NOVASCORE : A New Automated Metric for Evaluating Document Level Novelty

Lin Ai[1,2], Ziwei Gong[1,2], Harshsaiprasad Deshpande[1], Alexander Johnson[1], Emmy Phung[1], Ahmad Emami[1], Julia Hirschberg[2]
[1]Machine Learning Center of Excellence, JPMorgan Chase & Co.
[2]Department of Computer Science, Columbia University
{lin.ai, sara.ziweigong, julia}@cs.columbia.edu,

Abstract

The rapid expansion of online content has intensified the issue of information redundancy, underscoring the need for solutions that can identify genuinely new information. Despite this challenge, the research community has seen a decline in focus on novelty detection, particularly with the rise of large language models (LLMs). Additionally, previous approaches have relied heavily on human annotation, which is time-consuming, costly, and particularly challenging when annotators must compare a target document against a vast number of historical documents. In this work, we introduce **NOVASCORE** (**No**velty E**v**aluation in **A**tomicity **Score**), an automated metric for evaluating document-level novelty. NOVASCORE aggregates the novelty and salience scores of atomic information, providing high interpretability and a detailed analysis of a document's novelty. With its dynamic weight adjustment scheme, NOVASCORE offers enhanced flexibility and an additional dimension to assess both the novelty level and the importance of information within a document. Our experiments show that NOVASCORE strongly correlates with human judgments of novelty, achieving a 0.626 Point-Biserial correlation on the TAP-DLND 1.0 dataset and a 0.920 Pearson correlation on an internal human-annotated dataset.

1 Introduction

Textual novelty detection has long been a key challenge in information retrieval (IR) (Soboroff and Harman, 2005), focusing on identifying text that introduces new, previously unknown information. With the rapid expansion of online content, this issue has become more significant, as redundant information increasingly obstructs the delivery of critical, timely, and high-quality content (Ghosal et al., 2022). Schwartz (2022) reveals that 60% of internet content is duplicated. The rise of Large Language Models (LLMs) has further contributed to the generation of artificial and semantically redundant information. Detecting whether a document provides new, relevant, and salient information is crucial for conserving space, saving time, and maintaining reader engagement.

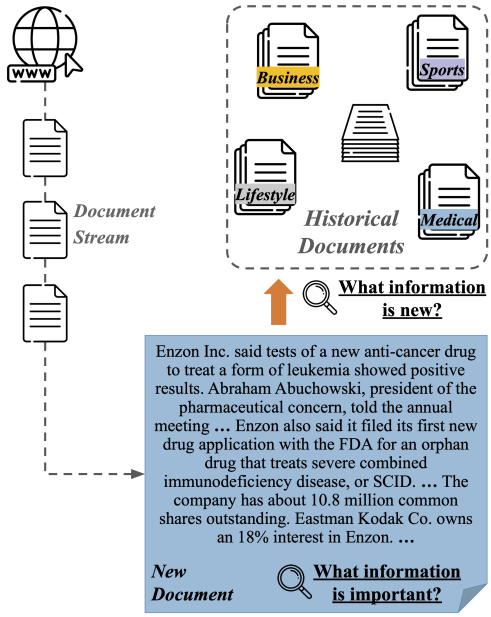

Figure 1: Conceptual illustration of novelty and salient information retrieval in real-world applications.

In addition, Li et al. (2024) introduce novelty as a key metric for benchmark design, noting that performance on existing benchmarks is often highly correlated (Liu et al., 2023a; Perlitz et al., 2024; Polo et al., 2024). Novelty helps uncover hidden performance patterns and unexpected model behaviors, enabling more dynamic evaluations and the development of higher-quality benchmarks that push the limits of model improvement.

Despite the increasing issue of information redundancy and the growing need for novelty in benchmarking, focus on novelty detection has declined, especially since the rise of LLMs after 2022. Most prior efforts in document-level novelty detection rely on single categorical classification,

Proceedings of the 31st International Conference on Computational Linguistics, pages 3479–3494
January 19–24, 2025.

lacking detailed analysis of what is genuinely new within a document. Additionally, previous work has overlooked the salience of information – how important each piece is and how it contributes to assessing a document's overall novelty and value. These methods also heavily depend on human annotation, which is time-consuming, costly, and challenging, especially when comparing a target document against many historical documents (Ghosal et al., 2018a), as illustrated in Figure 1.

Our motivation is twofold: **(a)** to develop a new metric for document-level novelty that offers granular analysis and incorporates the salience of information, and **(b)** to provide an automated solution that reduces the costs and time associated with manual labeling. Our contributions are as follows:

1. We introduce **NovAScore**, short for **No**velty Evaluation in **A**tomicity **Score**, an automated metric for evaluating document-level novelty. NovAScore aggregates the novelty and salience scores of atomic content units, providing high interpretability and demonstrating strong correlation with human judgments of novelty.
2. We release NovAScore as an open-source tool[1], encouraging further research to expand its applicability and enhance its scalability.

2 Related Work

Novelty Detection Textual novelty detection has its roots in early IR research, particularly through the Topic Detection and Tracking (TDT) campaigns. These efforts focused on new event detection by clustering news stories based on similarity thresholds (Wayne, 1997; Brants et al., 2003). The task gained further prominence during the Text Retrieval Conferences (TREC) from 2002 to 2004, where sentence-level novelty detection became a focal point (Soboroff et al., 2003; Clarke et al., 2004; Soboroff and Harman, 2005; Schiffman and McKeown, 2005). While sentence-level detection was well-researched, it is insufficient for addressing the vast amount of document-level information available on the web today (Ghosal et al., 2022).

At the document level, Yang et al. (2002) pioneered the use of topical classification for detecting novelty in online document streams. Zhang et al. (2002) introduced redundancy measures to assess document novelty. More recent approaches have explored information entropy measures (Dasgupta and Dey, 2016), deep neural networks (Ghosal et al., 2018a), multi-source textual entailment (Ghosal et al., 2022), and unsupervised approaches (Nair, 2024) for detecting novelty in documents.

Information Similarity Evaluation Directly assessing the novelty of information is challenging. However, numerous metrics exist for evaluating *semantic similarity* between pieces of information. A common approach involves using cosine similarity between contextual embeddings, as seen in methods like BertScore (Zhang et al., 2020), MoverScore (Zhao et al., 2019), and BartScore (Yuan et al., 2021). Additionally, Natural Language Inference (NLI) is widely recognized for evaluating information similarity and consistency. It is frequently employed in novelty detection (Dagan et al., 2022; Ghosal et al., 2022), summarization evaluation (Liu et al., 2023c; Laban et al., 2022), and factuality assessment (Min et al., 2023; Zha et al., 2023; Ji et al., 2023). Beyond these one-stage metrics, two-stage approaches, such as QA-based methods, are extensively used to evaluate information overlap and faithfulness in both summarization and factuality evaluations (Deutsch et al., 2021; Zhong et al., 2021; Goyal et al., 2022; Fabbri et al., 2022). In our work, we utilize and assess all three categories of approaches as a close approximation for identifying semantic-level non-novelty.

3 NovAScore

We introduce NovAScore, a new automated method for evaluating the novelty of a target document compared to a series of historical documents. Unlike previous methods that assign a categorical value to the entire document, NovAScore offers an interpretable and granular analysis at the atomicity level. As shown in Figure 2, the process starts by decomposing the target document into *Atomic Content Units* (ACUs). We define an ACU similarly to Min et al. (2023) (*atomic facts*) and Liu et al. (2023b), but with a more holistic perspective – an elementary information group that combines the minimal number of atomic facts necessary to convey a single message. Each ACU is then evaluated for novelty by comparing it to an *ACUBank* of historical documents and assessed for its salience within the document's context. The overall NovAScore is computed by aggregating the novelty and salience scores of all ACUs. After processing, the target document ACUs can be stored in the *ACUBank* for future analysis. This approach allows for a high-level assessment of the document's nov-

[1]Our code is available at this GitHub repository.

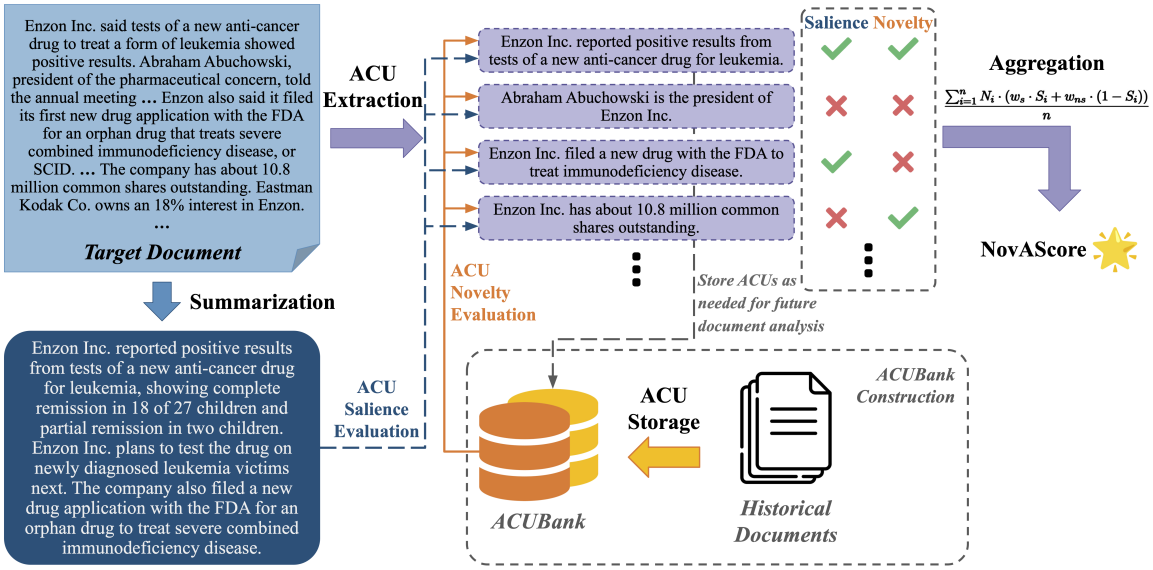

Figure 2: The NOVASCORE framework. The target document is first decomposed into ACUs. ACU-level novelty is assessed by comparing each ACU against the *ACUBank* of historical documents, while salience is determined by whether the ACU is included in the document's summary. The final NOVASCORE is calculated by aggregating the scores of the ACUs. ACUs can be stored in the *ACUBank* for future analysis if necessary.

elty while precisely identifying new and important information, providing fine-grained interpretability.

3.1 ACU Extraction and *ACUBank*

We build on the ideas from Liu et al. (2023b) and Min et al. (2023) to extract abstractive ACUs, but unlike their methods, which break down sentences into highly fine-grained units, we extract document-level ACUs directly. This approach better suits our task of novelty detection, which requires a holistic evaluation of new information rather than overly fine details. We frame automatic document-level ACU extraction as a sequence-to-sequence problem (Sutskever et al., 2014): $m(D) \rightarrow A$, where D is the input document, m is a language model, and A is the set of generated document-level ACUs.

While previous research has focused on sentence-level novelty (Schiffman and McKeown, 2005; Ghosal et al., 2022), we choose ACU-level analysis to better handle complex, information-dense sentences and to maintain context by considering messages that may span multiple sentences.

To efficiently evaluate the novelty of target ACUs, we construct an ***ACUBank*** – a collection of databases that store ACUs from historical documents – allowing for quick similarity searches at minimal computational cost without the need for real-time relevant content retrieval. The databases are built by indexing ACUs using SentenceBERT embeddings (Reimers and Gurevych, 2019). For each ACU, the most relevant historical ACUs are rapidly retrieved via semantic cosine similarity to assess novelty. To speed up searches, the *ACUBank* is organized into multiple databases, each containing ACUs from specific *clusters*, so a target document is only searched within its cluster, significantly narrowing the search scope.

3.2 ACU Novelty Evaluation

We approximate non-novel information assessment using three common information similarity evaluators: embedding cosine similarity, NLI, and QA, as discussed in Section 2. These evaluators assess ACU novelty, treating it as a binary task – determining whether the information is new or not.

Cosine similarity provides a straightforward approach to evaluating ACU novelty. We compare each target ACU with historical ACUs from the *ACUBank*. If any historical ACU exceeds a set similarity threshold with the target ACU, it is classified as non-novel, indicating likely repetition; and vice versa. This method efficiently assesses overlap, making it a practical tool for novelty detection.

NLI is based on the principle that a premise P entails a hypothesis H if a typical human reader would conclude that H is most likely true after reading P (Dagan et al., 2005). In the context of novelty detection, this means that *if one or more entailing premises are found for a given hypothesis, the content of that hypothesis is considered not new* (Bentivogli et al., 2011). To evaluate ACU-level

novelty, we concatenate the most relevant historical ACUs into a single premise and compare it against the target ACU as the hypothesis. If the historical content entails the target ACU, it is classified as non-novel; otherwise, it is considered novel.

QA-based approach is widely used to evaluate information overlap by representing reference information as question-answer pairs, with recall assessed by how well the candidate text answers these questions (Deutsch et al., 2021; Zhong et al., 2021). We adapt this method in reverse: for each target ACU, we generate questions where the target ACU itself is the answer. If any historical ACUs can answer these questions, the target ACU is considered non-novel; otherwise, it is novel. We generate three questions per ACU, focusing on *named entities* and *noun phrases* (Deutsch et al., 2021). The answers derived from historical ACUs are consolidated into a single sentence and compared to the target ACU. If the consolidated answer has a cosine similarity of 0.85 or higher with the target ACU, it is classified as non-novel. The rationale for this threshold is detailed in Appendix A.2.2.

3.3 ACU Salience Evaluation

Not all information in a document is equally important. For instance, as shown in Figure 2, the primary focus of the target document is Enzon Inc.'s positive results for a new medication, while the company's ownership structure, briefly mentioned later, is less significant within the pharmaceutical domain. Therefore, when evaluating novelty, it is essential to prioritize the most important content to ensure an accurate assessment.

To determine the salience of each ACU, we compare it to the document's summary. The underlying assumption is that a high-quality summary should include all and only the essential information from the document. Therefore, we formulate ACU salience evaluation as a binary classification problem: whether or not an ACU is mentioned in the document's summary.

3.4 ACU Scores Aggregation

When aggregating ACU scores to compute the overall NOVASCORE of a document, it is essential to assign higher weights to salient ACUs to accurately reflect their importance. To achieve this, we implement a **dynamic weight adjustment** scheme based on the following principles:

1. **Salience Emphasis at Low Salience Ratio:** When the ratio of salient ACUs is low, each salient ACU is assigned a significantly higher weight compared to non-salient ACUs. This ensures that the final score is not overly influenced by the novelty of less important content.
2. **Non-Salience Boost at High Salience Ratio:** When the proportion of salient ACUs is high, the weights of non-salient ACUs are increased slightly to ensure they still contribute meaningfully to the overall score.
3. **Consistent Prioritization:** Salient ACUs consistently receive higher weights than non-salient ACUs, regardless of their proportion.

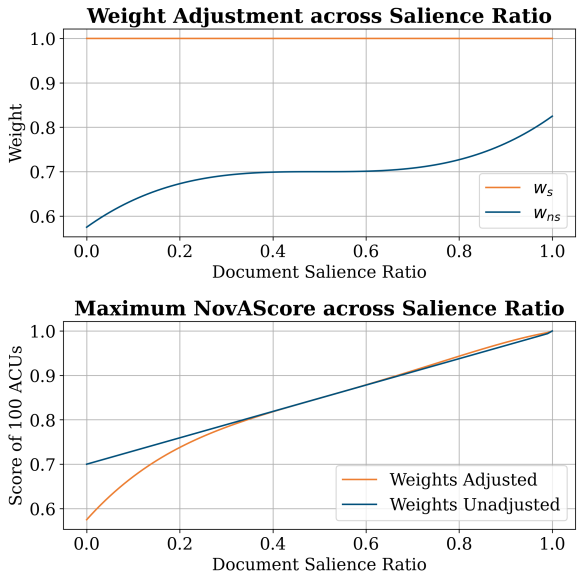

Figure 3: The top plot shows the weights for salient (w_s) and non-salient (w_{ns}) ACUs across different salience ratios with dynamic weight adjustment. The bottom plot compares the maximum NOVASCORE of 100 ACUs, with and without weight adjustment. Both plots utilize $\alpha = 1$, $\beta = 0.5$, and $\gamma = 0.7$.

To implement these principles, we set the weight for salient ACUs as $w_s = 1$ and dynamically adjust the weight of non-salient ACUs using a cubic function: $w_{ns} = min(w_s, \alpha(p_s - \beta)^3 + \gamma)$, where p_s is the salience ratio of the document. The hyperparameters α, β, and γ shape the cubic curve, with α controlling steepness and β and γ adjusting the midpoints on the x and y axes. These hyperparameters determine the devaluation of non-salient ACUs and the adjustment for extreme salience ratios, which can vary depending on the datasets and applications. Further details are in Appendix A.2.1. Figure 3 shows the impact of the weight adjustment scheme: as the salience ratio shifts to very low or high, non-salient ACU weights adjust more

rapidly. This adjustment ensures that documents with low salience ratios have a lower maximum NOVASCORE, giving less value to documents with less salient information.

The NOVASCORE of a document is then:

$$\text{NOVASCORE} = \frac{\sum_{i=1}^{n} N_i \cdot (w_s \cdot S_i + w_{ns} \cdot (1 - S_i))}{n}$$

where N_i and S_i represent the binary novelty and salience of the i-th ACU, respectively, and n denotes the total number of ACUs.

4 Experiments

In this section, we evaluate the effectiveness of NOVASCORE by examining its correlation with human judgments of novelty.

4.1 How Well Does NOVASCORE Align with Human Judgments of Novelty?

We begin by analyzing how closely NOVASCORE aligns with human judgments of novelty on a broad scale, specifically by examining its correlation with human-annotated document-level novelty.

#	TAP-DLND 1.0	APWSJ
Novel	250	259
Non-Novel	250	241
Total	*500*	*500*

Table 1: Statistics of dataset used for experiments.

Datasets We utilize the following two datasets, which, to the best of our knowledge, are among the few publicly available in the news domain:

1. **TAP-DLND 1.0** (Ghosal et al., 2018b): This dataset contains 2,736 human-annotated novel documents and 2,704 non-novel documents, all clustered into specific categories. Each novel or non-novel document is annotated against three source documents.
2. **APWSJ** (Zhang et al., 2002): This dataset comprises 10,833 news articles from the Associated Press (AP) and Wall Street Journal (WSJ) corpora, covering 33 topics. The documents are chronologically ordered and annotated into three categories: *absolutely redundant*, *somewhat redundant*, and *novel*. Of these, 7,547 are novel, 2,267 are somewhat redundant, and 1,019 are absolutely redundant.

For both datasets, we sample 500 documents. In TAP-DLND 1.0, we randomly select 500 documents across clusters. For APWSJ, where documents are chronologically sorted and annotated for novelty relative to earlier ones, we sequentially select a balanced set of 500 non-novel and novel documents. Table 1 provides the dataset statistics used in our experiments.

Setup and Implementation We utilize GPT-4o across all modules, including ACU extraction, document summarization, salient ACU selection, and both NLI and QA-based novelty evaluation. Details on the prompts are provided in Appendix A.1.

The *ACUBank* is implemented using FAISS[2] for fast similarity search and efficient indexing. Each ACU is indexed by its sentence embedding from the pre-trained SentenceBERT[3]. In TAP-DLND 1.0, we create separate databases for each of the 223 clusters. For APWSJ, documents are processed chronologically into a unified database.

To retrieve relevant historical ACUs from the *ACUBank*, we select the top-5 ACUs with a cosine similarity of 0.6 or higher. The rationale for this threshold is detailed in Appendix A.2.2. If any meet this threshold, the ACU is considered non-novel when using the cosine similarity novelty evaluator. For NLI or QA novelty evaluators, these similar ACUs are concatenated to form the premise for NLI or the context for QA, further assessing the ACU's novelty.

We use different hyperparameters for each dataset in the dynamic weight adjustment to account for ACU salience when calculating the overall NOVASCORE. These parameters control the devaluation of non-salient ACUs and adjust for extreme salience ratios, varying by dataset. For TAP-DLND 1.0, we use $\alpha = 0$, $\beta = 0.5$, and $\gamma = 1$ (no adjustment). For APWSJ, we use $\alpha = 1$, $\beta = 0.5$, and $\gamma = 0.7$. The rationale for these choices is detailed in Appendix A.2.1.

Metrics We employ **Point-Biserial**, **Spearman**, and **Kendall** to evaluate the relationship between the NOVASCORE and human annotations of document-level novelty. Point-Biserial is a special case of Pearson correlation that compares continuous variables with binary variables. For the TAP-DLND 1.0 dataset, where human annotations are binary, we assign a label of 1 to *novel* and 0 to *non-novel* for calculating correlations. In contrast, the APWSJ dataset contains three classes: we assign *novel* a value of 1, *somewhat redundant* a value of 0.5, and *absolute redundant* a value of 0

[2] https://ai.meta.com/tools/faiss/

[3] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Dataset →	**TAP-DLND 1.0**			**APWSJ**		
Novelty Evaluator →	CosSim	NLI	QA	CosSim	NLI	QA
Correlation ↓						
Point-Biserial	$\underline{0.545}_{(4.9e-40)}$	$\mathbf{0.626}_{(9.2e-56)}$	$0.508_{(2.3e-33)}$	$\underline{0.447}_{(6.5e-26)}$	$\mathbf{0.476}_{(1.2e-29)}$	$0.422_{(2.2e-22)}$
Spearman	$\underline{0.555}_{(1.1e-41)}$	$\mathbf{0.622}_{(9.2e-55)}$	$0.497_{(8.1e-32)}$	$\underline{0.446}_{(9.0-26)}$	$\mathbf{0.482}_{(1.8e-30)}$	$0.439_{(2.9e-24)}$
Kendall	$\underline{0.460}_{(2.8e-35)}$	$\mathbf{0.510}_{(8.0e-44)}$	$0.409_{(5.4e-28)}$	$\underline{0.358}_{(2.6e-24)}$	$\mathbf{0.395}_{(4.8e-27)}$	$0.353_{(5.2e-23)}$

Table 2: The correlations (statistics$_{(p\text{-value})}$) between NOVASCORE and human annotations on the TAP-DLND 1.0 and APWSJ datasets, using different novelty evaluators and correlation metrics.

for the Spearman and Kendall correlations. For Point-Biserial correlation on APWSJ, we set both *absolute redundant* and *somewhat redundant* to 0.

Results As shown in Table 2, across both datasets and using all novelty evaluators, NOVASCORE demonstrates a moderate to strong correlation with human judgments of document-level novelty, as indicated by the correlation cutoffs detailed in Appendix A.3. These correlations are statistically significant, with p-values ranging from 10^{-22} to 10^{-56}, indicating the robustness of NOVASCORE in aligning with human perceptions of novelty.

Among the different evaluators, the NLI novelty evaluator consistently outperforms the others, showing a particularly strong correlation with human annotations. Notably, on the TAP-DLND 1.0 dataset, the NOVASCORE with the NLI novelty evaluator achieves a Point-Biserial of 0.626, a Spearman of 0.622, and a Kendall of 0.510, all signifying a strong alignment with human judgment.

4.2 Can NOVASCORE Capture Granular Insights at the ACU Level?

In addition to broad document-level novelty analysis, we also assess NOVASCORE's reliability on a granular scale by examining its alignment with human judgments of novelty at the ACU level.

Human Annotation Since existing public datasets only provide single categorical labels at the document level without fine-grained annotations, we curate and annotate a new dataset for this purpose. We manually select 32 news articles, clustered into 8 topics. Within each topic, the documents are sorted in chronological order, and we extract ACUs using GPT-4o, following the strategy described in Sections 3.1. Human annotators evaluate each ACU based on four labels: correctness (logical and factual consistency), redundancy (non-informativeness or intra-document non-novelty), novelty, and salience. The full annotation instructions and label schema are provided in Appendix B.1. Two annotators independently perform the entire annotation task. After completing the annotations, the annotators meet to discuss and resolve any conflicting annotations, ensuring consensus on the final labels. Further discussion on annotation quality is presented in Appendix B.2.

Novelty Evaluator Performance We compare the performance of each novelty evaluator against human annotations of ACU-level novelty. As shown in Table 3, all novelty evaluators achieve strong classification results, with the NLI-based evaluator leading with an accuracy of 0.94.

Novelty Evaluator →	CosSim	NLI	QA
Metric ↓			
Accuracy	0.83	**0.94**	$\underline{0.91}$
Macro F1	0.71	**0.84**	$\underline{0.80}$

Table 3: Novelty evaluator performance.

NOVASCORE vs Human Judgments We aggregate ACU-level scores to compute the document-level novelty score, resulting in the following NOVASCORE variants: **(1)** NOVASCORE$_{\text{human}}$, using human-annotated novelty and salience labels, and **(2)** NOVASCORE$_{\text{CosSim}}$, NOVASCORE$_{\text{NLI}}$, and NOVASCORE$_{\text{QA}}$, which are fully automated versions utilizing their respective novelty evaluators and GPT salience evaluator. For all variants, we apply weight adjustment parameters of $\alpha = 1$, $\beta = 0.5$, and $\gamma = 0.7$, as used for the APWSJ dataset. We also compute all variants without incorporating salience to better assess the performance of the novelty evaluators.

We then use Pearson correlation to evaluate the relationship between NOVASCORE$_{\text{human}}$ and the three automated variants, as all produce continuous scores with nearly linear distributions. As shown in Table 4, all automated variants demonstrate strong to very strong correlations, with NOVASCORE$_{\text{NLI}}$ achieving the highest Pearson correlation of 0.920 without salience and 0.835 with salience. The discrepancy between human-annotated and GPT-

selected salient ACUs, as discussed in Section 5.1, slightly degrades the correlation statistics when salience is incorporated, which is expected. Despite this, the fully automated NOVASCORE with salience included still shows a very strong correlation with human judgments, indicating that GPT-4o is a reasonable estimator of information salience. The full correlation results between the three automated variants and $\text{NOVASCORE}_{\text{human}}$ are detailed in Table 8 in Appendix A.4.

	$\text{NOVASCORE}_{\text{CosSim}}$	$\text{NOVASCORE}_{\text{NLI}}$	$\text{NOVASCORE}_{\text{QA}}$
Salience ↓			
w/o	$0.748_{(8.5e-07)}$	$\mathbf{0.920}_{(9.6e-14)}$	$\underline{0.843}_{(2.9e-09)}$
w/	$0.722_{(3.1e-06)}$	$\mathbf{0.835}_{(2.9e-09)}$	$\underline{0.779}_{(2.4e-07)}$

Table 4: The Pearson correlations (statistics$_{(p\text{-value})}$) between $\text{NOVASCORE}_{\text{human}}$ and NOVASCORE with different novelty evaluators.

These results underscore the effectiveness of NOVASCORE in capturing document-level novelty while also providing fine-grained interpretability at the atomic level, making it a reliable tool for assessing document-level novelty.

4.3 How Does Dynamic Weight Adjustment Enhance Novelty Evaluation?

	$\text{NOVASCORE}_{\text{NLI}}$	$\text{NOVASCORE}_{\text{NLI}}$ w/o WA
Correlation ↓		
Point-Biserial	$\mathbf{0.476}_{(1.2e-29)}$	$0.442_{(2.2e-25)}$
Spearman	$\mathbf{0.482}_{(1.8e-30)}$	$0.468_{(1.4e-28)}$
Kendall	$\mathbf{0.395}_{(4.8e-27)}$	$0.393_{(1.4e-25)}$

Table 5: The correlations (statistics$_{(p\text{-value})}$) between $\text{NOVASCORE}_{\text{NLI}}$ and human annotations on APWSJ with and without weight adjustment (WA).

As discussed in Section 4.1 and Appendix A.2.1, the dynamic weight adjustment scheme is designed to ensure that the overall NOVASCORE reflects both the novelty and importance of the information. The magnitude and rate of adjustment, controlled by the hyperparameters, vary depending on the specific dataset and its standards. In datasets like APWSJ, where document-level novelty is determined with nuanced considerations of redundancy and individual perception differences, the concept of information salience is implicitly included in the final novelty label. As shown in Table 5, incorporating weight adjustment on APWSJ consistently results in higher correlation values across all metrics. Conversely, on datasets like TAP-DLND 1.0, where novelty labels are strict binary cutoffs reflecting only whether there is sufficient new information, adding weight adjustment does not necessarily improve the correlation.

The strength of this weight adjustment scheme lies in its flexibility to emphasize both important and less critical information when evaluating a document's overall novelty, tailored to the needs of the specific application. This provides NOVASCORE with an additional dimension, enabling it to assess not only the level of novelty but also the worthiness of the information within a target document.

5 Discussion

5.1 GPT-4o Performance and Reliability

ACU Extraction As introduced in Section 4.2, we collect correctness and redundancy labeled on GPT-generated ACUs during human annotation. Results reveal that none of the GPT-4o generated ACUs are labeled as incorrect by the annotators, and only 0.1% are considered redundant. These findings indicate that GPT-4o is highly reliable in generating high-quality ACUs.

Currently, no public datasets are designed for abstractive document-level ACU extraction. The closest are summarization datasets, which focus on key information and miss non-salient ACUs. Thus, our evaluation prioritizes precision over recall, leaving full ACU extraction and novelty recall for future work (see Limitations section).

Salience Evaluation Recent studies suggest that LLM-generated summaries are often on par with human-written ones (Zhang et al., 2024), supporting our confidence in GPT-4o's ability to evaluate salience. We compare GPT-selected salient ACUs with human-annotated salience labels as outlined in Section 4.2, and find that GPT achieves a macro F1 score of 0.6. This discrepancy may result from the different conditions: human annotators determine salience in real-time without summaries, making the task more challenging, while GPT-4o can reference the generated summary. Additionally, salience is inherently subjective, making it difficult to standardize. Despite these factors, GPT-4o's performance in salience evaluation is satisfactory.

5.2 Cost

We report the token usage of NOVASCORE during GPT-4o calls, as shown in Table 6. "ACU Extraction & Salience" refers to the one-pass call that handles both tasks, with the detailed prompt in Section A.1. The embedding cosine similarity evaluator

Dataset →	APWSJ	TAP-DLND 1.0
Module ↓		
ACU Extraction & Salience	1.7M	1.5M
NLI Novelty Evaluator	0.8M	0.9M
QA Novelty Evaluator	2.0M	1.7M

Table 6: Tokens Utilized in GPT-4o Calls for Each Module. "ACU Extraction & Salience" refers to the one-pass call that performs both ACU extraction and salience evaluation.

doesn't require GPT calls, making it cost-effective, especially for large-scale evaluations.

The QA novelty evaluator consumes about twice as many tokens as the NLI evaluator due to its two-step process: question generation followed by question answering. In addition, although QA-based methods are effective in other tasks like summarization evaluation (as discussed in Sections 2 and 3.2), they don't perform as well as NLI and sometimes even embedding cosine similarity on novelty detection. Therefore, **if budget is not a concern, we recommend using the NLI novelty evaluator for its strong performance**. Alternatively, embedding cosine similarity offers a good balance between cost and effectiveness.

5.3 Scalability

We examine the time required to search for similar ACUs across different *ACUBank* sizes. As shown in Figure 4, search time increases linearly with the size of the *ACUBank*. To improve scalability, **clustering documents or ACUs and creating separate databases within the *ACUBank* for each cluster would reduce search space and time**.

We do not report the average latency of GPT API calls, as various factors – such as usage time and network conditions – can affect this. However, we acknowledge that potential API lags could increase the framework's runtime. Replacing some modules with locally hosted smaller models, like fine-tuned open-source NLI models, could mitigate these delays and enhance efficiency.

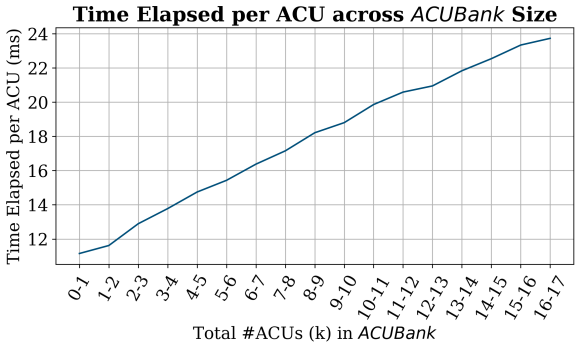

Figure 4: Search time for similar ACUs per ACU at varying *ACUBank* sizes with a single database.

5.4 Applications

Novelty detection in NLP has broad applications across various tasks, including plagiarism detection (Gipp et al., 2014), news event tracking (Ghosal et al., 2018b), scientific novelty detection (Gupta et al., 2024; Kelty et al., 2023), and misinformation detection (Qin et al., 2016; Ai et al., 2021). Furthermore, recent work (Li et al., 2024) introduces novelty as a key metric for benchmark design, revealing hidden performance patterns and unexpected model behaviors, which enhances evaluations and drives the creation of higher-quality benchmarks that advance model development. Despite its wide-ranging utility, this area has not received sufficient attention. Our work aims to address this gap and push forward the research on novelty detection.

6 Conclusions and Future Work

In this work, we introduce **NOVASCORE**, an automated metric for evaluating document-level novelty. NOVASCORE considers novelty and salience at the atomic level, providing high interpretability and detailed analysis. By incorporating information salience and a dynamic weight adjustment scheme, NOVASCORE offers enhanced flexibility and an additional dimension, allowing it to assess not only the level of novelty but also the worthiness of the information when evaluating the overall novelty of a document. Our experiments on both public datasets and an internal human-annotated dataset demonstrate that NOVASCORE strongly correlates with human judgments of novelty, validating its effectiveness and reliability.

Looking ahead, we aim to explore ways to improve the cost and scalability of NOVASCORE by integrating open-source LLMs and smaller models to replace GPT-4o in each module. This would reduce dependency on proprietary systems and enhance the accessibility of our framework.

Additionally, as outlined in Section 5.4, NOVASCORE serves as a foundation for various tasks and applications. We encourage further research to expand its use across more fields, and believe its potential in novelty detection and model evaluation will have a strong impact on the research community.

Limitations

Internal Human Annotated Data Restriction One limitation of our study is that the human-annotated data discussed in Section 4.2 is internal and proprietary, which means we cannot provide additional information about the specific content or characteristics of this data, nor can we release it for public use. However, we do provide complete annotation instructions and schema in Appendix B.1. Looking ahead, we plan to construct a publicly available human-annotated dataset to address this limitation and support future research in this area.

ACU Extraction Recall Evaluation Another limitation of our current approach is the challenge of evaluating the completeness of extracted ACUs in terms of covering the entire content of articles. Currently, there are no public datasets specifically designed for abstractive document-level ACU extraction. The most relevant datasets available are those used for summarization, where documents are paired with human-written summaries. However, these datasets are not ideal for evaluating non-salient ACUs, as summaries typically focus only on the most important information. Similar to Min et al. (2023), we rely on machine-generated atomic information as part of our pipeline. Consequently, our evaluation emphasizes precision rather than recall. We acknowledge this limitation and plan to address it in future work, where we aim to conduct a more comprehensive assessment of ACU extraction and novelty recall.

GPT-4o Reliance Another constraint of our work is its reliance on GPT-4o for evaluation, which presents two main challenges: **(1)** the non-deterministic nature of LLMs complicates reproducibility, as the same conditions may yield different results due to inherent variability, and **(2)** LLMs can be financially and computationally expensive, posing scalability issues.

We acknowledge these concerns but note that LLM-based evaluation methods are increasingly adopted in various applications (Zhang et al., 2024; Min et al., 2023) because their strong capabilities often outweigh these limitations. To address reproducibility, we conduct experiments under controlled conditions and report average results across multiple runs to mitigate the impact of variability. Furthermore, as highlighted in Sections 5.3 and 6, we are exploring the use of smaller, deterministic, open-source LLMs that offer finer control over sampling parameters, such as temperature and decoding, to ensure more consistent outputs.

Our choice of GPT-4o is motivated by its demonstrated performance across a variety of NLP tasks, including summarization, question answering, and natural language inference. While the higher cost of GPT-4o is a limitation, it provides a reliable benchmark for evaluating NOVASCORE's effectiveness and establishes a strong foundation for future refinements. Moving forward, we plan to leverage smaller, fine-tuned, open-source models to enhance scalability and cost efficiency. Recent studies suggest that such task-specific models can match or even surpass LLMs like GPT-4 in certain domains, bolstering our confidence that future evaluations using optimized models will maintain NOVASCORE's effectiveness while improving accessibility and affordability.

References

Lin Ai, Run Chen, Ziwei Gong, Julia Guo, Shayan Hooshmand, Zixiaofan Yang, and Julia Hirschberg. 2021. *Exploring New Methods for Identifying False Information and the Intent Behind It on Social Media: COVID-19 Tweets*. ICWSM, virtual.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2011. The seventh pascal recognizing textual entailment challenge. In *TAC*.

Thorsten Brants, Francine Chen, and Ayman Farahat. 2003. A system for new event detection. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 330–337.

Charles LA Clarke, Nick Craswell, Ian Soboroff, et al. 2004. Overview of the trec 2004 terabyte track. In *TREC*, volume 4, page 74.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Ido Dagan, Dan Roth, Fabio Zanzotto, and Mark Sammons. 2022. *Recognizing textual entailment: Models and applications*. Springer Nature.

Tirthankar Dasgupta and Lipika Dey. 2016. Automatic scoring for innovativeness of textual ideas. In *Workshops at the thirtieth AAAI conference on artificial intelligence*.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Tirthankar Ghosal, Vignesh Edithal, Asif Ekbal, Pushpak Bhattacharyya, George Tsatsaronis, and Srinivasa Satya Sameer Kumar Chivukula. 2018a. Novelty goes deep. a deep neural solution to document level novelty detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2802–2813, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tirthankar Ghosal, Tanik Saikh, Tameesh Biswas, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Novelty detection: A perspective from natural language processing. *Computational Linguistics*, 48(1):77–117.

Tirthankar Ghosal, Amitra Salam, Swati Tiwari, Asif Ekbal, and Pushpak Bhattacharyya. 2018b. TAP-DLND 1.0 : A corpus for document level novelty detection. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Bela Gipp, Norman Meuschke, and Corinna Breitinger. 2014. Citation-based plagiarism detection: Practicability on a large-scale scientific corpus. *Journal of the Association for Information Science and Technology*, 65(8):1527–1540.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Komal Gupta, Ammaar Ahmad, Tirthankar Ghosal, and Asif Ekbal. 2024. Scind: a new triplet-based dataset for scientific novelty detection via knowledge graphs. *International Journal on Digital Libraries*, pages 1–21.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Sean Kelty, Raiyan Abdul Baten, Adiba Mahbub Proma, Ehsan Hoque, Johan Bollen, and Gourab Ghoshal. 2023. Don't follow the leader: Independent thinkers create scientific innovation. *arXiv preprint arXiv:2301.02396*.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Xiang Lisa Li, Evan Zheran Liu, Percy Liang, and Tatsunori Hashimoto. 2024. Autobencher: Creating salient, novel, difficult datasets for language models. *arXiv preprint arXiv:2407.08351*.

Nelson F. Liu, Tony Lee, Robin Jia, and Percy Liang. 2023a. Do question answering modeling improvements hold across benchmarks? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13186–13218, Toronto, Canada. Association for Computational Linguistics.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Yixin Liu, Alexander Fabbri, Yilun Zhao, Pengfei Liu, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023c. Towards interpretable and efficient automatic reference-based summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16360–16368, Singapore. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Binesh Nair. 2024. Predicting document novelty: an unsupervised learning approach. *Knowledge and Information Systems*, 66(3):1709–1728.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. 2024. Efficient benchmarking (of language models). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2519–2536, Mexico City, Mexico. Association for Computational Linguistics.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*.

Yumeng Qin, Dominik Wurzer, Victor Lavrenko, and Cunchen Tang. 2016. Spotting rumors via novelty detection. *arXiv preprint arXiv:1611.06322*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Barry Schiffman and Kathleen McKeown. 2005. Context and learning in novelty detection. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 716–723, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Patrick Schober, Christa Boer, and Lothar A Schwarte. 2018. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768.

Barry Schwartz. 2022. Google Says 60% Of The Internet Is Duplicate. `https://www.seroundtable.com/google-60-percent-of-the-internet-is-duplicate-34469.html`. Accessed 09-08-2024.

Ian Soboroff and Donna Harman. 2005. Novelty detection: The TREC experience. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 105–112, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Ian Soboroff, Donna Harman, et al. 2003. Overview of the trec 2003 novelty track. In *TREC*, pages 38–53.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Charles L Wayne. 1997. Topic detection and tracking (tdt). In *Workshop held at the University of Maryland on*, volume 27, page 28. Citeseer.

Yiming Yang, Jian Zhang, Jaime Carbonell, and Chun Jin. 2002. Topic-conditioned novelty detection. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 688–693.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Yi Zhang, Jamie Callan, and Thomas Minka. 2002. Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 81–88.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

A Experiment Details

A.1 GPT Prompt Templates

We provide the detailed prompt templates we use in the GPT calls in this section.

A.1.1 ACU Extraction and Salient ACU Selection Prompt

INSTRUCTION:
1. First, extract the list of all atomic content units (ACUs) from a given document. An ACU is an elementary information group that conveys a single message without further division. When identifying any named entity, temporal entity, location entity, or attribute, avoid using indirect references. Instead, specify the actual entity, attribute, or noun directly. For example, replace 'this company' with the actual name of the company, 'this location' with the actual location name, 'it' with the actual subject being referred, etc.
2. Then, summarize the given document.
3. Finally, using the summary, identify the most salient ACUs from the full list of ACUs. The salient ACUs should be those explicitly mentioned in the summary.
Output the response in JSON format:
{"all_acus": "array of ACU strings", "summary": "document summary", "salient_acus": "array of salient ACU strings"}

Example 1:
###Document: {example document}
###Output: {example output}

###Document: {input document}
###Output:

A.1.2 NLI Novelty Evaluator Prompt

INSTRUCTION: For each given premise-hypothesis pair, perform Natural Language Inference (NLI) to determine whether the hypothesis should be classified as 'entailment', 'contradiction', or 'neutral' based on the information provided in the premise.
Output the response in JSON format:
{"nli_results": "array of NLI results in the following format: [{{"id": int, "nli": "entailment"|"contradiction"|"neutral"}}]"}

==========
EXAMPLE:
###Premise 1: ABC Bank reported a significant drop in profits for the second quarter due to rising loan defaults. The bank's CEO mentioned the challenging economic environment as a key factor.
###Hypothesis 1: ABC Bank's profits declined in the second quarter because of increased loadn defaults.
###Premise 2: Global oil prices surged by 5% on Monday following geopolitical tensions in the Middle East. Analysts predict that the prices may continue to rise if the situation escalates.
###Hypothesis 2: Oil price decreased despite tensions in the Middle East.
###Premise 3: The ECB decided to maintain its current monetary policy stance, keeping interest rates unchanged.
###Hypothesis 3: The ECB's decision will impact the foreign exchange rates of the Euro.

###Output:
{"nli_results": [{"id": 1, "nli": "entailment"}, {"id": 2, "nli": "contradiction"}, {"id": 3, "nli": "neutral"}]}

==========
{premise (similar ACUs) hypothesis (target ACU) pairs}
###Output:

A.1.3 QA Novelty Evaluator Prompt

Question Generation Prompt

INSTRUCTION: For each given sentence, generate three distinct questions that correspond to the named-entities and noun phrases found in this sentence, and use the sentence as the answer.
Output the response in JSON format:
{"questions_list": "list of question arrays in the format: [[question_str, ...], [question_str, ...], ...]"}

==========
EXAMPLE:
###Sentences:
1: The stock market experienced a sharp decline due to economic uncertainty.
2: Albert Einstein, a theoretical physicist, developed the theory of relativity.

###Output:
{"questions_list": [["What sector faced a significant downturn because of economic uncertainty?", "Why did the stock market show a sudden decrease

recently?", "What caused the sharp decline in the financial markets?"], ["Who is credited with developing the theory of relativity?", "What field was Albert Einstein associated with?", "What significant scientific theory did Albert Einstein develop?"]]}

==========
###Sentences:
{target ACUs}
###Output:

Question Answering Prompt

INSTRUCTION: For each context-questions pairs, follow these steps:
1. Given the context, answer the following questions.
2. Consolidate all responses into a single concise sentence.

==========
EXAMPLE:
Context 1: The stock market experienced a sharp decline due to economic uncertainty.
Q1: What sector faced a significant downturn because of economic uncertainty?
Q2: Why did the stock market show a sudden decrease recently?
Q3: What caused the sharp decline in the financial markets?
Context 2: Albert Einstein, a theoretical physicist, developed the theory of relativity.
Q1: Who is credited with developing the theory of relativity?
Q2: What field was Albert Einstein associated with?
Q3: What significant scientific theory did Albert Einstein develop?

###Output:
{"answers": ["The stock market experienced a sharp decline due to economic uncertainty.", " Albert Einstein, a theoretical physicist, developed the theory of relativity."]}

==========
{context (similar ACUs) questions (generated questions) list}
###Output:

A.2 Hyper-Parameter Selection

We describe the rationale of the hyperparameter selection in this section.

A.2.1 Dynamic Salient Weight Adjustment

As introduced in Section 3.4, we adjust the weight of non-salient ACUs using a cubic function: $w_{ns} = \textit{min}(w_s, \alpha(p_s - \beta)^3 + \gamma)$, where p_s represents the salience ratio of the document. This adjustment is designed to ensure that the overall NOVASCORE accurately reflects both the novelty and importance of the information within the document.

The parameter $\boldsymbol{\alpha}$ controls the steepness of the cubic function, determining how sensitive the weight adjustment is to the salience ratio. A higher α results in a more pronounced adjustment, causing the weight of non-salient ACUs to decrease or increase more rapidly in response to very low or very high salience ratios. This sensitivity allows us to fine-tune how much emphasis is placed on non-salient ACUs depending on the distribution of salient information within the document.

The parameter $\boldsymbol{\gamma}$ adjusts the midpoint on the y-axis, which corresponds to the general level of devaluation for non-salient ACUs, referred to as the "mean non-salience devaluation." For example, setting $\gamma = 0.7$ implies that, on average, non-salient ACUs are considered 70% as important as salient ACUs, with further adjustments based on the document's salience ratio, as controlled by α.

The parameter $\boldsymbol{\beta}$ shifts the midpoint on the x-axis, determining the salience ratio at which the "mean non-salience devaluation" is applied. For instance, if $\beta = 0.5$ and $\gamma = 0.7$, then in documents where the salience ratio is less than 0.5, non-salient ACUs are assigned a lower weight than the mean devaluation of 0.7, with the rate of adjustment dictated by α. Conversely, in documents with a salience ratio greater than 0.5, non-salient ACUs receive a higher weight than the mean devaluation, again with the rate of adjustment controlled by α.

The choice of α, β, and γ depends on the specific dataset and application requirements. For the TAP-DLND 1.0 and APWSJ datasets used in our experiments, we performed a grid search with $\alpha \in [0, 2]$, $\beta \in [0, 0.8]$, and $\gamma \in [0.5, 1]$. The optimal hyperparameters for TAP-DLND 1.0 are found to be $\alpha = 0$, $\beta = 0.5$, and $\gamma = 1$, indicating no weight adjustment was necessary. For APWSJ, the optimal values are $\alpha = 1$, $\beta = 0.5$, and $\gamma = 0.7$. This discrepancy arises from the different standards and annotation approaches used in the two datasets.

Statistics →	Pearson	Spearman	Kendall
Strength ↓			
Negligible	0.00	0.00	0.00
Weak	0.10	0.10	0.06
Moderate	0.40	0.38	0.26
Strong	0.70	0.68	0.49
Very Strong	0.90	0.89	0.71

Table 7: Cutoff values for the correlation statistics.

The advantage of this weight adjustment scheme lies in its flexibility to control and incorporate both important and less important information when evaluating the overall novelty of a document. This provides NOVASCORE with an additional dimension, allowing it to assess not only the level of novelty but also the worthiness of the information within a target document.

A.2.2 Similarity Thresholds

We choose a threshold of 0.85 for embedding cosine similarity to determine whether two ACUs are almost identical because a higher threshold ensures that the two units are very close in semantic content. At this level, the embeddings are nearly overlapping, indicating that the ACUs convey virtually the same information with minimal variation. Conversely, a lower threshold of 0.6 is used to decide whether two ACUs are similar but not necessarily identical. This threshold allows for some semantic variation while still capturing a significant level of similarity, making it suitable for identifying ACUs that share related content or themes without being exact duplicates. These thresholds are selected based on empirical results, which demonstrate that they provide the best performance in distinguishing between near-duplicates and related content, thereby enabling a more nuanced analysis of a document's novelty and relevance.

A.3 Correlation Statistics Interpretation

Table 7 details the cutoff values for the rank-based correlation statistics, which are based on the recommendations for the Pearson correlation by Schober et al. (2018). Note that Point-Biserial statistics is a special case of Pearson correlation.

A.4 Full NOVASCORE Correlation Results on Internal Data

Table 8 details the full results of the correlations between fully automated NOVASCORE and NOVASCORE$_{\text{human}}$ on our annotated data, using different novelty evaluators.

B Human Annotation

We provide the details of our human annotation process in this section.

B.1 Annotation Instruction and Label Schema

Following is the comprehensive annotation instruction and label schema we provide to the annotators.

Instruction: Articles are clustered and sorted by date within each cluster. Annotate the articles cluster by cluster, completing one cluster before moving on to the next. When annotating, reach each article in sequential order within its cluster. Memorize all information from the articles as you read. This is necessary for accurately judging the novelty of each ACU in subsequent articles within the same cluster. Novelty is only considered within the same cluster, not across different clusters.

First, read the news article carefully to understand the content and context of the entire article. Then label each ACU by the following steps.

Step 1: Assessing Correctness and Redundancy
Evaluate each ACU within the context of the article to determine the correctness. Determine the redundancy of the ACU by comparing it with the previous ACUs within the same article.

Label Schema
Correctness
correct: The ACU is accurate and logically consistent within the context of the article.
incorrect: The ACU contains incorrect information, errors, illogical, or LLM hallucinations.

Redundancy
redundant: The ACU

(a) is a direct repeat or rephrase of a previous ACU within the current article.

(b) does not convey any meaningful information. This is usually the case where the ACU describes the metadata of the article. For instance, an ACU such as "The article is written by xxx" or "The publish date of the article is xxx" should be marked as redundant.

not-redundant: The ACU provides new unique and meaningful information within the current article. If an ACU is partially new, it is also considered not-redundant.

Dataset →	w/ Salience			w/o Salience		
Score →	NOVASCORE$_{\text{CosSim}}$	NOVASCORE$_{\text{NLI}}$	NOVASCORE$_{\text{QA}}$	NOVASCORE$_{\text{CosSim}}$	NOVASCORE$_{\text{NLI}}$	NOVASCORE$_{\text{QA}}$
Correlation ↓						
Pearson	$0.722_{(3.1e-06)}$	$\mathbf{0.835}_{(2.9e-09)}$	$\underline{0.779}_{(2.4e-07)}$	$0.748_{(8.6e-07)}$	$\mathbf{0.920}_{(9.6e-14)}$	$\underline{0.843}_{(2.7e-09)}$
Spearman	$\mathbf{0.758}_{(5.2e-07)}$	$\underline{0.567}_{(7.3e-04)}$	$0.562_{(1.0e-04)}$	$\mathbf{0.836}_{(2.6e-09)}$	$0.782_{(1.2e-07)}$	$\underline{0.798}_{(7.5e-08)}$
Kendall	$\mathbf{0.559}_{(2.4e-05)}$	$\underline{0.423}_{(1.4e-03)}$	$0.409_{(2.5e-03)}$	$\mathbf{0.690}_{(8.6e-07)}$	$\underline{0.687}_{(1.9e-06)}$	$0.643_{(1.2e-05)}$

Table 8: The correlations (statistics$_{(p\text{-value})}$) between automated NOVASCORE and NOVASCORE$_{\text{human}}$ computed from our internal annotated data, using different novelty evaluators.

Step 2: Assessing Novelty and Salience (Only for Correct and Not Redundant ACUs)

Evaluate the novelty of each ACU by comparing it with the previous articles in the same clusters to check if all information in the ACU is already known. For each ACU, you will be shown the top 5 similar ACUs from previous articles. For the first article within the cluster, no similar ACUs will be shown as we assume there are no older articles to compare with. Therefore, all correct and not redundant ACUs in the first article should be considered novel. Use similar ACUs only as a reference.

Situation 1: Information not in the top 5 similar ACUs does not necessarily mean that it is not mentioned in previous articles. Try your best to memorize what you've read and always go back to the original article to verify if you recall something you've read but is not in the top 5 similar ACUs.

Situation 2: If an article is different in topic/domain than the previous articles, the top 5 similar ACUs might not be useful at all. Please always refer back to the original articles to check for detailed information. Assess whether the information in the ACU is crucial for understanding the main points of the article to determine the salience of the ACU.

Label Schema

Novelty

novel: The ACU introduces some new information that is not present in previous article. If an ACU is partially new, it is also considered novel.

not-novel: The ACU does not introduce any new information in the sense that all information mentioned in this ACU has been mentioned in older articles within the same cluster. Only consider inter-article novelty, not intra-article novelty – an ACU should only be annotated as "not-novel" if all information has been mentioned in previous articles within the same cluster. If an ACU introduces the exact same information as an earlier ACU within the same article, it should be labeled as "redundant".

Salience

salient: The ACU contains the essential information that you would include in a summary of the article – label the ACU as salient if you think it is an essential information to convey the main point of the article.

non-salient: If the ACU does not contain essential information for the summary.

B.2 Annotation Quality

Metric →	Precision	Recall	F1-Score	Support
Class ↓				
Non-Novel	0.00	0.00	0.00	0
Novel	1.00	0.99	1.00	222
Accuracy			0.99	222
Weighted Avg	1.00	0.99	1.00	222

Table 9: The classification report of human annotation on **expected novel** ACUs.

Metric →	Precision	Recall	F1-Score	Support
Class ↓				
Non-Novel	1.00	0.82	0.90	22
Novel	0.00	0.00	0.00	0
Accuracy			0.82	22
Weighted Avg	1.00	0.82	0.90	22

Table 10: The classification report of human annotation on **expected non-novel** ACUs.

We have two annotators independently perform the entire annotation task. After completing their annotations, they meet to discuss and resolve any conflicting labels, ensuring consensus on the final results. To further ensure the quality of the annotations, we discreetly create and insert three synthetic articles as quality control samples without informing the annotators. Two of these articles are complete paraphrases of previous articles within a cluster and are added to the end of the cluster; for these, all ACUs are expected to be non-novel. The

third article is manually written as a completely new piece, unrelated to any other articles in the cluster, where all ACUs are expected to be novel. Additionally, for the first article in each cluster, all ACUs are also expected to be novel. As shown in Tables 9 and 10, the human annotation achieves a weighted F1 score of 1.0 on the expected novel ACUs and 0.9 on the expected non-novel ACUs, indicating the high quality of the annotation process.