Entailment Progressions: A Framework for Identifying Author Usage of Logic

Anonymous ACL submission

Abstract

Textual entailment, or the ability to deduce whether a proposed hypothesis is logically supported by a given premise, has historically been applied to the evaluation of language modeling efficiency in tasks like question answering and text summarization. However, we believe that these zero-shot entailment evaluations can extend to a sequential evaluation of entailment on a sentence-by-sentence basis within a larger text. We refer to this approach as "entailment progressions". Additionally, entailment progressions shed light on the intentional logical approaches authors typically employ to construct their arguments, illustrating the points at which authors choose to integrate contradiction and entailment. Our results suggest that entailment progressions can both identify consistency in logical structures and establish a connection between this consistency and how humans typically author texts, as opposed to more formulaic approaches.

1 Introduction

As Large Language Models (LLMs) expand and evolve to accommodate more complex language generation tasks, model developers and researchers have leveraged validation mechanisms to ensure the accuracy of text outputs. These mechanisms, ranging from active feedback mechanisms informed by human input (Christiano et al., 2017; MacGlashan et al., 2017) to passive benchmarking designed to test LLMs using metrics indicative of human linguistic capability (Wang et al., 2018; Lin, 2004; Papineni et al., 2002), work towards the primary goal of bridging the gap between model capability and human language.

Textual Entailment – the ability to deduce whether a proposed hypothesis is logically supported by a given premise (Bentivogli et al., 2009) – has helped modelers understand the inferential capabilities of a given language model. Its origins lie in the belief that for a model to conduct specific Natural Language Processing (NLP) tasks, it must be capable of the elementary logical inference that underlies these tasks (Zaenen et al., 2005). However, we believe that entailment can potentially describe stylistic choices made by the author of the evaluated text through an examination of how the author chooses to introduce new information or support information they previously provided.

While RTE primarily focuses on the logical relationship between two statements, logical approaches require examining multiple statement-to-statement relationships for coherence. For example, statement (1) alone does not entail (2). However, if we introduce statement (3) in between the previous two statements, a new logical relationship emerges.

(1) *Blue light has the shortest wavelength in the electromagnetic spectrum.*

(2) *The sky is blue.*

(3) *Gas and particles in the sky reflect light with the shortest wavelength.*

In the context of RTE, the former case would yield a single outcome, indicating neutral entailment. In contrast, the latter scenario would generate two outcomes: neutral entailment between (1) and (3); and positive entailment between the first two statements and the final premise. Extending RTE to encompass intermediary logical relationships between a text and its final hypothesis not only enhances our understanding of the employed logical approach, but also clarifies why such an approach was chosen by the author. This raises a pair of interesting and, as far as we are aware, unexplored questions: *is logical reasoning inherently linked to the traits of the author employing it?* And if it is, *can this relationship be identified?*

Our work builds upon previously established definitions and modeling approaches for RTE, demonstrating its applications beyond traditional use cases. In this paper, we:

(1) Introduce entailment progressions, a framework in which a given piece of text can be represented as a series of values, with each value representing the level of textual entailment between two consecutive sentences in a text. This entailment progression describes the logical flow of the text, identifying how new information is introduced in relation to the preceding content (in support, in rejection, or with no relation).

(2) We show that analyzing the **entailment progressions in a set of documents** written by a specific author **can unveil similarities in their logical style, that can be attributed to the intentions of the author**.

(3) We show that entailment progressions are capable of assessing whether automatically generated human-like text, under specific prompted conditions, **adhere to an underlying structure** that modulates entailment at key points in the text.

The remainder of the paper is organized as follows. Section 2 presents the current state of the art, Section 3 describes our methodology, while Section 4 details the results of our experiment. We conclude by providing some perspectives for future work.

2 Related Work

The first notable definition of textual entailment was formulated as follows: *"T textually entails H if, typically, a human reading T would infer that H is most probably true"* (Dagan et al., 2010). This definition refined the RTE approach to specifically focus on the logical relationship between T and H on the basis of human evaluation rather than preexisting notions of implication, in which the plausibility of T and H could potentially misconstrue entailment (Korman et al., 2018).

Korman (Korman et al. 2018) later expanded upon Dagan's inferential approach to accommodate for edge cases associated with human inference like irrelevant trivialities, unexpressed conclusions, and potential disagreements in human interpretation, leading to an alternate understanding of RTE: *"a text T textually entails a hypothesis H relative to a group of end users G just in case, typically, a member of G reading T would be justified in inferring the proposition expressed by H from the proposition expressed by T"*. Korman's definition differs from Dagan's in three important ways. First, RTE must be restricted to a group G due to differing RTE approaches associated with variability in background knowledge, linguistic proficiency, and other human traits that can affect interpretation of logical relationships (Bos and Markert, 2005). Second, RTE requires justifiable inference in order to allow readers to assume logical transfer without believing the plausibility of T or H (Feldman, 2003). Finally, RTE should limit the scope of T and H to the literal expression of T and H in order to condition for inferential effects associated with differing grammar, semantic, and syntactical choices (Braun, 2001).

While textual entailment could be philosophically outlined in the broader context of logical inference, the practical application of RTE approaches was constrained by a limited understanding of the linguistic underpinnings and specific criteria that govern how expressions entail one another (Amoia, 2009). This became particularly evident in the academic analyses of the Pascal RTE challenges (Dagan et al., 2005; Giampiccolo et al., 2007, 2008; Bentivogli et al., 2009, 2011), a series of competitions in which participants tested RTE approaches against datasets comprised of premise-hypothesis pairs, along with *"support"* and *"reject"* labels (Dagan et al., 2005). While these datasets attempted to capture entailment through the binary classification of these premise-hypothesis pairs, this approach limits the broader scope of entailment, which can vary depending on factors such as world knowledge and negation (De Marneffe et al., 2008). Additionally, researchers found that sentence structure and general syntax played a significant role in improving entailment predictions, thus furthering the notion that RTE datasets should encompass the requisite linguistic diversity to comprehensively map entailment (Vanderwende and Dolan, 2005; Blake, 2007). Although the subsequent development of the Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) and Multi Natural Language Inference (MNLI) (Williams et al., 2017) corpora significantly improved the general recognition of entailment by incorporating human annotations of entailment across different genres and varying real-world conversations, more specialized RTE datasets, such as the Diverse Natural Language Inference Collection (Poliak et al., 2018) and the "NLI Stress Tests" (Naik et al., 2018), advocate for the recasting of datasets pertaining to specific linguistic phenomena such as event factuality, sentiment analysis, and numerical reasoning into RTE challenges (White et al., 2017). [1] Not only does this approach extend RTE approaches to intrinsically identify diverse logical structures expressed in various linguistic styles, but it also extends the practical application of these RTE methods to various NLP applications, such as question-answering (Khot et al., 2018; Harabagiu and Hickl, 2006), text summarization (Lloret et al., 2008; Naserasadi et al., 2019), and machine translation (Padó et al., 2009).

Current RTE modelling approaches require two main steps. First, the features of premise `T` and hypothesis `H` are extracted in order to represent the statements in accordance with relevant linguistic mechanisms associated with textual entailment (Li et al., 2020). These mechanism-oriented approaches include lexical approaches that leverage part-of-speech tagging, stopword removal, and named entity recognition to represent statements through word choice (Lan and Jiang, 2018), syntactic approaches that utilize parse trees and dependency graph representations to represent statements through sentence structure (Iftene and Moruz, 2009), and semantic approaches that use semantic converters like the Universal Natural Language and cross-referenced paraphrasing to represent statements in a fuller definitive context. These feature extraction processes can be hybridized accordingly and are often represented through word embeddings (Basak et al., 2018). Second, the statements are fed into a supervised multi-class classification model which predicts whether a premise-hypothesis pair possesses *positive* (the hypothesis can be inferred to be true if the premise is true), *negative* (the hypothesis can be inferred to be false if the premise is true), or *neutral* (the hypothesis' truth is not sufficiently conditional upon the premise being true) entailment. This step has been greatly facilitated by the development of robust RTE corpora like the previously mentioned datasets.

3 Methodology

3.1 Hypothesis

We incorporate Korman's RTE approach into our methodology due to its emphasis on refined human inference, which has several important implications (Korman et al., 2018). Consider a scenario in which an individual is tasked with crafting an argument in 10 sentences. Under these circumstances, according to the Korman approach, the opening sentence of the individual's argument cannot directly entail the concluding sentence. This is due to the presence of intermediate premises, which are necessary to make the argument convincing and logical, thus enabling readers to justifiably infer entailment within the argument.

In each successive sentence, new information is presented, which can be either affirmative, contradictory, or neutral in relation to the preceding premises. This incremental accumulation of information ultimately leads to the conclusion asserted by the final sentence. Thus, unless there are no immediate prerequisites for logical or textual coherence in presenting a set of claims, individuals should continuously incorporate textual entailment to ensure that the overarching message is effectively delivered. This is reinforced when taking into account Korman's stipulation of identifying where the individual stands in relationship to their group `G`. If another individual, not part of the original group, was tasked with crafting the same 10-sentence argument in their unique writing style, they might incorporate textual entailment by structuring intermediary propositions differently or using distinct modes of textual expression compared to the first individual. Ultimately, while these two individuals may hold identical viewpoints and employ the same set of evidence to support their stance, what sets them apart are the variations in the structure of their intermediary propositions and their respective modes of expression. We suggest that in such a scenario, assessing fluctuations in textual entailment on a sentence-by-sentence basis can quantitatively illustrate the differences in argumentative approaches between these two individuals.

The intentional fluctuation of textual entailment on behalf of the author adds a new dimension to Korman's RTE approach, one which examines RTE on behalf of the communicator rather than the audience. In the aforementioned scenario, if an evaluator were to assess the textual entailment of each communicator's argument sentence-by-sentence, disparities in RTE could be attributed to the variations in the evaluator's background knowledge or linguistic proficiency, which might hinder the evaluator from making valid inferences based on the modified expression. Additionally, these differences may also stem from the distinct approaches employed by the communicators in structuring their

[1] For an in depth analysis of existing corpora, see (Poliak, 2020).

messages in a logically coherent manner. If the former circumstance is adequately conditioned, then the evaluator's RTE approach can provide quantifiable insight into the logical reasoning of the communicators. This becomes more evident when the evaluator assesses multiple instances from both communicators. The *entailment 'progressions'* derived from the evaluator's RTE approach can effectively illustrate key patterns in the communicators' logical approach to crafting a cohesive argument or message.

The formal definition of the entailment progressions of a given text can be expressed as follows:

$$EP_{3\times n} = \begin{bmatrix} c_1 & c_2 & \cdots & c_n \\ p_1 & p_2 & \cdots & p_n \\ n_1 & n_2 & \cdots & n_n \end{bmatrix}$$

where *EP* is an entailment progression matrix composed of *c*, *p*, *n* row vectors representing the contradiction, positive, and neutral entailment probabilities between two sentences in a given text. To compute these values at a given point in a text, we introduce the following equation:

$$EP_{c_i,p_i,n_i} = E(s_{i-1}, s_i)$$

where *E* represents the entailment model used for calculating entailment between two sentences, and *s* represents a sentence at a given point in the text.

Drawing from our analysis of the existing RTE literature, we propose the following. Given two texts composed of an equal number of sentences, denoted as $\mathtt{T}_1$ and $\mathtt{T}_2$, which are equal in length and are composed of a series of similarly presented statements that serve to further the same logical premise, and an evaluator $\mathtt{E}$, an individual tasked with recognizing the textual entailment on a statement-by-statement basis for $\mathtt{T}_1$ and $\mathtt{T}_2$, if $\mathtt{E}$ identifies sufficient differences in the entailment progressions of $\mathtt{T}_1$ and $\mathtt{T}_2$, then $\mathtt{T}_1$ and $\mathtt{T}_2$ can be distinguished based on the distinct logical approaches employed by their respective communicators.

3.2 Experimental Design

To ensure that our hypothesis is satisfied, we design an experimental setup that effectively accounts for potential confounding limitations that may arise during the evaluator's analysis.

First, both $\mathtt{C}_1$ and $\mathtt{C}_2$ must employ similar linguistic mechanisms for presenting their premises. This is to avoid potential deterrents caused by limited semantic or syntactical knowledge, which could hinder the evaluator's ability to accurately assess the truth value of statements within $\mathtt{C}_1$ and $\mathtt{C}_2$, a necessary prerequisite for RTE.

Second, both $\mathtt{C}_1$ and $\mathtt{C}_1$ must *"further the same logical premise"* by pertaining to the highly similar, if not identical, domains in which the evaluator possesses a sufficient and equal understanding. This is to ensure that the evaluator possesses the necessary background knowledge to proficiently implement their RTE approach with *"justifiable inference"*.

Third, both $\mathtt{C}_1$ and $\mathtt{C}_2$ must be equal in length, or possess an equal amount of statements. This final condition is in direct reference to the previously mentioned issue of inferential distance, in which the difficulty of reasoning from one statement to another in a piece of text is associated to the number of intermediary propositions required to effectively connect the statements. If $\mathtt{C}_1$ and $\mathtt{C}_2$ both advance the same logical premise, but $\mathtt{C}_1$ is significantly longer than $\mathtt{C}_2$, with both covering an equal total inferential distance from their initial to final statements, then, on average, the statement-by-statement fluctuations in textual entailment for $\mathtt{C}_1$ would be lesser than those of $\mathtt{C}_2$, simply by dividing the total inferential distance by the number of statements. While this explanation may not fully account for cases where $\mathtt{C}_1$ offers a richer explanation of its relevant topic (and subsequently exhibits a higher degree of variability in the textual entailment it employs), it does hold true in a qualitative sense. If a communicator was tasked with crafting a 5-sentence argument on a subject that typically requires 15 statements for a cohesive exposition, the condensed length will force the communicator to emphasize specific points with greater urgency. In turn, this could lead to larger logical jumps and thus greater deviations in RTE.

When controlling for these conditions, we design an experimental setting in which the evaluator is capable of qualitatively and quantitatively distinguishing between the logical structures employed in $\mathtt{C}_1$ and $\mathtt{C}_2$. Please note that this comparison can be made not only between texts from the same communicator but also between texts from different communicators, and the conclusions drawn from each approach may carry different implications. If both $\mathtt{C}_1$ and $\mathtt{C}_2$ are authored by the same communicator, analysing the textual entailment progressions of both texts can serve as a stress test. This stress test helps evaluate the robustness of the communicator's logical approach relative to minor topical differences that still adhere to the second condition. However, when C_1 and C_2 originate from different communicators, evaluating the textual entailment progressions of both texts allows for a comparison of intention in logical approaches between these communicators. By combining these *"intracommunicator"* (i.e., evaluating the logical approaches pertaining to a single author) and *"intercommunicator"* (i.e., comparing logical approaches across multiple authors) approaches, we can create a third type of comparison. This approach enables us to evaluate whether the differences in intended logical approaches between communicators remain distinguishable and robust across various examples.

3.3 Data and Models

The data used to evaluate our hypothesis is sourced from mutual fund evaluations written in English (hereafter *narratives*) provided by a leading financial analysis firm. These narratives supplement the firm's assessments of mutual fund performance metrics by elaborating on key qualitative aspects of the fund's performance.

The narratives are primarily categorized based on authorship or the specific method used for their creation. The narratives for the top-performing 25% of the examined funds are authored by human analysts (referred to as `Analyst narratives`), while the narratives for the remaining 75% are generated using a proprietary algorithm called Smart Text (referred to as `Quant narratives`). Smart Text was first launched in the spring of 2021 and is generated through a deterministic rules-based process. The firm employs a rule-based approach to group funds into various mental models, where each mental model has an associated template structure that helps to assign text branches to. These branches contain embedded data points (such as fund name and expense ratio) and also a synonym bank of words to ensure text variation. It is important to note that this process is entirely rules-based and doesn't utilize any recent generative AI techniques.[2]

While `Analyst narratives` are categorized by the authoring analyst, `Quant narratives` are categorized by subject-specific templates used to evaluate a fund in relation to a specific area in investment management research that is most relevant to its performance. Our dataset contains 26 different templates, each covering a specific aspect of a fund's performance relative to its core characteristics. For example, the `Active Allocation` template is structured to evaluate funds that are both actively managed by a portfolio manager and are diversified across different asset classes, and is primarily focused on the fund's active allocation strategy rather than other aspects that the fund may possess. To examine the relationship between authorship and logical approach, we divided the sets by analyst (for `Analyst narratives`) and by template type (for `Quant narratives`).

Several constraints were applied when implementing the proposed experimental approach. Firstly, as cosine similarity served as the metric used for evaluating the similarity of entailment progressions within evaluated groups, and given that entailment progressions often varied in length due to differences in the number of sentences across sets of `Quant` and `Analyst narratives`, we grouped each set into subsets based on the number of sentences. Additionally, since cosine similarity is inflated in lower dimensional settings, we only considered subsets with 10 or more sentences in our final analysis. Furthermore, to ensure the accuracy of our analysis and prevent potential issues that could arise due to a low sample size, we only considered subsets with 10 or more narratives. The final cosine similarity scores were computed as a weighted average of the cosine similarity scores within each subset, with weights determined by the number of narratives within the subset belonging to the respective set.

Table 1 provides an overview of the dataset used in our experiments categorized by authorship (i.e., whether the narrative originates from an analyst or the Smart Text model) as well as the number of categories within each type of narrative.

NARRATIVE TYPE	# OF AUTHOR CATEGORIES	TOTAL
`Analyst`	5	1000
`Quant`	26	4000

Table 1: Total count of narratives and categories within each narrative type.

To calculate the textual entailment on a sentence-by-sentence basis for the narratives in our dataset, we rely on **RoBERTa-base** (Liu et al., 2019), an optimized BERT-like (Devlin et al., 2019) encoder

[2]The reason for employing a rule-based model is the firm's focus on generating precise and reliable narratives.

Transformer. We specifically leverage a version of this model adapted to the domain of RTE by fine-tuning on datasets designed for Natural Language Inference (Reimers and Gurevych, 2019). For performing the experiments, we relied on the Hugging-Face transformers library (Wolf et al., 2020).

4 Results and Discussion

Table 2 presents the cosine similarity scores for the entailment progressions of the `Quant` and `Analyst narratives`. These results highlight that `Quant narratives` exhibit, on average, higher cosine similarity scores within sets of entailment progressions compared to `Analyst narratives`. These results align with our *"intra-communicator"* experimental design, where higher cosine similarity scores indicate greater robustness in the stylized logical approach characterizing the narrative set. Since the `Quant narratives` are generated through the Smart Text algorithm, this consistency within the narratives can be attributed to the (more) formulaic communication style embedded into the narratives' delivery method. In contrast, `Analyst narratives` are prone to display greater variability within the narrative sets, reflecting the less structured approach often taken by human analysts when crafting their narratives.

Table 3 presents the cosine similarity scores for the entailment progressions of the `Quant` and `Analyst narratives` between different author groups, an example of the previously noted *"intercommunicator"* approach. While we anticipate the cosine similarity across different authors to be lower than the *"intracommunicator"* comparison, our analysis suggests that this is dependent upon the specific contextual factors that shape the writing styles of authors within a group. For example, the cosine similarities of different analyst pairs do not differ from the results presented in Table 2. This can be attributed to the fact that all the evaluated analysts are expected to adhere to the writing standards set by the financial services firm employing them. This trend is not upheld when examining the cosine similarities between different template pairs, where we observe not only a significant decrease in the cosine similarities as outlined in Table 2, but also a greater variance, ranging from -0.148 to 0.370. This aligns with the manner in which the templates are generated (i.e., on the basis of specific rules associated with the subject matter they address). Therefore, a higher cosine similarity is observed when evaluating these templates in comparison to those that utilize a very similar logical structure. The most notable instance of this is phenomenon is evident when comparing a template to itself (cf. Table 2)).

The results presented in Tables 2 and 3 suggest that the entailment progressions of a given author's work are not only influenced by the authors themselves, but also by the stylistic constraints imposed upon them. The first notable constraint observed in our analysis pertains to *structure*. Both the `Quant` and `Analyst narratives` adhere to the structural constraint in terms of length (i.e., the analysts and the Smart Text algorithm are required to condense their analysis of a given mutual fund into an informative and concise format). Were the narratives written without this constraint, and thus allowing for more sentences to be used for covering a given fund, then their entailment progressions would change to accommodate more intermediary propositions. This would reduce the inferential distance between the initial and final claims (Korman et al., 2018). The second notable constraint pertains to the subject matter of the text. A broader, more complex subject matter can create variability in the logical approaches employed to effectively support the claim at hand. This can be observed in both the `Quant` and `Analyst narratives`; while the former specializes in narrower subject matters (e.g., Active Equity), the latter focuses on multiple aspects of a fund that can be used to evaluate performance. While Table 2 shows how this subject constraint renders the cosine similarity within narrative groups higher for `Quant narratives` than for `Analyst narratives`, Table 3 highlights how the cosine similarity between `Quant narratives` groups experiences a more pronounced decrease compared to `Analyst narratives` groups by virtue of differing subject matters altogether.

This becomes even more evident when we visualize the entailment progressions associated with specific sets of `Quant` and `Analyst narratives` (cf. Figure 1). While a high cosine similarity generally indicates a similarity in trends between two vectors, this distinction is particularly pronounced when assessing the points at which significant fluctuations (or lack thereof) occur in the entailment progressions of the narrative set.

Depending on the template under examination,

Analyst	Cosine Similarity	Template	Cosine Similarity
Analyst A	0.285	Active Allocation	0.601
Analyst B	0.237	Passive Allocation	0.584
Analyst C	0.234	Active Equity	0.579
Analyst D	0.230	Long Term	0.554
Analyst E	0.201	Active Fixed Income	0.530

Table 2: Cosine similarity scores of entailment progressions within the evaluated narrative sets. A higher score indicates a greater similarity between entailment progressions and suggests a closer alignment in the underlying logical structure employed by the respective authors.

Analyst Pairs	Cosine Similarity	Template Pairs	Cosine Similarity
Analyst A, Analyst B	0.213	Active Fixed Income, Long Term	0.370
Analyst A, Analyst C	0.237	Active Fixed Income, Total Risk	0.112
Analyst A, Analyst D	0.189	Active Fixed Income, Passive Allocation	-0.148
Analyst B, Analyst C	0.244	Passive Allocation, Active Equity	0.192
Analyst B, Analyst D	0.188	Passive Allocation, Total Risk	0.199

Table 3: Cosine similarity scores of entailment progressions between different evaluated narrative sets. Analyst and Quant narratives are directly compared in pairs by their respective analyst or template.

we observe that the Quant narratives exhibit fluctuations in entailment at key points. Particularly, in the case of Active Allocation Quant narratives, we can observe spikes of contradictory entailment occurring around the 40% and 90% marks. In contrast, Analyst narratives are less uniform due to their non-formulaic nature – human analysts may not purposefully introduce contradictory entailment at consistent points across all their narratives. Additionally, while entailment progressions in Quant narratives exhibit a clearer dominance of neutral entailment (with occasional spikes in positive or negative entailment), Analyst narratives integrate both positive and negative entailment throughout the narrative, a distinction that can be attributed to the difference in communicative quality between the method used for generating the Quant narratives and human authorship.

We can link our analysis results to our initial hypothesis in three key ways. First, we identify the similarity in entailment progressions for template-based sets of Quant narratives as a result of the rigidity in the logical approaches employed within templates. If one were to read such narratives within a specific template, they would be able to recognize the similar logical structure and shared logical processes used for their generation. Second, we note the lower similarity in entailment progressions for the Analyst narratives sets, indicating a less rigid logical structure in their crafting. The variability in logical approaches used by the analysts when rating mutual funds can make it challenging for a reader to identify the analyst solely based on logical flow. However, this does not preclude the reader from identifying the analyst on the basis of word choice, tone, and other stylistic indicators that are separate from RTE. Third, the varying levels of similarity between sets of Quant and Analyst narratives in their entailment progression subtly highlight the distinction between human and automatically-generated communication. Quant narratives tend to follow a a more structured and explicit logical structure compared to Analyst narratives. Consequently, if a reader were to read a combined set of Quant narratives (from a specific template) and Analyst narratives (written by a specific author), they could discern whether a narrative was automatically generated or crafted by a human analyst based on its resemblance to narratives they previously encountered. While this may be more implicit than the previous two points, it underlies the sentiment expressed by the firm's clients, who find Quant narratives to be "too robotic" and lacking the qualitative aspects of an analyst's narrative. In all three situations, our hypothesis holds, confirming the benefits of leveraging entailment progressions.

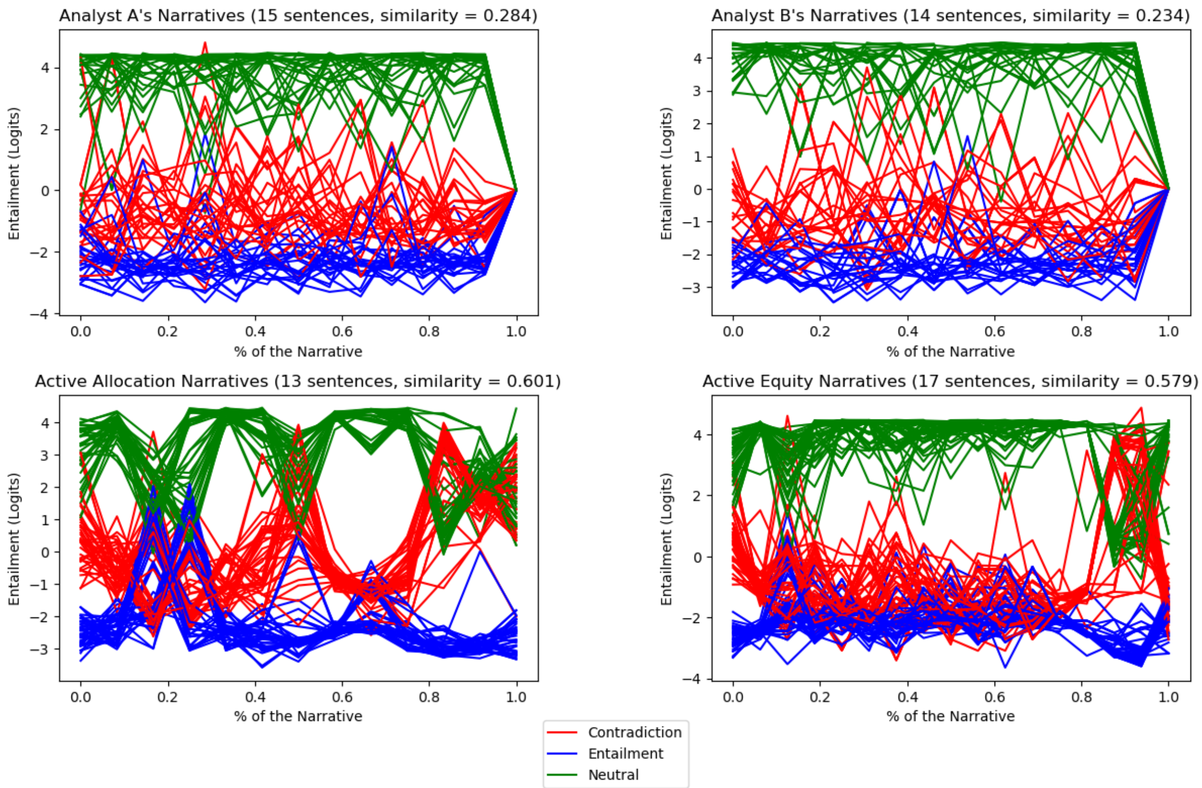

Figure 1: Line plots displaying the entailment progressions of narratives within selected sets. Entailment is measured in logits calculated by the RoBERTa model used to calculate the entailment scores for the progression.

5 Conclusion

In this paper we introduce entailment progressions, a framework which serves in identifying the logical flow of a text and highlights the way in which different authors choose to integrate nuance and affirmation on a sentence-by-sentence basis. This deviation from traditional RTE literature seeks to characterize textual entailment similarly to author style, which is often linked to lexical and semantic choices, as opposed to being seen as a purely objective or impersonal linguistic phenomenon. As demonstrated through the difference in cosine similarity between the `Quant` and `Analyst narrative` entailment progressions, the author's stylistic rigidity influences the robustness of the underlying logical structures. These findings suggest that entailment progressions can potentially distinguish between human and "non-human" logical approaches. In future work, we will explore the logical approaches taken by LLMs and assess to what degree they align with the humanistic logical approaches. This exploration can play a crucial role in identifying whether these models possess an inherent logical structure to adhere to, which, in turn, can potentially contribute to the larger area of LLM interpretability. Second, we will explore whether entailment progressions can serve as a benchmark for defining the similarity between two logical approaches, extending the analysis presented in Table 3 to encompass other authorship styles. If two sets of texts' entailment progressions, both of which are conditioned in accordance with our defined methodology, were compared on the basis of cosine similarity, a high cosine similarity could indicate that both texts adhere to similar logical structures and that the authors of the texts have similarly integrated logic into how they have devised their overall texts. This concept can apply to both human and non-human approaches, where comparisons between human authors, human and model-based authors, and model-based authors can be analysed using entailment progressions as a heuristic to assess whether similar logical processes are at play.

Ethics Statement

The data that was used for conducting the experiments was provided by a leading financial analysis firm and is not publicly available. Both the `Quant` and `Analyst narratives` were collected internally by members of the firm's team, and are exclusively available to clients who have subscribed to service offerings that grant access to these narratives. Although the firm has given explicit approval to our methodology and use of their data, the proprietary nature of this information warrants selective discretion when describing and disseminating this data.

Limitations

Entailment progressions serve to identify the logical approach employed by a given author, to which we characterize it as an aspect of the author's style that can be leveraged for tasks that involve author classification or author style transfer. Since our analysis extends to assessing the internal logical structures employed by non-human approaches in relation to typical human logic approaches, the former task can aid tools seeking to distinguish between model-generated and human-generated texts when necessary, a problem that is newly arising in areas concerning intellectual property and fraud detection (Yu et al., 2023; Májovskỳ et al., 2023). Although entailment progressions can help identify stylistic differences between model and human-generated outputs, they can simultaneously improve these models in their closeness in inferential capability to humans, potentially rendering the task of differentiating between human and model outputs more difficult.

In this work, we acknowledge a number of limitations and discuss their implications.

(1) Our analysis leverages cosine similarity to compare entailment progressions, but this is limited by specific metric requirements. First, cosine similarity requires that the vectors being compared be of the same length, thus limiting a comprehensive analysis. Although we address this issue within our methodology, this limitation significantly hinders our ability to perform an in-depth analysis of entailment progressions that fall outside this length constraint. Second, cosine similarity is inflated in low-dimensional settings, which can hinder the comparison of entailment progressions derived from shorter texts. These limitations can obscure entailment progression analysis to the extent that it artificially inflates the similarity scores between entailment progressions in every instance.

(2) As entailment progressions are generated using a pre-existing language model from the literature, the quality of these entailment progressions is directly dependent on the performance of the model. In scenarios like ours, where the true entailment is unknown, assessing the performance of the model can be challenging. Additionally, it is important to note that the model is trained on corpora containing general language premise-hypothesis pairs, which can limit its performance in more specialized domains, such as investment management research. Given that our methodology relies on the model's capability to accurately infer textual entailment, this requirement may not always be satisfied when applied to areas outside the scope of general language understanding. For example, a model might construe (4) and (5) as negatives, when in fact they are positives.

(4) The expensive ratio for this fund is low.

(5) Manager turnover for this firm is below average.

This can be mitigated through ensemble modelling, where multiple entailment models are deployed, and the corresponding agreement/disagreement can be leveraged to determine the final entailment probabilities at a given point in the progression.

(3) Given that entailment progressions adhere to the Markov property, wherein entailment between two statements in an entailment progression is sequential, entailment progression approaches should evaluate whether sufficient context is captured by examining only the immediately preceding sentence, rather than a larger set of previous sentences. Examining only the preceding sentence when calculating entailment may lead to an inflation of contradiction and neutral probability scores, as it overlooks prior sentences which may sufficiently entail the sentence under evaluation. This limitation is more procedural than the other limitations, as entailment progression generation can be adapted to account for different Markov assumptions.

References

Marilisa Amoia. 2009. Linguistic-based computational treatment of textual entailment recognition.

Rohini Basak, Sudip Kumar Naskar, and Alexander Gelbukh. 2018. A simple hybrid approach to recognizing textual entailment. *Journal of Intelligent & Fuzzy Systems*, 34(5):2873–2885.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. *TAC*, 7:8.

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

Catherine Blake. 2007. The role of sentence structure in recognizing textual entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 101–106.

Johan Bos and Katja Markert. 2005. Recognising textual entailment with logical inference. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 628–635.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

David Braun. 2001. Indexicals.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. Recognizing textual entailment: Rational, evaluation and approaches–erratum. *Natural Language Engineering*, 16(1):105–105.

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

Marie-Catherine De Marneffe, Anna N Rafferty, and Christopher D Manning. 2008. Finding contradictions in text. In *Proceedings of acl-08: Hlt*, pages 1039–1047.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Richard Feldman. 2003. Epistemology. *Tijdschrift Voor Filosofie*, 68(2).

Danilo Giampiccolo, Hoa Trang Dang, Bernardo Magnini, Ido Dagan, Elena Cabrio, and Bill Dolan. 2008. The fourth pascal recognizing textual entailment challenge. In *TAC*.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and William B Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9.

Sanda Harabagiu and Andrew Hickl. 2006. Methods for using textual entailment in open-domain question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 905–912.

Adrian Iftene and Mihai Alex Moruz. 2009. Uaic participation at rte5. In *TAC*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Daniel Z Korman, Eric Mack, Jacob Jett, and Allen H Renear. 2018. Defining textual entailment. *Journal of the Association for Information Science and Technology*, 69(6):763–772.

Yunshi Lan and Jing Jiang. 2018. Embedding wordnet knowledge for textual entailment. ACL.

Peiguang Li, Hongfeng Yu, Wenkai Zhang, Guangluan Xu, and Xian Sun. 2020. Sa-nli: A supervised attention based framework for natural language inference. *Neurocomputing*, 407:72–82.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*.

Elena Lloret, Oscar Ferrández, Rafael Munoz, and Manuel Palomar. 2008. A text summarization approach under the influence of textual entailment. In *NLPCS*, pages 22–31.

James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L Roberts, Matthew E Taylor, and Michael L Littman. 2017. Interactive learning from policy-dependent human feedback. In *International conference on machine learning*, pages 2285–2294. PMLR.

Martin Májovskỳ, Martin Černỳ, Matěj Kasal, Martin Komarc, and David Netuka. 2023. Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened. *Journal of Medical Internet Research*, 25:e46924.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. *arXiv preprint arXiv:1806.00692*.

Ali Naserasadi, Hamid Khosravi, and Faramarz Sadeghi. 2019. Extractive multi-document summarization based on textual entailment and sentence compression via knapsack problem. *Natural Language Engineering*, 25(1):121–146.

Sebastian Padó, Michel Galley, Dan Jurafsky, and Christopher D Manning. 2009. Robust machine translation evaluation with entailment features. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 297–305.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Adam Poliak. 2020. A survey on recognizing textual entailment as an NLP evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 92–109, Online. Association for Computational Linguistics.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. *arXiv preprint arXiv:1804.08207*.

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.

Lucy Vanderwende and William B Dolan. 2005. What syntax can contribute in the entailment task. In *Machine Learning Challenges Workshop*, pages 205–216. Springer.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhiyuan Yu, Yuhao Wu, Ning Zhang, Chenguang Wang, Yevgeniy Vorobeychik, and Chaowei Xiao. 2023. Codeipprompt: Intellectual property infringement assessment of code language models.

Annie Zaenen, Lauri Karttunen, and Richard Crouch. 2005. Local textual inference: can it be defined or circumscribed? In *Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment*, pages 31–36.