Mondrian: A Framework for ABT-Logic Abstract (Re)Structuring

Elizabeth Orwig **Shinwoo Park** **Hyundong Jin** **Yo-Sub Han**⋆
Yonsei University, Seoul, Republic of Korea
orwig@yonsei.ac.kr, pshkhh@yonsei.ac.kr, tuzi04@yonsei.ac.kr, emmous@yonsei.ac.kr

Abstract

The well-known rhetorical framework, ABT (And, But, Therefore), mirrors natural human cognition in structuring an argument's logical progression - apropos to academic communication. However, distilling the complexities of research into clear and concise prose requires careful sequencing of ideas and formulating clear connections between them. This presents a quiet inequitability for contributions from authors who struggle with English proficiency or academic writing conventions. We see this as impetus to introduce: **Mondrian**, a framework that identifies the key components of an abstract and reorients itself to properly reflect the ABT logical progression. The framework is composed of a deconstruction stage, reconstruction stage, and rephrasing. We introduce a novel metric for evaluating deviation from ABT structure, named *EB-DTW*, which accounts for both ordinality and a non-uniform distribution of importance in a sequence. Our overall approach aims to improve the comprehensibility of academic writing, particularly for non-native English speakers, along with a complementary metric. The effectiveness of Mondrian is tested with automatic metrics and extensive human evaluation, and demonstrated through impressive quantitative and qualitative results, with organization and overall coherence of an abstract improving by an average of 27.71% and 24.71%.

1 Introduction

The ABT structure is a well-established rhetorical framework often used to organize arguments in a clear, logical progression - where an initial scenario ("And") is presented, followed by a challenge or contradiction ("But"), and ultimately resolved through a conclusion ("Therefore") (Olson, 2020). The ABT structure is particularly effective in academic writing, where clarity and logical coherence

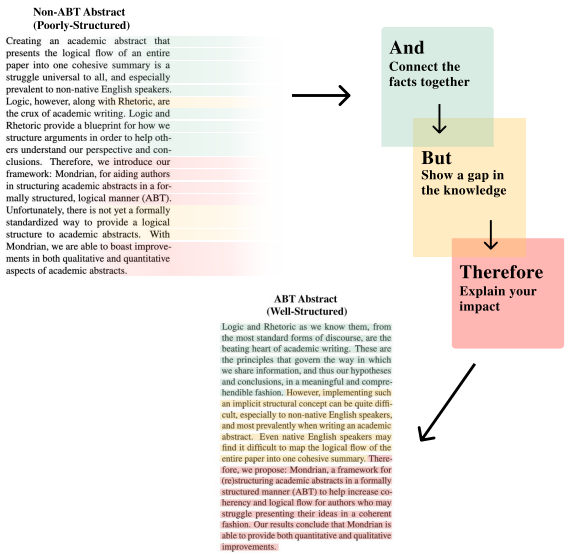

Figure 1: Reordering an abstract into ABT format to increase logical cohesion.

are crucial to reader comprehension (Swales, 2014; Thompson, 2001). By guiding readers through the complexities of research in a structured manner, ABT enhances the impact of academic writing (Moffett, 2019). Our research explores the application of this structure to academic abstracts, aiming to improve their logical flow and optimize reader comprehension.

In addition to this ABT task, we introduce our proposed Mondrian framework, inspired by the abstract art of Piet Mondrian, known for his emphasis on structure while employing the three primary colors. This framework is applied to the organization of academic abstracts, to ensure that the three key components (ABT) of the author's argument are arranged logically. The Mondrian framework guides the segmentation and reordering of text based on the ABT structure, helping to maintain proportionality and coherence throughout the abstract. By leveraging the principles of the Mondrian framework, we aim to create a paradigm to help aca-

⋆ Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 33675–33690

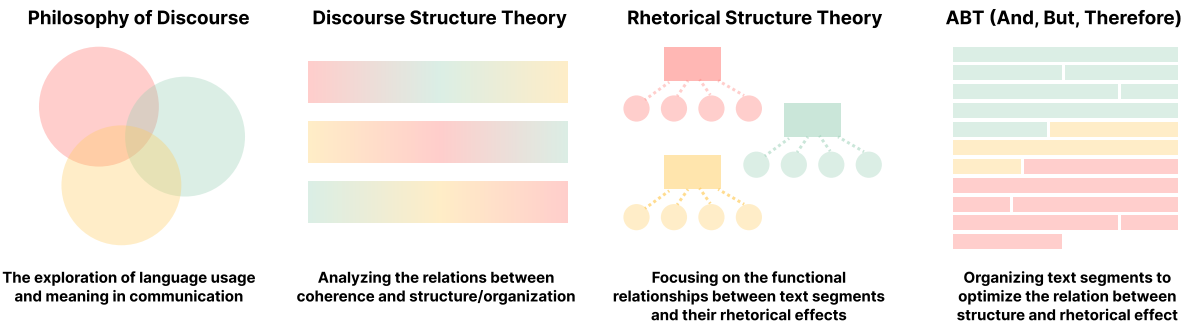

Figure 2: The ABT (And, But, Therefore) framework employed in Mondrian is rooted in the philosophy of discourse.

demic authors create structured, balanced text that not only adheres to logical patterns but also clearly, and justly, explains the novelty and impact of a research work - something that can be more difficult for non-native English speakers.

To quantitatively evaluate how well a text adheres to the ABT structure, we employ a novel metric known as Extrema-Bounded Dynamic Time Warping (EB-DTW). Traditional alignment metrics, such as Dynamic Time Warping (DTW) and Levenshtein distance, are limited in their ability to account for adherence to a specific structure or unequal distribution of importance in a sequence. EB-DTW addresses this limitation by introducing weights that reflect how much an abstract deviates from the ABT pattern, with penalties assigned to deviations in the key segments associated with A B and T. The result is a metric that provides a more nuanced evaluation of the abstract's logical flow and structural coherence.

By introducing the ABT task, the Mondrian framework, and the EB-DTW metric, we aim to contribute to the ongoing research in the domain of NLP & academia. This approach not only supports the development of more effective communication strategies in academic writing but also provides a robust framework for future research in discourse analysis and text structuring. . Our results demonstrate that our Mondrian framework is able to increase overall quality and comprehension of academic abstracts, proven through both quantitative and qualitative metrics, and especially for abstracts that our human evaluators unanimously agreed were most likely written by non-native English speakers.

2 Background & Preliminaries

Examining the argumentative structure behind persuasive speaking (rhetoric), logical argumentation (dialectic), and rational inquiry (reason) is a crucial aspect of many NLP tasks, as it allows us to understand the "why" coinciding with the "what" that is being communicated (Barnes, 1984; Hobbs et al., 1985; Degand and Simon, 2009). However, we must start from the origins of these three aspects - philosophy of discourse - to understand the coinciding impact in computational linguistics.

2.1 The Philosophy of Discourse

The Philosophy of Discourse refers to the foundational principles governing how meaning is constructed, communicated, and interpreted in a language (Webber and Joshi, 2012; Hobbs et al., 1985). In its more fine-grained forms, it analyzes how coherence and cohesion are maintained in discourse (the logical connections between different parts of a text or conversation that make it comprehensible) as well as the role of rhetorical strategies (metaphors, analogies, and narrative structures) in shaping the effectiveness and persuasiveness of communication (Degand and Simon, 2009; Macagno, 2016; Hyland, 2000, 2005).

In Computational Linguistics Pattern matching involves identifying and aligning text structures according to a specific sequence - in the context of our research, we refer to pattern matching as the alignment of segments in an academic abstract with the ABT structure. This approach is grounded in the idea that certain rhetorical patterns, like ABT, are inherently more effective in organizing and communicating complex ideas. We hypothesize that adherence to ABT structure corresponds to enhanced clarity and persuasiveness in academic writing, as a culmination of the philosophy of dis-

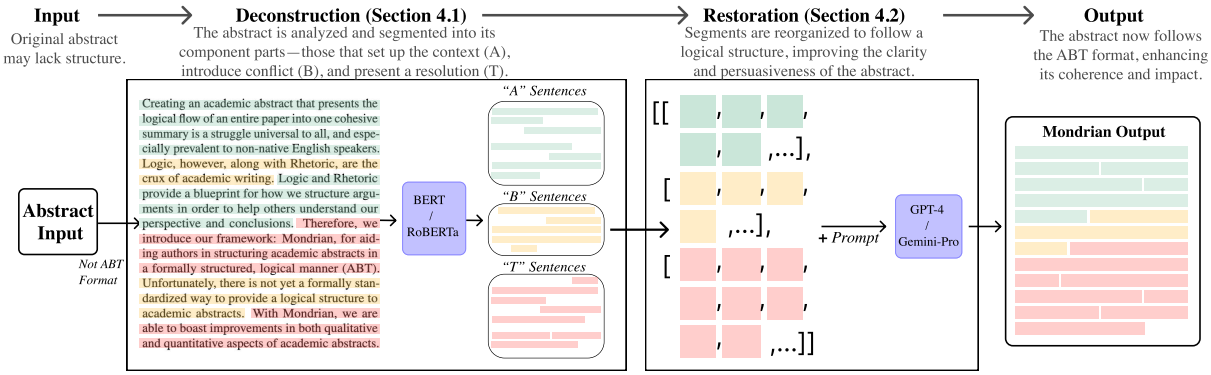

Figure 3: Mondrian Framework. An abstract is taken as input to the framework, the classifier determines if the abstract is in ABT format, and if not, it separates the sentences accordingly. Those groups of sentences are then fed into an LLM with a prompt guiding it to modify the sentences so they flow coherently in the ABT structure.

course sub-ideologies: Rhetorical Structure Theory (RST) and Discourse Structure Theory (DST), as seen in Figure 2.

Both theories, while similar, concentrate on different aspects. DST focuses on the relations between discourse segments - how the organization of elements, such as topic shifts, and the use of rhetorical devices impact overall coherence and cohesion (Kuppevelt, 1995; Althusser et al., 2006; Torfing, 2005). RST further refines the analysis of DST by focusing on the specific rhetorical relationships between different parts of a given text, then categorizes these relationships (e.g., cause-effect, contrast, and elaboration) and examines how they contribute to the coherence of the text(Jasinskaja and Karagjosova, 2020; Degand and Simon, 2009; Hobbs et al., 1985; Taboada, 2006; Mann and Thompson, 1988).

ABT & Similar Structures The ABT structure draws upon the principles of the philosophy of discourse - logical flow and concise arguments - to create a framework for structuring academic communication (Abdulmajid, 2021; Belcher, 2019; Kirkman, 2012; Hyland, 2019). The ABT structure can be seen as a practical application of DST and RST principles, as it uses clear rhetorical strategies to guide the audience through a coherent narrative (Abdulmajid, 2021). Also, ABT's clear delineation of premise, conflict, and resolution inherently weights the structure, so algorithms like our proposed EB-DTW can more precisely quantify the importance of each segment (Day and Gastel, 2024; Swales, 2014; Swales and Feak, 2009; Birkenstein and Graff, 2018).

3 Related Works

Paul Thompson (2001) highlights how structuring content in a systematic fashion can enhance the readability of research papers. He argues that a structured approach, by combining syntax and semantics, provides a clearer and more intuitive form with which to convey dense academic arguments. Ken Hyland (2019), in his work on second language (L2) writing, advocates for the use of structured text templates as pedagogical tools. The templates are designed to help L2 writers organize their ideas and arguments in a manner that adheres to academic conventions, ensuring that the they maintain clarity and logical flow.

Similarly, Graff and Birkenstein (2009) discuss the role of structured rhetorical templates in academic writing. They argue that such templates focus writers' attention on the rhetorical forms that structure their content, thereby enhancing their awareness of the rhetorical patterns in use. While there is some debate about the potential drawbacks of this "Mad Libs" (Nelson, 2011) style of writing, Graff and Birkenstein assert that these templates are crucial for organizing arguments in a clear and effective manner, especially in academic contexts.

4 Methodology

The Mondrian Framework is composed of two distinct stages: deconstruction and restoration. In the deconstruction stage, we intend to uncover the key points the author makes in an effort to restore them in the next stage as anchors for the ABT argument. In the restoration stage, we leverage these anchors to reorganize the rest of the abstract into ABT format. To mitigate the grammatical disjoint

associated with sentence re-ordering, we use an LLM to connect the ABT segments, while preserving as much of the original abstract as possible. The framework is visualized in Figure 3.

4.1 Deconstruction

In the first stage of Mondrian, if the original abstract is not already in ABT format, then it is split into its sentences. If the original abstract is already in ABT form, then the process is ended - there is no need for Mondrian. We train a sentence classifier using abstracts from academic papers, where each sentence has been annotated as one of 'And', 'But', or 'Therefore' by native English speakers. The trained sentence classifier[1], when given an abstract as input, classifies each sentence in the abstract as either 'And', 'But', or 'Therefore'. That is, whether the sentence explains background information, explains a gap in the knowledge, or displays the author's impact, respectively. Each category maintains a sorted list of its corresponding sentences (so as to maintain semblance of structure). These are then used as the foundational components for ABT structure in the second stage.

4.2 Restoration

With the categorical segments and their sentences uncovered from the initial abstract, the second stage of the Mondrian framework focuses on reconstruction. As previously noted, identifying the ABT segments and pasting them together in the correct order is not sufficient for rebuilding an abstract - it could lead to varying levels of incoherency, given how different the initial abstract's structure is from the ABT format. We employ GPT-4 and Gemini-Pro to mitigate this issue and provide them with a prompt and the abstract segments. The prompt specifies that there are a certain number of following sections[2], and they should be pieced together in a coherent and logically sound way, while maintaining and preserving as much of the original text and inputted structure as possible.

4.3 Self-Training

The accuracy of the sentence classifier that categorizes each sentence in a paper abstract is critical for the Mondrian framework to work effectively. To improve the performance of the sentence classifier, we utilize self-training. We call the version of Mondrian that leverages self-training as Mondrian-ST. We scraped OpenReview for papers that were still under review/anonymous to create a dataset of about 1000 abstracts. We train the sentence classifier by performing self-training using these 1000 unlabeled abstracts. Specifically, we use the sentence classifier trained on labeled abstracts to classify the sentences in the 1,000 abstracts (pseudo-labeling). Then, we retrain the sentence classifier using these 1,000 abstracts.

5 ABT Task Metric

5.1 Dynamic Time Warping

DTW generally measures the similarity between two sequences that potentially vary in time or speed (Bellman and Kalaba, 1959). The general form for DTW is as follows:

$$DTW(org, ref) = \sqrt{\gamma(n, m)} \tag{1}$$

where n is the length of org and m is the length of ref. DTW allows for a flexible matching between elements of sequences, and does not require them to have equal length. It finds the optimal alignment between the two sequences by warping one in "time" to match the other, while minimizing the total distance between corresponding elements. The optimal warping path is the path with the lowest cumulative cost:

$$\begin{aligned}\gamma(i, j) = |org[i] - ref[j]| + \\ \min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\}.\end{aligned} \tag{2}$$

However, DTW as well as Levenshtein distance, are unable to distinguish between structurally in/significant segments (and deviations) in text - evidenced by the fact that both metrics produce a final cost wherein each node is weighted **equally**. As ABT is designed to guide logical flow, each segment is independently significant to the overall textual coherence. DTW may indicate that two sequences are *dissimilar* in terms of their overall alignment, but fails to capture the significance of each dissimilarity.[3]

Weighted Dynamic Time Warping (WDTW) is a concept introduced by Jeong et al. (2011) to address the equal weighting issue of traditional DTW. WDTW takes ordinality (sequential phases) into

[1] We employ BERT or RoBERTa as base models for.

[2] Three sections will be sent to the LLM if all ABT segments are present, however, if one of the segments is not present in the original abstract, then only the existing 2 will be sent

[3] DTW may assign the same cost to two abstracts with minor reordering of sentences, even if one of those abstracts significantly disrupts the logical flow of the argument.

account, however it was not designed to emphasize a certain structure, an inherent characteristic of the ABT task. As such, we propose: EB-DTW.

5.2 Proposed Metric: Extrema-Bounded Dynamic Time Warping (EB-DTW)

Our proposed metric, EB-DTW, introduces segment weights that reflect adherence to structural patterns, as seen in Algorithm 1. In the context of ABT, these weights penalize deviations based on how heavily they disrupt the rhetorical flow. We purposely chose to create a tangential metric of DTW due to its graphability for weight calculation.

Algorithm 1 Algorithm for EB-DTW.

```
function COMPUTE EB-DTW DISTANCE(abstract)
    org = abstract
    ref = [0, -1, 1]                          ▷ A, B, T
    EBDTW ← 0
    warping_path = DTW(org, ref)
    segment_weights =
FIND-SEGMENT-WEIGHT(org, ref, warping_path)
    for weight ∈ segment_weights do
        EBDTW ← EBDTW + weight
    end for
     return EBDTW
end function
```

We consider an abstract as a sequence of sentences, and each sentence is represented by a single 'A', 'B', or 'T' character depending on its classification - so representing an abstract as a sequence looks like: 'AABBATTTBBT'. The number of elements in each segment is trivial, as long as it matches the ordinality of 'ABT'. Therefore, we choose our reference sequence, ref, as simply: 'ABT'.

For graphing purposes, we use the character's index in the sequence as its "time", plotted along the X-axis, and our Y-axis is an arbitrary set of 3 sequential, numerical values implicitly corresponding to the ABT values. The value in the middle corresponds to 'A', the value on the bottom corresponds to 'B', and the value on top corresponds to 'T'. Therefore, when we graph only ref, we define the extrema for the desired structure, which we can measure deviation from after computing the warping path with the original abstract sequence, org. For sake of example, we will define A→0, B→-1, T→1.

The warping path wp in Algorithm 2 is an array of indices discerning which points in org are mapped to which ABT index in ref—that is, the optimal warping alignment. The $org[j]$ is mapped to the $wp[i]$-th element of ref. If there exists a segment wherein one of the elements does not match the reference point, this indicates that there is a discordant pair, and that it affects the cost matrix. For instance, 'AABBATTTBBT' divides into three segments: A: 'AA', B: 'BBA', T: 'TTTBBT'. There are discordant pairs in both the B and T segments. For each segment with discordant pairs, the points in that segment are interpolated using Cubic Spline, for a function that accounts for all points in the segment.

To find the weight of each segment we take the integral between the reference point as a y-intercept line ($y =$: 0, -1, or 1) and the Cubic Spline function for the corresponding segment, as seen in Algorithm 2. We then apply the weight to the corresponding segment, and re-combine to achieve the weighted final cost. By doing so we posthumously update the DTW cost by applying weight to the optimal warp path of each segment. For evaluation, we use this to measure how aligned the textual structure is to the desired structure. For optimization, we propose that one may find points with higher penalties and individually evaluate them to redistribute information to other sentences - further grouping similar information and minimizing incurred costs. However, we leave this optimization to future work.

Algorithm 2 Calculating Segment Weight.

```
function FIND-SEGMENT-WEIGHT(org, ref, wp)
    for j ∈ [1..|ref|] do
        seg ← {1 ≤ i ≤ |org| | wp[i] = j}
        if ∃i ∈ seg such that org[i] ≠ ref[j] then
            curve ← {(i, org[i]) | i ∈ seg}
            f_spline ← CUBIC-SPLINE(curve)
            x_0 ← min{x | (x, y) ∈ curve}
            x_1 ← max{x | (x, y) ∈ curve}
            seg_w[j] ← ∫_{x_0}^{x_1} |f_spline(x) − wp[j]| dx
        end if
    end for
    return seg_w
end function
```

6 Experimental Setup

6.1 Dataset

Our dataset consists of abstracts from papers on OpenReview that were under review/"Anonymous" at the time. We did this to simulate the LLM portion of our framework assisting in the re-phrasing of abstracts for *novel* ideas and papers. If the paper had already been published, the LLM could add information to the abstract that was not provided in the prompt.

Initially, we collected approximately 450 Anonymous submissions from the ICLR and TMLR venues. Each abstract was manually split into "And", "But", and "Therefore" sentences. We split the dataset into training, validation, and test sets in a ratio of 8:1:1. We train the sentence classifier using the training and validation sets. During this process, we apply synonym replacement-based data augmentation to the training set to increase the amount of training data.

6.2 Evaluation Settings & Employed Metrics

We evaluate on four settings: zero-shot, one-shot, Mondrian, and Mondrian-ST. For zero-shot, we send the original abstract to ChatGPT or Gemini-Pro, prompting them to re-organize it to follow ABT structure. For one-shot, we use the same protocol as in the zero-shot setting, but also provide an example to the LLM[4]. For Mondrian and Mondrian-ST, we first classify each sentence in the paper abstract as one of "And", "But", or "Therefore." Then, we reorder all the sentences to follow the ABT structure and use an LLM to naturally refine the reordered abstract. The prompts we used can be found in the Appendix D.

The quantitative metrics include both DTW and EB-DTW, for precision comparison. The qualitative metrics include human evaluation results for the original and outputted abstracts with regard to: content, organization, language, and overall (coherence). Further details are in in Appendix B.

7 Experimental Results & Analyses

An optimal DTW and EB-DTW value is 0, so as the Δ of either decreases, the returned results conform more to ABT structure. Our qualitative metrics are shown as the average percentage of improvement between the original and the output of the experimental setting, which for the zero- and one-shot settings is original abstract/LLM output and for the Mondrian and Mondrian-ST settings, is original abstract/Mondrian output.

7.1 Results Overview

Automatic Evaluation Results The general trend of the data from our zero-shot setting to our Mondrian-ST setting, reflected in Table 1, indicates that our proposed Mondrian framework is able to improve academic abstract quality in both quantitative and qualitative aspects.

An optimal DTW and EB-DTW score is 0, corresponding to perfect conformity to ABT pattern. Therefore, the trend of both quantitative metrics decreasing to larger negative values reflects the ability of Mondrian and Mondrian-ST to transform an abstract from non-ABT to ABT format, successfully. In both zero- and one-shot settings, GPT-4 is able to somewhat modify the abstract to conform with ABT while Gemini-Pro actively deviates the abstract further from ABT format. In turn, the qualitative results corresponding to the zero- and one-shot experiments indicate that the output, ultimately, suffers in terms of quality from a human evaluation aspect.

	ΔDTW	ΔEB-DTW
Zero-Shot		
GPT-4	-0.35	-0.30
Gemini-Pro	+0.15	+0.31
One-Shot		
GPT-4	-0.35	-0.12
Gemini-Pro	+0.55	+0.75
Mondrian		
BERT +		
GPT-4	-1.57	-1.79
Gemini-Pro	-1.35	-1.50
RoBERTa +		
GPT-4	-1.57	-2.07
Gemini-Pro	-1.52	-1.87
Mondrian-ST		
BERT +		
GPT-4	-1.62	-1.98
Gemini-Pro	-1.38	-1.65
RoBERTa +		
GPT-4	**-2.00**	**-2.18**
Gemini-Pro	-1.24	-1.81

Table 1: Quantitative results are expressed as the average change in accumulated cost between the original and the resulting abstract. An optimal EB/DTW score refers to a sequence that is perfectly aligned, and thus has the value of 0. The higher the EB/DTW, such as 3 or larger, the less aligned the sequence is to the ABT structure. Mondrian+ST, on average, lowers the EB/DTW score of the original abstract by approximately **2.00** - a significant decrease in score and increase in alignment.

7.2 Results Analyses

Abstract Preference. Human evaluators were asked to pick which abstract they preferred be-

[4]The example consists of an academic abstract not present in the dataset that does not reflect the ABT structure, we then re-organize it in a coherent and cohesive form so that it reflects the ABT structure.

	Qualitative (%)			
	Content	Org.	Lang.	Overall
Zero-Shot				
GPT-4	-12.1	-9.4	-7.9	-6.4
Gemini-Pro	-17.9	-12.0	-16.4	-19.0
One-Shot				
GPT-4	-9.4	+11.3	-9.6	-2.1
Gemini-Pro	-7.1	-9.4	-0.1	-6.0
Mondrian				
BERT +				
GPT-4	+10.7	+22.3	+9.9	+11.0
Gemini-Pro	+1.4	+4.3	+4.9	+3.7
RoBERTa +				
GPT-4	+8.9	+24.1	+18.7	+21.3
Gemini-Pro	+6.7	+11.3	+12.1	+11.6
Mondrian+ST				
BERT +				
GPT-4	**+16.1**	+26.7	**+27.9**	**+24.7**
Gemini-Pro	+11.1	+20.4	+15.4	+11.9
RoBERTa +				
GPT-4	+13.0	**+27.7**	+21.1	+22.0
Gemini-Pro	+9.7	+21.4	+12.7	+16.6

Table 2: Qualitative results are expressed as the average percentage of change between the original and the resulting abstract. If the percentage is negative, the output is deemed worse in these categories by human evaluators. If the percentage is positive, then the output is deemed better than the original by that percentage in the corresponding categories by the human evaluators.

tween the two (without knowledge of which was original and which was the output). As the experiments progressed from zero-shot to Mondrian-ST, there is a consistent trend of improved qualitative scores from human evaluators (visualization in Appendix C). There was a similar trend between the human evaluators and which abstract they preferred in each setting. In the zero- and one-shot settings, most evaluators preferred the original abstract to the LLM-generated abstract. In the Mondrian setting, the evaluators tended to prefer the Mondrian output, and the output from the Mondrian-ST was the overwhelming choice in that setting. We attribute this as evidence of correlation between increased logical structure and organization in academic abstracts and reader preference.

Human Evaluation Results Table 2 shows the human evaluation results. The qualitative metrics were measured by our human evaluators on a scale from 1-7 (7 indicating the best possible score) for both the input and output abstract for each setting. The percentage in the table reflects the average amount of increase or decrease for each qualitative facet of the respective experimental output in relation to its input. Therefore, a higher percentage indicates increased quality for the given qualitative measure. The zero- and one-shot settings generally trend downward in their qualitative aspects, however notably less so in their organizational aspect. For the specific case of One-Shot GPT-4, there is actually an increase in organization - to which we draw the conclusion that even adding ABT format at a one-shot level, without the Mondrian framework, still lends to its overall benefit.

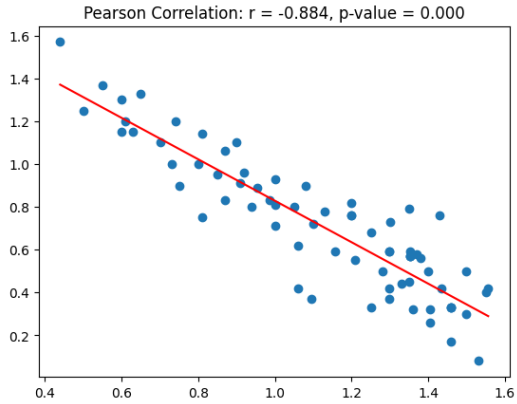

Figure 4: The precision of EB-DTW correlates with the entropy of the sequence. This indicates EB-DTW is highly effective in maintaining structure and detecting deviations in sequences where ordinality and structure are non-trivial.

Results from the Mondrian experimental settings prove beneficial in both quantitative and qualitative measures, effectively decreasing the DTW & EB-DTW while increasing the qualitative metrics. However, the largest improvement comes from the Mondrian-ST setting, boasting stable quantitative and qualitative improvements, as well as the best metrics across all experimental settings. We ascribe this as evidence reflecting the merit and effectiveness of the Mondrian framework.

Justification of EB-DTW. To display the precision and usefulness of our proposed metric (Figure 4), we randomly sample 70 abstracts from our evaluation set and calculate the entropy associated with their original abstract sequence and plot these values on the X-axis. The entropy is particularly meaningful in this case because a sequence that does *not* need to adhere to a particular structure, would have a higher entropy (randomness) than a sequence with a more predictable structure, such as one that minds ABT structure. The Y-axis contains the absolute value of the difference between the EB-DTW and the DTW (isolating the precision of EB-DTW).

The precision from EB-DTW and the entropy of a given sequence have a Pearson Correlation

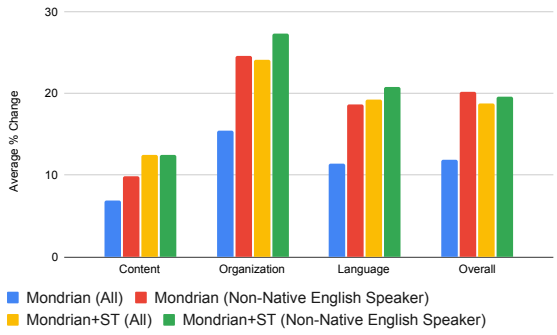

Figure 5: The average improvement for abstracts written by non-native English speakers outranks the average improvement of all abstracts.

Rank of -0.884. This indicates that the precision provided by EB-DTW over DTW correlates non-linearly with the entropy of the given sequence. So, EB-DTW's effectiveness rises with ordinality and structure of a sequence. However, for sequences with higher entropy that are not bounded to ordinality and structure, EB-DTW may not be the appropriate metric.

Non-Native English Abstracts. One of our primary motivations was to provide a framework for non-native English speakers or those who struggle with academic writing conventions. As such, we compared the overall data against the abstracts that the human evaluators unanimously agreed were likely written by non-native English speakers. Our results find that in both the Mondrian and Mondrian-ST settings, the improvements for abstracts written by non-native English speakers are greater than the overall improvements, as seen in Figure 5. The two categories with the greatest dichotomy in scores are organization and overall (coherence). Similarly, for this subset of abstracts, human evaluators preferred the output from Mondrian over the original abstract for 86.49% of them.

8 Generalizability & Future Work

8.1 Mondrian Framework

We believe cross-linguistic text alignment could benefit from the Mondrian framework. In multilingual settings, a direct one-to-one sentence alignment between languages might not always be possible due to differences in syntactic structures and contextual nuances. Instead of aligning the texts sentence-by-sentence, the Mondrian framework could be applied to align larger segments that represent logical units of meaning. It could also ensure that each part of the aligned text is given the appropriate weight and representation, as sometimes source and target translations differ in length or structure. By focusing on segment-based alignment and proportional representation, Mondrian could be generalized to cross-linguistic text alignment tasks to assure that the aligned texts retain their logical flow and meaning.

8.2 EB-DTW

An NLP task we believe might benefit from the EB-DTW metric is plagiarism detection. Traditionally, plagiarism tools rely on string matching. While this is useful for explicit plagiarism, more sophisticated forms of plagiarism often go undetected.

In cases of more subtle plagiarism, individuals might try and disguise copied content by rearranging the order of sentences, paragraphs, or entire sections - but still maintaining much of the original text. This method of reordering content can make it difficult for traditional plagiarism detection systems, because they often rely on linear or sequential text matching. In cases of partial plagiarism, individuals might copy portions of a source document, and try to blend the copied content with original work to evade detection. This could be beneficial in scenarios where authors want to use key arguments from a separate source while still appearing to produce original content. Because it doesn't rely on string matching, but rather analyzes the structure and flow of a text, it could potentially help detect plagiarism that involves rewording, re-ordering, or cherry-picking content.

9 Conclusion

We introduce the formalization of the ABT task, our Mondrian framework for ABT-logic abstract restructuring, and the EB-DTW metric - the latter two we believe to be easily generalizable. Based in the philosophy of discourse, our results prove that using ABT structure for abstracts improves their readability and coherence, as evidenced by significant positive results in qualitative metrics. Both DTW and EB-DTW prove to be sufficient metrics for the ABT task, however EB-DTW - as it has correlation with sequence entropy - provides a more precise and nuanced quantitatve evaluation. Furthermore, our Mondrian framework proved to be especially helpful for abstracts written by non-native English speakers, providing an equitable tool for academic communication.

Limitations

Some limitations that potentially hinder the optimality of the Mondrian framework, as well as a proper evaluation and analysis of it primarily come from our sources of data and human evaluation. All our abstracts were taken from OpenReview and in English. As such, they have bias toward language as well as domain. Also, since the entire dataset is composed of anonymous abstracts, even if the human evaluators unanimously agreed that an abstract was written by a non-native English speaker, there is possibility that it was, in fact, written by a native English speaker.

Acknowledgments

This research was supported by the National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) (RS-2025-00562134), and by the AI Graduate School Program (RS-2020-II201361).

References

Adib Abdulmajid. 2021. *Discourse: Theory and Practice*, pages 97–130. Springer International Publishing, Cham.

Louis Althusser et al. 2006. Ideology and ideological state apparatuses (notes towards an investigation). *The anthropology of the state: A reader*, 9(1):86–98.

Jonathan Barnes, editor. 1984. *Complete Works of Aristotle, Volume 1: The Revised Oxford Translation*. Princeton University Press.

Wendy Laura Belcher. 2019. *Writing your journal article in twelve weeks: A guide to academic publishing success*. University of Chicago Press.

Richard Bellman and Robert E. Kalaba. 1959. On adaptive control processes. *Ire Transactions on Automatic Control*, 4:1–9.

Cathy Birkenstein and Gerald Graff. 2018. *They say/I say: The moves that matter in academic writing*. WW Norton & Company.

Robert A Day and Barbara Gastel. 2024. *How to write and publish a scientific paper*. Cambridge University Press.

Liesbeth Degand and Anne Catherine Simon. 2009. On identifying basic discourse units in speech: theoretical and empirical issues. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (4).

G. Graff, C. Birkenstein, and R.K. Durst. 2009. *"They Say/I Say": The Moves that Matter in Academic Writing : with Readings*. W.W. Norton & Company.

Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, et al. 2023. Fabric: Automated scoring and feedback generation for essays. *arXiv preprint arXiv:2310.05191*.

Jerry R Hobbs et al. 1985. *On the coherence and structure of discourse*, volume 208. CSLI Stanford, CA.

K Hyland. 2000. Disciplinary discourses: Social interactions in academic writing. longman.

Ken Hyland. 2005. Metadiscourse: Exploring interaction in writing. continuum. *London, UK*.

Ken Hyland. 2019. *Second language writing*. Cambridge university press.

Katja Jasinskaja and Elena Karagjosova. 2020. Rhetorical relations. *The Wiley Blackwell companion to semantics*, 4:2645–2673.

Young-Seon Jeong, Myong K Jeong, and Olufemi A Omitaomu. 2011. Weighted dynamic time warping for time series classification. *Pattern recognition*, 44(9):2231–2240.

John Kirkman. 2012. *Good style: writing for science and technology*. Routledge.

Jan Van Kuppevelt. 1995. Discourse structure, topicality and questioning. *Journal of Linguistics*, 31(1):109–147.

Fabrizio Macagno. 2016. Argument relevance and structure. assessing and developing students' uses of evidence. *International Journal of Educational Research*, 79:180–194.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Noran L Moffett. 2019. Creating a framework for dissertation preparation: Emerging research and opportunities: Emerging research and opportunities.

Jennie Nelson. 2011. Review of they say/i say: The moves that matter in academic writing. *Journal of Teaching Writing*, 26(1):107–116.

Randy Olson. 2020. *The narrative gym: Introducing the ABT framework for messaging and communication*. Prairie Starfish Press.

John Swales and Christine B Feak. 2009. Abstracts and the writing of abstracts. *(No Title)*.

John M Swales. 2014. 1990. genre analysis: English in academic and research settings. cambridge: Cambridge university press, selected 45–47, 52–60. In *The Discourse Studies Reader*, pages 306–316. John Benjamins.

Maite Taboada. 2006. Discourse markers as signals (or not) of rhetorical relations. *Journal of pragmatics*, 38(4):567–592.

Paul Thompson. 2001. *A pedagogically-motivated corpus-based examination of PhD theses: Macrostructure, citation practices and uses of modal verbs*. University of Reading Reading, UK.

Jacob Torfing. 2005. *Discourse Theory: Achievements, Arguments, and Challenges*, pages 1–32. Palgrave Macmillan UK, London.

Bonnie Webber and Aravind Joshi. 2012. Discourse structure and computation: Past, present and future. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 42–54, Jeju Island, Korea. Association for Computational Linguistics.

A Appendix

This section of the Appendix provides further data and analyses on the Zero- and One-Shot Evaluation Results.

B Appendix

This section of the Appendix is intended to provide further information regarding our evaluation metrics.

While we have high confidence in our framework, our gathered data, and thus our classifiers, we also acknowledge that in the two step process of our framework, an error could potentially occur in the sentence classification, or through an unexpectedly errant response from the LLM. Thus, we take great care to employ human evaluation for each LLM-returned abstract.

For evaluation, we employ four subgroups of human evaluators, combinations of the categories: Native English Speakers & Non-Native English Speakers and Evaluators with a Background in AI & Evaluators without a Background in AI.

1. **Native English Speakers + Background in AI**. These human evaluators are those who are native English speakers and knowledgeable in the majority of AI topics discussed in each abstract. These evaluators are presumed to have higher sensitivity as to whether the LLM generates grammatically correct output, while maintaining the correct subject and details.

2. **Native English Speaker + No Background in AI**. These human evaluators are those whose native language is English, but do not have an understanding of AI topics beyond that of a layman. These evaluators are presumed to have a higher sensitivity as to whether the LLM generates grammatically correct output, while also providing insight on how well the LLM coherently clarifies unfamiliar topics.

3. **Non-Native English Speaker + Background in AI**. These human evaluators are those who are not native English speakers, but are knowledgeable in the majority of AI topics discussed in each abstract. These evaluators are presumed to have a higher sensitivity to the impact of content, organization, and language between the original and LLM-generated abstracts, as well as whether the LLM-generated abstract maintains the correct subject and details.

4. **Non-Native English Speaker + No Background in AI.** These human evaluators are those who are not native English speakers, and also do not have an understanding of AI topics beyond that of a layman. These evaluators are presumed to have a higher sensitivity to the impact of content, organization, and language between the original and LLM-generated abstracts, as well as providing insight on how well the LLM coherently clarifies unfamiliar topics.

Additionally, we would like to evaluate the correlation between non-native English speakers and their abstract preference at each evaluation setting, as well as native English speakers and their abstract preference at each evaluation setting.

In order to simulate an environment similar to a blind conference submission review, we tried to vary our human evaluators by numerous factors outside of Native English Speakers and those with a Background in AI. The specific profiles of the human evaluators are as follows:

- **Native English Speaker + Background in AI**: Two human evaluators belonged to this category and tackled various parts of the evaluation. One of the evaluators in this category is a female PhD student from the United States, who speaks English fluently. The other is a male Master's student from South Korea, who grew up in the United States and considers English his native language.

- **Native English Speaker + No Background in AI**: Two human evaluators belonged to this category and covered various parts of the evaluation. One of the evaluators in this category is a female high school teacher with a Master's degree in Education, and a Bachelor's degree in English from the United States. The other is a practicing female lawyer with a JD in the United States and the equivalent in the United Kingdom, also a native English speaker.

- **Non-Native English Speaker + Background in AI**: Two human evaluators belonged to this category and covered various parts of the evaluation. One of the evaluators is a male software engineer with a Master's Degree in Artificial Intelligence and a native German speaker.

***Zero-Shot* GPT-4**	Avg. $\Delta_{Content}$	Avg. Δ_{Org}	Avg. Δ_{Lang}	Avg. $\Delta_{Overall}$ (%)
Native English Speaker				
AI Background	-10.71	-9.29	-4.29	-2.86
No AI Background	-22.86	-14.29	+6.00	-9.86
Non-Native English Speaker				
AI Background	-12.14	-11.86	-14.29	-11.43
No AI Background	-2.86	-2.00	+4.71	-1.58

Table 3: Zero-Shot GPT-4 qualitative results with respect to the human evaluators' respective native languages and their amount of background knowledge in artificial intelligence.

***Zero-Shot* Gemini-Pro**	Avg. $\Delta_{Content}$	Avg. Δ_{Org}	Avg. Δ_{Lang}	Avg. $\Delta_{Overall}$ (%)
Native English Speaker				
AI Background	-24.14	-1.14	-9.86	-14.29
No AI Background	-14.29	-9.57	-12.00	-24.43
Non-Native English Speaker				
AI Background	-20.00	-14.29	-23.86	-20.00
No AI Background	-12.71	-22.86	-20.00	-17.29

Table 4: Zero-Shot Gemini-Pro qualitative results with respect to the human evaluators' respective native languages and their amount of background knowledge in artificial intelligence.

The other is also a male software engineer (currently working in AI) with a Master's Degree in Computer Science and a native Czech speaker.

- **Non-Native English Speaker + No Background in AI**: Two human evaluators belonged to this category and covered various parts of the evaluation. One of the evaluators is a male frontend engineer with experience in software engineering, but not in artificial intelligence or machine learning, and is fluent in English, but a native Mandarin speaker. The other is a female author, with no computer science background, who speaks fluent English but is a native Spanish speaker.

For conducting the human evaluation, we provide the evaluator with both the original abstract, and the final outputted abstract (both unmarked and in a mixed order to prevent bias). For further bias prevention, the evaluators were given the following information beforehand:

> "For some segments, one abstract is the original abstract (written by a human) taken from the original academic paper, and the other is re-written by an LLM (Large Language Model, such as ChatGPT) - they are mixed so as to not implicate which is the human-written original and which is the LLM re-written version. For other segments, both Abstract 1 and Abstract 2 are written by humans. One of the abstracts is the original academic abstract taken from the original paper, the other is a human-rewritten form. These are also mixed so as to not implicate which is the original and which is the re-written."

While this was an incorrect statement to give to the evaluators, it was done to ensure that human evaluators were not actively seeking for indicators of LLM-generated material in each abstract comparison. The human evaluators were informed after evaluation that one of the abstracts was, indeed, generated by an LLM and which one it was.

We ask the evaluator to rate each abstract on a scale of 1-7 in terms of four categories: content, organization, and language, as defined in Han et al.(2023), as well as an overall readability score.

We employ the same metrics, defined as "rubrics", in Han et al.(2023)'s work on Automated Essay Scoring. The three metrics, or rubrics, are defined as follows:

Content: The abstract is well-developed and relevant to the argument. It is sufficiently supported with evidence/examples.

Organization: The abstract is effectively structured, making it easy for the reader to follow

***One-Shot* GPT-4**	Avg. $\Delta_{Content}$	Avg. Δ_{Org}	Avg. Δ_{Lang}	Avg. $\Delta_{Overall}$ (%)
Native English Speaker				
AI Background	-7.86	+14.29	-3.86	-1.29
No AI Background	-7.14	+15.43	-2.43	0.00
Non-Native English Speaker				
AI Background	-13.43	+8.00	-3.86	-4.43
No AI Background	-9.00	+7.43	+2.14	-3.00

Table 5: One-Shot GPT-4 qualitative results with respect to the human evaluators' respective native languages and their amount of background knowledge in artificial intelligence.

***One-Shot* Gemini-Pro**	Avg. $\Delta_{Content}$	Avg. Δ_{Org}	Avg. Δ_{Lang}	Avg. $\Delta_{Overall}$ (%)
Native English Speaker				
AI Background	-8.86	-5.43	+6.00	-3.29
No AI Background	-11.86	-20.00	-8.57	-17.14
Non-Native English Speaker				
AI Background	-4.71	-2.86	+3.86	+0.86
No AI Background	-3.00	-9.00	-2.00	-4.14

Table 6: One-Shot Gemini-Pro qualitative results with respect to the human evaluators' respective native languages and their amount of background knowledge in artificial intelligence.

the intended logical flow.

Language: The writing displays a wide range of vocabulary, as well as correct usage of it. The essay follows grammar and usage rules throughout the paper. Spelling and punctuation are correct throughout the paper.

Lastly, we ask the evaluators which abstract is better, and to indicate if either or both abstracts appear to be written by non-native English speakers or an LLM. We evaluate the output from the LLM in each setting in comparison to the abstract in its original form. The quantitative metrics include DTW and our EB-DTW. The qualitative metrics include average percent change in: "content" score, "organization" score, "language" score, and "overall" score as reported by human evaluators.

C Appendix

Abstract Preference for all experimental settings. As the experimental output moved from zero-shot to Mondrian-ST, there is a linearly increasing trend of Mondrian output being chosen as the better abstract, as opposed to the original. Visualization can be found in Figure 6.

D Appendix

Zero-Shot Prompt.

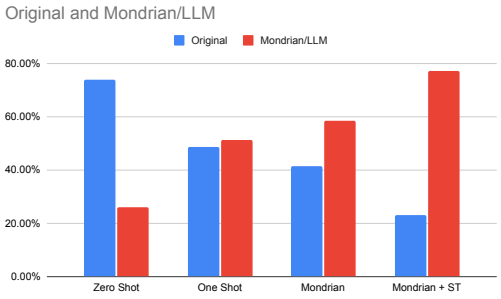

Figure 6: The progression of which abstract human evaluators preferred across experimental settings.

ABT format (And, But, Therefore) is an informal structure for bodies of text, following a narrative paradigm. Please restructure this following paragraph into ABT format:
[Input Abstract]

One-Shot Prompt.

ABT format (And, But, Therefore) is an informal structure for bodies of text, following a narrative paradigm. Restructuring a paragraph into ABT format would look like this:

Original: Despite significant advancements in medical technology and treatment options, the prevalence of chronic diseases continues to rise, posing a sig-

***Gemini-Pro* + BERT**	Avg. $\Delta_{Content}$	Avg. Δ_{Org}	Avg. Δ_{Lang}	Avg. $\Delta_{Overall}$ (%)
Native English Speaker				
AI Background	-0.71	0.00	+7.86	+5.00
No AI Background	-3.00	+9.57	+5.86	+1.86
Non-Native English Speaker				
AI Background	+6.14	+8.14	0.00	+4.14
No AI Background	+3.00	-0.71	+5.43	+3.71

Table 7: Mondrian qualitative results for BERT + Gemini-Pro with respect to the human evaluators' respective native languages and their amount of background knowledge in artificial intelligence.

***GPT-4* + BERT**	Avg. $\Delta_{Content}$	Avg. Δ_{Org}	Avg. Δ_{Lang}	Avg. $\Delta_{Overall}$ (%)
Native English Speaker				
AI Background	+10.43	+22.57	+6.43	+8.71
No AI Background	+11.29	+23.43	+13.29	+12.29
Non-Native English Speaker				
AI Background	+9.14	+20.29	+5.14	+11.00
No AI Background	+11.71	+23.00	+14.71	+12.14

Table 8: Mondrian qualitative results for BERT + GPT-4 with respect to the human evaluators' respective native languages and their amount of background knowledge in artificial intelligence.

nificant burden on healthcare systems worldwide. Chronic conditions such as diabetes, cardiovascular diseases, and respiratory illnesses account for a substantial portion of healthcare expenditures and contribute to morbidity and mortality rates. Addressing the complex challenges associated with chronic disease management requires a multifaceted approach that goes beyond traditional medical interventions. While medical treatments play a crucial role in symptom management and disease progression, addressing the underlying social determinants of health is equally essential. Factors such as socioeconomic status, access to healthcare services, and health literacy significantly influence health outcomes and disparities in chronic disease management. Therefore, healthcare systems must adopt comprehensive strategies that integrate medical interventions with social determinants of health approaches to effectively manage chronic diseases and improve patient outcomes. By implementing evidence-based interventions that address the root causes of health disparities and promote health equity, healthcare providers can enhance the quality of care and mitigate the adverse effects of chronic conditions on individuals and communities.

ABT: Addressing the complex challenges associated with chronic disease management requires a multifaceted approach that goes beyond traditional medical interventions. While medical treatments play a crucial role in symptom management and disease progression, addressing the underlying social determinants of health is equally essential. Factors such as socioeconomic status, access to healthcare services, and health literacy significantly influence health outcomes and disparities in chronic disease management. Despite significant advancements in medical technology and treatment options, the prevalence of chronic diseases continues to rise, posing a significant burden on healthcare systems worldwide. Chronic conditions such as diabetes, cardiovascular diseases, and respiratory illnesses account for a substantial portion of healthcare expenditures and contribute to morbidity and mortality rates. Therefore, healthcare systems must adopt comprehensive strategies that integrate medical interventions with social determinants of health approaches to effectively man-

***Gemini-Pro* + RoBERTa**	Avg. $\Delta_{Content}$	Avg. Δ_{Org}	Avg. Δ_{Lang}	Avg. $\Delta_{Overall}$ (%)
Native English Speaker				
AI Background	+7.86	+9.86	+14.29	+10.43
No AI Background	+2.00	+15.43	+9.00	+13.00
Non-Native English Speaker				
AI Background	+6.00	+10.43	+12.14	+11.71
No AI Background	+10.71	+9.14	+13.29	+11.29

Table 9: Mondrian qualitative results for RoBERTa + Gemini-Pro with respect to the human evaluators' respective native languages and their amount of background knowledge in artificial intelligence.

***GPT-4* + RoBERTa**	Avg. $\Delta_{Content}$	Avg. Δ_{Org}	Avg. Δ_{Lang}	Avg. $\Delta_{Overall}$ (%)
Native English Speaker				
AI Background	+9.00	+26.00	+20.71	+23.43
No AI Background	+9.86	+23.14	+19.71	+22.00
Non-Native English Speaker				
AI Background	+7.86	+23.29	+20.57	+19.71
No AI Background	+8.86	+24.29	+14.00	+20.14

Table 10: Mondrian qualitative results for RoBERTa + GPT-4 with respect to the human evaluators' respective native languages and their amount of background knowledge in artificial intelligence.

age chronic diseases and improve patient outcomes. By implementing evidence-based interventions that address the root causes of health disparities and promote health equity, healthcare providers can enhance the quality of care and mitigate the adverse effects of chronic conditions on individuals and communities.
Using the example above, please restructure the following paragraph into ABT format:
[Input Abstract]

Mondrian Prompt.

I have a series of text segments with varying sentiments that need to be combined into one cohesive narrative. The first segment is neutral in sentiment and provides preliminary information. The second segment is generally negative, highlighting a gap or problem related to the information introduced in the first segment. The third segment is positive, offering a solution or positive outcome that addresses the issues mentioned in the second segment.

Could you please create a seamless narrative that transitions smoothly between these segments? Ensure that:

Each segment remains intact without altering its core content. Include explicit transition sentences that logically connect each segment, guiding the reader from neutral to negative, and finally to a positive resolution. Maintain thematic continuity throughout, ensuring that the transition from the problem to the solution feels natural and directly addresses the issues raised earlier. The narrative should follow a specific sentiment trajectory: starting neutral, moving to negative, and concluding on a positive note. The goal is to weave these segments into a single paragraph, where the flow of ideas and sentiments is coherent and fluid, without needing detailed knowledge of each segment's content in advance.
1: **A Segment**
2: **B Segment**
3: **T Segment**

ST + *Gemini-Pro* + BERT	Avg. $\Delta_{Content}$	Avg. Δ_{Org}	Avg. Δ_{Lang}	Avg. $\Delta_{Overall}$ (%)
Native English Speaker				
AI Background	+12.29	+21.43	+16.29	+12.29
No AI Background	+9.86	+21.86	+16.14	+13.00
Non-Native English Speaker				
AI Background	+12.14	+19.71	+14.29	+9.86
No AI Background	+10.43	+18.43	+15.14	+12.14

Table 11: Mondrian-ST qualitative results for BERT + Gemini-Pro with respect to the human evaluators' respective native languages and their amount of background knowledge in artificial intelligence.

ST + *GPT-4* + BERT	Avg. $\Delta_{Content}$	Avg. Δ_{Org}	Avg. Δ_{Lang}	Avg. $\Delta_{Overall}$ (%)
Native English Speaker				
AI Background	+14.29	+32.43	+27.26	+24.71
No AI Background	+17.29	+29.71	+30.71	+25.71
Non-Native English Speaker				
AI Background	+13.86	+24.86	+25.71	+23.71
No AI Background	+19.00	+20.00	+27.86	+24.43

Table 12: Mondrian-ST qualitative results for BERT + GPT-4 with respect to the human evaluators' respective native languages and their amount of background knowledge in artificial intelligence.

ST + *Gemini-Pro* + RoBERTa	Avg. $\Delta_{Content}$	Avg. Δ_{Org}	Avg. Δ_{Lang}	Avg. $\Delta_{Overall}$ (%)
Native English Speaker				
AI Background	+10.43	+22.14	+14.29	+18.14
No AI Background	+8.86	+19.14	+9.71	+17.43
Non-Native English Speaker				
AI Background	+9.71	+21.86	+13.00	+15.71
No AI Background	+9.57	+22.29	+13.86	+14.86

Table 13: Mondrian-ST qualitative results for RoBERTa + Gemini-Pro with respect to the human evaluators' respective native languages and their amount of background knowledge in artificial intelligence.

ST + *GPT-4* + RoBERTa	Avg. $\Delta_{Content}$	Avg. Δ_{Org}	Avg. Δ_{Lang}	Avg. $\Delta_{Overall}$ (%)
Native English Speaker				
AI Background	+15.29	+34.86	+23.86	+25.43
No AI Background	+13.14	+22.86	+20.57	+22.43
Non-Native English Speaker				
AI Background	+11.57	+29.29	+21.72	+35.25
No AI Background	+11.86	+23.86	+18.29	+19.86

Table 14: Mondrian-ST qualitative results for RoBERTa + GPT-4 with respect to the human evaluators' respective native languages and their amount of background knowledge in artificial intelligence.