Extracting a Prototypical Argumentative Pattern in Financial Q&As

Giulia D'Agostino
Università della Svizzera italiana
Switzerland
giulia.dagostino@usi.ch

Michiel van der Meer
Universiteit Leiden
The Netherlands
m.t.van.der.meer@liacs.leidenuniv.nl

Chris Reed
University of Dundee
Scotland
c.a.reed@dundee.ac.uk

Abstract

Argumentative patterns are recurrent strategies adopted to pursue a definite communicative goal in a discussion. For instance, in Q&A exchanges during financial conference calls, a pattern called Request of Confirmation of Inference (ROCOI) helps streamline conversations by requesting explicit verification of inferences drawn from a statement. Our work presents two ROCOI extraction approaches from interrogative units: sequence labeling and text-to-text generation. We experiment with multiple models for each task formulation to explore which models can effectively and robustly perform pattern extraction. Results indicate that machine-based ROCOI extraction is an achievable task, though variation among metrics that are designed for different evaluation dimensions makes obtaining a clear picture difficult. We find that overall, ROCOI extraction is performed best via sequence labeling, though with ample room for improvement. We encourage future work to extend the study to new argumentative patterns.

1 Introduction

An argumentative pattern is a recurrent and identifiable structure with a specific function in an argumentative discussion. Such a pattern offers valuable insights into the reasoning processes and dialectical strategies employed by interlocutors in argumentative discourse.

Extracting argumentative patterns from natural discourse presents a significant challenge in the field of Argument mining (AM) (Lawrence and Reed, 2019). Typically, AM involves three stages: (1) the identification, segmentation, and classification of argumentative discourse units (ADUs) (Ghosh et al., 2014), (2) the characterization of the relations between ADUs (Peldszus and Stede, 2013), and (3) the identification of argument schemes, which denote implicit and explicit inferential relations within and across ADUs (Macagno and Walton, 2014). This area of research is often challenged by the idiosyncrasies of spoken language. For instance, in Earnings Conference Call (ECC) Q&A sessions, argumentative content is often embedded in complex statements aimed at maximizing information content while minimizing exchanges (Keith and Stent, 2019). Instead of employing a typical end-to-end AM pipeline, leveraging linguistic patterns that are clearly identifiable as part of argument schemes could be useful for locating argumentative moves, unraveling the complexities in such dialogues.

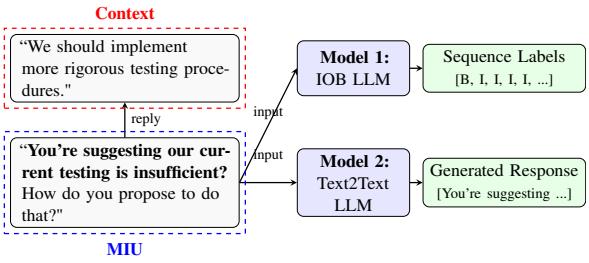


Figure 1: Example ROCOI and the two extraction approaches.

In this paper, we present a novel task and approach to the extraction of a prototypical argumentative pattern called the **Request Of Confirmation Of Inference** (ROCOI). Our work focuses on this argumentative pattern that emerges in questions, which presents some easily identifiable surface elements that complement the underlying argumentative function.

The ROCOI pattern is a structure that signals the presence of a reasoning process, of which the interrogative instance represents the conclusion (or, more accurately, the request for confirmation thereof). By carefully bridging lexical and syntactic recurrent features with the pragmatic role that such a pattern plays, ROCOIs constitute a unique pattern whereby we shed light on the reasoning process behind strategic inquiry. In the extraction

Proceedings of the 2025 CLASP Conference on Language models And RePresentations (LARP), pages 51–64

process, we move beyond the analysis of entire discourse units, instead allowing us to localize ROCOIs inside dialogues. This approach allows us to maintain precise control over pattern detection while dealing with the inherent complexity of argumentative texts. In this sense, the current approach moves beyond pattern identification–a task already tackled by D'Agostino and Rocci (2024)–towards pattern extraction.

We specifically focus on the ROCOI pattern as it represents an ideal proto-pattern for exploring how well NLP methods can extract argumentative patterns from text. These patterns exhibit characteristics that make them readily identifiable by trained human annotators, including their interrogative nature on an inferential conclusion, and explicit marking of prior reasoning. This clarity provides an excellent starting point for developing and evaluating automated extraction methods. At the same time, such characteristics are neither regular nor exclusive to this task, so they must be paired with recognition of the argumentative reasoning that the patterns reside in Since LLMs may not behave like human annotators in argumentative reasoning (de Wynter and Yuan, 2024), with this study, we probe the limits of pragmatic pattern recognition by means of surface elements.

The selection of our domain of application is mainly utilitarian: ECCs are strategic exchanges that constrain discourse in a way that makes the ROCOI pattern relatively frequent and prominent–an optimal environment for an exploratory study.

Our experiments encompass two task formulations, comparing a classification (sequence labeling) and generation (text-to-text extractive generation) paradigm. Comparing these two approaches allows us to bridge traditional boundary-marking techniques (Eger et al., 2017; Kuribayashi et al., 2019; Bao et al., 2021) with state-of-the-art language modeling approaches (Raffel et al., 2020; Gorur et al., 2024).

This work represents a crucial step toward the broader goal of comprehensive argumentative analysis, laying the groundwork for future exploration of more complex patterns, as well as the incorporation of contextual features in detecting argumentative patterns. Furthermore, our models can support humans in locating argumentation in financial contexts (van der Meer et al., 2024), with potential applications in areas such as investor relations, corporate communication, and financial analysis.

2 Related Work

2.1 The Request of Confirmation of Inference: An argumentative pattern in ECCs

The ROCOI (previously introduced and qualitatively studied by Rocci and Raimondo, 2018) is an argumentative pattern in ECCs that is originated in question units. It is *relevant* to the discussion in the sense that it creates an argumentative confrontation (van Eemeren and Grootendorst, 2004).

A ROCOI is an assertive question, i.e., in which a stance is asserted by the questioner. As a consequence, when it is formulated directly, a ROCOI is a closed question. Moreover, ROCOIs make explicit by lexical means the fact that the stance asserted is the result of an inferential process, the conclusion of which is expected to be (dis)confirmed by the interlocutor. This results in the ROCOI being a challenging question, regardless of the degree of semantic indirectness of its formulation.

Example 1 shows some ROCOIs; underlined, the lexical elements that indicate the inferential process, which constitutes the keystone of the pattern.

(1) a. Does that mean that customers are reluctant to term out these sort of prices?

b. Should we think of the capital commitment has a hard cap now.

c. Is it fair to say that you've maxed out on what was pre-approved at the AGM and that any incremental issue from here would require AGM approval?

Previous studies on the ROCOI (Rocci and Raimondo, 2018; D'Agostino and Rocci, 2024) identify subcategories of the pattern. This article considers the class that D'Agostino and Rocci (2024) call Type 1, that is, ROCOIs in which the inferential conclusion–of which the questioner asks confirmation–is part of the interrogative unit, as shown in all questions of Example 1. On the other hand, Type 2 ROCOIs correspond to patterns in which the questioner's inference is not explicitly part of the interrogative sentence, such as in the following example (ROCOI in italics): "And looking at the take rate in the fourth quarter, might have slowed down a little bit. *So just – is that true?*". The reason behind the choice of experimenting with Type 1 only is twofold: on the one hand, Type 1 ROCOIs are more compact, in the sense that the conclusion and the question pertain to the same

unit, and are therefore more easily identifiable; on the other hand, they are the most frequent ones.

Easiness in identification should not be mistaken with triviality of the task, and so the retrieval of this pattern cannot be simply achieved via rule-based search of key-phrases such as those emphasized in Example 1. The reason is twofold. The typical lexical signals that indicate the presence of a ROCOI belong to the domain of knowledge management (epistemicity and evidentiality) (Musi and Rocci, 2017; Miecznikowski, 2020; Lucchini et al., 2024); however, while not all ROCOIs display them, these indicators also have a much wider scope of application than introducing this question type. Conversely, the retrieval of such key-phrases does not ensure the extraction of the entire pattern, as it does not provide any indication about its extension–which is not predefined.

3 Method

We outline the dataset, task formulation, and evaluation setup for the ROCOI extraction approaches.

3.1 Dataset

Our work focuses on a dataset that comprises 60 Earnings Conference Calls (ECCs) between 2020–2023 for companies Airbnb (ABNB), British Petroleum (BP), Credit Suisse (CS), Door Dash (DASH), Hasbro (HAS), Shell (SHEL), Exxon Mobil (XOM) and Zillow (Z), for a total of 1377 question units. Manual annotation identified 180 question units featuring ROCOIs, in total containing 193 unique ROCOI patterns. Of these, 134 were Type 1 ROCOIs, and thus represent the final corpus for this study.

The annotation was first carried out by trained student assistants. Annotators were MA students selected on the basis of their joint background in linguistics/languages and financial communication. Whereas financial literacy supports domain knowledge and text comprehension, higher impact in annotation quality is credited to linguistic awareness: ROCOIs are, in fact, a linguistic phenomenon that happens to be frequent in this context, but whose form is not influenced by the content.

Each document was analyzed by two to four annotators. The agreement on the request type labeling task (for which the ROCOI is one out of eight possible values) is $\alpha = 0.79$ (Krippendorff, 1970)[1]; agreement on ROCOI span length is $\Gamma = 0.52$ (Mathet et al., 2015). Two PhD students subsequently curated the annotation until reaching a shared gold standard. This was followed by an additional round of dictionary-based search of (potential) remaining instances, performed by the first contributor of the current paper. Further information about the annotation guidelines for request types is provided in Lucchini and D'Agostino (2023, p. 15-19); ROCOI type classification is borrowed from D'Agostino and Rocci (2024). In total, 18% of tokens in the dataset are part of a ROCOI, whereas 82% of tokens are non-ROCOI tokens. At the present stage and to the best of our knowledge, this is the most extensive collection of manually annotated ROCOIs.

3.2 Task formulation

We compare two task formulations for ROCOI extraction: (1) sequence labeling and (2) text generation, applied to interrogative units that were previously identified as exhibiting the pattern in question. These two tasks allow us to compare the results obtained from applying a *classification* and a *generation* paradigm. Classification, where we mark the boundaries between the presence and absence of a ROCOI, represents the standard method of identifying a substructure. However, such an approach usually requires ample training data. In contrast, text-to-text extractive generation, which involves generating the part of the input text that contains the ROCOI pattern, is similar to more recent state-of-the-art LLMs. We aim to investigate which approach works better given our relatively small dataset. We describe each task formulation separately and provide details about hyperparameters and training settings for all models in Appendix A. The results of both these experiments are compared against an LLM-generated baseline (decoder-only architecture), obtained by prompting the GPT-4o API. Seven-shot in-context learning was adopted as prompting technique for sequence labeling, five-shot for text-to-text generation.[2]

(1) Sequence labeling Sequence labeling for ROCOI extraction is formulated as a task whereby we mark the boundaries between the presence and ab-

[1]Disagreements relate to the selection of a different request type, mostly biased by the *content* of the inference: a ROCOI may be tagged as a request for opinion if, for instance, the inference is about an opinion that the management may hold. No trends were found upon disagreement analysis.

[2]Models GPT-4o, GPT-4o-mini, and GPT-4.1 were compared across 0-, 3-, 5-, and 7-shot contexts; reported as "LLM baseline" is the combination that performed best on average across metrics for the task.

sence of the pattern. Each token is classified as whether pertaining to the sequence (tags "B" and "I"), or not (tag "O"). Such a format represents the standard method of identifying a substructure in a text–for instance, for Named Entity Recognition (NER). Unlike traditional NER, we only consider one type of pattern and thus do not need to specify the class to which the tags pertain.

We experiment with 5 open-source models in total; three of those are encoder-only models:

TinyBERT The smallest model to gauge task complexity. If the smallest model can learn it well, we do not need to train a more capable model (Jiao et al., 2020).

Vanilla BERT Since it is commonly used as a baseline (Devlin et al., 2019).

SpanBERT As a version of BERT that is optimized to represent spans of text, since ROCOIs are often single contiguous spans (Joshi et al., 2020).

In addition, we also experiment with two encoder-decoder models:

T5 Strong empirical results indicate that this model may be used across contexts and tasks (Raffel et al., 2020).

FlanT5 Updated version of T5 that includes a wider array of tasks, the model may generalize better to unseen tasks (Longpre et al., 2023).

(2) Text-to-text generation For this task, the pattern is considered a substring of the question unit given as an input; hence, the output corresponds to a *verbatim* generation of a portion of the wider unit (similar to the use of the text-to-text architecture already intended by Raffel et al. (2020)). Therefore, particular attention must be devoted to the quality of the generation and, specifically, that the fine-tuned model reports an exact portion of the original text (and not, for instance, a summarization of it) and learns that a pattern is a continuous sequence within the text.

This portion of the study is carried out on two text-to-text model families:

BART serves as the encoder-decoder counterpart to our BERT baseline for sequence labeling. We use the base and large varieties (Lewis et al., 2020) to further investigate the impact of model size.

T5 in the small, base, and large varieties, again to see whether a more versatile text-to-text training procedure benefits performance (Raffel et al., 2020).

3.3 Evaluation

We outline how we evaluate models on each task formulation.

3.3.1 Sequence labeling

We initially aimed to adopt a similar evaluation approach as Named Entity Recognition (NER), as it shares the IOB tagging setup (Li et al., 2020). Performance in NER and similar tasks is traditionally evaluated at the **token level** (Tjong Kim Sang and Buchholz, 2000). However, tagging is typically performed (a) on short sequences, (b) in multiclass classification, and (c) featuring multiple units in a text; none of these characteristics strictly hold for ROCOIs. Even in the NER extraction domain, however, there has been a propensity towards evaluation at the full entity level, especially if the prediction is aimed at downstream tasks (Segura-Bedmar et al., 2013). Since ROCOIs are long and complex spans of text with potentially variable boundaries, we additionally adopt **span-level evaluation** and compare it to individual token-level evaluation.

Token-level evaluation At the token level, we first provide an overview of the accuracy in the prediction by individual tags ('O', 'I', 'B'). Then we aggregate the tags and provide a measure of precision, recall, and F1 score, alongside the calculation of token-based Krippendorff's α (Krippendorff, 1970).

Span-level evaluation To evaluate the entire span over which the ROCOI develops and not only the individual tokens that constitute it, we make use of the ROUGE-L metric, to determine the longest matching string, as well as the Gamma (Γ) method for inter-annotator agreement measure and alignment (Mathet et al., 2015)[3] in a basic, one-label, positional dissimilarity detection configuration.

3.3.2 Text-to-text generation

For the text-to-text generation evaluation, we use various metrics to investigate the quality of the extracted pattern. Each model is evaluated according to six metrics, clustered into three classes, each of which corresponds to a different way of interpreting the nature of the task: *syntactic* (pattern matching), *semantic* (embedding similarity), or *annotation* (inter-annotator agreement). The rationale behind such a three-fold choice lies in the nature of generative models: on the one hand, they tend to

[3]Taken from the Python library `pygamma-agreement` (`https://github.com/bootphon/pygamma-agreement`)

be too creative despite being nudged to extract verbatim text. This would not be captured by semantic metrics but is counterbalanced by syntactic metrics. On the other hand, syntactic evaluation cannot capture whether some slightly shifted boundary still correctly identifies the core of the pattern–which can however be reintegrated into the equation to some extent by the use of semantic similarity (although not entirely, since such metrics are not specialized in ROCOI *core meaning* detection, similar to sequence labeling). Inter-annotator agreement metris works as a sanity check that decidedly signals the presence of ill-formed sequences in generated patterns.

Syntactic evaluation In this view, the extraction performance is evaluated in terms of string matching. The first naïve evaluation that establishes the baseline consists of checking whether the pattern is present in the extracted string. We call this evaluation "pattern matching" and its most obvious flaws are that (a) over-extraction to the point of reporting the entire original string is a hit and (b) even slight under-extraction is a complete miss. The three possible values are 'full match' (if the retrieved string contains exactly the correct pattern), 'partial match' (if the retrieved string contains at least the full correct pattern), and 'no match' otherwise; reported are the frequency distributions across the three classes. This is paired with a more refined version of such an evaluation, that is, the calculation of the ROUGE score (Lin, 2004); specifically the ROUGE-L metric, which identifies the longest co-occurring sequence.

Semantic evaluation In this case, what is evaluated is the semantic distance between the predicted and the actual pattern. This is achieved by (1) calculating a simple Euclidean distance between the embedding representation of the patterns and (2) applying some well-established evaluation methods that are typically used for text generation and summarization: notably (a) BERTScore (Zhang et al., 2020) and (b) Sentence-BERT (SBERT) (Reimers and Gurevych, 2019).[4]

Annotation agreement evaluation The true pattern can be considered a gold standard annotation and the extracted pattern a machine-generated annotation; in this perspective, the two are compared with a tool designed to capture the inter-annotator agreement and the dissimilarity in span boundaries.

[4]SBERT in its base configuration measures cosine similarity.

	Accuracy		
Model	O	I	B
BERT (base)	<u>0.93</u>	<u>0.61</u>	**0.70**
TinyBERT	0.92	0.39	0.40
SpanBERT	0.89	<u>0.61</u>	0.60
T5 (base)	**0.95**	**0.67**	<u>0.65</u>
FlanT5 (base)	0.92	**0.67**	**0.70**
GPT-4o	*0.83*	*0.40*	*0.05*

Table 1: Sequence labeling accuracy by tag. The best models are shown in bold, second best underlined. The LLM baseline is in italic.

In particular, we use the Gamma (Γ) method for inter-annotator agreement measure and alignment (Mathet et al., 2015). The metric cannot compute on instances in which the extracted pattern is not a lexical match to a substring of the input text, and thus tells us that the generated string is ill-formed.

4 Results and discussion

We describe our results after training the models on the two tasks: sequence labeling and text-to-text generation respectively.

4.1 Sequence labeling

Table 1 reports the accuracy values by individual tag. As reported in Section 3.1, 82% of tokens in the dataset are non-ROCOI elements; these are identified by 'O' tags. Therefore, since they represent the most frequent type, as expected 'O' tokens reach a higher accuracy across models. On the contrary, 'B'-type tokens understandably are the least frequent ones in the corpus but its accuracy levels are not far from that of 'I' tokens overall–if not better. It is worth noticing that SpanBERT appears to be performing badly despite being optimized for encoding contiguous spans of texts. It achieves the lowest accuracy on the 'O' tag, indicating it most strongly mislocates ROCOI patterns in the text. The LLM baseline confirms the accuracy trends but uniformly scores lower than any other model. At this stage, the best performing models seem to be the two belonging to the T5 family (both best in two out of three accuracy values), followed by vanilla BERT (second best in two out of three accuracy values).

Further classification results aggregated over the three tag categories are displayed in Table 2, both at the token level (former four columns) and span level (latter two columns). Token-level evaluation

appears to favor FlanT5, which achieves the highest results in three out of four metrics and is second best in the remaining one. Surprisingly, SpanBERT performs below par in full span detection, according to span-level evaluation results, which are instead dominated again by T5 (ROUGE-L = 0.90) and FlanT5 ($\Gamma = 0.63$). To conclude, the performance exhibited by T5, FlanT5, and TinyBERT on sequence labeling at the span level compares with or exceeds human agreement (i.e, $\Gamma \geq 0.52$). This indicates that we may use automatic ROCOI extraction for machine annotation for new samples in the future. However, the machine annotations fail in a way that is not captured by this metric, or disagree with human annotators in novel ways. Hence, we set out to further understand the limitations of the automatic ROCOI extraction approach in Section 5.1.

The LLM baseline confirms weak over both token- and span-level labeling, displaying for most metrics below average to nearly zero agreement. Tag sequences appear to be well formed, but not labeling the pattern correctly; moreover, the returned sequence is shorter than the reference in 98% of cases. This indicates that the model is unfit for the job, even though sequence labeling is a generation task in the linguistic domain–which is supposedly the type of task at which these models excel.

4.2 Text-to-text generation

We present the evaluation results sorted by evaluation approach type (*syntactic*, *semantic*, *annotation*), each of which is presented in a dedicated table.

Table 3 reports *syntactic* evaluation. For both evaluation methods, the two BART models appear to be by far the best-performing ones, particularly the *large* configuration–with best results across all metrics. *Semantic* measures are reported in Table 4. The baseline metric represented by raw Euclidean distance between the true and predicted pattern favors BART models; moreover, both SBERT and BERTScore, again identify BART-large as the best-performing model, reaching $F1 = 0.94$. Similar outcomes are shown in Table 5, which displays surprisingly bad results for the T5 models on the *inter-annotator agreement* metrics. This will be appropriately discussed in Section 5.2.

Different metrics capture different aspects of the ROCOI extraction task in a text-to-text generation setup, For instance, syntactic pattern matching informs us of the capability to lexically overlap with the ground truth patterns, while semantic evaluation allows us to observe how well the model captures the underlying meaning and intent of the ROCOI spans. We observe that BART models achieve good performance along all three dimensions for this task.

It is worth noting that for the text-to-text generation task, in contrast with what previously observed for sequence labeling, the tested models are often outperformed by the LLM baseline; this is especially evident in the semantic and annotation agreement metrics. This practically means that pattern boundaries detection may not be extremely accurate in the majority of cases, but the core content of the reference sequence is often included in the generated one. The good level of annotation agreement, moreover, ensures that the generated text is sufficiently well-formed with reference to the original pattern string.

5 Error analysis

In addition to our previous results, we present a qualitative analysis of the predicted patterns. Specifically, we observe the onset point and length of all extracted patterns in the test set to identify whether models tend to make consistent mistakes.

Further, we also present an overview of the distribution of ill-formed sequences in the prediction. In the sequence labeling task, this corresponds to cases in which a sequence onset is not correctly followed by the next element in the sequence: a 'B' tag is immediately followed by an 'O' (not possible in a well-formed ROCOI). In the text generation task, this corresponds to token sequences that are inconsistent with the original text.

While the results here summarize the findings, Tables 9 and 10 for sequence labeling and text generation respectively–available in Appendix B–report by row the measures over each instance in the test set.

5.1 Sequence labeling

We compare performance between T5 and FlanT5.

As for T5, perfect alignment with the start of the pattern occurs in 60% of cases, while the majority of predicted patterns appear to be shorter than expected (55%). Worth noting is the near-perfect acquisition of the IOB-tagging rules, which is reflected in a single instance of ill-formed sequence.

As for FlanT5, the right starting point is detected in 70% of cases; extraction of exact right length

Model	Token-level Precision	Recall	F1	α	Span-level ROUGE-L	Γ
BERT (base)	0.22	0.25	0.23	0.58	0.87	0.49
TinyBERT	0.09	0.05	0.06	0.37	0.82	0.60
SpanBERT	0.27	**0.30**	0.29	0.51	0.83	0.47
T5 (base)	0.17	0.15	0.16	**0.67**	**0.90**	0.56
FlanT5 (base)	**0.32**	**0.30**	**0.31**	0.61	0.87	**0.63**
GPT-4o	*0.03*	*0.02*	*0.02*	*0.28*	*0.77*	*0.39*

Table 2: Additional results for the sequence labeling approaches. The best models are shown in bold, second best underlined. The LLM baseline is in italic.

Model	Pattern matching Full match	Partial match	No match	ROUGE-L
BART (base)	**0.20**	0.50	0.30	0.63
BART (large)	**0.20**	**0.60**	**0.20**	**0.67**
T5 (small)	0.00	0.45	0.55	0.43
T5 (base)	0.15	0.50	0.35	0.54
T5 (large)	0.00	0.15	0.85	0.31
GPT-4o	***0.40***	*0.44*	***0.16***	***0.72***

Table 3: Syntactic evaluation for text-to-text generation. For pattern matching, results must be read as "the higher the better" for full and partial match, and "the lower the better" for no match. The best models are shown in bold, second best underlined. The LLM baseline is in italic; additionally, baseline results are in bold if their value is equal or better than the best result.

spurts to 30%. However, 100% of predicted patterns contain 1 to 4 ill-formed sequences. If the extraction process was integrated in a pipeline, this would easily result in error propagation.

Following, a test instance misclassified by both models (ROCOI pattern in italics): "And secondly, on U.S. gas, you're very well-positioned with I believe pretty much fully hedged production for this year, but *I'm wondering if at $2 per MCF gas, you're actually starting to see the opportunity to perhaps take away some of the rigs* and refocus them in the Permian where you keep strongly growing the activity. Thank you."

In this example, FlanT5 recognizes three starting points (underlined the tokens corresponding to a 'B' tag in the predicted sequence) and one well-formed sequence roughly corresponding to the true pattern (in bold the tokens corresponding to 'I' tags): "And secondly, on U.S. gas, you're very well-positioned with I believe pretty much fully hedged production for this year, but I*'m wondering if at $2 per MCF gas,* you*'re actually starting to see the opportunity to perhaps take away some of the rigs and refocus them in the* Permian where you keep strongly growing the activity. Thank you.". FlanT5 therefore not only marks multiple onset points, but those may also interrupt ongoing sequences.

In brief, T5 is the most reliable model for onset position prediction (offset mean $= 0.93$); FlanT5 is the best at predicting pattern length (offset mean $= -6.2$), as confirmed by similar length distribution in Figure 2f compared to the gold standard of Figure 2a. For further insight, we refer to the overview of Figure 2 (Appendix B).

5.2 Text-to-text generation

The two varieties of BART models were the best performing across metrics. They show similar behavior and the *large* configuration mostly hits some of the misses of the *base* configuration (cf. Table 10). The right starting point is detected in 45% of cases by BART base, increasing to 60% for BART large. The distribution of predicted lengths was the same across both varieties; this means that the *base* configuration is already powerful enough to pick up such a feature to the best that this model family allows given the quantity of training data available. In conclusion, both BART models learned to identify the start of the pattern in the vast majority of

Model	Euclidean distance	SBERT similarity	BERTScore Precision	Recall	F1
BART (base)	**0.42**	0.07	0.91	0.95	0.93
BART (large)	0.46	**0.08**	**0.92**	**0.96**	**0.94**
T5 (small)	0.46	0.05	0.86	0.93	0.90
T5 (base)	0.59	0.06	0.89	0.94	0.91
T5 (large)	0.54	0.07	0.84	0.90	0.87
GPT-4o	*0.35*	***0.78***	***0.93***	*0.95*	***0.94***

Table 4: Semantic evaluation for text-to-text generation. The best models are shown in bold, second best underlined. The LLM baseline is in italic; additionally, baseline results are in bold if their value is equal or better than the best result.

Model	Γ
BART (base)	**0.56**
BART (large)	0.54
T5 (small)	0.07
T5 (base)	0.26
T5 (large)	—
GPT-4o	***0.69***

Table 5: Annotation agreement evaluation for text-to-text generation. The best models are shown in bold, second best underlined. The LLM baseline is in italic; additionally, baseline results are in bold if their value is equal or better than the best result.

cases; remaining errors, however, greatly diverge from the gold standard and both models tend to considerably overgenerate in the majority of cases (by 77 tokens on average for BART base, 73 for BART large).

T5-large is, conversely, a case of extremely flawed generation: despite all safeguards implemented, none of the retrieved patterns corresponds to a substring of the original text–hence hindering the calculation of the Γ metric in Table 5. For example, compared to the true pattern "Are you suggesting that you could potentially ship to Russia later this year?", the corresponding generation reads: "- And then my follow, as it is in terms of Europe. I just want to clarify that? So this has the potential risk from Russia for approximately 100 million.".

6 Conclusions and future work

This paper introduces a prototypical argumentative pattern that originates in the questions asked during the Q&A sessions of financial dialogues, called the Request Of Confirmation Of Inference (ROCOI). Since argumentation is a pivotal aspect of human communication, the identification and extraction of argumentative patterns is argued to be fundamental in the study of language in interaction. Particularly, given that the identification of argumentative patterns is a challenging yet doable task for trained humans, this study seeks to answer the question of whether language models can perform this task as well.

We adopted two concurrent ML approaches to the extraction of ROCOIs from a wider interrogative unit: sequence labeling and text-to-text generation. The sequence labeling approach, evaluated both at the token- and span-level, shows that FlanT5 is the best-performing model. Qualitative observation of the results, however, marks its outputs as potentially unreliable. T5 is therefore the best-performing model both for accuracy and reliability of the output. The text-to-text generation approach identifies BART-large as the best-performing model across syntactic, semantic, and annotation agreement evaluation measures.

GPT-4o was identified as the state-of-the-art representative for the decoder-only category of language models: appropriate for the task due to its power and despite its limited reasoning abilities–as pattern extraction is not formulated as a reasoning task. The LLM performed poorly in the sequence labeling task in terms of pattern identification, whereas it represents state-of-the-art for extractive generation. The increment in performance, however, does not appear to hold effective positive correlation with model size and its use price. Consequently, we do not deem the LLM baseline as the winning model–especially due to its unreliability across tasks.

In conclusion, this task can be carried out by language models. At the present stage, results suggest that sequence labeling is still the most trustworthy method to approach the task. While results would improve with a larger training dataset, gathering additional samples containing ROCOIs is difficult due to their low absolute frequency (although it represents a relatively frequent argumentative pattern in the ECC context).

Further work may include the insertion of intermediate steps to fine-tune for similar tasks (such as argumentative sequence labeling) before applying them to ROCOI extraction (van der Meer et al., 2022), alongside cross-domain extraction and cross-pattern comparison in extraction performance. Additionally, ROCOI retrieval may enhance current argument mining techniques (D'Agostino, 2025). Type 1 ROCOIs, in fact, always explicitly include the conclusion of a reasoning instance. Even if the rest of the inferential process was omitted from the conversation (i.e., they are enthymemic), the acknowledgment of the ROCOI functions as a placeholder that marks where an inference was drawn in the conversation–thus supporting the retrieval of argumentative instances.

Limitations

Our work has several limitations to consider. While we carefully selected the models for fine-tuning that are open source and accepted baselines among related work in Argument Mining literature, our choice of model architecture remains limited. Further, our relatively limited dataset size affects the generalizability of our results, especially in cases of context shift. Training models with more data or increasing the size of the evaluation set may paint a different image of the relative performance among models. Despite using fixed model checkpoints and consistent dataset splits, we observed that T5's generation outputs exhibit high predictive variability. In addition, we found that FlanT5 has a systematic tendency to overpredict multiple ROCOI spans within individual samples, potentially inflating our metrics.

Consideration must also be given to the inherent limitations in the formulation of the current task. All instances fed to the models did contain at least one ROCOI by design–as the experimental setup assumes the availability of candidate question units, and considers their identification an upstream task (see D'Agostino and Rocci, 2024). However, it is true that the current study neither accounts for guardrails against potential error propagation, nor explicitly handles cases that would entail empty generation (alternatively, fully "O" labeling) as the correct output. This can be addressed with the development of a pipeline that performs pattern identification before its extraction.

Lastly, the span length of the gold standard annotations over which a ROCOI develops is not a settled matter–in fact, IAA is only fair with value $\Gamma = 0.52$. The gold standard against which models are tested in this experimental setup is a pairwise expert curation of such annotations until almost perfect agreement was achieved. Additionally, two ROCOI configurations were defined: the *minimal* and *maximal* extension of the pattern–the latter typically including a phrase or sentence that contains a premise to the conclusion that constitutes the core of a ROCOI (e.g., *minimal ROCOI*: "should we assume that Premier Agent revenue growth should be more muted for the remainder of the year?", *maximal ROCOI*: "But given both of those are likely to remain challenges for at least the remainder of 2022, should we assume that Premier Agent revenue growth should be more muted for the remainder of the year?"). Experiments were conducted on both settings; this study only reports on the minimal setting, as it was the one consistently achieving better results. Further studies will also include refinement of the characterization of the maximal ROCOI extension and a comparison of the retrieval of the two varieties.

Ethical Considerations

Recognizing argumentative content can be biased to the content of the training set. This may result in predictions that are poor in novel contexts or edge cases. Responsible implementations of an extraction system, especially in the financial domain, should always be checked by a human. Our work is a first attempt at creating a system for analyzing argumentative patterns for financial dialogues. Situating our approach in an ecosystem that contains checks and balances will not only ensure responsible use of the predictive model but also may yield valuable insights into the actual use of the model.

Supplementary materials availability statement

The dataset on which these experiments were conducted is freely available on GitHub: https://github.com/dagosgi/ROCOIs/tree/main/LARP2025

Acknowledgments

The work in this paper was in part supported by the Swiss National Science Foundation under the project "Mining argumentative patterns in context. A large scale corpus study of Earnings Conference Calls of listed companies" (grant n. 200857).

References

Jianzhu Bao, Chuang Fan, Jipeng Wu, Yixue Dang, Jiachen Du, and Ruifeng Xu. 2021. A neural transition-based model for argumentation mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6354–6364, Online. Association for Computational Linguistics.

Giulia D'Agostino. 2025. *A framework for the large-scale analysis of argumentative patterns in financial discourse*. PhD thesis, Università della Svizzera italiana, Lugano, Switzerland.

Giulia D'Agostino and Andrea Rocci. 2024. Argumentative patterns in the context of dialogical exchanges in the financial domain. In *Proceedings of the 24th Edition of the Workshop on Computational Models of Natural Argument (CMNA 24)*, Hagen, Germany.

Adrian de Wynter and Tangming Yuan. 2024. "I'd Like to Have an Argument, Please": Argumentative Reasoning in Large Language Models. In *Computational Models of Argument*, pages 73–84. IOS Press.

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.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22, Vancouver, Canada. Association for Computational Linguistics.

Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. 2014. Analyzing argumentative discourse units in online interactions. In *Proceedings of the first workshop on argumentation mining*, pages 39–48.

Deniz Gorur, Antonio Rago, and Francesca Toni. 2024. Can large language models perform relation-based argument mining? *arXiv preprint arXiv:2402.11243*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics*, 8:64–77.

Katherine Keith and Amanda Stent. 2019. Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 493–503, Florence, Italy. Association for Computational Linguistics.

Klaus Krippendorff. 1970. Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educational and Psychological Measurement*, 30(1):61–70.

Tatsuki Kuribayashi, Hiroki Ouchi, Naoya Inoue, Paul Reisert, Toshinori Miyoshi, Jun Suzuki, and Kentaro Inui. 2019. An empirical study of span representations in argumentation structure parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4691–4698, Florence, Italy. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2019. Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering*, 34(1):50–70.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.

Costanza Lucchini, Rocci Andrea, and Elena Battaglia. 2024. Epistemic and Evidential Expressions as Context-Specific Argumentative Indicators in Institutional Dialogues: A Corpus Study of Interactions in the Financial Domain. In *Proceedings of the Tenth Conference of the International Society for the Study of Argumentation*, pages 590–603.

Costanza Lucchini and Giulia D'Agostino. 2023. Good answers, better questions. Building an annotation scheme for financial dialogues. Technical report.

Fabrizio Macagno and Douglas Walton. 2014. Argumentation schemes and topical relations. In Giovanni Gobber and Andrea Rocci, editors, *Language, reason and education*, pages 185–216. Peter Lang, Bern.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The Unified and Holistic Method Gamma (Γ) for Inter-Annotator Agreement Measure and Alignment. *Computational Linguistics*, 41(3):437–479.

Johanna Miecznikowski. 2020. At the juncture between evidentiality and argumentation. *Journal of Argumentation in Context*, 9(1):42–68.

Elena Musi and Andrea Rocci. 2017. Evidently epistential adverbs are argumentative indicators: A corpus-based study. *Argument & Computation*, 8(2):175–192.

Andreas Peldszus and Manfred Stede. 2013. From Argument Diagrams to Argumentation Mining in Texts. *International Journal of Cognitive Informatics and Natural Intelligence*, 7(1):1–31.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

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*. Association for Computational Linguistics.

Andrea Rocci and Carlo Raimondo. 2018. Dialogical Argumentation in Financial Conference Calls: The Request of Confirmation of Inference (ROCOI). In *Argumentation and Inference: Proceedings of the 2nd European Conference on Argumentation*, pages 699–715. College Publications.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 Shared Task Chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

Michiel van der Meer, Enrico Liscio, Catholijn Jonker, Aske Plaat, Piek Vossen, and Pradeep Murukannaiah. 2024. A hybrid intelligence method for argument mining. *Journal of Artificial Intelligence Research*, 80:1187–1222.

Michiel van der Meer, Myrthe Reuver, Urja Khurana, Lea Krause, and Selene Báez Santamaría. 2022. Will it blend? mixing training paradigms & prompting for argument quality prediction. In *Proceedings of the 9th Workshop on Argument Mining*, pages 95–103.

Frans van Eemeren and Rob Grootendorst. 2004. *A Systematic Theory of Argumentation: The Pragma-dialectical Approach*. Cambridge University Press.

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

A Experimental details

A.1 Training parameters

We present additional details regarding the usage of pretrained models for the two formulations of the ROCOI extraction. We present an overview of the initial model checkpoints and their parameter counts in Table 6. The hyperparameters to train the models on the sequence labeling task are given in Table 7, and the ones for text-to-text generation are given in Table 8. Training a single model generally takes up to one hour at most on modern hardware (one RTX3090 or A100 GPU).

Model	Checkpoint	Size
BERT (base)	`google-bert/bert-base-uncased`	109M
SpanBERT	`SpanBERT/spanbert-base-cased`	108M
TinyBERT	`huawei-noah/TinyBERT_General_4L_312D`	14M
T5 (base)	`google-t5/t5-base`	110M
Flan-T5 (base)	`google-t5/flan-t5-base`	110M
BART (base)	`facebook/bart-base`	139M
BART (large)	`facebook/bart-large`	406M
T5 (small)	`google-t5/t5-small`	61M
T5 (base)	`google-t5/t5-base`	223M
T5 (large)	`google-t5/t5-large`	738M

Table 6: Description of each model and the specific checkpoint we used.

Sequence labeling For the sequence labeling models, we train on the training set (75% of total available samples) while observing metrics on a validation set (10% of samples). We pick the model iteration with the highest token-level F_1 score and evaluate that model on the test set (15% of samples) to obtain the results reported in Tables 1 and 2. We use the same split for each experiment.

Model	Parameter	Value
BERT (base)	learning rate	2e-05
SpanBERT	learning rate	2e-05
TinyBERT	learning rate	2e-05
T5 (base)	learning rate	4e-04
Flan-T5 (base)	learning rate	4e-04
all	batch size	16
all	max sequence length	256
all	max epochs	100

Table 7: Hyperparameters for the sequence labeling approaches.

Text-to-text sequence generation For the text-to-text generation models, we train on the training set (75% of total available samples) while observing metrics on a validation set (10% of samples). We optimized hyperparameters and picked the best model iteration with the lowest loss value, and evaluated that model on the test set (15% of samples) to obtain the results reported in Tables 3, 4, and 5. We use the same split for each experiment.

B Error analysis

We present additional details upon which we based our qualitative observations of Section 5. Particularly, we display the the raw numerical data for each test instance, which in the body of the paper was instead merged in the form of percentage over the total. Table 9 refers to the sequence labeling task and reports

Model	Parameter	Value
all	learning rate	6e-06
BART *all*	batch size	4
T5 *(all)*	batch size	6
all	max sequence length	256
all	max epochs	100

Table 8: Hyperparameters for the text-to-text approaches.

begin- and length- offsets of the predicted patterns with respect to the gold standard, alongside the number of ill-formed sequences in the tag sequence. Table 10 presents begin- and length- offset numbers only, from the text-to-text generation task. Finally, Figure 2 displays the differences in predicted ROCOI lengths across models for the sequence labeling approach, compared to gold standard.

begin offset	length offset	ill-formed sequences
0	-2	0
n.a.	n.a.	0
0	0	0
0	0	0
n.a.	n.a.	1
0	-5	0
0	-9	0
0	-33	0
0	-2	0
0	0	0
-41	-16	0
n.a.	n.a.	0
0	-34	0
n.a.	n.a.	0
0	3	0
1	-41	0
0	3	0
73	-5	0
-18	-6	0
0	-15	0

(a) T5 (base)

begin offset	length offset	ill-formed sequences
0	-1	4
0	0	4
0	0	3
0	0	1
0	10	2
0	0	3
0	4	2
0	-33	3
0	0	2
0	0	3
n.a.	n.a.	2
0	-1	2
0	-34	3
n.a.	n.a.	2
0	5	3
1	-14	3
12	-5	2
69	-1	2
-18	-6	2
0	-26	4

(b) FlanT5 (base)

Table 9: Qualitative error analysis: sequence labeling approach. Reported the two best performing models. For each sub-table, the first two columns indicate offsets (predicted-true) and the third one indicates the absolute number of instances. The best value is zero for all features.

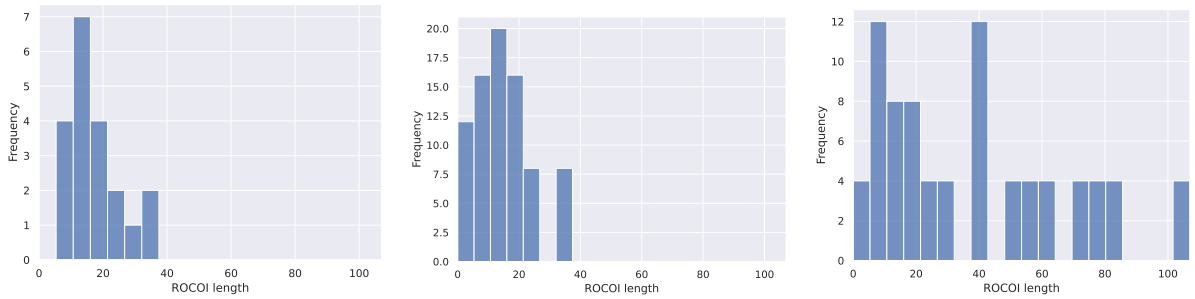
begin offset	length offset
-218	187
0	0
n.a.	n.a.
0	0
0	78
158	5
0	71
-47	47
0	0
-160	77
-179	34
-196	196
0	0
n.a.	n.a.
0	33
n.a.	n.a.
0	70
-4	87
-74	74
0	38

(a) BART (base)

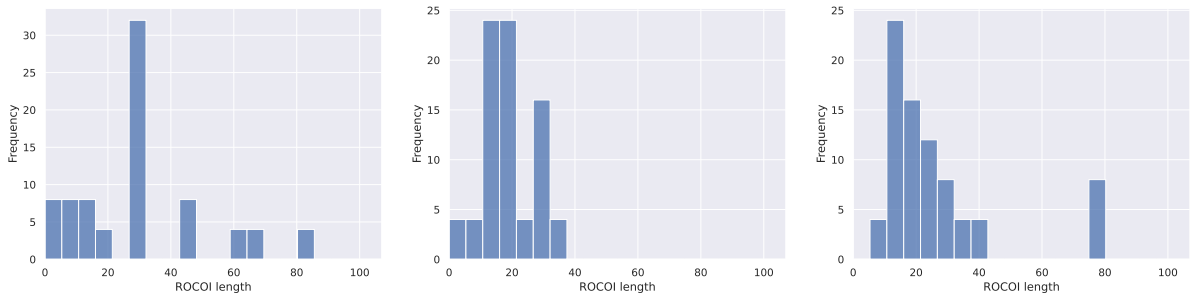
begin offset	length offset
0	182
0	228
n.a.	n.a.
0	0
0	78
158	5
0	22
-47	47
0	0
0	35
-179	34
0	0
0	0
-186	85
0	33
n.a.	n.a.
0	70
-4	87
n.a.	n.a.
0	38

(b) BART (large)

Table 10: Qualitative error analysis: text-to-text sequence generation approach. Reported the two best performing models. For each sub-table, the two columns indicate offsets (predicted-true). The best value is zero for all features.



(a) Labeled data (b) TinyBERT (c) BERT



(d) SpanBERT (e) T5 (f) FlanT5

Figure 2: Error analysis: sequence labeling approach. True (upper left) and predicted (others) ROCOI lengths.