary features, yet they appear less frequently in actual slogans than their theoretical significance would suggest.

Recall	Threshold (Frequency)	Data Size
1	5	236
0.889	93	86
0.778	217	66
0.667	571	38
0.556	661	31
0.444	1,005	16
0.333	1,115	9
0.111	1,192	2
0	1,196	1

Table 6: Recall and Data Size Across Frequency Thresholds

(2) **Mid-level Exclusions (<1000)**: *3. Pleasant tone* (moderate importance) and *9. Include brand name* (lower importance) are filtered out at moderate thresholds, aligning with their moderate importance categorization.
(3) **Late Exclusions**: Multiple core features like *4. Convey the sales idea clearly* (critical importance) and *1. Direct and straightforward* (high importance) has high frequency, confirming that these features are both theoretically critical and practically prevalent. *7. Clear positioning through comparison* (lower importance) shows a notable deviation between theoretical and practical implementation.

E Threshold Analysis on Research Titles

Similar to our analysis on slogans (Appendix D), the results suggest that DSAI can effectively capture the nuanced reality of title construction, where practical implementation patterns may differ from theoretical guidelines.

E.1 Analysis of Expert-defined Criteria

We begin by categorizing expert-defined criteria based on their contribution to a title's primary function: effective delivery of research content to readers.

Core Information Delivery features (Critical Importance). These features are fundamental as they directly affect a title's ability to enable readers quickly grasp the paper's topic and contributions.

- 3. The title needs to be informative and specific
- 4. The title needs to be functional (with essential scientific "keywords")
- 6. The title should include the main theme of the paper

Structural and Format features → Readability (High Importance). These features optimize the title's readability and clarity. While they do not directly affect content, they are crucial for successful information delivery.

- 1. The title needs to be simple in terms of format
- 2. The title needs to be direct
- 5. The title should be concise and precise
- 7. The title should not be too long or too short

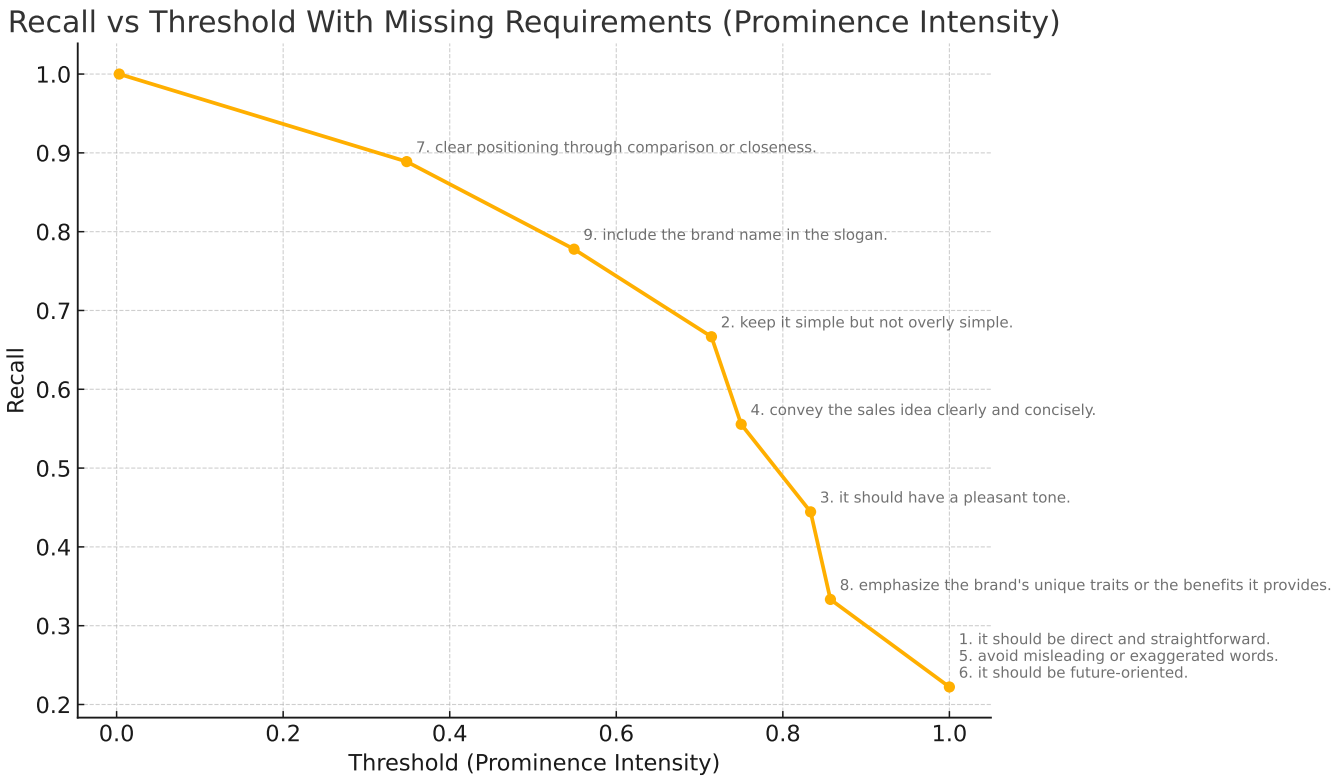


Figure 6: Dropped criterion as Prominence threshold increases

Linguistic Expression features (Moderate Importance). These features maintain academic professionalism while ensuring accessibility. They enhance the title's effectiveness without being critical to its basic function.

- 8. The title should avoid whimsical or amusing words
- 9. The title should avoid non-standard abbreviations and unnecessary acronyms (or technical jargon)

Supplementary Guidelines (Lower Importance). These guidelines can be advantageous in specific contexts but are not universally essential.

- 10. Place of the study and sample size should be mentioned only if it adds to the scientific value of the title
- 11. Important terms/keywords should be placed at the beginning of the title
- 12. Descriptive titles are preferred to declarative or interrogative titles

E.2 Prominence Intensity Threshold Analysis

Table 7 demonstrates the impact of increasing the prominence threshold on the recall and data size.

The majority of features are retained even at a high threshold of 0.692, suggesting that our methodology is robust in covering essential features. Analysis of the exclusion pattern demonstrates a general relationship between feature importance and retention, although some show divergence from their importance categorization:

Recall	Threshold (Prominence)	Data Size
0.833	0	199
0.750	0.692	40
0.667	0.778	29
0	1	0

Table 7: Recall and Data Size Across Prominence Thresholds

(1) **Early Exclusions (<0.6)**: None.
(2) **Mid-Level Exclusions (<0.8)**: Style-related, moderate-importance features including *7. Not too long or short* and *8. Avoid whimsical words* and are excluded at moderate thresholds.
(3) **Late (No) Exclusions**: Features of higher importance related to core message delivery and readability remain intact until the highest thresholds, underscoring their fundamental nature. Notable exceptions are *9. Avoid non-standard abbreviations* and *12. Descriptive type*, which are retained despite their mid to lower importance.

E.3 Frequency Threshold Analysis

We observe recall changes across varying 'Frequency' thresholds.

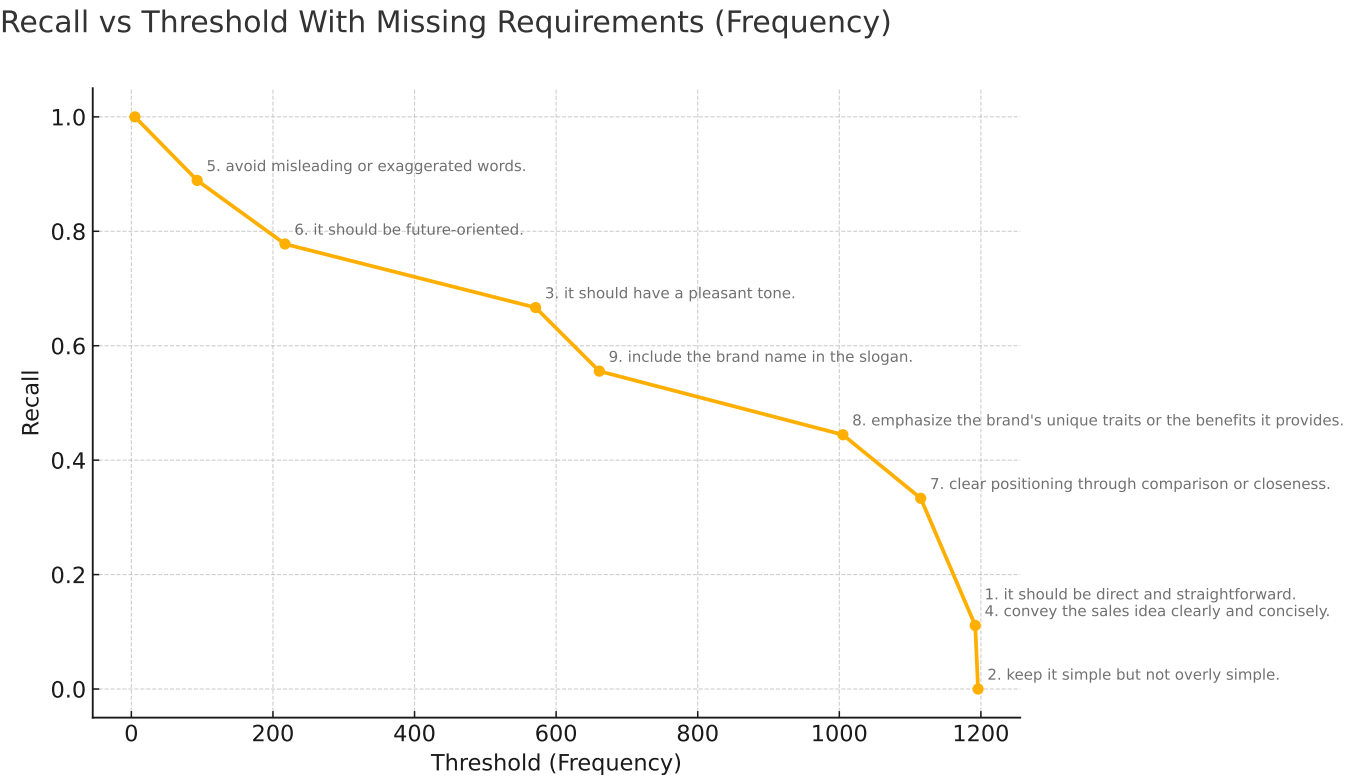


Figure 7: Dropped criterion as Frequency threshold increases.

Recall	Threshold (Frequency)	Data Size
0.833	1	284
0.750	791	44
0.667	969	32
0.583	1,233	26
0.500	1,294	20
0.417	1,296	19
0.333	1,326	13
0.250	1,333	11
0.167	1,344	4
0.083	1,345	1

Table 8: Recall and Data Size Across Frequency Thresholds

Recall remains stable until a threshold of 791. The relationship between theoretical importance and retention patterns again reveals some unexpected deviations: a high-importance feature drops out early, while other features of similar importance persist until high thresholds. Additionally, while some supplementary features are not recalled at all, the recalled one (*12. Descriptive type*) shows remarkably strong retention, suggesting that practical title construction may prioritize certain features differently from theoretical guidelines.

(1) **Early Exclusions (<500)**: None.
(2) **Mid-Level Exclusions (<1000)**: Features ranging from moderate (*9. Avoid non-standard abbreviations*) to critical importance (*6. Include the main theme*) are excluded at this stage.
(3) **Late Exclusions**: Structural features like *1. Simple format* (high importance), *5. Concise and precise* (high importance) persist until the highest thresholds, aligning with their theoretical importance. Interestingly, *12. Descriptive type* (low importance) shows the highest retention despite its low theoretical importance.

F DSAI-Generated Top/Bottom Prominence Features of Expert-Driven Annotation Dataset

As discussed in Section 5.1, the features generated using the slogan and title datasets demonstrate high recall values when compared to expert-defined requirements. This holds true even when applying a high threshold for prominence, indicating that the generated features effectively capture the key characteristics outlined by experts. However, as highlighted in Section 5.2, while features with high prominence exhibit strong discriminative power and reliability, those with lower prominence tend to have relatively lower reliability.

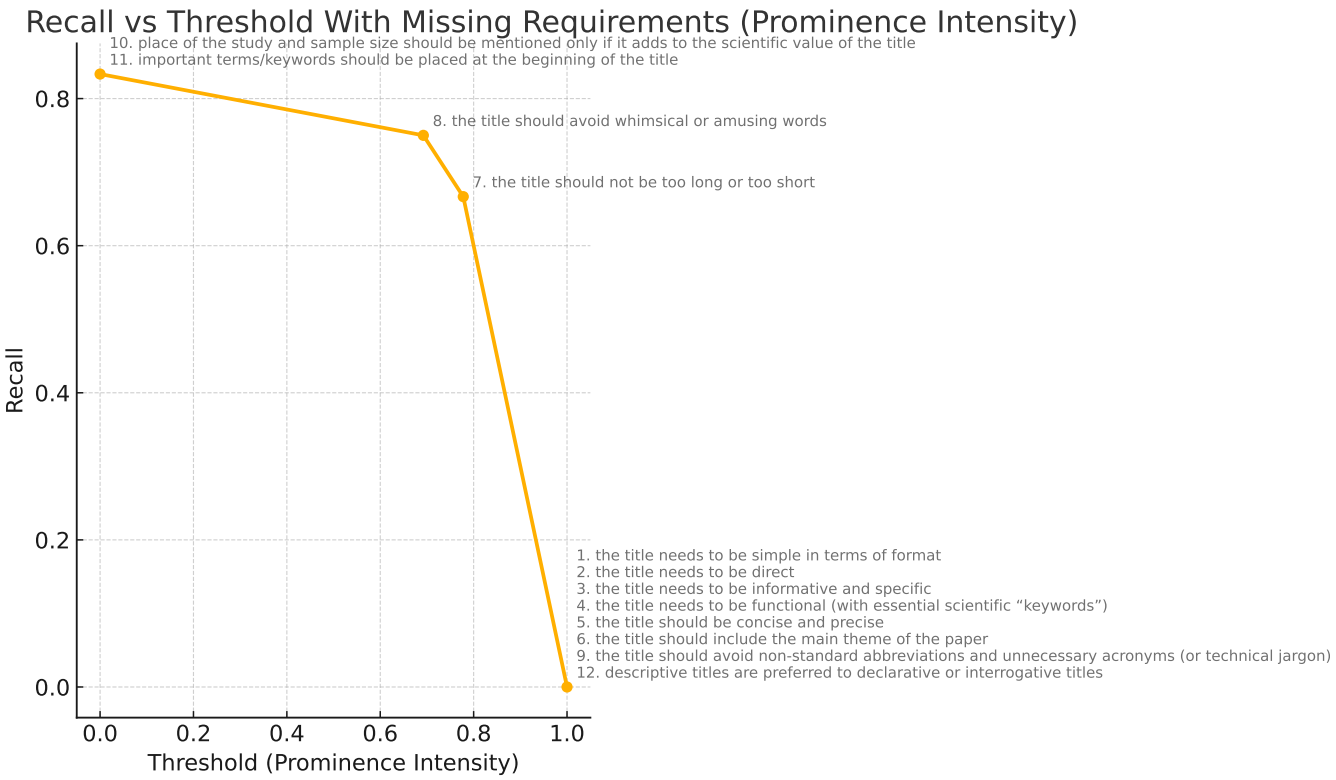


Figure 8: Dropped criterion as Prominence threshold increases.

To provide further insights, we present the top and bottom 20 features for each dataset along with their prominence scores and frequencies. The results for the slogan dataset are detailed in Table 10, and those for the title dataset are outlined in Table 11.

G DSAI-Generated Top/Bottom Prominence Features of Industry Dataset

G.1 Prominence Distribution

The following analysis evaluates the prominence score distribution of criteria across three datasets: Reddit, MIND, and SPAM. These scores indicate how reliably each feature identifies the latent characteristics within the datasets.

MIND. The MIND dataset shows a highly concentrated prominence distribution, with most scores reaching 1.0. Few features fall below 0.85, reflecting strong alignment between the generated features and the structured nature of the dataset. These consistently high scores validate the methodology's effectiveness in extracting meaningful patterns from structured data with CTR rates.

Spam. The Spam detection dataset exhibits a broader prominence distribution than MIND. Still, a large portion are above 0.9, indicating that the methodology is generally effective for SPAM. However, a small subset scoring below 0.5 indicates features that may be less suitable for classification purposes, suggesting the need for careful filtering of specific features for robust classification.

Reddit. The Reddit dataset displays a broad and uneven distribution of prominence scores, peaking at 0.67 with a gradual decline toward lower scores. A significant number of features cluster below 0.4, with many falling below 0.1. This distribution reflects the unstructured and highly diverse nature of Reddit content, making consistent pattern identification challenging. The broad range of topics and content variability suggests that individual features may be highly specific to certain data subsets while falling to generalize across others.

G.2 General vs Specific Features

DSAI-generated features capture both general characteristics shared across the dataset and highly specific features. As shown in Table 9, some features represent broad, overarching traits that are present across a significant portion of the dataset, with high frequencies (900+ instances). These include common patterns such as "lack of logical reasoning" or "absence of historical reflection," which are applicable across diverse contexts.

In contrast, more specific features capture nuanced and detailed attributes that apply to smaller data subsets, appearing in fewer samples ($\tilde{1}$00 instances), such as "sarcastic undertone" or "emotionally intense". These characteristics necessitate fine-grained, data-driven analysis and are typically challenging for LLMs to identify due to their limited presence in pre-training data and the source dataset. This capability to extract such specific patterns beyond general

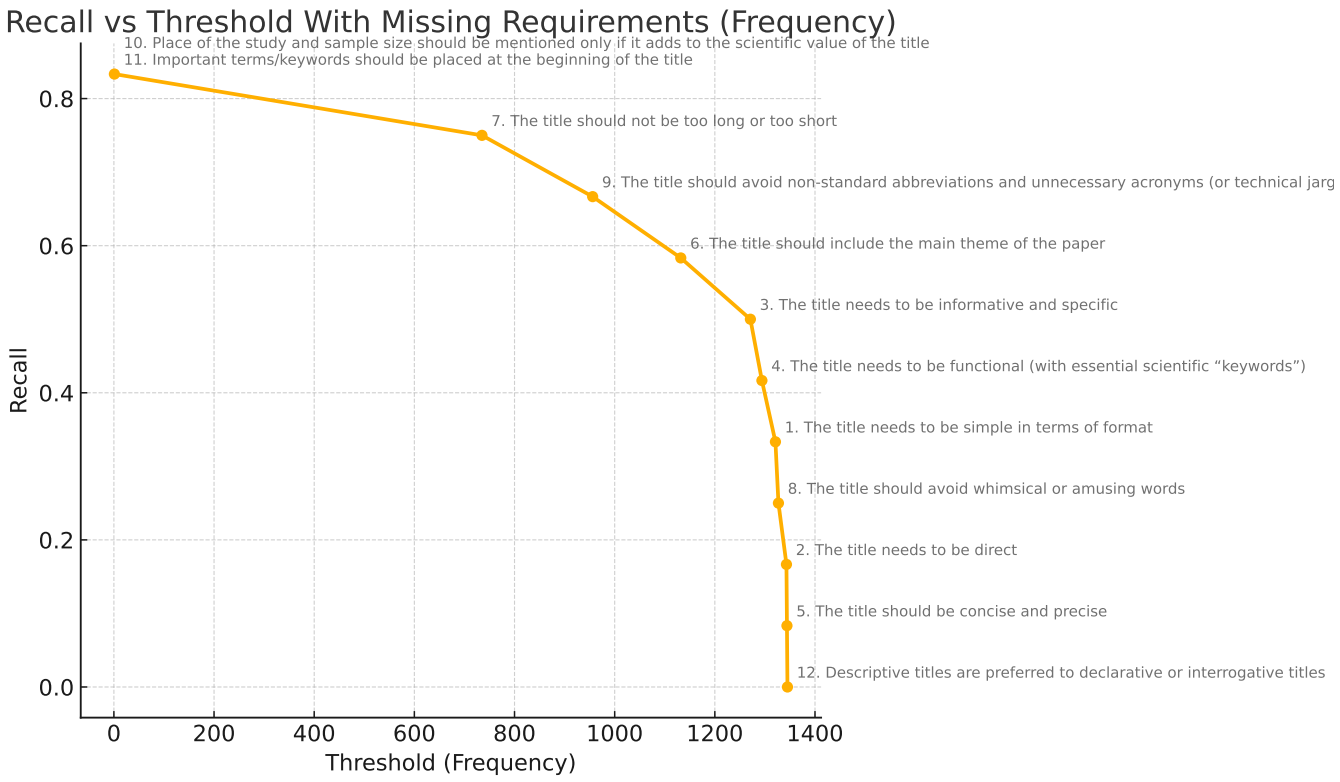


Figure 9: Dropped criterion as Frequency threshold increases

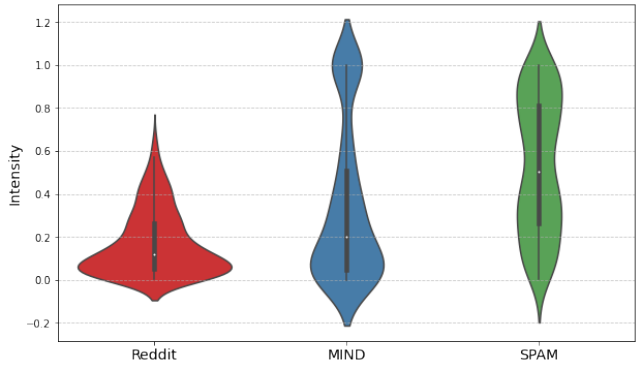


Figure 10: Distribution of each dataset based on prominence scores.

trends demonstrate DSAI's unique strength over direct feature extraction from LLMs.

The top and bottom 20 prominence fearures for MIND, SPAM and Reddit dataset are provided in Table 12, 13, 14, respectively.

Perspective	Label	Type	Frequency
General Features			
reasoning	Lack of Logical Reasoning	NEGATIVE	1,038
fictional_reference	Absent	NEGATIVE	1,033
gender_roles	no_reference	POSITIVE	1,014
archaic_reference	no_reference	NEGATIVE	1,004
euphemistic_language	euphemism_absent	NEGATIVE	986
historical_reflection	Absent	NEGATIVE	966
partisan_tone	absent	POSITIVE	964
legal_context	Absence	POSITIVE	962
objective_evaluation	Subjective	NEGATIVE	950
validity	subjective	NEGATIVE	926
scientific_reference	No Scientific Reference	NEGATIVE	918
profanity	Profanity Absent	POSITIVE	910
Specific Features			
specific_undertone	Sarcastic	NEGATIVE	116
theme_recognition	Justice and Morality	NEGATIVE	116
cultural_sensitivity	high_sensitivity	POSITIVE	114
efficacy	Ineffectiveness	POSITIVE	114
comparison	Indirect Comparison	NEGATIVE	114
pragmatic	Criticism	NEGATIVE	110
audience_engagement	Indirect Engagement	POSITIVE	110
irony	Intensity_Irony	NEGATIVE	110
conflict	Presence	POSITIVE	104
demographic_target	Youth	NEGATIVE	102
emotionally_charged	Emotionally Intense	NEGATIVE	102
length	two_digit_high	POSITIVE	102

Table 9: Features and their corresponding labels, types, and frequency.

Requirement	Prominence	Frequency
Top Requirements		
The advertising tone should convey a focus on quality, using optimistic and aspirational language.	0.8571	14
The sentence should convey an optimistic advertising tone that encourages engagement.	0.8333	108
The sentence should incorporate cultural references that align with consumptive themes.	0.8333	12
The sentence should employ indirect methods to engage the audience effectively.	0.7857	28
Ensure the sentence contains a component that emotionally appeals to the reader.	0.7831	83
The sentence should effectively incorporate figurative language to enhance meaning.	0.7556	90
The sentence should convey a high level of prominence using strong and emotive language.	0.7215	79
Ensure the sentence conveys its message concisely, avoiding unnecessary words or lengthy expressions.	0.7143	14
The sentence should clearly present a promise of quality, assuring trust and excellence.	0.6406	217
The sentence should employ direct engagement techniques to capture the audience's attention.	0.6250	128
The sentence should avoid losing its identity by being overly generic or broad in purpose.	0.6190	21
Ensure grammatical accuracy throughout the sentence.	0.6154	26
Avoid using the ampersand (&) in formal writing or titles.	0.5882	34
The sentence should clearly define and communicate a distinct value proposition.	0.5870	184
Ensure the sentence effectively conveys its intended message.	0.5789	19
The sentence should convey a strong and clear emotional tone.	0.5709	275
The sentence should provoke thought and engage the audience in a meaningful way.	0.5556	18
The sentence should avoid presenting the main idea in a way that misaligns with the intended topic or domain.	0.5556	18
The sentence should use persuasive language to effectively encourage action or belief.	0.5528	407
Ensure the sentence includes the mention of a relevant company or organization.	0.5493	213
Bottom Requirements		
The sentence should include a clear and compelling call to action.	0.0294	1,088
The sentence should fully and effectively communicate its intended message.	0.0267	1,159
The sentence should avoid merely providing information without an intended action or emotion.	0.0250	1,122
Ensure the sentence includes references to cultural significance.	0.0248	1,009
Avoid using imperative or overly complex grammatical structures in titles.	0.0244	41
Ensure the use of inclusive language in the sentence.	0.0225	1,109
Avoid sentences that lack necessary cultural references.	0.0224	1,115
The sentence should not be overly specific, limiting its relevance to a narrow audience.	0.0205	537
The sentence should maintain a general level of specificity to appeal to a wide audience.	0.0166	661
Ensure that sentences maintain grammatical correctness.	0.0162	1,171
The sentence should focus on delivering specific and detailed information.	0.0148	741
Avoid using general language and focus on providing specific details in the sentence.	0.0131	458
The sentence should be concise, conveying its message briefly and efficiently.	0.0110	1,185
The sentence should maintain clear and direct language.	0.0095	1,157
The sentence should avoid overly focusing on the business aspect at the expense of other dimensions.	0.0070	572
The sentence should explicitly mention the presence of a specific service being offered.	0.0048	1,043
The sentence should effectively convey clear and useful information.	0.0043	1,163
The sentence should maintain simplicity in structure and composition.	0.0034	,192
The sentence should be composed in a single language for consistency.	0.0033	1,196
Avoid emotionally neutral language that fails to connect with readers on an emotional level.	0.0032	315

Table 10: Top and Bottom Requirements of Slogan Dataset based on Prominence.

Requirement	Prominence	Frequency
Top Requirements		
Avoid using list structures that compromise grammatical integrity and clarity.	1.0000	11
Avoid vague and nonspecific language in the sentence.	1.0000	11
Avoid the use of pronominal subjects in sentences.	1.0000	12
Ensure the sentence provides a clearer focus on the topic.	1.0000	16
Avoid using an informal narrative style in the sentence.	1.0000	33
Ensure the sentence maintains a strong technical focus without shifting to a non-technical direction.	1.0000	41
Ensure the inclusion of phrasal nouns in the sentence where appropriate.	0.9333	30
The sentence should clearly convey its purpose or intent.	0.9259	27
Ensure the sentence addresses the topic clearly without any ambiguity.	0.8854	192
Ensure the sentence clearly specifies its domain of application.	0.8750	16
The sentence should incorporate technical adjectives to enhance precision and clarity in describing nouns or pronouns.	0.8333	12
The sentence should clearly highlight learning methods as the main topic.	0.8182	11
Avoid using interrogative sentences.	0.8182	33
Ensure complete grammatical structures in the sentence.	0.8095	21
Avoid excessive use of concrete nouns in the sentence.	0.7857	28
Avoid overly complex or frequent use of interrogative structures in the content.	0.7838	37
The sentence should avoid being overly verbose and aim for syntactic compression.	0.7739	115
Ensure sentences are concise and avoid unnecessary length.	0.7500	40
The sentence should clearly use either passive or active voice to avoid ambiguity.	0.7333	30
Avoid focusing on concrete, tangible items in the sentence.	0.7000	40
Bottom Requirements		
Ensure modal verbs are absent in the sentence if not necessary for conveying meaning.	0.0840	917
Limit the length of main noun phrases to enhance readability.	0.0833	96
The main noun phrase in the sentence should ideally consist of two words.	0.0833	144
Avoid using sentence fragments to ensure complete and meaningful sentences.	0.0805	174
Avoid using definite articles when introducing new or less familiar concepts.	0.0748	147
The sentence should effectively use past tense to communicate events or actions that have already occurred.	0.0667	15
The sentence should embrace high complexity in its structure and vocabulary.	0.0581	172
Avoid sentences with overly simplistic structures or vocabulary.	0.0575	435
Ensure sentences utilize participles effectively to enhance clarity and detail.	0.0529	473
Ensure the sentence includes infinitive verb forms for clarity and action orientation.	0.0526	57
Avoid reliance on non-verbal tense constructs in the sentence.	0.0500	40
The sentence should clearly identify and be applicable to the educational domain.	0.0476	21
The sentence should maintain a neutral tone by balancing between passive and active voice.	0.0476	84
Limit the use of multiple verbs in a single sentence.	0.0455	88
Avoid sentences that inappropriately mix technical terms outside the humanities and social sciences domain.	0.0400	25
Limit the use of multiple adjectives in a sentence to maintain clarity and precision.	0.0357	56
The sentence should clearly articulate its relationship to the field of Health and Medicine.	0.0323	31
Ensure the sentence includes well-chosen adjectives to enhance description and clarity.	0.0121	330
Ensure the presence of a clear tense in the sentence.	0.0076	264
Ensure the presence of finite verbs to convey clear action or timing in the sentence.	0.0039	510

Table 11: Top and Bottom Requirements of Title Dataset based on Prominence.

Requirement	Prominence	Frequency
Top Requirements		
The sentence should effectively convey a sense of distress through emotional language.	1.0000	12
Avoid language that implies assistance or support for actions that should be independent.	0.8462	13
Avoid statements that may have an unintended negative impact on the audience.	0.7895	19
Avoid overloading the sentence with complex or unnecessary details related to technology and cybersecurity.	0.7241	29
The sentence should effectively convey a sense of frustration through emotional language.	0.7143	14
The sentence should aim to surprise or alarm the audience to elicit a strong reaction.	0.6667	24
Avoid using language that only implies future outcomes without clear details.	0.6667	18
Avoid using domain-specific language related to the military and defense.	0.6364	11
The sentence should convey information of lesser importance effectively.	0.6364	11
The sentence should avoid focusing solely on technology and innovation when identifying the audience.	0.6364	11
The sentence should clearly specify the legal or criminal aspects of a particular field.	0.6000	35
Avoid using language that fails to motivate or inspire the audience.	0.5625	32
Ensure the main subject of the sentence pertains to sports.	0.5556	18
The sentence should clearly reference past legal cases using appropriate legal terminology.	0.5385	26
Avoid using quality-based adjectives that may imply subjective judgment or bias.	0.5000	24
The sentence should not overly highlight negative issues that could alarm the audience.	0.5000	12
Avoid referencing military actions, personnel, or events in the content.	0.4737	38
The sentence should avoid focusing solely on social and lifestyle aspects within its domain.	0.4737	19
The sentence should avoid using action verbs that emphasize creation-related actions.	0.4545	11
Avoid using ambiguous modal verbs when expressing plans or intentions.	0.4444	18
Bottom Requirements		
Ensure the inclusion of specific numerical data or statistics in the sentence.	0.0078	766
Avoid overly clear or explicit language that might be inappropriate or too revealing in certain contexts.	0.0073	961
Ensure the sentence does not lack attention to potential contradictions.	0.0071	989
The sentence should exclude references to social media platforms.	0.0062	966
The sentence should effectively communicate its message without relying on figurative language.	0.0059	678
The sentence should avoid the use of separators such as dashes, colons, or slashes.	0.0054	742
Ensure the inclusion of sensory details in the sentence to enhance vividness.	0.0051	973
The sentence should clearly provide instruction or guidance to the reader.	0.0051	979
The sentence should present concrete, specific ideas and details.	0.0042	946
Ensure that the sentence is clear and needs no further clarification.	0.0042	952
Ensure that scientific theories or principles are present in the sentence.	0.0041	976
The sentence should avoid the use of hashtags.	0.0040	994
The sentence should provide sufficient context to be understood independently.	0.0035	865
Avoid or limit the use of figurative language in the sentence.	0.0032	313
Avoid sentences that lack necessary negations to clarify intended meaning.	0.0031	977
Avoid using ambiguous or unsuitable speech acts in sentence construction.	0.0030	989
The sentence should avoid hypothetical scenarios to maintain factual clarity.	0.0022	898
The sentence should avoid content that lacks relevance or significance, especially for a news context.	0.0020	980
The sentence should clearly identify the language being used, ensuring linguistic features are consistent with the specified language.	0.0020	998
The sentence should be free from unnecessary repetition, ensuring clarity and conciseness.	0.0010	985

Table 12: Top and Bottom Requirements of MIND dataset based on Prominence.

Requirement	Prominence	Frequency
Top Requirements		
Avoid using list structures that compromise grammatical integrity and clarity.	1.0000	11
Avoid vague and nonspecific language in the sentence.	1.0000	11
Avoid the use of pronominal subjects in sentences.	1.0000	12
Ensure the sentence provides a clearer focus on the topic.	1.0000	16
Avoid using an informal narrative style in the sentence.	1.0000	33
Ensure the sentence maintains a strong technical focus without shifting to a non-technical direction.	1.0000	41
Ensure the inclusion of phrasal nouns in the sentence where appropriate.	0.9333	30
The sentence should clearly convey its purpose or intent.	0.9259	27
Ensure the sentence addresses the topic clearly without any ambiguity.	0.8854	192
Ensure the sentence clearly specifies its domain of application.	0.8750	16
The sentence should incorporate technical adjectives to enhance precision and clarity in describing nouns or pronouns.	0.8333	12
The sentence should clearly highlight learning methods as the main topic.	0.8182	11
Avoid using interrogative sentences.	0.8182	33
Ensure complete grammatical structures in the sentence.	0.8095	21
Avoid excessive use of concrete nouns in the sentence.	0.7857	28
Avoid overly complex or frequent use of interrogative structures in the content.	0.7838	37
The sentence should avoid being overly verbose and aim for syntactic compression.	0.7739	115
Ensure sentences are concise and avoid unnecessary length.	0.7500	40
The sentence should clearly use either passive or active voice to avoid ambiguity.	0.7333	30
Avoid focusing on concrete, tangible items in the sentence.	0.7000	40
Bottom Requirements		
Ensure modal verbs are absent in the sentence if not necessary for conveying meaning.	0.0840	917
Limit the length of main noun phrases to enhance readability.	0.0833	96
The main noun phrase in the sentence should ideally consist of two words.	0.0833	144
Avoid using sentence fragments to ensure complete and meaningful sentences.	0.0805	174
Avoid using definite articles when introducing new or less familiar concepts.	0.0748	147
The sentence should effectively use past tense to communicate events or actions that have already occurred.	0.0667	15
The sentence should embrace high complexity in its structure and vocabulary.	0.0581	172
Avoid sentences with overly simplistic structures or vocabulary.	0.0575	435
Ensure sentences utilize participles effectively to enhance clarity and detail.	0.0529	473
Ensure the sentence includes infinitive verb forms for clarity and action orientation.	0.0526	57
Avoid reliance on non-verbal tense constructs in the sentence.	0.0500	40
The sentence should clearly identify and be applicable to the educational domain.	0.0476	21
The sentence should maintain a neutral tone by balancing between passive and active voice.	0.0476	84
Limit the use of multiple verbs in a single sentence.	0.0455	88
Avoid sentences that inappropriately mix technical terms outside the humanities and social sciences domain.	0.0400	25
Limit the use of multiple adjectives in a sentence to maintain clarity and precision.	0.0357	56
The sentence should clearly articulate its relationship to the field of Health and Medicine.	0.0323	31
Ensure the sentence includes well-chosen adjectives to enhance description and clarity.	0.0121	330
Ensure the presence of a clear tense in the sentence.	0.0076	264
Ensure the presence of finite verbs to convey clear action or timing in the sentence.	0.0039	510

Table 13: Top and Bottom Requirements of Spam Detection Dataset based on Prominence.

Requirement	Prominence	Frequency
Top Requirements		
The sentence should avoid unnecessary criticism and focus on constructive feedback.	0.6727	110
The sentence should avoid using strong partisan language or political bias.	0.6129	124
The sentence should include elements that evoke a surreal or dream-like quality.	0.6104	154
The sentence should avoid being perceived solely as a review.	0.6082	194
The sentence should effectively convey a positive emotion.	0.5699	186
The sentence should effectively convey information and facilitate informative sharing with the audience.	0.5333	120
Avoid using a critical tone in the sentence.	0.5313	128
The sentence should convey a positive sentiment.	0.5294	136
Avoid language that inaccurately places the sentence within the politics and media domain.	0.5238	126
Ensure the sentence does not include language indicating complaints or grievances.	0.5130	230
The sentence should align with public opinion and mainstream consensus.	0.5062	162
The sentence should address topics in a non-sensitive manner, avoiding contentious or inflammatory language.	0.5044	226
Avoid displaying a dominant interpersonal stance in relational dynamics.	0.4907	161
The sentence should convey a positive emotional tone.	0.4815	108
The sentence should convey information with high intensity and detail to maximize audience knowledge gain.	0.4737	190
The content should maintain an objective, unbiased perspective throughout.	0.4729	129
The sentence should use words with optimistic or positive connotations.	0.4516	124
The language used should maintain a neutral tone, avoiding extreme politeness or rudeness.	0.4476	210
Avoid using circular or flawed reasoning in arguments.	0.4468	188
The sentence should avoid rhetorical devices and maintain a straightforward approach.	0.4430	237
Bottom Requirements		
The content should maintain a moderate level of complexity.	0.0116	518
Avoid non-impressionistic language and incorporate more vivid or subjective descriptions.	0.0110	182
The sentence should avoid legal language or references.	0.0104	962
The sentence should avoid suggesting a lack of consensus or collective agreement.	0.0099	202
The sentence should appropriately reference hierarchical authority to establish credibility or significance.	0.0093	214
Avoid using American regional dialects or cultural references in the sentence.	0.0090	222
The sentence should avoid depicting any specific gender roles or influences.	0.0081	867
Ensure the sentence includes relevant jargon where necessary to convey expertise and precision.	0.0080	754
The sentence should avoid overly intense expressions of real-world relevance.	0.0062	1,129
Ensure the sentence length is concise, ideally with a word count in the single digits.	0.0061	330
The sentence should avoid making explicit factual assertions without sufficient evidence or context.	0.0060	668
The sentence should be inclusive and cater to a broad demographic audience.	0.0056	354
The sentence should clearly demonstrate an understanding of various sentence types and structures.	0.0053	1,126
The sentence should not restrict or misrepresent opinion sharing.	0.0047	426
The sentence should be intellectually demanding but still accessible at a medium level of complexity.	0.0045	446
Avoid using exclusive language or sentiments in the sentence.	0.0043	235
Minimize the use of overly dramatic or theatrical language in the sentence.	0.0027	375
The sentence should maintain a concise length, ideally within a low two-digit word count.	0.0026	389
The sentence should intentionally exclude hashtags.	0.0025	1,191
Ensure language identification and classification are accurate within the sentence.	0.0017	1,194

Table 14: Top and Bottom Requirements of Reddit dataset based on Prominence.

SLOGAN

Expert-Defined Features

1. **It should be direct and straightforward.**
2. **Keep it simple, but not overly simple.**
3. **It should have a pleasant tone.**
4. **Convey the sales idea clearly and concisely.**
5. **Avoid misleading or exaggerated words.**
6. **It should be future-oriented.**
7. **Clear positioning through comparison or closeness.**
8. **Emphasize the brand's unique traits or the benefits it provides.**
9. **Include the brand name in the slogan.**

TITLE

Expert-Defined Features

1. **The title needs to be simple in terms of format**: Avoid complex or overly structured titles. A clear, single-sentence structure is ideal.
2. **The title needs to be direct:** Go straight to the point without embellishment or unnecessary words. Emphasize what the research is about, making it easy to understand.
3. The title needs to be informative and specific: Clearly state the subject, scope, and purpose of the study. Include key variables, interventions, or outcomes being studied.
4. **The title needs to be functional (with essential scientific "keywords")**: Include important scientific terms that are likely to be used in search engines or databases.
5. **The title should be concise and precise**: Each word should add value-if it's not essential, consider removing it. Focus on delivering the core message without sacrificing clarity.
6. **The title should include the main theme of the paper:** Highlight the central idea or concept of the research, ensuring it's prominent in the title. If your research has a novel angle or contribution, make it clear in the title. Avoid distracting details that take away from the main topic.
7. **The title should not be too long or too short:** Keep the length balanced: avoid single-word titles or excessively long sentences. Aim for a length that is easy to read but informative enough to cover the subject. Try to stay within 12–15 words to maintain readability.
8. **The title should avoid whimsical or amusing words:** Scientific titles should maintain professionalism and avoid humor, puns, or colloquial expressions. The tone should be neutral and objective.
9. **The title should avoid nonstandard abbreviations and unnecessary acronyms (or technical jargon):** Avoid using abbreviations that aren't widely understood by the target audience. Use only well-known acronyms, such as DNA, or none at all. If technical terms are necessary, ensure they are clear to the intended readership.
10. **Place of the study and sample size should be mentioned only if it adds to the scientific value of the title.** Only mention specific locations (e.g., "in Africa" or "in urban China") if they significantly impact the results or generalizability of the study.: Sample size should only be included if it adds notable scientific value (e.g., "meta-analysis of 50,000 participants").
11. **Important terms/keywords should be placed in the beginning of the title**: Place the most relevant words/phrases at the start of the title to grab attention quickly. Consider what keywords are essential for indexing and searchability, and prioritize them. Ensure the first few words convey the core subject of the study.
12. **Descriptive titles are preferred to declarative or interrogative titles**: Focus on describing the study rather than making a statement or posing a question. Declarative titles ("X is the cause of Y") or interrogative ones ("Is X the cause of Y?") are often less effective in scientific communication. Descriptive titles ("Investigation of X as a potential cause of Y") are clearer and more widely accepted.

Figure 11: Expert-Defined Criteria

SLOGAN

(NoData) Result

1. Short and concise: A slogan should be easy to remember, so it needs to be brief and to the point.
2. Memorable: A slogan should stick in people's minds.
3. Unique: A slogan should differentiate from competitors.
4. Emotionally impactful: It should evoke a feeling in the audience.
5. Relevant to the brand: Reflect the core values and mission of the company.
6. Easy to pronounce: It should be simple to say aloud.
7. Timeless: The slogan should remain effective for years to come.
8. Flexible: Can be adapted to different contexts or campaigns.
9. Clear message: It should be easy to understand without explanation.
10. Catchy: A rhythmic or playful slogan can be more appealing.
11. Positive tone: A positive message resonates better with audiences.
12. Target audience-focused: Tailored to the interests and needs of your target demographic.
13. Action-oriented: Encourages consumers to take action.
14. Reflective of the product: Should directly connect to the product or service offered.
15. Original: Avoids clichés and overly used phrases.
16. Cultural fit: Respects and fits the cultural context of the target audience.
17. Flexible in translation: If used internationally, it should maintain meaning when translated.
18. Direct and assertive: A clear call to action can make it more compelling.
19. Humor: Using humor can make a slogan more relatable.
20. Inspiring: Motivates the audience toward a specific idea or goal.

(...)

80. Sincere: Authentic and truthful, not overhyped.
81. Global appeal: Resonates with audiences from various cultures or backgrounds.
82. Cohesive with visuals: Matches the imagery and color palette of the brand.
83. Surprising: Takes an unexpected angle that grabs attention.
84. Confidence-building: Inspires the consumer to feel confident about their choice.
85. Balanced tone: Neither too serious nor too casual; fits the brand's voice.
86. Encourages sharing: Motivates consumers to spread the message or talk about the brand.
87. Puns or wordplay: Clever use of language can make it stand out.
88. Geared toward action: Motivates the consumer to do something, whether buying or engaging.
89. Simple but profound: Deep meaning expressed in simple words.
90. First-person appeal: Speaks as if directly from the brand's voice to the consumer.
91. Value-driven: Emphasizes the brand's core values, such as sustainability or social responsibility.
92. Focuses on results: Highlights what the consumer will achieve by choosing the brand.
93. Contrasts competitors: Subtly shows how the brand differs from competitors.
94. Heartfelt: Evokes a genuine emotional connection.
95. Smart and witty: Intellectually stimulating while being enjoyable.
96. Exciting promise: Offers an exciting, bold future for the consumer.
97. Relevant to current issues: Ties into major societal or cultural issues.
98. Encourages exploration: Suggests that there's more to discover with the brand.
99. Play on words: Cleverly manipulates language for a memorable twist.
100. Energetic rhythm: The cadence of the words should have a natural flow, making it catchy and fun.

Figure 12: NoData Result for slogan

SLOGAN

(PosData) Result

0. Concise and Memorable Phrase: Use short, impactful phrases that are easy to remember, conveying the core value of the brand.
1. Emphasize Clear Benefits: Highlight the benefits or advantages that users will gain.
2. Form an Emotional Connection: Use emotional words to evoke a connection with the audience.
3. Emphasize Brand Uniqueness: Clearly state what makes the brand unique or its competitive advantage.
4. Directly Address the Target Audience: Mention the target audience explicitly so they feel related to the brand.
5. Focus on Problem-Solving: Emphasize the brand's solution to the problems that users face.
6. Convey Trust and Expertise: Use words that convey reliability and expertise to instill confidence.
7. Present a Clear Offering: Clearly explain what the brand offers without ambiguity.
8. Present Specific, Measurable Outcomes: Mention specific, measurable outcomes to build trust in the brand's offering.
9. Emphasize Local Connection: Highlight the brand's connection to a specific region to strengthen community ties.
10. Emphasize User Experience: Highlight the positive experience users will gain by using the brand.
11. Strive Beyond Problem-Solving Towards a Goal: Go beyond solving a problem, and present a vision of a better future or goal.
12. Emphasize Leadership in Expertise: Convey the brand's leading role in the industry.
13. Use Encouraging Expressions for Users: Use a tone that encourages and supports users to create a positive brand image.
14. Highlight Trustworthy Track Record: Mention the brand's long-standing experience or trustworthy track record.
15. Message of Hope and Change: Use messages that provide hope or promote positive change.
16. Use Imperative Expressions to Drive Engagement: Use imperative phrases that directly call users to action.
17. Emphasize Innovation and Progress: Highlight the brand's innovative approach or technological advancement.

(FlippedPosData) Result

0, Clarity & Conciseness: Keep the slogan short and easy to understand.
1, Targeted Audience Appeal: Tailor the message to the needs or desires of the target audience.
2," Emotional Connection: Evoke feelings, whether of safety, excitement, or reliability."
3, Unique Selling Proposition (USP): Highlight what sets the business apart from competitors.
4, Branding Consistency: The slogan should be consistent with the brand's image or mission.
5, Positive Tone: Emphasize positivity to create a favorable impression.
6, Call to Action or Benefit Statement: Suggest an action or communicate a benefit to customers.
7, Memorability: Use catchy words or phrases to ensure the slogan is memorable.
8, Relevance to Core Services: Include keywords related to the business's services.
9, Problem-Solving Approach: Identify the problem and position the brand as a solution.
10, Use of Powerful Adjectives: Employ strong adjectives to convey a sense of superiority.
11, Industry Specificity: Use language specific to the industry to show expertise.
12, Promise of Quality or Reliability: Emphasize quality or reliability to build trust.
13, Innovation Highlight: Showcase innovation or unique technology.
14, Wordplay or Rhyme: Use clever wordplay or rhyme to enhance recall.
15, Geographical Relevance: Mention location to appeal to local markets.
16, Future-Oriented Perspective: Provide a vision for progress or innovation.
17," Focus on Core Values: Reflect core values like sustainability, customer focus, or excellence."
18, Implying Exclusivity or Luxury: Convey exclusivity for high-end products or services.
19," Adaptability for Multiple Media: Design the slogan to fit different platforms, from social media to print."

Figure 13: PosData and FlippedPosData Results for slogan

SLOGAN

(MixedData) Result

0. Clear and Intuitive Message: A slogan should be short, concise, and easy to understand
1. Conveys Brand Value Specifically: It should clearly convey the value of the service or product
2. Positive and Inspirational Tone: It should be positive and inspire customers to feel good about the brand
3. Clearly Targeted Audience: The target audience should be easily identifiable from the slogan
4. Offers a Solution to Customer Needs: It should include content that solves a customer's problem or meets a need
5. Simple and Easy to Remember: It should be simple and easy to recall, using short phrases
6. Use of Active and Action-Oriented Words: It should use words that encourage customer participation or action
7. Reflects Brand Identity: It should reflect the brand's identity, philosophy, and values
8. Originality: It should use language that is unique and cannot be easily replicated by others
9. Use of Relevant Keywords: It should be composed of keywords related to the brand
10. Avoid Vague and Unclear Expression: It is difficult for customers to understand exactly what is being offered
11. Avoid Too Generic and Unattractive: Uses too generic or clichéd words, lacking brand differentiation
12. Avoid Overly Complex or Uses Technical Terms: Uses jargon or technical terms that are difficult for customers to understand
13. Avoid Unnecessary Redundancy: Repeats words or phrases unnecessarily, making the slogan long and complicated
14. Avoid Inappropriate Tone: The slogan does not match the brand's image or may give a negative impression
15. Avoid Lacks Perceived Benefits for Customers: The benefits that customers can gain from the brand are not clearly stated
16. Avoid Fails to Elicit Interest or Emotion: Does not evoke emotional connection from customers, making it dry and unengaging

(FlippedMixedData) Result

1. Clear Message Delivery: The core message should be easily understood at a glance.
2. Engaging Question Format: Using questions can pique curiosity and encourage customer engagement.
3. Directly Mentioning Customer Benefits: Clearly state the benefits or value the customer can expect.
4. Highlighting Real-world Applications or Trustworthiness: Mentioning specific examples or industry standing can build trust.
5. Concise and Memorable Expression: A short and catchy slogan is easier to remember.
6. Clear Proposal for Problem-solving: Identify the customer's problem and offer a solution.
7. Unique and Differentiated Ideas: Present a unique service or concept that stands out from competitors.
8. Emphasizing Specific Regions or Target Audiences: Mentioning specific regions or target demographics gives a sense of personalized service.
9. Using Action-Oriented Verbs: Use expressions that encourage immediate action from customers.
10. Strengthening Emotional Connection: Use phrases that connect emotionally with the customer.
11. Focus on Core Competency or Specialty: Highlighting what makes the business unique or what they do best helps establish expertise.
12. Creating a Sense of Urgency or Timeliness: Using words that create urgency can prompt quicker customer actions.
(...)
26. Playing with Alliteration for Catchiness:Alliteration can make slogans easier to remember and more pleasant to read.
27. Using Analogies or Comparisons:Drawing comparisons helps make abstract concepts easier to understand.
28. Using Rhymes for Memorability:Rhymes make slogans catchy and easier to remember.

Figure 14: MixedData and FlippedMixedData Results for slogan

SLOGAN

(NoContext) Result

1. Purposefulness and Consistency: The data maintains a clear theme aligned with its purpose and is structured to meet specific language processing objectives, such as semantic analysis, syntax parsing, or translation quality improvement.
2. Practical Applicability: The data provides information that can be effectively used in real-world language processing applications, such as translation, information retrieval, text generation, or meaning inference.
3. Well-defined Research Problems: It addresses a well-defined research question with specific aims, such as morpheme analysis or syntax tree construction for a particular language.
4. Accurate and Consistent Labeling: The data is thoroughly labeled, with clear and consistent labeling criteria. For example, data aimed at identifying biological interactions through pattern learning would have precise labeling aligned with its objective.
5. Consideration for Multiple Languages and Domains: It includes data that can be beneficial for multiple languages or specific domains (e.g., biomedical texts, medical data, legal documents). For instance, data optimized for applications like multilingual translation or medical record analysis.
6. Sophisticated Structure and Use of Labels: The data provides rich information through syntactic structures, semantic layers, or multi-level labels. This might include data with syntax trees and semantic frames that offer detailed structure.
7. Accuracy and Reliability: The data is collected from reliable sources or processed in a way that guarantees high accuracy, often using high-quality linguistic resources.
8. Clear Copyright and Data Source Information: The data is clearly labeled with copyright and source information, ensuring ethical usage conditions.

Figure 15: NoContext Result for slogan

TITLE

(NoData) Result

1. Clearly Specify Research Subject or Application Area: The research subject or application area should be clearly mentioned.
2. Concise and Specific: Keep the title brief but detailed enough to convey the essence of the study.
3. Use Descriptive Words: Include descriptive keywords that capture the core of the research.
4. Avoid Jargon: Minimize the use of technical jargon that might not be understood by a broader audience.
5. Indicate the Study Type: Mention the type of study, e.g., case study, analysis, experiment.
6. Incorporate Key Findings: Hint at the main findings or conclusions of the research.
7. Use Active Voice: Employ an active voice to make the title more dynamic and direct.
8. Be Accurate: Ensure the title accurately reflects the content and scope of the paper.
9. Avoid Abbreviations and Acronyms: Use full terms instead of abbreviations to avoid confusion.
10. Highlight the Value or Impact: Suggest how the research contributes to the field or practical applications.
11. Incorporate Relevant Keywords: Include keywords that are relevant and likely searched by the target audience.
12. Balance General and Specific Elements: Provide enough specific detail while keeping the title applicable to a broader field.
13. Attract the Intended Audience: Tailor the language and content of the title to appeal to the intended readers.
14. Make it Memorable: Aim for a title that is catchy and memorable, where appropriate.
15. Avoid Misleading Terms: Don't use terms or phrases that might mislead readers about the study's scope or findings.
16. Check for Unintentional Humor or Puns: Ensure the title doesn't inadvertently create confusion or amusement unless intended.
17. Use Numerical Results: If applicable, include quantitative results (like percentages or statistical significance).
18. Reflect the Tone of the Writing: Match the formality or informality of the title to the tone of the paper.
19. Consult Co-Authors: Get input from all authors to ensure the title effectively encompasses everyone's views.
20. Incorporate Geographical Scope: If relevant, include the geographic focus of the research.
(...)
86. Clarify the Main Variables: Make sure to specify the main variables of interest.
87. Mention the Study's Limitations: If limitations are a critical aspect of the discussion, they can be hinted at in the title.
88. Highlight the Type of Intervention: Detail the type of intervention if the research focuses on this aspect.
89. Use Terms Reflecting the Research Outcome: Use terms that directly relate to the main findings or outcomes.
90. Include the Nature of the Study Sample: If the sample nature is crucial, include this detail.
91. Highlight the Control Measures Used: If applicable, note the control measures employed in the study.
92. Specify Any Mathematical Models Used: If mathematical models are central, mention them.
93. Mention the Key Chemicals or Compounds: If specific chemicals or compounds are critical, include them.
94. Highlight the Research Framework: If a particular framework guides the research, mention it.
95. Clarify the Study's Core Purpose: Make sure the title reflects the primary purpose of the research.
96. Indicate the Main Data Sources: If particular data sources are vital, mention them.
97. Use Impactful Adjectives: Selectively use adjectives that add clarity and interest.
98. Mention the Research Environment: If the environment is a significant factor, include it.
99. Clarify the Main Contribution: Specify what new contribution the paper makes to its field.
100. Be Original: Ensure the title stands out by being original and not merely imitative of existing works.

Figure 16: NoData Result for title

TITLE

(PosData) Result

1. Conciseness and Clarity: The title should be short, concise, and clearly convey the main idea. For example, "BestCut: A Graph Algorithm for Coreference Resolution" effectively explains the core topic in a succinct manner.
2. Specificity: Mentioning the specific topic or methodology of the paper allows the reader to understand the research's subject at a glance. "Cross-lingual Lexical Triggers in Statistical Language Modeling" clearly specifies an element of the language model.
3. Emphasis on Research Methods: Highlighting specific techniques or algorithms used in the study helps clarify the approach. For example, "Approximate PCFG Parsing Using Tensor Decomposition" includes the methodology, enabling the reader to quickly understand the research approach.
4. Description of Results or Goals: Mentioning the outcome or goal of the research helps readers understand what problem the study aims to solve. Titles like "Improving Translation Quality by Discarding Most of the Phrasetable" make the goal of the research easy to grasp.
5. Mention of Application Area: When the research is applied to a specific field, mentioning it in the title helps specify the study's purpose. For example, "Exploring the Use of NLP in the Disclosure of Electronic Patient Records" clearly describes the practical objective of the research.
6. Specify Data or Language: When the research is limited to a specific language or dataset, specifying it helps define the scope of the study. "Chinese Named Entity Recognition with Conditional Random Fields" clearly states the target language.
7. Problem and Solution Structure: A title that explicitly mentions the problem and its solution is effective. For example, "Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation" succinctly expresses the problem and solution.
8. Use of Quantitative Expressions: Representing the scope of research with numbers or quantitative expressions enhances specificity. For instance, "Million-scale Derivation of Semantic Relations" emphasizes the large-scale data used in the research.
9. Creative Combination: Creatively combining well-known concepts can emphasize the uniqueness of the study. For example, "Grounding Strategic Conversation: Using Negotiation Dialogues to Predict Trades in a Win-Lose Game" creatively shows connections to other fields.
10. Use of Algorithm or Technique Name: Including specific methodologies or techniques in the title emphasizes the key tools of the research. For example, "Convolutional Neural Networks for Sentence Classification."
(...)
16. Specify the Research Domain: It is beneficial to clarify the subject or phenomenon being studied. For example, "Language Modeling for Determiner Selection" mentions the specific linguistic phenomenon.
17. Mentioning New Datasets: If a new dataset was built or used, mentioning it in the title highlights the contribution, such as "Creating a Test Corpus of Clinical Notes Manually Tagged for Part-of-Speech."
18. Emphasize Theoretical Approach: It is also effective to emphasize theoretical models or approaches proposed in the study, e.g., "Towards Creating Precision Grammars from Interlinear Glossed Text."
19. Use of Adjectives and Adverbs: Using adjectives or adverbs that emphasize the nature of the research can be effective, such as "Efficient Dynamic Programming Search Algorithms for Phrase-Based SMT."
20. Include Key Keywords: Including important keywords of the study in the title helps make the core subject clear, e.g., "Paraphrase Lattice for Statistical Machine Translation."
21. Comparison and Contrast: When the study compares or contrasts multiple approaches, mentioning it in the title can clarify the scope of the research, e.g., "Comparing Phrase-Based and Syntax-Based Paraphrase Generation."
22. Emphasizing Outcomes or Contributions: Highlighting the outcomes or contributions of the research can explain its value, e.g., "Treebank Translation for Cross-Lingual Parser Induction."
23. Explicitly Stating Key Research Questions: It can also be helpful to explicitly mention the key questions addressed in the research, e.g., "Understanding Differences in Perceived Peer-Review Helpfulness Using Natural Language Processing."

Figure 17: PosData Result for title

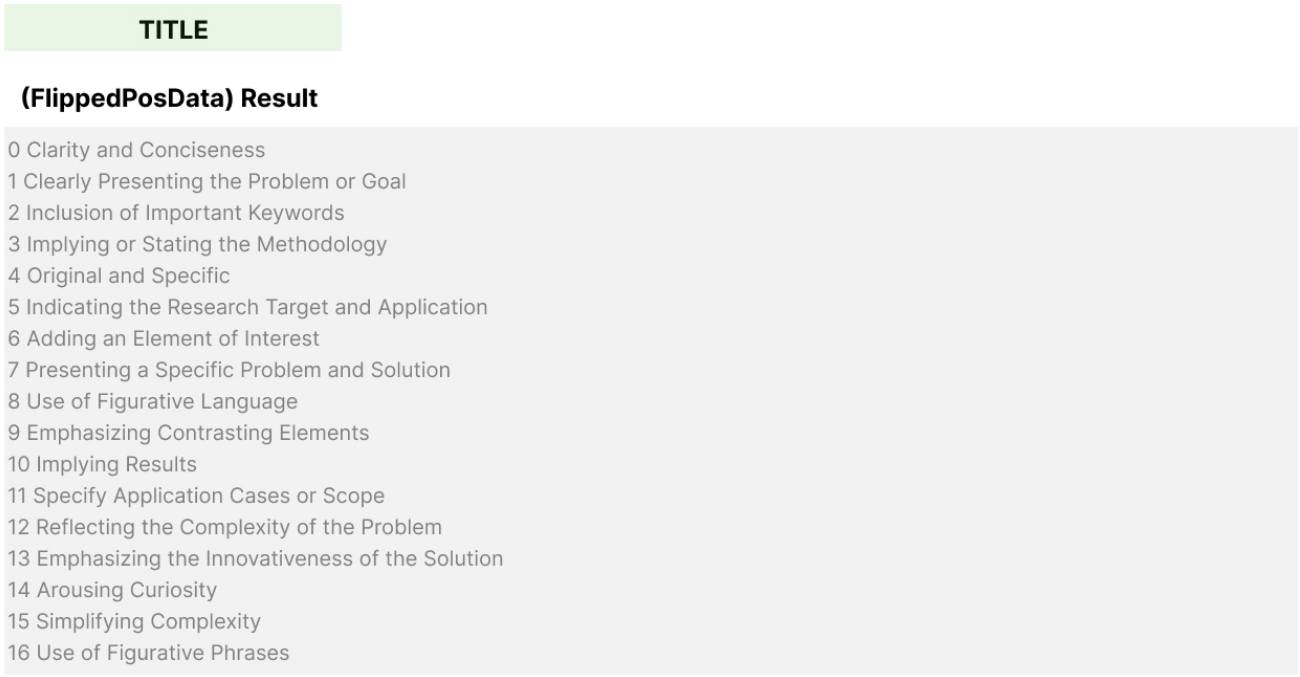


Figure 18: FlippedPosData Result for title

TITLE

(MixedData) Result

1. Clear and Specific Topic: The title should clearly express the subject of the paper and indicate the specific research objective or methodology.
2. Include Key Terms: The title should contain key terms related to the main technique, methodology, subject, or data used in the research.
3. Descriptive Subtitle: The title should include a descriptive subtitle that is informative without being overly long.
4. Emphasize Methodology or Dataset: It is beneficial to emphasize the primary methodology or dataset used in the research.
5. Clearly Specify Research Subject or Application Area: The research subject or application area should be clearly mentioned.
6. Indicate Research Objective or Outcome: If possible, convey the objective or outcome of the research in the title, making it easier for readers to understand.
7. Use Precise Vocabulary: Avoid ambiguous or exaggerated expressions, and use terminology that accurately reflects the research.
8. Minimize Use of Abbreviations: Avoid abbreviations that are difficult for readers to understand or are rarely used. If necessary, include the full name.
9. Emphasize Contribution: Highlighting the contribution of the research in the title makes it easier for readers to assess its value.
10. Proper Grammar and Expression: Ensure there are no grammatical errors or awkward expressions, and structure the sentence clearly.
11. Appropriate Length: The title should be around 10-15 words, sufficient to convey the content of the research without being too long or complicated.
12. Highlight Novel Approaches: If the research involves a novel methodology or innovative approach, it should be highlighted to emphasize its value.
13. Precise Focus: The research should focus on a specific problem and convey a clear focus in the title.

(FlippedMixedData) Result

0 Specific and Clear Topic
1 Mention of Methodology or Techniques
2 Application Area or Case Study Mentioned
3 Emphasis on Novelty or Approach
4 Presentation of Outcomes or Impact
5 Concise and Direct Expression
6 Appropriate Use of Technical Terms
7 Balanced Length
8 Focus on Problem-Solving
9 Highlighting Interdisciplinary Connections
10 Using Action-Oriented Words
11 Indicating Evaluation Metrics or Criteria
12 Including Scope or Scale
13 Clarity in Comparative Studies
14 Avoiding Redundancy
(...)
74 Avoiding Overly Detailed Descriptions
75 Avoiding Lack of Substantial Information
76 Avoiding Overuse of Difficult Jargon
77 Avoiding Missing Goals or Expected Outcomes
78 Avoiding Repetitive Phrasing

Figure 19: MixedData and FlippedMixedData Results for title

TITLE

(NoContext) Result

1. Technical NLP Approaches: Group A prominently includes papers focusing on technical NLP methodologies, particularly for complex tasks like entity linking, morphosyntactic parsing, probabilistic models, machine translation quality, and the computational structure of discourse. Group B, in contrast, includes general linguistic applications or dialogue systems with a more applied focus.
2. Multilingual Processing and Cross-Linguistic Analysis: Group A's examples frequently address multilinguality (e.g., multilingual translation lexicons, Arabic phonetic adjustments) and cross-linguistic techniques (e.g., Hindi, Bengali, Korean, Chinese), highlighting a strong interest in processing diverse languages and dialects. Group B also includes multilingual aspects but not with the same emphasis on cross-linguistic computational challenges.
3. Advanced Computational Linguistics Concepts: Group A shows a particular emphasis on computational linguistics at a deep level, with topics like dependency-based word spaces, corpus-based semantic lexicon induction, approximate parsing, structural models, and entity tracking within complex frameworks. Group B generally focuses on language processing and semantic similarity but lacks the depth of structured computational concepts found in Group A.
4. Corpus and Data-Driven Approaches: Group A has a notable focus on leveraging extensive corpora and structured data (e.g., domain adaptation in SMT, machine translation evaluation frameworks) and the technical constraints of processing large datasets for training or lexical derivations. Group B examples lean more toward user-based or interface-focused applications without extensive corpus analysis.
5. Higher-Level NLP Techniques and Algorithms: Group A frequently references complex algorithms and specialized techniques such as tensor decomposition, latent variable models, tree kernels, and dynamic programming, often oriented towards optimizing and refining natural language understanding and machine learning tasks. Group B's focus is generally less computationally intensive and leans more toward general ML approaches like sentiment analysis or relation extraction.
6. Focus on Language-Specific Challenges: Group A highlights challenges tied to specific languages or linguistic properties, such as morphology, syntactic variations, and regional language dialects, as seen in work on Inuktitut morphology, Arabic pronunciation, and Hindi root extraction. Group B does not emphasize language-specific intricacies to the same extent.
7. Entity, Event, and Relation Extraction Emphasis: Group A includes a significant focus on extracting complex linguistic elements, such as events, semantic relations, or discourse structures, from text (e.g., role extraction, semantic coherence, noun-verb structures). Group B examples show a broader focus on general-purpose NLP tasks rather than in-depth extraction of intricate linguistic elements.
8. Theoretical or Abstract Language Processing: Group A presents research in abstract concepts within NLP, including discourse modeling, lexical cohesion, and sentence-level semantic abstractions. Group B appears more practically focused, addressing sentiment classification, translation accuracy, or dialogue systems with fewer abstract linguistic models.

Figure 20: NoContext Result for title