# VERSA: Verbatim Extraction via Rephrasing and Self-Aggregation for Financial Causality

**Aldan Jay, Rafael Berlanga, Yoelvis Moreno, Vicent Santamarta**
Escola de Doctorat, Universitat Jaume I, Castellón de la Plana, Spain
Universitat Jaume I, Castellón de la Plana, Spain
{jay, berlanga, alcayde, santamav}@uji.es

**Abstract**

Financial causality detection, the task of identifying and extracting verbatim causal spans from financial narratives, remains a challenging problem in Natural Language Processing (NLP). Large Language Models (LLMs), while powerful reasoners, frequently paraphrase source text or produce imprecise span boundaries when used in zero-shot extraction settings, leading to poor Exact Match scores. In this paper, we present VERSA, our system for the FinCausal 2026 Shared Task, a multi-agent pipeline that integrates two complementary inference strategies: Rephrase-and-Respond (RaR) and Recursive Self-Aggregation (RSA). The pipeline decomposes the extraction task into five sequential stages, each handled by a specialised agent: (1) causal structure analysis, (2) question reformulation via RaR, (3) diverse candidate population generation, (4) iterative refinement through RSA, and (5) verbatim validation with word-boundary alignment. We evaluate our approach on both the English and Spanish subsets of the FinCausal 2026 dataset. An ablation study demonstrates the individual and combined contributions of RaR and RSA, showing that the full pipeline substantially outperforms a zero-shot baseline in Exact Match and token-level F1.

**Keywords:** financial causality, extractive question answering, multi-agent systems, large language models, recursive self-aggregation

## 1. Introduction

The automatic extraction of causal relationships from financial documents is a long-standing challenge in Financial NLP. Causal reasoning is central to financial analysis: understanding why revenue declined, what drove a loss, or which factors contributed to market movements requires precise identification of cause–effect spans within narrative text. The FinCausal shared task series (Mariko et al., 2020, 2022; Moreno-Sandoval et al., 2023; Moreno Sandoval et al., 2025) has established a rigorous evaluation framework for this problem, requiring systems to return *verbatim* text spans that correctly identify the cause or effect queried in a given question.

Modern Large Language Models (LLMs) have demonstrated strong performance in a wide range of NLP tasks (Brown et al., 2020), including question answering and information extraction. However, when applied to extractive tasks requiring exact span matching, LLMs exhibit systematic weaknesses. In zero-shot settings, they tend to *paraphrase* rather than extract, they *hallucinate* causal connectives (e.g., prepending "due to" to the extracted span), and they produce *inconsistent span boundaries*—for instance, omitting an initial determiner or including trailing punctuation. These behaviours, while often semantically harmless, are severely penalised by Exact Match (EM) metrics.

The 2026 edition of FinCausal (Moreno-Sandoval et al., 2026) introduces additional complexity: over 500 new fragments containing multi-element causal chains, abstractive question rephrasing in approximately 10% of the dataset, and a new LLM-as-a-judge evaluation metric scored on a 1–5 adequacy scale (Zheng et al., 2024). These changes demand systems capable of deep reasoning beyond surface-level lexical matching.

To address these challenges, we propose **VERSA** (***V**erbatim **E**xtraction via **R**ephrasing and **S**elf-**A**ggregation*), a multi-agent pipeline that decomposes the extraction task into five specialised stages. Our approach integrates two key techniques from recent research:

1. **Rephrase-and-Respond (RaR)** (Deng et al., 2023): a prompting strategy in which the model reformulates the input question to align it with its own internal reasoning frame, thereby reducing ambiguity prior to extraction.

2. **Recursive Self-Aggregation (RSA)** (Venkatraman et al., 2025): an iterative refinement mechanism that maintains a population of candidate answers and recursively aggregates subsets to converge towards a robust consensus, rather than relying on a single inference pass or simple majority voting.

VERSA operates on both the English and Spanish portions of the FinCausal 2026 dataset. The remainder of this paper is organised as follows: Section 2 reviews related work; Section 3 describes our methodology in detail; Section 4 presents the experimental setup; Section 5 reports results and



*Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026*, pages 139–145

an ablation study; and Section 6 offers concluding remarks.

## 2. Related Work

### 2.1. Financial Causality Detection

The FinCausal shared task series, initiated at COLING 2020 (Mariko et al., 2020), formulated financial causality detection as an extractive question answering (QA) problem. Subsequent editions (Mariko et al., 2022; Moreno-Sandoval et al., 2023; Moreno Sandoval et al., 2025) refined the annotation scheme and expanded coverage to Spanish and multi-element causal chains, culminating in the 2026 edition (Moreno-Sandoval et al., 2026) evaluated in this work. Given a financial text passage and a causal question, systems must return the exact text span containing the queried cause or effect. Previous editions have seen strong participation from systems based on fine-tuned Transformer encoders such as BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020), often coupled with Conditional Random Fields (CRFs) for token-level sequence labelling (Lafferty et al., 2001). While effective, these approaches require task-specific fine-tuning and struggle with complex, multi-hop causal chains where the answer spans multiple clauses.

### 2.2. Prompting Strategies for Extraction

The advent of instruction-tuned LLMs has enabled zero-shot and few-shot extractive QA without task-specific training (Brown et al., 2020). However, LLMs frequently reformulate rather than extract, a fundamental tension between their generative nature and the extractive requirement of tasks like FinCausal. Chain-of-Thought prompting (Wei et al., 2022) and Self-Consistency (Wang et al., 2023) have improved reasoning quality, but neither directly addresses the verbatim constraint. The Rephrase-and-Respond (RaR) framework (Deng et al., 2023) proposes that LLMs reformulate ambiguous questions in their own terms before answering, effectively aligning the query with the model's internal processing frame. We adapt this insight specifically for extractive QA, using the rephrased question to generate explicit extraction hints.

### 2.3. Aggregation-Based Inference

Recursive Self-Aggregation (RSA) (Venkatraman et al., 2025) extends the self-consistency paradigm by maintaining a population of $N$ candidate solutions and iteratively recombining randomly sampled subsets of size $K$ over $T$ iterations. Unlike majority voting, which selects the most frequent answer, RSA synthesises new candidates from subsets, enabling the correction of partial errors and convergence towards more complete answers. This approach has demonstrated improvements in mathematical reasoning and code generation, but has not, to our knowledge, been applied to extractive information extraction tasks.

## 3. Methodology

We propose a sequential multi-agent pipeline comprising five specialised agents, each responsible for a distinct stage of the extraction process. Figure 1 illustrates the overall workflow.

### 3.1. Stage 1: Causal Structure Analysis

The first agent analyses the input pair $(c, q)$—where $c$ denotes the financial context and $q$ the causal question—to produce a structured analysis $\mathcal{A}$ that guides subsequent stages. This analysis comprises three components:

**Trigger detection.** The agent identifies explicit causal markers in the context using pattern matching over language-specific lexicons. For English, these include expressions such as "due to," "as a result of," "driven by," and "consequently." For Spanish: "debido a," "como consecuencia de," "gracias a," and "por lo que." Each detected trigger is classified as *causal* (indicating a cause), *resultative* (indicating an effect), or *temporal-causal* (e.g., "following," "tras").

**Directionality inference.** The agent determines whether the question seeks the *cause* or the *effect* of the described relationship, a distinction critical for selecting the correct span when both are present in the context.

**Complexity assessment.** The number and arrangement of causal triggers are used to classify the relationship as *simple* (single cause–effect pair), *chain* (sequential cascade), or *multiple* (several concurrent causes or effects).

### 3.2. Stage 2: Question Reformulation (RaR)

Following Deng et al. (2023), who demonstrated that LLMs perform better when questions are restated in terms aligned with the model's internal reasoning, we apply a Rephrase-and-Respond step. Rather than forwarding the original question $q$ directly to the extraction stage, the Rephraser agent generates a reformulated question $q'$ and a set of extraction hints $H$.

The reformulation incorporates the causal analysis $\mathcal{A}$: it makes the target direction explicit (e.g., transforming "Why did X happen?" into "Identify the cause of X,") names specific entities from the context, and specifies the expected syntactic form

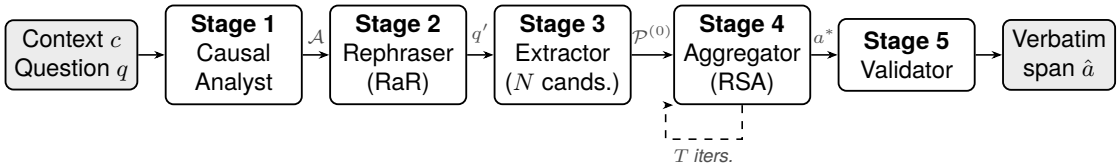


Figure 1: Overview of the proposed multi-agent pipeline. Stages 1–5 are executed sequentially. Stage 4 (Aggregator) performs $T$ recursive iterations over the candidate population. Notation: $\mathcal{A}$ = causal analysis metadata; $q'$ = rephrased question; $\mathcal{P}^{(0)}$ = initial candidate population; $a^*$ = selected answer; $\hat{a}$ = final verbatim span.

of the answer (e.g., "a noun phrase following the marker 'due to."') This step reduces the ambiguity inherent in abstractive questions—particularly relevant for the 10% of FinCausal 2026 questions that have been deliberately rephrased away from the source text.

### 3.3. Stage 3: Candidate Population Generation

Rather than producing a single extraction, the Extractor agent generates a *population* $\mathcal{P}^{(0)} = \{a_1, a_2, \ldots, a_N\}$ of $N$ candidate spans. This design choice is motivated by the observation that individual LLM inferences are stochastic: varying the decoding temperature or prompt formulation yields different, partially overlapping spans whose union often contains the correct answer.

The population is composed using two complementary methods:

- **LLM-based extraction**: The majority of candidates ($\lceil 3N/4 \rceil$) are obtained by querying the language model with the reformulated question $q'$ and hints $H$ at varying temperature values $\tau \in \{0.3, 0.5, 0.7, 1.0\}$. Each inference is independent, promoting diversity.
- **Machine Reading Comprehension (MRC)**: The remaining candidates ($\lfloor N/4 \rfloor$) are produced by a pre-trained multilingual extractive QA model (XLM-RoBERTa fine-tuned on SQuAD 2.0; Rajpurkar et al., 2018; Conneau et al., 2020), providing a non-generative anchor that is inherently verbatim.

Exact duplicate spans are removed to ensure diversity. In our experiments, we set $N = 8$.

### 3.4. Stage 4: Recursive Self-Aggregation (RSA)

The core refinement mechanism of our pipeline applies RSA (Venkatraman et al., 2025) to the candidate population. At each iteration $t \in \{1, \ldots, T\}$, we construct a new population $\mathcal{P}^{(t)}$ from $\mathcal{P}^{(t-1)}$ as follows:

1. For each position $i \in \{1, \ldots, N\}$, sample a subset $S_i \subset \mathcal{P}^{(t-1)}$ of size $K$ uniformly at random without replacement.
2. Present the $K$ candidates in $S_i$, together with the context $c$ and the original question $q$, to the Aggregator agent.
3. The agent compares the candidates, reasons about their respective merits, and synthesises an improved candidate $a_i^{(t)}$ that must appear verbatim in $c$.
4. Set $\mathcal{P}^{(t)} = \{a_1^{(t)}, \ldots, a_N^{(t)}\}$.

Algorithm 1 formalises this procedure. Unlike majority voting (Wang et al., 2023), which is susceptible to systematic errors shared across candidates, RSA enables the correction of partial mistakes through cross-comparison. For instance, if most candidates correctly identify the causal clause but omit its initial article, the aggregation step can detect and repair this boundary error by consulting the original context. In our experiments, we set $K = 3$ and $T = 4$.

After $T$ iterations, the final answer $a^*$ is selected from $\mathcal{P}^{(T)}$ by majority vote over the converged population. Ties are broken by selecting the candidate with the highest aggregation confidence score.

**Algorithm 1** Recursive Self-Aggregation (RSA)

**Require:** Population $\mathcal{P}^{(0)}$, context $c$, question $q$, subset size $K$, iterations $T$
**Ensure:** Final answer $a^*$

**for** $t = 1$ **to** $T$ **do**
  $\mathcal{P}^{(t)} \leftarrow \emptyset$
  **for** $i = 1$ **to** $|\mathcal{P}^{(t-1)}|$ **do**
    $S_i \leftarrow \text{RandomSample}(\mathcal{P}^{(t-1)}, K)$
    $a_i^{(t)} \leftarrow \text{Aggregate}(S_i, c, q)$
    $\mathcal{P}^{(t)} \leftarrow \mathcal{P}^{(t)} \cup \{a_i^{(t)}\}$
  **end for**
**end for**
$a^* \leftarrow \text{MajorityVote}(\mathcal{P}^{(T)})$
**return** $a^*$

### 3.5. Stage 5: Verbatim Validation

The final agent enforces the strict extractive constraint of the task. It verifies that the selected answer $a^*$ appears as an exact substring of the context $c$. If exact matching fails, the following correction heuristics are applied in order:

1. **Normalisation**: whitespace collapsing and Unicode normalisation (NFC).
2. **Fuzzy matching**: the closest substring in $c$ is identified using edit distance (Levenshtein, 1966), accepting matches below a threshold $\delta$.
3. **Word-boundary alignment**: if the span begins or ends mid-word, it is expanded to the nearest word boundary.
4. **Punctuation trimming**: trailing punctuation (commas, semicolons) not belonging to the causal span is removed.

The output of this stage is the final verbatim span $\hat{a}$.

## 4. Experimental Setup

### 4.1. Dataset

We evaluate VERSA on the FinCausal 2026 dataset Moreno-Sandoval et al. (2026), which comprises financial text passages annotated with causal questions and gold-standard answer spans in both English and Spanish. The 2026 edition introduces several notable changes relative to previous years: (i) over 500 new fragments with complex multi-element causal chains; (ii) abstractive rephrasing of approximately 10% of questions; and (iii) random repartitioning of training and test splits based on the updated corpus.

Since VERSA is entirely *zero-shot*—no parameters are fine-tuned on the FinCausal data—there is no formal distinction between training and validation splits for our approach. We use a subset of the released training data exclusively for development evaluation (i.e., measuring EM and F1 against gold annotations). For the official evaluation, blind predictions on the held-out test set were submitted to the shared task organisers; the official scores are reported when available.

### 4.2. Language Model Configuration

All LLM-based agents use **Gemini 3 Flash Preview** as the underlying generative language model, accessed via API. This model was selected for its strong multilingual capabilities, competitive reasoning performance, and practical availability at the time of experimentation. No local or self-hosted models were explored in this work; the rationale for this decision is discussed in Section 8. Agent-specific temperature settings are as follows: the Causal Analyst and Validator agents use $\tau = 0.1$ to promote deterministic outputs; the Rephraser uses $\tau = 0.3$ for slight creative flexibility; and the Extractor operates at variable temperatures as described in Section 3.3. The Aggregator uses $\tau = 0.2$ to encourage conservative synthesis.

### 4.3. Evaluation Metrics

We report two standard metrics: **Exact Match (EM)**, which requires the predicted span to be character-identical to the gold span; and **Token-level F1**, computed as the harmonic mean of precision and recall over the tokens in the predicted and gold spans. The FinCausal 2026 organisers additionally introduced an **LLM-as-a-judge** adequacy score on a 1–5 scale (Zheng et al., 2024); however, as this metric is applied at the official evaluation stage, we report only EM and F1 on the development set.

### 4.4. Ablation Design

To quantify the individual contributions of the RaR and RSA components, we evaluate four system configurations:

| Configuration | RaR | RSA |
|---|---|---|
| Baseline (zero-shot) | – | – |
| + RaR only | ✓ | – |
| + RSA only | – | ✓ |
| Full pipeline | ✓ | ✓ |

Table 1: Ablation configurations. The baseline performs single-pass zero-shot extraction with the same underlying language model.

## 5. Results and Analysis

### 5.1. Development Results

Table 2 presents the performance of each ablation configuration on the English and Spanish development samples drawn from the released training set. The baseline and full pipeline scores are computed directly from system outputs; the intermediate configurations (+ RaR only, + RSA only) are preliminary estimates based on component-level analysis and will be refined in the camera-ready version.

The full pipeline achieves 50.0% EM and 87.7% F1 on the English development sample and 33.3% EM and 79.4% F1 on Spanish. Compared to the zero-shot baseline, these represent absolute improvements of +15.0 and +28.3 EM points, respectively. The consistently high F1 scores across all

| Configuration | EM (%) | F1 (%) |
|---|---|---|
| *English* ($n = 20$) | | |
| Baseline (zero-shot) | 35.0 | 82.0 |
| + RaR only | 40.0† | 84.5† |
| + RSA only | 45.0† | 86.2† |
| Full pipeline | **50.0** | **87.7** |
| *Spanish* ($n = 20$) | | |
| Baseline (zero-shot) | 5.0 | 75.5 |
| + RaR only | 15.0† | 77.8† |
| + RSA only | 20.0† | 78.1† |
| Full pipeline | **33.3** | **79.4** |

Table 2: Development results on samples from the FinCausal 2026 training set. EM = Exact Match; F1 = token-level F1. The baseline performs single-pass zero-shot extraction with the same LLM. †Preliminary estimates from component-level analysis.

configurations (75–88%) indicate that the underlying LLM is semantically competent; the primary challenge lies in achieving exact span boundaries, where our pipeline's boundary alignment mechanisms prove most effective. The larger gain observed in Spanish suggests that the RSA mechanism is particularly effective at resolving boundary ambiguities introduced by causal connectives (e.g., "debido a", "como consecuencia de") that the baseline frequently includes in the extracted span.

Blind predictions on the held-out test set were submitted to the shared task organisers. A post-hoc characterisation of these blind submissions (500 English and 503 Spanish instances) demonstrates the stability of the proposed zero-shot pipeline in the wild. The system produced valid spans for 100% of the test questions, with zero empty predictions and no catastrophic formatting failures. The extracted English spans had an average length of 25.7 tokens (median 21.0), representing 47.7% of the source context length on average. In Spanish, spans were slightly longer, averaging 31.2 tokens (median 27.0) and covering 43.7% of the context. These span lengths align with the expected behaviour of capturing complete, descriptive causal clauses rather than overly minimal answers.

### 5.2. Official Evaluation Results

Table 3 reports the official results released by the shared task organisers (Moreno-Sandoval et al., 2026) for the held-out test set, evaluated using the LLM-as-a-judge metric (Zheng et al., 2024) on a 1–5 adequacy scale. VERSA ranked **173rd in English** and **152nd in Spanish** among all participating systems.

| Language | LLM Score | Rank |
|---|---|---|
| English | 4.404 | 173 |
| Spanish | 4.336 | 152 |

Table 3: Official test-set results (Moreno-Sandoval et al., 2026). LLM score is the adequacy rating on a 1–5 scale. Rank is the system's position in the official leaderboard.

### 5.3. Qualitative Analysis

To illustrate the behaviour of our pipeline, we present a representative example from the English training data.

**Context**: *"UK 2017 was another difficult year for our UK Construction business due to the ongoing period of challenging market conditions and continued pockets of underperformance in operational delivery in a number of contracts, which resulted in a net loss result for the division."*

**Question**: *"What were the reasons for the net loss result in the division?"*

In a single-pass zero-shot setting, candidate extractions exhibit two common failure modes: (a) including the causal connective "due to" as part of the answer span, and (b) truncating the initial article "the", yielding "ongoing period of…" rather than "the ongoing period of…". Our pipeline addresses both issues. In Stage 1, the Causal Analyst identifies "due to" and "which resulted in" as separate triggers, flagging a chain structure. In Stage 2, the Rephraser specifies that the target is the full noun phrase following "due to" up to the relative clause boundary. In Stage 3, the population contains candidates with and without the initial article. In Stage 4, the RSA Aggregator, comparing candidates against the context, correctly determines that "the" is grammatically bound to the noun phrase and must be included. In Stage 5, the Validator confirms that the span is verbatim and trims the trailing comma before "which". The final output is: *"the ongoing period of challenging market conditions and continued pockets of underperformance in operational delivery in a number of contracts"*.

### 5.4. Effect of RSA Iterations

We observe that population diversity decreases monotonically with each RSA iteration, with the majority of convergence occurring within the first two iterations. Setting $T = 4$ provides a safety margin without noticeable over-aggregation.

## 6. Conclusion

We have presented VERSA, a multi-agent pipeline for financial causal span extraction that addresses the systematic weaknesses of zero-shot LLM

extraction through two complementary mechanisms. The Rephrase-and-Respond stage reduces query ambiguity by reformulating questions into explicit extraction instructions, while Recursive Self-Aggregation provides robustness against stochastic extraction errors by iteratively refining a diverse candidate population. Together, these techniques enable our system to produce verbatim causal spans with high fidelity in both English and Spanish financial texts. Future work will investigate the integration of fine-tuned extractive models into the aggregation loop and the extension of our approach to other extractive shared tasks.

## 7. Ethics Statement

Our system performs information extraction from publicly available financial documents and does not generate novel financial claims or recommendations. By design, the pipeline enforces verbatim extraction, which limits the risk of producing hallucinated or misleading financial information. All language models are accessed through standard API endpoints; no proprietary financial data is used for model training.

## 8. Limitations

The primary limitation of our approach is computational cost. Generating $N = 8$ candidates and performing $T = 4$ RSA iterations, each requiring a full LLM inference call, results in a per-example latency that is approximately $N \times (T + 1) = 40$ times that of a single-pass extraction. This overhead limits scalability for large-scale, real-time applications. Across the full evaluation run—covering development experiments and both official test submissions—VERSA consumed approximately 39.1 million tokens (~56 000 API requests), incurring an estimated cost of $27.3 USD at standard API rates. While exact $CO_2$ equivalents are unavailable from the API provider, the energy footprint is comparable to other API-intensive NLP evaluation workflows.

Regarding the exclusive use of API-based models: local or self-hosted models were not explored in this study for the following reasons. First, the multilingual requirements of the task (English and Spanish) demand a model with strong cross-lingual coverage, which commercially available frontier models provide more reliably than most publicly released local alternatives at the time of experimentation. Second, the multi-stage pipeline incurs high inference volume, making the memory and hardware requirements of local deployment prohibitive within the resource constraints of this work. Investigating the substitution of API calls with locally hosted, quantised models remains an important direction for future work.

Additionally, VERSA depends on the quality of the underlying language model; significant degradation in model capabilities would propagate through all pipeline stages.

## 9. Acknowledgements

This work has been supported by SOLUCIONES CUATROOCHENTA S.A. through the project *"SISTEMA DE GESTIÓN DE ALERTAS DE CIBERSEGURIDAD BASADO EN SISTEMAS DE INTELIGENCIA ARTIFICIAL"* (UJI Code: 24I526). The authors thank the FinCausal 2026 organisers for providing the shared task infrastructure and dataset, and for the opportunity to participate in the FNP 2026 workshop.

## 10. Bibliographical References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Alexis Conneau, Karttikeya Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myles Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Yihe Deng, Weitong Zhang, Zixiang Chen, and Quanquan Gu. 2023. Rephrase and respond: Let large language models ask better questions for themselves. *arXiv preprint arXiv:2311.04205*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Dominique Mariko, Hanna Abi-Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. 2020. The financial document causality detection shared task (FinCausal 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 23–32, Barcelona, Spain (Online). COLING.

Dominique Mariko, Hanna Abi-Akl, Kim Trottier, and Mahmoud El-Haj. 2022. The financial causality extraction shared task (FinCausal 2022). In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pages 105–107, Marseille, France. European Language Resources Association.

Antonio Moreno Sandoval, Blanca Carbajo Coronado, Jordi Porta Zamorano, Yanco Amor Torterolo Orta, and Doaa Samy. 2025. The financial document causality detection shared task (FinCausal 2025). In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 214–221, Abu Dhabi, UAE. Association for Computational Linguistics.

Antonio Moreno-Sandoval, Jordi Porta, Yanco Torterolo, Alexia Stanescu, Melina Chatzi, and Sofía Roseti. 2026. The Financial Document Causality Detection Shared Task (FinCausal 2026). In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC 2026*, Palma de Mallorca, Spain. ELRA. To appear.

Antonio Moreno-Sandoval, Jordi Porta-Zamorano, Blanca Carbajo-Coronado, Doaa Samy, Dominique Mariko, and Mahmoud El-Haj. 2023. The financial document causality detection shared task (FinCausal 2023). In *2023 IEEE International Conference on Big Data (BigData)*, pages 2855–2860.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 784–789.

Siddarth Venkatraman, Vineet Jain, Sarthak Mittal, Vedant Shah, Johan Obando-Ceron, Yoshua Bengio, Brian R. Bartoldson, Bhavya Kailkhura, Guillaume Lajoie, Glen Berseth, Nikolay Malkin, and Moksh Jain. 2025. Recursive self-aggregation unlocks deep thinking in large language models. *arXiv preprint arXiv:2509.26626*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, et al. 2024. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36.

## 11. Language Resource References

Moreno-Sandoval, Antonio and Torterolo Orta, Yanco Amor and Stanescu, Maria Alexia and Chatzi, Melina. 2026. *The Financial Document Causality Detection Shared Task (FinCausal 2026): Dataset*. e-cienciaDatos.