JaCorpTrack: Corporate History Event Extraction for Tracking Organizational Changes

Yuya Sawada[1,3], Hiroki Ouchi[1,3], Yuichiro Yasui[2], Hiroki Teranishi[1,3], Yuji Matsumoto[3], Taro Watanabe[1], Masayuki Ishii[2],
[1]Nara Institute of Science and Technology [2]Nikkei Inc.
[3]RIKEN Center for Advanced Intelligence Project
{sawada.yuya.sr7,hiroki.ouchi,taro}@is.naist.jp
{hiroki.teranishi,yuji.matsumoto}@riken.jp
{yuichiro.yasui,masayuki.ishii}@nex.nikkei.com

Abstract

Corporate history in corporate annual reports includes events related to organizational changes, which can provide useful cues for a comprehensive understanding of corporate actions. However, extracting organizational changes requires identifying differences in companies before and after an event, raising concerns about whether existing information extraction systems can accurately capture the relations. This work introduces *JaCorpTrack*, a novel event extraction task designed to identify events related to organizational changes. *JaCorpTrack* defines five event types related to organizational changes, and is designed to identify the company names before and after each event, as well as the corresponding date. Experimental results indicate that large language models (LLMs) exhibit notable disparities in performance across event types. Our analysis reveals that these systems face challenges in identifying company names before and after events, and in interpreting event types expressed under ambiguous terminology. We will publicly release our dataset and experimental code at https://github.com/naist-nlp/JaCorpTrack.

1 Introduction

Corporate annual reports are comprehensive summaries of a company's financial performance and corporate actions. These reports are regarded as one of the most important sources to evaluate the financial and business condition of a company, thereby enabling investors to formulate more accurate projections of earnings (Sugiura et al., 2025) and potential risk factors (Huang and Li, 2008; Fujii et al., 2022).

Corporate annual reports include corporate history that provides an overview of a company's activities since its founding. Corporate history includes events on organizational changes, such as mergers and name changes, as shown in Figure 1. Such information provides insights into transitional phases, including rebranding and business reconstruction. Furthermore, this history also includes information about subsidiaries and group companies, covering a broader range of organizational changes over time. The history contributes to a deeper understanding of corporate actions and plays a key role in demonstrating the company's value and credibility to investors.

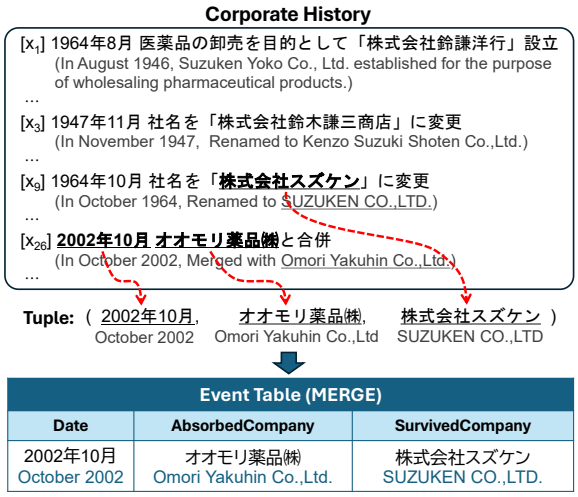

Figure 1: Example of Corporate History Event Extraction

In this study, we focus on extracting temporal events related to organizational changes from corporate histories. Corporate annual reports include corporate history using a timeline format and free-form narratives, and update their history whenever key events occur within a fiscal year. To extract the events, there are two technical challenges regardless of the formats: (1) Identifying relationships between multiple companies: This requires identifying company names both before and after an event. As illustrated in Figure 1, when predicting company names before the event, it is also necessary to take into account any name changes that occurred before that point. (2) Distinguishing events

under ambiguous terminology: While the same type of events can follow different processes, they are sometimes described using the same term regardless of these differences. For example, mergers can involve absorption or consolidation processes, but they are often simply referred to as "merge" in context. Multiple terms are sometimes also used to refer to a single event. Existing event extraction (EE) systems are evaluated on datasets that define business- (Walker et al., 2006) and finance- (Yang et al., 2018; Zheng et al., 2019; Li et al., 2022; Han et al., 2022; Zhang et al., 2024a) related events, and they demonstrate reasonably good performance (Du and Cardie, 2020; Lu et al., 2021; Hsu et al., 2022; Peng et al., 2023b). However, since the event argument roles defined in these datasets do not take organizational changes into account, it remains unclear whether these EE systems can extract temporally consistent events in the face of such challenges.

To investigate these challenges, we propose a novel event extraction task, called *JaCorpTrack*, designed to extract events from Japanese corporate history. In this task, we define five event types associated with organizational changes: `ESTABLISH`, `CHANGE`, `MERGE`, `SPLIT`, and `LIQUIDATION`. Each event is associated with event arguments annotated across 210 corporate histories. For example, `MERGE` notes an event where companies are integrated into one, annotating three arguments: *AbsorbedCompany*, *SurvivedCompany*, and *MergedDate*.

In the experiments, we reproduce several conventional EE systems as baselines and evaluate large language models (LLMs) using both in-context learning and finetuning. The experimental results show that while the EE systems and LLMs can achieve an overall F1-scores of 78.5, there exists a performance gap of up to 61.3 points across different event types, indicating that they still face the challenges. Detailed analysis reveals that these systems often predict incorrect arguments for events involving temporal relations and erroneously detect events in the presence of ambiguous terminology.

2 JaCorpTrack

JaCorpTrack is a task of detecting events related to changes in company names and identifying the companies involved in the events, as well as the corresponding dates. We adopt Japanese corporate histories formalized timelines, an example of this task is illustrated in Figure 1.

Event type	Argument roles
`ESTABLISH`	EstablishedDate, EstablishedCompany
`CHANGE`	ChangedDate, BeforeCompany, AfterCompany
`MERGE`	MergedDate, AbsorbedCompany, SurvivedCompany
`SPLIT`	SplitDate, ParentCompany SeparatedCompany
`LIQUIDATION`	LiquidatedDate, LiquidatedCompany

Table 1: Event Schema in our dataset

2.1 Task Definition

In this work, we formalize JaCorpTrack as an event table filling task, i.e., filling arguments into predefined event tables (Yang et al., 2018; Zheng et al., 2019). Specifically, JaCorpTrack is designed to extract these events in the form of an n-tuple, consisting of argument roles of each event. By extracting tuples from corporate history, we can automatically generate structured tables of events as illustrated in Figure 1. Formally, given a corporate history X comprising n timelines x_i $(1 \leq i \leq n)$ as an input, an EE system is required to extract triplets Y_e $(|Y_e| \geq 0)$ for each event type $e \in \mathbb{E}$, each of which comprises a set of event arguments $\{a_r \mid r \in \mathbb{R}_e\}$ for the argument roles $\mathbb{R}_e$.

2.2 Target Entities

The target entities that can be filled as event arguments in this task fall into two categories: company names and date expressions. Entity mentions for company names and date expressions are defined according to Sekine's Extended Named Entity Hierarchy (Sekine et al., 2002). Detailed definitions of the target entities are provided in Appendix A.

2.3 Event Types and Argument Roles

Table 1 shows the definition of event types and argument roles. In this study, we define five event types: `ESTABLISH`, `CHANGE`, `MERGE`, `SPLIT` and `LIQUIDATION`. (a) `ESTABLISH` refers to an event where a company is established. (b) `CHANGE` specifies an event where a company is renamed to another. (c) `MERGE` notes an event where companies are merged into one. (d) `SPLIT` specifies an event where a company spin-off a part of its businesses as a separate company. Note that `SPLIT` is distinct from a sell-off, where a company sells a section to another company or firm in exchange for cash or securities. (e) `LIQUIDATION` refers to an event where a company is liquidated or dissolved. For `MERGE` and `SPLIT`, multiple companies can be integrated

or divided, thus *AbsorbedCompany* and *SeparatedCompany* can be used more than once. For example, "In April 1969, Kawasaki Aircraft Co., Ltd., and Kawasaki Rail Car Inc. merged to Kawasaki Heavy Industries Ltd." indicates "Kawasaki Heavy Industries Ltd" absorbed two companies, thus the *AbsorbedCompany* is " Kawasaki Aircraft Co., Ltd." and "Kawasaki Rail Car Inc."

3 Dataset Annotation

3.1 Data Collection

We used Japanese corporate annual reports published on EDINET[1] between April 2023 and March 2024. We extracted the corporate history section from the reports and constructed the timelines by removing HTML tags. More details are discussed in Appendix B. In addition, the reports may include timelines with few events of company name change such as newly established companies, or with other companies such as subsidiaries. To eliminate the history, we randomly sampled 210 corporate histories that comprised 30–80 timelines and more than 10 words[2] suggesting that the specific events appear in the context. On average, each document describes about 70 years, with the range spanning from 12 to 151 years.

3.2 Human Annotation

In the annotation step, we hired three annotators and asked them to conduct the annotation process as follows. (a) **Entity Mention**: Identify entity mentions representing company names or date expressions. We employed automatic entity mention annotation tool with with GPT-4o (OpenAI, 2024b) and heuristic postprocessing to reduce the annotation cost[3], and then, corrected or added the annotations if the predictions had boundary errors or omissions. (b) **Event Trigger**: Determine whether timelines are relevant to one or multiple predefined events. If relevant, assign the event type label to the trigger word that best represents the occurrence of an event in timelines.(c) **Event Argument**: Search entity mentions corresponding to the predefined event arguments of each event type. If any of the entity mentions are found, assign an argument role.

[1] https://disclosure2.edinet-fsa.go.jp/WEEK0010.aspx

[2] 設立 (establish), 創立 (found), 合併 (merge), 変更 (rename), 解散 (dissolve), 清算 (liquidate), 独立 (spin off), 分社 (split)

[3] The implementation details are in Appendix C

	Train	Test	Total
#Docs	50	160	210
#Timelines	2,431	7,349	9,780
#Date	2,502	7,582	10,084
#Company	2,094	6,764	8,858
#Events	959	3,265	4,224
#Arguments	2,282	7,986	10,268

Table 2: Statistics of our dataset.

If there are cases where some arguments of an event are missing, we allow to ignore the arguments.

We adopted the brat annotation tool (Stenetorp et al., 2012) for mention and event annotation. For an understanding of the annotation process and event schema, we provided 10 demo samples randomly sampled from the documents in preliminary studies. When it is not possible to determine whether an entity mention corresponds to an event argument based on the context, we consult official company websites, press releases, or relevant news. If an annotator had difficulty making a clear judgement, we asked them to apply temporary labels which were finally reviewed and corrected by ourselves. To assess dataset quality, we re-annotated 20 corporate histories in our dataset and calculated the agreement of event arguments with their annotations using F1 score. The F1-score is 93.6%, indicating a satisfactory level of annotation agreement. More details of the annotation agreement are shown in Appendix D.

3.3 Data Statistics

The main statistics of our dataset are shown in Table 2, in which each document comprises an average of 46.6 timelines with 2,128 characters in Japanese. Compared to existing document-level EE datasets, the average size of timelines is twice the size of CFinDEE (Zhang et al., 2024a), and ten times the size of RAMS (Ebner et al., 2020). The average number of tokens is shown in Appendix B.

The distribution of event types in Table 3 shows that `ESTABLISH` has the highest proportion for event types, appearing in all of the documents. Moreover, `CHANGE` and `MERGE` appear in more than three times the number of documents and about 90% of all documents, indicating that most companies have experienced multiple name changes or mergers. On the other hand, `SPLIT` and `LIQUIDATION` are relatively rare, which is likely because company splits and bankruptcies are primarily carried out by large companies with subsidiaries or group companies.

	Total	Ratio	Average	Distance
ESTABLISH	2,400	100.0	11.4	0.0
CHANGE	799	94.3	4.0	1.7
MERGE	736	89.0	3.9	4.5
SPLIT	136	37.1	1.7	5.9
LIQUIDATION	153	31.9	2.3	0.1

Table 3: Event Type Distribution of our dataset. Ratio represents the ratio of documents containing the event, Average represents the average frequency of each event type in the documents containing the event type, and Distance represents the average distance of sentences between arguments.

4 Experiment

We conducted an empirical study to investigate the performance of existing EE systems and LLMs in JaCorpTrack, using supervised finetuning and in-context learning. As shown in Table 2, we evaluate the models with 160 documents, and use 50 documents as the training set for supervised finetuning or demonstrations for in-context learning. The impact of training data sizes is discussed in Appendix E, indicating that the performance in our task stabilizes around 40 training instances.

4.1 Experiment Setup

Supervised Finetuning We adopt two pipeline event extraction systems, DMBERT (Wang et al., 2019a) and BERT+CRF (Wang et al., 2020). DMBERT and BERT+CRF are BERT-based approaches for token classification and sequential labeling, respectively. For LLMs, we select two open-source LLMs, Llama-3-Meta[4] (Dubey et al., 2024) and Llama-3-Swallow[5] (Fujii et al., 2024; Okazaki et al., 2024), which adopt the 8B parameters models for all series and use LoRA (Hu et al., 2022) for finetuning. The implementation details of the EE systems and LLMs are in Appendix F.1, and the average runtime is reported in Appendix F.3.

In-context Learning We add two closed LLMs, GPT-4o (OpenAI, 2024b) and GPT-4o-mini (OpenAI, 2024a). Since there are documents where no event record appears for event types other than ESTABLISH as shown in Table 3, we embed two demonstration examples in the instruction to illustrate the format for no event. The implementation details are in Appendix F.2.

[4] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[5] https://huggingface.co/tokyotech-llm/Llama-3-Swallow-8B-Instruct-v0.1

Evaluation Metrics We used three evaluation metrics to investigate the challenges for tracking organizational changes: (1) **Event Detection**: A result is considered correct if at least one event argument matches between the correct record and the predicted record. (2) **Event Table Filling**: We count the number of matched arguments between predicted and correct records and aggregate these statistics among all documents.(Yang et al., 2018; Zheng et al., 2019). The correct record is selected as the most similar record to the prediction in each event table. (3) **Event Record Identification**: A result is considered correct if all event arguments match between the correct record and the predicted record. For each metric, we count the number of matched records to calculate the model's precision, recall, and F1 score. Considering the performance fluctuations caused by different seeds, we report the average of the scores with five different seeds.

4.2 Main Results

The experimental results are shown in Table 4. While finetuned Llama-3 and GPT-4o underperformed DMBERT and BERT+CRF in Event Detection and Event Table Filling, the LLMs outperformed them by up to 10.7 points in Event Record Identification. Since DMBERT and BERT+CRF are limited to processing sentences with a maximum length of 512 tokens, the results indicate that they overlooked the arguments that are farther from the trigger. For LLMs, Llama-3 models with in-context learning perform poorly on all metrics. One possible reason for the performance drop is that Llama-3 struggles to understand the event scheme from demonstrations. Previous studies have also reported that LLMs with in-context learning perform poorly in specification-heavy information extraction tasks (Peng et al., 2023a; Li et al., 2023; Han et al., 2024; Wang et al., 2024), indicating this finding is consistent. GPT-4o models achieve performance comparable to finetuned Llama-3 models, however, the performances may rely on memorizing of corporate histories. Further analyses presented in Appendix G demonstrate that GPT-4o's performance declined markedly when companies in the timelines are anonymized.

4.3 Results for Each Event Type

Table 5 shows the performance of LLMs for each event type. While all models perform over 85 points for ESTABLISH, there is a performance gap of more than 11.4 points and up to 61.3 points

	Event Detection			Event Table Filling			Event Record Identification		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
DMBERT (Wang et al., 2019a)	**94.5**	**95.3**	**94.9**	**84.2**	**75.0**	**79.3**	67.5	68.1	67.8
BERT+CRF (Wang et al., 2020)	94.4	91.0	92.6	85.3	67.4	75.3	57.8	55.7	56.7
Llama-3-Swallow (In-context)	23.8	52.5	32.8	8.0	16.7	10.8	16.5	36.4	22.7
Llama-3-Meta (In-context)	39.6	80.0	52.9	19.3	40.5	26.1	28.6	57.9	38.3
Llama-3-Swallow (Finetune)	90.2	86.0	88.0	73.5	70.4	71.9	77.4	73.8	75.6†
Llama-3-Meta (Finetune)	88.9	85.7	87.1	72.0	69.2	70.2	73.3	70.5	71.7†
GPT-4o-mini (In-context)	87.5	77.8	82.3	71.8	63.4	67.3	73.5	65.3	69.1†
GPT-4o (In-context)	91.0	88.0	89.5	75.1	72.1	73.6	**79.9**	**77.2**	**78.5**†

Table 4: Experimental results of LLMs on our dataset. † indicates that the F1-score is significantly higher than the baselines under $p < 0.05$.

	ESTABLISH	CHANGE	MERGE	SPLIT	LIQUIDATION
Llama-3-Swallow (Finetune)	88.0 (90.5)	64.6 (86.6)	55.4 (88.5)	39.3 (61.7)	74.0 (79.9)
Llama-3-Meta (Finetune)	85.2 (89.3)	59.9 (86.6)	49.0 (87.3)	33.4 (60.3)	73.8 (80.6)
GPT-4o-mini (In-context)	86.0 (88.6)	52.2 (76.2)	41.5 (78.1)	24.7 (48.8)	58.3 (66.4)
GPT-4o (In-context)	92.0 (94.9)	60.0 (79.5)	62.4 (89.9)	33.8 (57.3)	74.0 (82.2)

Table 5: F1-score of LLMs for each event type in Event Record Identification. The scores in parentheses indicate the F1-score in Event Detection.

	CHANGE			MERGE			SPLIT		
	Date	*Before*	*After*	*Date*	*Absorbed*	*Survive*	*Date*	*Parent*	*Separated*
Llama-3-Swallow (Finetune)	84.5	75.7	84.5	88.6	81.1	69.1	65.3	58.1	57.3
GPT-4o (In-context)	79.8	68.0	75.0	90.4	80.6	74.0	60.2	60.8	42.8

Table 6: Difference of F1-score for each argument roles in CHANGE, MERGE, SPLIT. *Italic* texts represent the abbreviations of argument roles.

between ESTABLISH and the other event types, suggesting these models still struggle to track organizational changes. Additionally, the results indicate the following findings. (1) LLMs achieve higher scores for ESTABLISH and LIQUIDATION than for CHANGE, MERGE, and SPLIT, suggesting they often fail to identify the arguments of companies before and after a change. (2) Despite the similar record structure, LLMs perform worse for LIQUIDATION than for ESTABLISH. LIQUIDATION appears in only 31.9% of the documents and occurs at approximately one-sixteenth the frequency of ESTABLISH (as shown in Table 3). Given that precision exceeds recall in the overall Event Detection results, this suggests that LLMs are prone to false positives when detecting rare events in long texts. (3) F1-scores for MERGE and SPLIT are lower than those for CHANGE, except for GPT-4o, indicating that LLMs have difficulty identifying events based on the process types. In particular, LLMs perform poorly for SPLIT, suggesting LLMs struggle to distinguish between sell-offs and spin-offs.

4.4 Analysis on Argument Roles

To analyze the behavior of LLMs for identifying event records, we report argument role-level performance. Table 6 shows the performance for CHANGE, MERGE, and SPLIT, indicating the following findings. (1) The F1-scores for *Date* are comparable to the score in Event Detection, suggesting *Date* plays an alternative role to event triggers. (2) For CHANGE and MERGE, the F1-scores of the *Survive* and *Before* are lower than *Absorbed* and *Before*, respectively. *Before* and *Survive* appear far from *Date* than *Survive* and *Before*, indicating LLMs often misidentify the companies when identifying arguments that span across lines. (3) Conversely, for SPLIT, the F1-scores of *Parent*, which appear closer to Date, are better than *Separated*, suggesting LLMs memorize the root of the company names while reading timelines.

4.5 Error Analysis

We conducted an error analysis on the prediction results for SPLIT. Table 7 shows erroneous detection

Text	1939年12月 社名を川崎重工業株式会社と商号変更 (In December 1939, the company changed its name to Kawasaki Heavy Industries, Ltd) 2021年10月 車両事業を分離し、川崎車両株式会社(連結子会社)に承継 モーターサイクル&エンジン事業(現・パワースポーツ&エンジン事業)を分離し、 カワサキモータース株式会社(連結子会社)に承継 (In October 2021, the vehicle business was separated and transferred to Kawasaki Railcar Manufacturing Co., Ltd. (a consolidated subsidiary). The Motorcycle & Engine business (currently the Powersports & Engine business) was separated and transferred to Kawasaki Motors, Ltd. (a consolidated subsidiary)).
Ground-Truth (SPLIT)	None
Llama-3-Swallow (SPLIT)	Date: 2021年10月 (October 2021), Parent: 川崎造船所 (Kawasaki Dockyard Co., Ltd.) Separated: 川崎車両株式会社(Kawasaki Railcar Manufacturing Co., Ltd.)
GPT-4o (SPLIT)	Date: 2021年10月(October 2021), Parent: 川崎造船所 (Kawasaki Dockyard Co., Ltd.) Separated: 川崎車両株式会社(Kawasaki Railcar Manufacturing Co., Ltd.), カワサキモータース株式会社 (Kawasaki Motors, Ltd.)

Table 7: Error Analysis

by Llama-3-Swallow and GPT-4o. This example includes two terms that represent a spin-off and a sell-off, "分離 (separated)" and "承継 (transferred)", which makes identifying them more challenging. Furthermore, the official report on the company has reported the series of procedures as a spin-off [6], indicating the predictions can also be interpreted as correct. Spin-offs can be categorized into two types: (1) a business of the recipient company is transferred to a preparatory company, and (2) the new company is established at the time of completing the procedure. Therefore, SPLIT involves ambiguity due to differences in the spin-off procedures, and distinguishing the process types is a task-specific challenge.

From the above results, we conclude that LLMs still have room for improvement in the three factors: (1) detecting events in long texts, (2) identifying the argument related to the company before and after a change, and (3) identifying arguments based on the type of process.

5 Related Work

5.1 Event Extraction Dataset

Event extraction datasets define event types related to organizational changes in business (Walker et al., 2006) and finance (Yang et al., 2018; Zheng et al., 2019; Li et al., 2022; Han et al., 2022; Zhang et al., 2024a), such as organization establishment, mergers, bankruptcies, and dissolutions. Pioneering work on event extraction (Grishman and Sundheim, 1996) also defines organizational events such as company management successions and mergers, and presents specific template slots to facilitate a deeper understanding of the events, including executive information and the status of a job. Although our event types are similar to these datasets, we specifically designed them and their arguments to track organizational changes in longer documents. The average document length in JaCorpTrack is more than twice that of existing document-level EE datasets (Ebner et al., 2020; Zhang et al., 2024a), and the average distance between event arguments ranges from 1.7 to 5.9 timelines.

5.2 Event Argument Extraction

Event argument extraction models extract arguments with pre-extracted trigger words (Wang et al., 2019b, 2021; Li et al., 2020; Lu et al., 2021; Ma et al., 2022), or without trigger words (Yang et al., 2018; Zheng et al., 2019; Xu et al., 2021; Yang et al., 2021; Zhu et al., 2022; Liang et al., 2022; Wang et al., 2023b; Huang et al., 2023). Recently, leveraging LLMs for event argument extraction has been increasingly paid attention, which has been proposed some approaches (Wang et al., 2023a; Zhou et al., 2024; Sainz et al., 2024; Chen et al., 2024; Zhang et al., 2024b). This study investigated the performance of LLMs without triggers and organized the challenges of the models.

[6] https://www.global-kawasaki-motors.com/en/history/

6 Conclusion

In this paper, we propose an EE task to track the organizational changes. We defined five event types related to organizational changes and described the process of creating a dataset using corporate annual reports. Then, we evaluate the performance of existing EE systems and LLMs using our dataset and demonstrate that tracking organizational changes is still challenging, particularly when they identify company names before and after a change. In the future, we plan to extend the event as supplementary information for the KBs, developing an EL system that automatically updates shifts in target entities.

Limitations

Broaden Entities and Dataset JaCorpTrack focuses on evaluating the performance of tracking entity changes in text, thus, we adopt company names that change frequently over time and corporate histories that contain this information. Defined events in JaCorpTrack can be applied to other named entities such as person and location. For example, people can adopt a new name different from their current name, and geopolitical entities can merge with other geographical entities. In future work, we will try to investigate the name changes for broader entities and other domain texts.

Broaden Corporate Histories We used Japanese corporate annual reports, thus, the scope is limited to Japanese companies. Although the specifics of corporate actions abroad may vary subtly from those of Japanese companies, we believe that our findings are broadly applicable because (i) corporate histories in other countries are similarly long, (ii) organizational changes such as `SPLIT`, `MERGE`, and `CHANGE` commonly occur, and (iii) the nature of such processes depends on each country's corporate law. As the availability of corporate annual reports from other countries remains uncertain, we intend to explore the use of corporate histories in other resources, such as Wikipedia, in future work.

Model Performance on Event Extraction We performed LLMs with simple prompts, and did not investigate how to design a good prompt for event table filling. As shown in Section 5.2, recent event argument extraction models adopt some prompt strategies such as heuristic-driven prompt (Zhou et al., 2024) and code-style prompt (Wang et al., 2023a; Sainz et al., 2024), thus performing experiments with these strategies has potential for further improvement in our task. In addition, other document-level EE systems (Yang et al., 2018; Zheng et al., 2019; Wang et al., 2023b) have been proposed, but we have not shown the results of these models in this paper since these implementations do not work well in preliminary experiments.

Application to Other Tasks We intend to use the event information in our dataset for other tasks, such as Entity Linking, but we have not discussed how these events contribute to enhancing the task performance. In entity linking, we plan to use event information as an edit history of entity knowledge, expecting that leveraging this edit history will improve the accuracy of candidate selection. Developing an entity linking system that leverages events of company name change is a future work.

Ethical Considerations

License We used corporate annual reports published in EDINET. Information made available on EDINET may be freely used, copied, publicly transmitted, translated or otherwise modified on condition that the user complies with provisions of Public Data License (Version 1.0)[7]. Following the provisions, we confirmed that annual securities reports are contents on EDINET to which the terms of use, and we will provide the source citations at `https://github.com/naist-nlp/JaCorpTrack`.

Worker Treatments We provided 70 corporate histories per annotator and asked to conduct the event mention, trigger word, and event argument annotations. All the annotators are men and native Japanese speakers. The age range of annotators is in their 20s, and they hold undergraduate degrees. The payment amount to each annotator was about 8 USD per hour, and the worked time for annotation was about 15 minutes per document. The overall annotation cost is about 480 USD.

References

Ruirui Chen, Chengwei Qin, Weifeng Jiang, and Dongkyu Choi. 2024. Is a large language model a good annotator for event extraction? *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17772–17780.

[7] `https://www.digital.go.jp/en/resources/open_data/public_data_license_v1.0`

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 510 others. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In *Proceedings of the First Conference on Language Modeling*, COLM, page (to appear), University of Pennsylvania, USA.

Motomasa Fujii, Hiroki Sakaji, Shigeru Masuyama, and Hajime Sasaki. 2022. Extraction and classification of risk-related sentences from securities reports. *International Journal of Information Management Data Insights*, 2(2):100096.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Cuiyun Han, Jinchuan Zhang, Xinyu Li, Guojin Xu, Weihua Peng, and Zengfeng Zeng. 2022. Duee-fin: A large-scale dataset for document-level event extraction. In *Natural Language Processing and Chinese Computing: 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24–25, 2022, Proceedings, Part I*, page 172–183.

Ridong Han, Chaohao Yang, Tao Peng, Prayag Tiwari, Xiang Wan, Lu Liu, and Benyou Wang. 2024. An empirical study on information extraction using large language models.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. DEGREE: A data-efficient generation-based event extraction model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1890–1908.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Guanhua Huang, Runxin Xu, Ying Zeng, Jiaze Chen, Zhouwang Yang, and Weinan E. 2023. An iteratively parallel generation method with the pre-filling strategy for document-level event extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10834–10852.

Ke-Wei Huang and Zhuolun Li. 2008. A multilabel text classification algorithm for labeling risk factors in sec form 10-k. *ACM Trans. Manage. Inf. Syst.*, 2(3).

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness.

Fayuan Li, Weihua Peng, Yuguang Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu. 2020. Event extraction as multi-turn question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 829–838.

Guozheng Li, Peng Wang, Jiafeng Xie, Ruilong Cui, and Zhenkai Deng. 2022. Feed: A chinese financial event extraction dataset constructed by distant supervision. In *Proceedings of the 10th International Joint Conference on Knowledge Graphs*, page 45–53.

Yuan Liang, Zhuoxuan Jiang, Di Yin, and Bo Ren. 2022. RAAT: Relation-augmented attention transformer for relation modeling in document-level event extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4985–4997.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. 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 2795–2806.

Yubo Ma, Zehao Wang, Yixin Cao, Mukai Li, Meiqi Chen, Kun Wang, and Jing Shao. 2022. Prompt for extraction? PAIE: Prompting argument interaction for event argument extraction. In *Proceedings of the*

60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6759–6774.

Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. 2024. Building a large japanese web corpus for large language models. In *Proceedings of the First Conference on Language Modeling*, COLM, page (to appear), University of Pennsylvania, USA.

OpenAI. 2024a. GPT-4o mini: advancing cost-efficient intelligence.

OpenAI. 2024b. Hello GPT-4o.

Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. 2023a. When does in-context learning fall short and why? a study on specification-heavy tasks.

Hao Peng, Xiaozhi Wang, Feng Yao, Zimu Wang, Chuzhao Zhu, Kaisheng Zeng, Lei Hou, and Juanzi Li. 2023b. OmniEvent: A comprehensive, fair, and easy-to-use toolkit for event understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 508–517, Singapore.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2024. GoLLIE: Annotation guidelines improve zero-shot information-extraction. In *The Twelfth International Conference on Learning Representations*.

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. Extended named entity hierarchy. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Issa Sugiura, Takashi Ishida, Taro Makino, Chieko Tazuke, Takanori Nakagawa, Kosuke Nakago, and David Ha. 2025. Edinet-bench: Evaluating llms on complex financial tasks using japanese financial statements.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus.

Xiaozhi Wang, Xu Han, Zhiyuan Liu, Maosong Sun, and Peng Li. 2019a. Adversarial training for weakly supervised event detection. 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 998–1008.

Xiaozhi Wang, Hao Peng, Yong Guan, Kaisheng Zeng, Jianhui Chen, Lei Hou, Xu Han, Yankai Lin, Zhiyuan Liu, Ruobing Xie, Jie Zhou, and Juanzi Li. 2024. MAVEN-ARG: Completing the puzzle of all-in-one event understanding dataset with event argument annotation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4072–4091.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1652–1671.

Xiaozhi Wang, Ziqi Wang, Xu Han, Zhiyuan Liu, Juanzi Li, Peng Li, Maosong Sun, Jie Zhou, and Xiang Ren. 2019b. HMEAE: Hierarchical modular event argument extraction. 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 5777–5783.

Xingyao Wang, Sha Li, and Heng Ji. 2023a. Code4Struct: Code generation for few-shot event structure prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3640–3663.

Xinyu Wang, Lin Gui, and Yulan He. 2023b. Document-level multi-event extraction with event proxy nodes and hausdorff distance minimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10118–10133.

Ziqi Wang, Xiaozhi Wang, Xu Han, Yankai Lin, Lei Hou, Zhiyuan Liu, Peng Li, Juanzi Li, and Jie Zhou. 2021. CLEVE: Contrastive Pre-training for Event Extraction. 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 6283–6297.

Runxin Xu, Tianyu Liu, Lei Li, and Baobao Chang. 2021. Document-level event extraction via heterogeneous graph-based interaction model with a tracker. 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 3533—-3546.

Hang Yang, Yubo Chen, Kang Liu, Yang Xiao, and Jun Zhao. 2018. DCFEE: A document-level Chinese

financial event extraction system based on automatically labeled training data. In *Proceedings of ACL 2018, System Demonstrations*, pages 50–55.

Hang Yang, Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, and Taifeng Wang. 2021. Document-level event extraction via parallel prediction networks. 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 6298—-6308.

Tian Zhang, Maofu Liu, and Bingying Zhou. 2024a. Cfindee: A chinese fine-grained financial dataset for document-level event extraction. In *Companion Proceedings of the ACM Web Conference 2024*, page 1511–1520.

Xinliang Frederick Zhang, Carter Blum, Temma Choji, Shalin Shah, and Alakananda Vempala. 2024b. ULTRA: Unleash LLMs' potential for event argument extraction through hierarchical modeling and pairwise self-refinement. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8172–8185.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2019. Doc2EDAG: An end-to-end document-level framework for Chinese financial event extraction. 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 337–346.

Hanzhang Zhou, Junlang Qian, Zijian Feng, Lu Hui, Zixiao Zhu, and Kezhi Mao. 2024. LLMs learn task heuristics from demonstrations: A heuristic-driven prompting strategy for document-level event argument extraction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11972–11990.

Tong Zhu, Xiaoye Qu, Wenliang Chen, Zhefeng Wang, Baoxing Huai, Nicholas Yuan, and Min Zhang. 2022. Efficient document-level event extraction via pseudo-trigger-aware pruned complete graph. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4552–4558.

A Entity Definition

Company An entity mention for a company name refers to an organization engaged in economic activities for profit, including acquired foreign companies and group companies established abroad. The span of a company name includes legal entity terms, such as in "富士通信機製造株式会社" (Fuji Tsushinki Manufacturing Co., Ltd.) and "KOREA SHINKO MICROELECTRONICS CO., LTD.", and also covers abbreviated forms like "ヤンセン社" (Janssen Co.). Expressions that do not refer to specific companies, such as "an American sales company," and nested company names like "Amazon" within "Amazon Web Service Inc.", are not regarded as mentions. Expressions like "our company" are considered mentions only when they refer to the company of interest and no other company name appears in the text.

Date An entity mention for date expression directly indicates a specific year and month, such as "2003年4月" (April 2003), or refers to a specific date, such as "同年2月" (February of the same year). Ambiguous expressions that are not instantiated as specific dates, such as "現在の" (current), "後の" (later), and "上旬" (early in the month), are not regarded as mentions.

	Max	Min	Average
Japanese Word Tokenizer	2,859	537	1,079
BERT-Japanese	2,725	471	1,205
GPT-4o series	3,582	651	1,526
LLama-3 series	3,672	683	1,578

Table 8: The number of tokens per corporate history.

B Corporate History

Tables 14 show examples of the corporate history as presented in the annual securities report. Japanese corporate annual reports include a corporate history section, which typically comprises two-column chronological tables with dates and descriptions formatted in HTML. Since the date expressions in the table are almost standardized in a year-month format, there is no need to separate dates and descriptions using tags. We converted the table to timelines by removing HTML tags, and normalized the date expressions that only include the month to include the year using the previous timeline. Table 8 shows the average tokens per corporate history for each tokenizer.

C Implementation of Entity Mention Extraction

We implemented an automatic entity mention annotation tool to reduce annotation costs. For `Date`, we search the string that matches the regular expressions[8] and extract the span of the longest string. For `Company`, we generate the mentions by GPT-4o following the prompt as in Table 15, and extract

[8] "([0-9]{3}|[1-2][0-9]{3})年([1-9]|1[0-2])月", "(明治|昭和|平成|令和|同)([1-9][0-9]|[1-9])年([1-9]|1[0-2])月"

	Mentions	
	Date	Company
Regular Expressions	95.6	-
GPT-4o+Heuristics	-	97.3
Author	95.9	98.9

Table 9: IAA between two annotators and Preliminary annotation by GPT-4o for Entity Mention Annotation.

	Arguments		
	Precision	Recall	F1
ESTABLISH	98.3	94.5	96.3
CHANGE	97.7	92.6	95.1
MERGE	87.8	88.9	88.3
SPLIT	87.0	52.6	65.6
LIQUIDATION	95.8	95.8	95.8
Total	95.1	91.1	93.0

Table 10: IAA between two annotators for Event Annotation. Precision indicates the ratio dividing the number of matching cases by the number of annotations made by the annotator. Recall indicates the ratio dividing the number of matching cases by the number of annotations we made.

the span that matches the string of mentions. Since GPT-4o occasionally normalizes or abbreviates the legal entity term in corporations, we also search with the strings replaced by an abbreviation or a canonical form of the legal entity term. For example, when "トヨタ自動車株式会社 (Toyota Motor Corporation)" is generated by GPT-4o, predict the spans that match the two strings "トヨタ自動車株式会社 (Toyota Motor Corporation)" and "トヨタ自動車(株) (Toyota Motor Corp.)".

D Inter-Annotator Agreement

We measure the inter-annotator agreement (IAA) for the two annotation tasks, Entity Mention Annotation and Event Annotation. We re-annotated 20 documents in 210 documents by ourselves; we randomly selected them.

D.1 Entity Mention Annotation

To measure the agreement for entity mention annotation, we calculated F1 scores between the results of two annotators based on the exact match of both boundaries for entity mentions. Additionally, we compared the agreement with the result of the automatic entity mention annotation tool. Table 9 shows the F1 scores for each entity type and demonstrates that the F1 scores for all types were over 95.0 points. Interestingly, the F1 score of GPT-4o was 97.3 points, in which GPT-4o can correctly predict the company names from the Japanese documents. These documents include mentions of foreign companies, but this finding suggests GPT-4o can also correspond to legal entity terms by foreign countries such as "有限公司" and "Co., Ltd."

D.2 Event Annotation

For event annotation, we measure the inner-annotation agreement with the same evaluation criteria as those used in Zheng et al. (2019); Zhang et al. (2024a), comparing two annotators' event tables. Specifically, we select one event record of an annotator and the most similar record of a different annotator where the most arguments are matched. We count the matched arguments in the two selected records and aggregate these statistics among all documents. If some arguments in the record are missing, the output is considered correct if the argument is left empty.

Table 10 shows the F1 scores between two annotators' results for each IAA measure. The result indicates that the annotator could assign the same argument to the majority of event types. Among all event types, while ESTABLISH, CHANGE, and LIQUIDATION were over 95.0, MERGE, and SPLIT were relatively lower. This is due to the mismatch in two arguments related to company names before and after the event. For MERGE, as described in Section 3.2, some sentences can be interpreted in two ways: either the former company absorbs the latter, or both companies are absorbed. SPLIT has two types of splits, the incorporation-type split where a company is spun off, and the absorption-type split where a business is sold to an existing company. Disambiguating the relation between absorption and two types of splits demands domain-specific knowledge, and improving agreement on this type remains a challenge for future work.

E Dataset Representativeness

To better understand dataset representativeness in our dataset, we investigate the performance curve of a finetuned model for different amounts of the training set. Figure 2 shows the F1 scores of Event Record Identification by Llama-3-Swallow models trained on 10, 20, 30, 40, and 50 documents in the training set, where 0 represents the F1 score of Llama-3-Swallow model with initial parameters. We observe that the F1 score of ESTABLISH reaches a peak at 30 documents but the F1 scores for other event types continue to increase. This

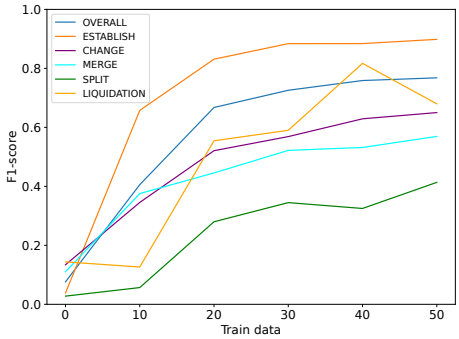

Figure 2: Performance of finetuned Llama-3-Swallow models for different amounts of the training set.

	Parameters
Seeds	[0, 21, 42, 63, 84]
Training Epochs	2
Batch size	2
Gradient accumulation steps	5
Scheduler	cosine
Optimizer	AdamW
Warmup	0.1
Max grad norm	0.3
LoRA r	128
LoRA α	128
LoRA Dropout	0.05

Table 11: Hyperparameters

result indicates that the LLMs can quickly adapt to the ESTABLISH events with 50 documents, but complexity remains for other events.

F Implementation of Event Extraction

We implemented the DMBERT and BERT+CRF using official codes from OmniEvent (Peng et al., 2023b), and adopted bert-base-japanese-v3 [9] as a backbone. Since BERT supports a maximum token length of 512, we divided the timelines into segments of up to 500 characters. When arguments are excluded from segments due to timeline splitting, the corresponding gold arguments are treated as missing.

For LLMs, we adopt a simple prompt strategy to benchmark the performance of our task. Figure 3, 4, 5, 6, 7 shows our instruction templates for each events. The instruction templates are designed that LLMs directly extract tuples from timelines. For Llama-3-Swallow, we apply the Japanese translation version of the instruction templates, since these models are continually pretrained with the Japanese corpus to enhance Japanese language capability. LLMs are presented with corporate history timelines, a query, and a description of each event type as input, then they generate a JSONLines text. In the queries, we request to generate a text following the format where a company name before the event is *subject* and a company name after the event is *object*. Specifically, the output format comprises some keys: *date*, *subject*, and *object*, and the values: the names of argument roles, and the name of argument roles are translated to phrase expressions (ex. *EstablishedCompany* to "established company name"). Subsequently, the extracted tuples are automatically inserted into the corresponding event tables, and if no event record for the event type in a document, LLMs generate "None." We perform the EE systems and Llama-3 models on a single NVIDIA RTX A6000 GPU with 48GB memory.

[9] https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

F.1 Supervised Finetuning of LLMs

We use a text y^e converted to JSONLine format as gold tokens and apply a language modeling objective on these tokens. Table 11 shows the hyperparameters used for supervised finetuning. We train Llama-3 models with cross entropy loss and use AdamW (Loshchilov and Hutter, 2019) optimizer with the default hyperparameters. We use cosine annealing schedule with warmup and the initial learning rate is 5e-5. Training the models takes less than 30 minutes on a single GPU, NVIDIA RTX A6000 GPU with 48GB memory.

F.2 In-Context Learning of LLMs

LLMs are presented with the template including two demo samples, then generate a text y^e_{input} converting the event records extracted from the document X_{input}. A demo sample comprises a query of an event type, a document in another document in the training set X_{demo}, and an answer y^e_{demo}. We randomly select two demo samples from 50 documents in the training set, but the selected samples always include a sample where the event record does not appear in the document. For ESTABLISH, we randomly select two documents from 50 documents since the event record always appears in the document. For GPT-4o models, we perform `gpt-4o-2024-08-06` and `gpt-4o-mini-2024-07-18` with OpenAI's Batch API. The temperature is fixed at 0.2 for all models, and the pricing for running GPT-4o ranges from USD 0.004 to USD 0.014 per example.

	Runtime
DMBERT (batch size=16(ED), 32(EAE))	3.0 s
BERT+CRF (batch size=32(ED), 64(EAE))	0.4 s
GPT-4o (BatchAPI)	7.9 s
Llama-3-Swalow (batch size=5)	33.4 s

Table 12: The average runtime per corporate history. ED and EAE mean Event Detection and Event Argument Extraction, respectively.

	Precision	Recall	F1
Meta (In-context)	28.6	57.9	38.3
+Anonymize	30.4	62.8	41.0†
Meta (Finetune)	73.3	70.5	71.7
+Anonymize	77.0	73.9	75.4†
GPT-4o (In-context)	79.9	77.2	78.5
+Anonymize	35.2	34.9	35.1†

Table 13: Comparison of Llama-3-8B and GPT-4o models in Event Record Identification by anonymizing company names in documents. † indicates that the F1-score significantly changes by anonymizing companies.

F.3 Inference Speed

The average runtime per corporate history for each model is shown in Table 12. Llama-3-Swallow requires approximately 33 seconds per document, which is substantially slower than DMBERT and BERT+CRF. Although such latency is not acceptable for online services, we consider it acceptable for our purpose, as the task is intended for offline data analysis. The Llama-3 models were implemented using HuggingFace's LlamaForCausalLM, thus, the throughput could be improved with alternative implementations such as vLLM (Kwon et al., 2023).

G Entity Extraction Performance with Anonymized Timelines

Considering the possibility that LLMs may memorize corporate history, we evaluate event extraction performance in a setting where company names appearing in the timelines are anonymized. In this setting, we assign some special token (ex. [COMPANY_1]) to each company name, and make to answer which special token is the correct argument. Table 13 shows the performance of LLMs on anonymized documents, indicating that while the F1-score of Llama-3 models increase by anonymizing, the F1-score of GPT-4o significantly drops to 35.1. This score is comparable to the F1-score of Llama-3-Meta with in-context learning, suggesting the GPT-4o's ability to identify organizational changes from context is equivalent to Llama-3 8B parameters models.

October 1946 Soichiro Honda established Honda R&D in Hamamatsu, Shizuoka Prefecture, Japan, and engages in research and manufacture of internal combustion engines and various machine tools.
September 1948 Established Honda Motor Co., Ltd. by inheriting Honda R&D
August 1949 Started production of motorcycles.
April 1952 Moved the headquarters to Tokyo.
September 1952 Started production of power products.
May 1953 Started operations at the Yamato Plant (Wako Plant of the Saitama Factory from January 1973).
April 1954 Started operations at the Aoi Plant of the Hamamatsu Factory (Transmission Manufacturing Department since April 2014).
December 1957 Listed shares on the Tokyo Stock Exchange.
June 1959 Established American Honda Motor Company, Inc. in the United States.
May 1960 Started operations at the Suzuka Factory
July 1960 Separated Honda R&D from our company and established Honda R&D Co., Ltd.
June 1963 Started production of four-wheeled vehicles.
October 1964 Established Asian Honda Motor Co., Ltd. in Thailand.
November 1964 Started operations at the Sayama Factory (Transmission Manufacturing Department since April 2014).
March 1969 Established Honda Canada Inc. in Canada.
September 1970 Separated the machinery division of the Sayama Factory's second plant from our company and established Honda Machinery Co., Ltd. (Honda Engineering Co. Ltd. since July 1974.).
December 1970 Started operations at the Mooka Plant (Powertrain unit Manufacturing Department since April 2014).
October 1971 Established Honda Motor do Brasil Ltda. in Brazil (Honda South America Ltda. since April 2000).
July 1975 Established Moto Honda da Amazônia, Ltda. in Brazil.
March 1976 Started operations at the Kumamoto Factory.
February 1977 Listed ADR (American Depositary Receipt) on the New York Stock Exchange.
March 1978 Established Honda of America Manufacturing, Inc. in the United States.
February 1980 Established American Honda Finance Corporation in the United States.
September 1985 Established Honda de México S.A. de C.V. in Mexico.
January 1987 Established Honda Canada Finance Inc. in Canada.
March 1987 Established Honda North America, Inc. in the United States with oversight functions for North American subsidiary operations.
August 1989 Established Honda Motor Europe Ltd. in the United Kingdom with oversight functions for European subsidiary operations.
July 1992 Established Honda Cars Manufacturing (Thailand) Co., Ltd. in Thai (Honda Automobile (Thailand) Co., Ltd. since December 2000)
May 1996 Established oversight functions for ASEAN subsidiary operations in Asia Honda Motor Co., Ltd.
April 1999 Established Honda Credit Co., Ltd. in Tokyo Metropolis (Honda Finance Co., Ltd. since July 2002).
December 1999 Established Honda Manufacturing of Arabama, LLC in the United States.
April 2000 Established oversight functions for North American subsidiary operations in Honda South America Ltda.
June 2002 Ended production of four-wheeled vehicle engines at the Wako Plant of the Saitama Factory and transferred the production functions to the Sayama Plant of the Saitama Factory (Saitama Factory since October 2002).
(The site of the Wako Plant of the Saitama Factory has been utilized as the Honda Wako Building since July 2004.)
January 2004 Established Honda Motor (China) Investment Co., Ltd in China with oversight functions for Chinese.
September 2009 Started operations at the Ogawa Plant of the Saitama Factory
July 2013 Started operations at the Yorii Plant of the Saitama Factory
July 2020 American Honda Motor Company, Inc. merged Honda North America, Inc. which has oversight functions for North American subsidiary operations.
April 2021 Honda of America Manufacturing, Inc. merged Honda Manufacturing of Arabama, LLC and six other other companies, and changed its name to Honda Development and Manufacturing of America, LLC.
December 2021 Ended production of four-wheeled vehicles at the Sayama Plant of the Saitama Factory.

Table 14: Example of Corporate history timelines. This example is created by editing the annual corporate report from EDINET (https://disclosure2.edinet-fsa.go.jp/WZEK0040.aspx?S100R1U9). We note that the original text is in Japanese, which we have translated into English by us to facilitate understanding.

Find company names from text, and tag each company name of using jsonline format.
Don't contain the input text in the output.

Text:
1943年3月水産統制令により、株式会社林兼商店の内地水産部門、大洋捕鯨株式会社及び遠洋捕鯨株式会社で、
捕鯨業、トロール漁業及び底曳網漁業を事業目的とした西大洋漁業統制株式会社(資本金6千万円)を下関市に設立
1945年3月水産物及び農畜産物の製造、加工、販売業並びに冷蔵倉庫業を事業目的に追加
12月西大洋漁業株式会社に商号変更
大洋漁業株式会社に商号変更
1947年10月大都魚類株式会社を設立(現・連結子会社)
1948年4月神港魚類株式会社を設立(現・連結子会社)
1949年6月本社を東京都に移転
1961年3月肥料・飼料事業を事業目的に追加
1983年7月宇都宮市に練り製品工場完成
1985年8月宇都宮市に調味料・薬品・健康食品工場完成
1990年9月つくば市に中央研究所完成
1993年9月マルハ株式会社に商号変更
2004年4月株式移転により完全親会社株式会社マルハグループ本社を設立
2007年10月株式会社マルハグループ本社と株式会社ニチロが経営統合、
株式会社マルハグループ本社が株式会社ニチロを株式交換により完全子会社化
株式会社マルハグループ本社は株式会社マルハニチロホールディングスに商号変更
2008年4月株式会社マルハニチロ水産に商号変更
当社食品事業を、株式会社マルハニチロ食品(株式会社ニチロより商号変更)を承継会社として吸収分割
当社畜産事業を、株式会社マルハニチロ畜産を承継会社として吸収分割
株式会社マルハニチロ食品の水産事業を、当社を承継会社として吸収分割
2014年4月マルハニチロ株式会社に商号変更
株式会社マルハニチロホールディングス、株式会社マルハニチロ食品、
株式会社マルハニチロ畜産、株式会社マルハニチロマネジメント及び株式会社アクリフーズを吸収合併
東京証券取引所市場第一部に上場
2022年4月東京証券取引所の市場区分の見直しにより、東京証券取引所市場第一部からプライム市場へ移行

Company names:
{"name": "西大洋漁業統制株式会社"}
{"name": "株式会社林兼商店"}
{"name": "大洋捕鯨株式会社"}
{"name": "大都魚類株式会社"}
{"name": "神港魚類株式会社"}
{"name": "マルハ株式会社"}
{"name": "株式会社マルハグループ本社"}
{"name": "株式会社ニチロ"}
{"name": "株式会社マルハニチロホールディングス"}
{"name": "株式会社マルハニチロ水産"}
{"name": "株式会社マルハニチロ食品"}
{"name": "株式会社マルハニチロ畜産"}
{"name": "マルハニチロ株式会社"}
{"name": "株式会社アクリフーズ"}

Document: X
Company names:

Table 15: Prompt for Company name extraction. The demonstration is created by editing the annual corporate report from EDINET (https://disclosure2.edinet-fsa.go.jp/WZEK0040.aspx?S100R5V0).

You are a sincere and excellent Japanese assistant. Unless otherwise specified, please always answer in Japanese.

Please analyze the following document to determine if it contains any of the events expressing that a company is established. If events are detected, kindly provide the following arguments: established date (“date”), an established company (“subject”), and format your response as jsonline:
{“date”: established date, “subject”: established company}
If no arguments are found, simply respond with “None”.

Document: {timelines}

Response:

Figure 3: The instruction template for ESTABLISH for experiments. Timelines are assigned to the curly brackets

You are a sincere and excellent Japanese assistant. Unless otherwise specified, please always answer in Japanese.

Please analyze the following document to determine if it contains any of the events expressing that a company name is changed to another name. If events are detected, kindly provide the following arguments: changed date (“date”), a previous company name (“subject”), a last company name (“object”), and format your response as jsonline:
{“date”: changed date, “subject”: previous company name, “object”: last company name}
If no arguments are found, simply respond with “None”.

Document: {timelines}

Response:

Figure 4: The instruction template for CHANGE for experiments. Timelines are assigned to the curly brackets

You are a sincere and excellent Japanese assistant. Unless otherwise specified, please always answer in Japanese.

Please analyze the following document to determine if it contains any of the events expressing that companies are merged or absorbed into one. If events are detected, kindly provide the following arguments: merged date (“date”), list of absorbed companies (“subject”), a survived company (“object”), and format your response as jsonline:
{“date”: merged date, “subject”: [list of absorbed companies], “object”: a survived company}
If no arguments are found, simply respond with “None”.

Document: {timelines}

Response:

Figure 5: The instruction template for MERGE for experiments. Timelines are assigned to the curly brackets

You are a sincere and excellent Japanese assistant. Unless otherwise specified, please always answer in Japanese.

Please analyze the following document to determine if it contains any of the events expressing that a company is spun off a section as a separate company. It is distinct from a sell-off, where a company sells a section to another company or firm in exchange for cash or securities. If events are detected, kindly provide the following arguments: split date (“date”), a parent company (“subject”), list of separated companies (“object”), and format your response as jsonline:
{“date”: split date, “subject”: a parent company, “object”: [list of separated companies]}
If no arguments are found, simply respond with “None”.

Document: {timelines}

Response:

Figure 6: The instruction template for `SPLIT` for experiments. Timelines are assigned to the curly brackets

You are a sincere and excellent Japanese assistant. Unless otherwise specified, please always answer in Japanese.

Please analyze the following document to determine if it contains any of the events expressing that a company is established. If events are detected, kindly provide the following arguments: established date (“date”), an established company (“subject”), and format your response as jsonline:
{“date”: liquidated date, “subject”: liquidated company}
If no arguments are found, simply respond with “None”.

Document: {timelines}

Response:

Figure 7: The instruction template for `LIQUIDATION` for experiments. Timelines are assigned to the curly brackets