Extracting Explanatory Rationales of Activity Relationships using LLMs - A Comparative Analysis

Research Paper

Kerstin Andree[0009−0007−6360−8661], Zahi Touqan,
Leon Bein[0000−0001−9064−7905], and Luise Pufahl[0000−0002−5182−2587]

Information Systems, CIT, Technical University of Munich, Heilbronn, Germany
{kerstin.andree, zahi.touqan, leon.bein, luise.pufahl}@tum.de

Abstract. Contextual information about business processes is essential for business process management techniques, such as business process redesign (BPR). In particular, the extraction of explanatory rationales of activity relationships – laws, business rules, and best practices – improves the results of BPR tasks. Their extraction from textual data, however, is usually done manually in an expensive and time-consuming manner. This paper investigates the use of Large Language Models (LLMs) to automate the extraction of explanatory rationales from textual data. By comparing four LLM prompting techniques (Vanilla, Few-Shot, Chain-of-Thought, and their combination), we evaluate their effectiveness in classifying relationships based on contextual origins. Our findings show that the Few-Shot and combined approaches significantly enhance precision, recall, and F1 scores. Furthermore, smaller, cost-effective LLMs, such as GPT 4o-mini, achieved excellent performance, making advanced classification accessible to organizations with limited resources.
Keywords: Activity Relationships Classification, Large Language Models, Explanatory Rationales, Process Context.

1 Introduction

Business Process Management (BPM) provides concepts, techniques, and methods to understand, analyze, and improve *business processes* (Weske 2019). For most BPM techniques, such as *Business Process Redesign* (BPR) (Mansar and Reijers 2005) or compliance checking (Groefsema et al. 2022), *contextual knowledge* about business processes is highly relevant (Rosemann et al. 2008; Suriadi et al. 2014; Saidani and Nurcan 2007). This includes the origin of activity relationships, such as the reasons behind specific sequences and dependencies which is specifically important for BPR (Adamo et al. 2018). For example, Adamo et al. (2021) show that the implementation of BPR tasks can be significantly improved if contextual information about the activity relationships is known to the designers beforehand. Andree and Pufahl (2024) present a more general perspective on contextual information specifically for activity relationships. They introduce a metamodel that lists a set of aspects that need to be taken into account when checking the feasibility of implementing change operations: the *vulnerability of relationships*, associated *risks*, *consequences*, and *explanatory rationales*.

The classification of an activity relationship into explanatory rationales is still done manually (Revoredo 2023) as part of process discovery (Dumas et al. 2018, Ch.5) and

20th International Conference on Wirtschaftsinformatik
September 2025, Münster, Germany

process annotation (Adamo et al. 2018). Since contextual information is usually only available in textual form, this results in a cost- and time-consuming task. (Re)designers have to analyze several documents, conduct interviews and carefully classify each relationship regarding its vulnerability and risks. Similarly, compliance checking requires a thorough investigation of which constraints originate from regulations.

Large Language Models (LLMs) are used as a technology that has proven particularly well in dealing with text-based data (Vidgof et al. 2023; Grohs et al. 2023). Literature has shown that LLMs can analyze vast amounts of textual data, such as process descriptions and regulatory documents, to identify activity relationships (Bellan et al. 2023; Klessascheck et al. 2024) and find relevant parts of textual process data (Sai et al. 2024). The classification of relationships according to their contextual origin, however, has not been investigated so far.

To address this research gap, this paper compares different LLM approaches and techniques to discuss the potential of automating the extraction of explanatory rationales of activity relationships based on natural language texts. Accordingly, we aim to study the following research question:

RQ How do different large language models and prompting approaches perform concerning extracting explanatory rationales for a given set of activity relationships from textual descriptions?

To answer this research question, we follow an iterative approach: the experimental setup of the comparative analysis was developed using a small, manually created use case and later evaluated through a comprehensive use case scenario at a German university. Overall, we provide the following contributions:

1. Comparison of four different LLM prompting configurations coupled with Retrieval Augmented Generation (RAG): Vanilla Prompting, Few-Shot learning, Chain-of-Thought Reasoning, and a combination of both
2. Assessment of the effectiveness of different LLMs for classifying activity relationships according to their contextual origin
3. Publicly available experimental setup that can be modified for other use cases or LLMs
4. Validation of the approach using a real-world use case scenario

The remainder of this paper is structured as follows. Sec. 2 introduces relevant terms and concepts used in this paper. Sec. 3 discusses related work before the experimental setup of this comparative analysis is introduced in Sec. 4. Results are presented in Sec. 5 and discussed in Sec. 6, which also concludes.

2 Background

2.1 Activity Relationships in Business Processes

Business process models are the fundamental artifact for organizations when redesigning their business processes (Adamo et al. 2018). Process behavior is defined by *activity relationships* (Kunze and Weske 2016), i.e., dependencies between two activities (Weske 2019). They are defined based on the business goal or contextual environment (Andree and Pufahl 2024; Adamo et al. 2018), e.g., governmental laws.

Andree and Pufahl (2024) introduce a metamodel for contextual information about activity relationships specifically needed for BPR. It combines the insights of Risk-Aware BPM (Suriadi et al. 2014) and Context-Aware BPM (Saidani and Nurcan 2007). In particular, the authors provide an overview of *explanatory rationales* of activity relationships explaining the origin and motivation of a relationship. They are illustrated using Fig. 1 showing a simplified BPMN diagram of a subset of activities involved in writing a thesis at a German university.

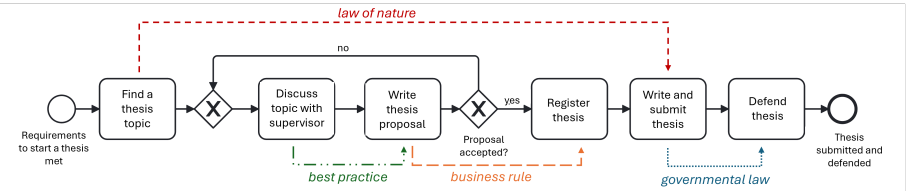

Figure 1. BPMN process model representing the behavior of a thesis process. Explanatory rationales are highlighted.

Best Practice. Procedures or techniques commonly accepted by the process participants to be superior to other alternatives. They are not required to be enforced legally or internally. Writing a thesis proposal after having discussed the topic with the supervisor is considered a best practice (see Fig. 1) as it improves the quality of the proposal. Yet, it is also possible to hand in a proposal first before discussing the topic in more detail.

Business Rule. Specific, actionable directives defined by the stakeholders of the process, the organization itself, or the suppliers that define or constrain some aspect of business. These rules stem from the business objectives and are dependent on business strategy. For example, writing a proposal before registering a thesis is considered to be a business rule (see Fig. 1), as the chair generally does not accept any theses without having seen a proposal. This constraint is defined for the entire chair and not by individual supervisors.

Governmental Law. Legal and regulatory obligations imposed by governmental bodies. These relationships must be strictly adhered to ensure compliance and avoid legal consequences. The laws are enforced by external governmental bodies. For example, examination regulations of universities include that a thesis defense can only take place after the thesis is submitted. This regulation is valid for the entire university, not only for a single chair or individual supervisor. Thus, the ordering of the activities shown in Fig. 1 is considered to be a governmental law.

Law of Nature. A law of nature is treated as a separate concept defining the vulnerability of the ordering between the two activities of a given relationship. We define a law of nature as an inviolable relationship where the second activity cannot precede the first activity due to either a deadlock occurring where you cannot proceed with the process or due to data, or resource dependencies. For example, the law of nature shown in Fig. 1 indicates that finding a topic has to occur before writing and submitting the thesis: writing a thesis without having a topic in mind is not possible. The class of law of nature is independent of the other rationales. While only one of the categories best practices, business rules, and governmental laws can apply to a relationship, a law of nature can be combined with each of them.

2.2 Large Language Models and Prompting Techniques

Large Language Models (LLMs) present a promising avenue for automating the contextualization of activity interrelationships. With their capacity to process and comprehend vast textual data, LLMs can analyze diverse sources of knowledge, such as process documentation, expert interviews, and external knowledge sources, to identify and classify the contextual origin of activity relationships (Zhu et al. 2024). The transformer architecture (Vaswani et al. 2017) used in modern LLMs is well-suited for this task. Its attention mechanism allows the model to dynamically focus on the most relevant parts of the input to recognize long-term dependencies used to extract relationships from complex and lengthy process descriptions.

LLMs have demonstrated capabilities such as reasoning, planning, decision-making, in-context learning, and answering in zero-shot settings (Naveed et al. 2023). For information and relation extraction tasks, optimization techniques, such as Retrieval Augmented Generation (RAG) (Lewis et al. 2020), Prompt Engineering (Few-Shot learning and Chain-of-Thought reasoning), and fine-tuning (P. Liu et al. 2023), can be used to improve the performance of vanilla prompting, i.e., simply prompting the LLM.

Retrieval Augmented Generation (RAG) is a technique that enhances the capabilities of LLMs by incorporating externally retrieved knowledge before generating a response (Lewis et al. 2020). Thus, the technique performs well in tasks requiring domain-specific knowledge on which the LLM may not have been explicitly trained.

Few-Shot learning provides several examples demonstrating the desired task to the LLM, enabling it to learn the pattern and generalize to new, unseen inputs. Thus, Few-Shot learning is especially well suited for tasks with limited training data.

Chain-of-Thought (CoT) explicitly prompts the model to articulate its thought process, which can generate more accurate and logical outputs, especially for tasks that require multi-step reasoning or problem-solving (Wei et al. 2022).

Inspired by the successful application of these techniques, we incorporated them into the approach for automated relationship classification.

3 Related Work

Since the topic of extracting contextual information about business processes is relevant for several BPM disciplines (cf. Sec. 2.1), initial approaches were already presented in the literature. For business process repair and redesign, the extraction of explanatory rationales is still done in a manual manner to differentiate between changeable and non-changeable process parts (Adamo et al. 2021; Revoredo 2023). Nevertheless, automated support in processing contextual information has been presented in the fields of compliance checking and discovery of activity relationships.

Dragoni et al. (2017) present an approach to extract permissions, obligations, and prohibitions from legal texts and convert them into machine-readable rules used for compliance checking (Governatori et al. 2016). Winter and Rinderle-Ma (2018) propose an NLP-based method to group previously identified constraints. By preprocessing textual data to find relevant parts, the authors group by frequency, structure of sentences, and external knowledge before presenting the results in a graph-based structure. The groupings

are not predetermined, providing a high degree of flexibility. However, this approach was developed using legal documents and has not been tested for other contextual settings. Similarly, Winter et al. (2020) focus only on the legal domain for presenting their approach to map process models to regulatory documents to enhance compliance checking. In contrast to the works mentioned before, Sai et al. (2024) show the potential of generative AI in processing textual process data. The authors present an approach to identify relevant requirements in the legal context. In our paper, we expand the spectrum and include not only legal constraints in the text classification but also business rules and best practices that arise due to business goals or workarounds.

Another path of automated extraction of contextual information in the context of BPM is the text-based discovery of activity relationships. For instance, van der Aa et al. (2019) use Natural Language Processing (NLP) techniques and focus on declarative constraints. Honkisz et al. (2018) present a method for discovering BPMN models from textual data. However, both works neglect the extraction of explanatory rationales so that the vulnerability of discovered relationships cannot be defined. Barrientos et al. (2023) propose an approach to automatically identify temporal constraints from textual process descriptions and their verification over event log data. Thus, the authors overcome the formalization into logic-based rules, e.g., linear temporal logic (LTL), but still neglect the classification according to the explanatory rationale of the constraint.

4 Experimental Setup

The general setup of the comparative analysis of different LLM models and techniques is shown in Fig. 2. As input, we use the process description, a list of activities that occur in the process, and the definitions of the explanatory rationales. A RAG pipeline preprocesses the process description so that only relevant information is passed to the LLM (N. F. Liu et al. 2024).

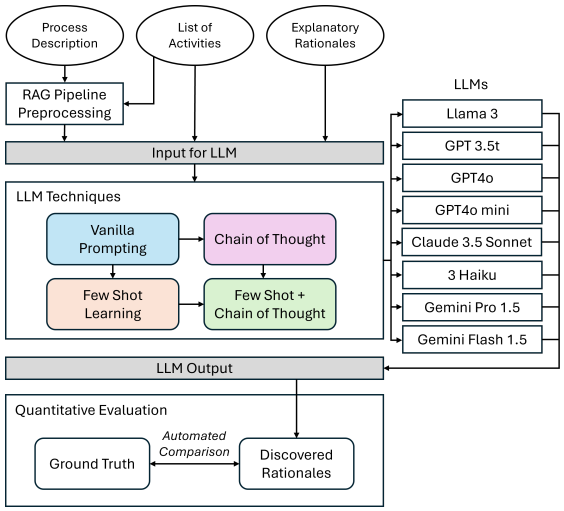

Figure 2. Experimental setup of the comparative analysis

For a given process use case, we run each LLM with each of the listed prompting techniques: Vanilla prompting, Few-Shot learning, Chain-of-Thought reasoning, and Chain-of-Thought combined with Few-Shot learning. The generated outputs are then automatically compared with the manually defined ground truth for a given process.

In the following, each step of the comparative analysis will be explained in detail. For further information, please refer to the GitHub repository[1]. By following the instructions given there, it is possible to replicate the results of our comparative analysis and extend the experimental setup. Used LLMs can be easily replaced by newer versions. Similarly, other textual process descriptions can be used as input. This makes the experimental setup provided in this paper generalizable to other use cases and models.

4.1 Dataset Generation

The dataset for running the experiments includes a process description, a list of activities, and the ground truth. Process descriptions can be based on interviews or publicly available texts. As additional sources, laws and regulation documents might be beneficial.

For our experiments, we use the process of a thesis at a German university. The process descriptions were created based on interviews with process participants (see Fig. 3). Additional information was also provided by extracts from the examination regulations. A validation of the results was done by conducting a second round of interviews which was also used to define the ground truth. Overall, the input process description is one text file concatenating the interviews and the excerpt from the general examination regulations.

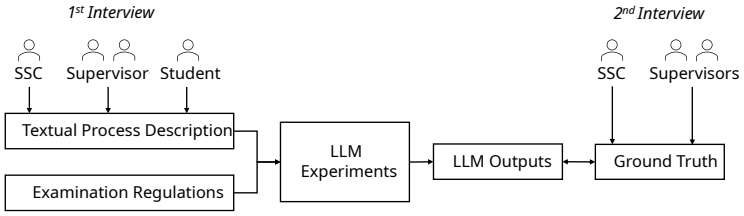

Figure 3. Generation of the experimental dataset through interviews with process participants

The first round of interviews was done with a student undertaking their thesis, two supervising researchers, and a representative of the department (SSC) that supervises the theses from a regulatory point of view. The different backgrounds and roles of the process participants ensure the quality of the resulting process description minimizing errors and lack of detail. Interviews were limited to 20 minutes each. Starting with a general question regarding the overall process, its activities, and goals, we continued with more detailed questions targeting their role in the process, the regulations they need to adhere to, and the reasoning behind their way of process execution. Audio recordings of each interview were transcribed and formatted.

The paragraph related to thesis regulations from the general academic and examination regulations [2] was added to the process descriptions. These regulatory documents represent

[1] `github.com/INSM-TUM/activity-relationship-classification`

[2] The paragraph can be found in the provided GitHub repository

laws defined by the university, with the responsible ministry governing the process. Thus, we define the following mapping between explanatory rationales and decision takers shown in Table 1. Laws, as defined in the official regulations, are defined by the university, whereas the school and the individual chairs decide on business rules. Best practices are defined by the supervisors themselves.

Table 1. Mapping between explanatory rationale and decision taker in the thesis process context

Explanatory Rationale	Decision Taker
Laws	German University
Business Rule	School, Chair
Best Practice	Supervisor

Based on the interviews, we identified the following nine activities: *Search for a topic, conduct informal meeting to explain topic, write and submit proposal, get registered by chair via system, student accepts thesis, start writing thesis, conduct regular catch-up meetings, submit thesis, present thesis in colloquium.*

The second round of interviews was done with the department representative and the two research supervisors to define the ground truth. The student was excluded from this round because of insufficient experience with the process, regulations, and possibilities of flexibility. Each interview was limited to 30 minutes. Having 36 pairs of activities for each we need to discuss the explanatory rationale with the interviewees, we decided to split the set of activity pairs into three overlapping groups resulting in at least two assessments for each pair to minimize errors. In case of conflicting answers, an expert group was consulted. In total, the ground truth covers seven governmental laws, 12 best practices, 15 business rules, and six laws of nature.

The derived ground truth was then used to validate the results and the performance of the LLMs coupled with the different prompting techniques.

4.2 RAG Pipeline Preprocessing

For the RAG pipeline, we employed out-of-the-box libraries. As preparation, the contextual process description was split into non-overlapping, sentence-preserving, token-size-based chunks, which was transformed into embeddings using a sentence transformer model, and stored in a vector database. This RAG technique is less cost-intensive while minimizing context loss within a sentence due to fixed-size chunking. For a query of two activities, relevant context was retrieved by embedding the pair using the same sentence transformer, performing a similarity search on the database, ranking the resulting chunks, and returning the topmost-ranked.

This setup facilitates the retrieval of relevant context from the process description based on the specific activities, enabling the LLM to generate more accurate and contextually relevant outputs. Especially for long text inputs, this leads to a significant improvement in answer conciseness. Moreover, it mitigates the issue of exceeding the context-window length of the LLM and reduces the costs of API calls.

4.3 LLM Prompting

For all prompting approaches, we used the same baseline system prompt with slight variations. It defines explanatory rationales used for classification as well as the expected JSON output structure. In the following, we provide a shortened version. For further details, please refer to the data provided in our repository.

> *You are an assistant business process re-designer. Your job is to explain the context behind the ordering of a pair of activities, given the pair of activities and the process description, by categorizing the reason of the specific order in zero or one of the three following categories:*
> *1- Governmental Law: [...]*
> *2- Best Practice: [...]*
> *3- Business Rule: [...]*
> *Separate from the categories, you need to decide if the relationship is due to a law of nature, which is an inviolable relationship where the second activity cannot precede the first activity due to either a deadlock occurring or due to a data [...], resource [...], or logical dependency from the first activity. Structure your answer in the following format without any additional text, and replace the placeholders with the correct values:*
>
> *{"First Activity": "-",*
> *"Second Activity": "-",*
> *"Category": "-",*
> *"Justification": "-",*
> *"Law of Nature": "-" }*
>
> *If none of the three contextual origin categories apply to the relationship, put a dash in both the Category and Justification fields and do not include any other text for justifying your decision. Otherwise, put the category you chose in the "Category" field, the justification for your choice in the "Justification" field. In the "Law of Nature" field, if the answer is yes, then you should put justification in the value, if it is no, then only put a single dash. You will receive the prompt as "which of the categories best describes the contextual origin of why [First Activity] occurs before [Second Activity]?", the first activity always occurs in time before the second activity. Return only the JSON response [...] .*

In the following, we describe the prompting setups we tested. Highlighted in angle brackets (< >) are placeholders filled out by the Python script.

Vanilla Prompting (VP). Our baseline *vanilla* prompt simply asks for a classification, using the activity names and retrieved context. All other prompts are built upon this baseline.

> *Based on the following context, decide which of the categories best describes the contextual origin of why <First Activity> occurs before <Second Activity>? Explain why you chose this category and not another one. If you think none of the categories fit the pair of activities, you do not have to choose a category. After discussing the contextual origin, discuss if the ordering is due to a law of nature.*
> *Context:*
> *<Top 5 retrieved contextual embeddings relating to query (RAG output)>*

Chain-of-Thought (CoT). For CoT, we allowed the LLMs to formulate full-text responses and sent the instructions to format the output as JSON only with a follow-up query. Further, we added the following statement to the system prompts:

"Before you make a decision, you should think logically about the classification task. Work it out in a step-by-step way to be sure to have the right answer."

Few-Shot Learning. For Few-Shot learning (FL), we provide three to five manually written examples of already classified activity pairs in addition to the normal vanilla prompt shown before. Examples were chosen to encompass a wide range of possible results. In the following, we present one example. For the complete set of examples, please refer to the GitHub repository.

> *Examples:*
> *What is the relationship between Conduct topic-introduction meeting and Start writing?*
> *{"First Activity": "Conduct topic-introduction meeting",*
> *"Second Activity": "Start writing",*
> *"Category": "Best Practice",*
> *"Justification": "In the interview with the supervisor, they say "I think it's not mandatory by the university, but it's necessary and what the supervisors usually do." therefore indicating that this introductory meeting is not mandated by the university or the examination regulation, but it is rather accepted by the thesis supervisors as a superior procedure, hence, a best practice.",*
> *"Law of Nature": "-"}*

Chain-of-Thought combined with Few-Shot Learning (CoT+FL). For the combination of the two prompting techniques, we provide a set of manually written examples that include a reasoning before classifying the relationship. The explanations describe step by step how the solution was derived to show the LLM what a logical justification and derivation looks like. For a detailed look at the examples, please refer to the GitHub repository.

5 Results

For each LLM and prompting technique used we present precision (P), recall (R), and F1-Score (F1) Tables 2 and 3. Further breakdown per rationale can be found on GitHub. Since classifying relationships into multiple explanatory rationales constitutes a multi-class classification problem, we first computed the metrics for each rationale and then took the average for best practice, business rule, and governmental law, as displayed in Table 2. As law of nature can be classified independently of the other three categories, its results have been collected separately and are shown in Table 3. The values highlighted in bold represent the best score for each model and metric. The LLMs and prompting techniques used for the comparative analysis are shown in Fig. 2.

Table 2. Average performance over contextual origin classes

Approach	gpt-3.5t			gpt-4o			gpt-4o-mini			haiku			sonnet			gem-flash			gem-pro			llama		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
VP	0.61	0.61	0.59	0.76	0.63	0.66	0.71	0.57	0.51	0.67	**0.69**	0.67	0.62	0.57	0.56	0.71	0.65	0.66	0.85	0.57	0.67	0.66	0.63	0.62
FL	0.69	0.67	0.67	**0.88**	**0.66**	**0.74**	**0.77**	**0.81**	**0.79**	**0.77**	0.68	**0.72**	0.70	0.53	0.58	**0.80**	0.70	0.74	**0.92**	0.53	0.67	0.62	0.54	0.57
CoT	0.63	0.59	0.58	0.71	0.51	0.51	0.67	0.63	0.64	0.59	0.59	0.54	0.69	0.26	0.34	0.75	0.64	0.66	0.83	**0.63**	**0.69**	0.73	**0.68**	**0.69**
CoT+FL	**0.76**	**0.73**	**0.72**	0.81	0.63	0.68	0.68	0.65	0.64	0.69	0.68	0.68	**0.80**	**0.59**	**0.68**	**0.80**	**0.77**	**0.78**	0.80	0.61	0.67	**1.00**	0.11	0.20

Prompting Approaches. Vanilla prompting (VP) generally yielded lower performance across all models. Thus, providing only the basic description of the task without any additional context or examples is insufficient for most LLMs to achieve consistently satisfactory classification accuracy. For instance, 4o-mini classified most relationships as best

practices when only tested with vanilla prompting. E.g., the execution ordering of "Write and submit proposal" before "Get registered by chair" is defined by the chair (i.e., business rule) but was classified as a *best practice* with the justification: *Writing and submitting a proposal is a common practice that ensures the chair has the necessary information to register the thesis. It is not mandated by law but is considered a best practice in the academic process.* The distinction between business rule and best practice was generally difficult for the models when VP was used, indicating that this classification is particularly challenging.

Few-Shot learning (FL), which provides a limited number of examples within the prompt, consistently improved performance compared to VP across almost all models. Notably, GPT 4o, GPT 4o-mini, Claude 3 Haiku, have the best scores using this prompting technique, with most others exhibiting close-to-best performance. This highlights the effectiveness of guiding the LLMs to better understand the desired task and generalize to unseen examples. With FL, 4o-mini was able to correctly classify the aforementioned example, providing the following justification: *The context indicates that the requirement to write a proposal before registration is dependent on the chair's policies, which can vary. This suggests that the relationship is governed by a business rule set by the chair, as it is not universally mandated by the university but rather at the discretion of each chair.*

Except for Gemini Pro, Llama 3, and GPT 4o-mini, Chain-of-Thought reasoning does not significantly enhance the performance of the models. In many cases, CoT results in a lower F1-score than the VP baseline. This could indicate that the effectiveness of CoT is dependent on the complexity of the target task. Missing examples that guide the LLM, as given in a Few-Shot learning technique, might reinforce the model's wrong assumptions.

Combining CoT and FL results in a high performance for GPT 3.5-turbo, Claude 3 Sonnet, and Gemini Flash, and higher than baseline results for all models except Llama 3. This shows that the collective effect of providing both demonstrations and chaining of intermediate reasoning steps allows LLMs to better grasp the complex task and to achieve improved classification outcomes. Summing up, Few-Shot learning has proven to be valuable in improving performance, with and without Chain-of-Thought. In contrast, Chain-of-Thought provided little improvement in isolation, potentially even harming performance. In comparison to FL-only, the combined approach sometimes shows better and sometimes worse performance.

Models. Interestingly, GPT 4o-mini, combined with Few-Shot learning without Chain-of-Thought, and Gemini Flash 1.5, combined with FL and CoT, demonstrated the best performance, suggesting that smaller models can achieve comparable or even better performance to larger models when provided with appropriate prompting techniques. Similarly, Claude 3 Haiku outperformed Sonnet, showing best performance with FL without CoT. Llama3, Gemini Pro, and Claude Sonnet achieved the lowest maximum performance, all notably being models that did not profit much from Few-Shot learning. This is further notable for the case of Llama, which in our setting had a comparatively tiny size and was not able to process CoT+FL, but still achieved comparable performance.

Law of Nature. For law of nature, the different models show very contrasting performances. For example, GPT 3.5-turbo shows very high precision while having a very low recall, i.e., it rarely classified a relationship as a law of nature, but if it did, it was often correct. In contrast, most other models were overall far less conservative, resulting in high

Table 3. Performance for law of nature labelling

Approach	gpt-3.5t			gpt-4o			gpt-4o-mini			haiku			sonnet			gem-flash			gem-pro			llama3		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
VP	**1.00**	0.14	**0.25**	0.29	**1.00**	0.45	0.26	0.29	0.27	0.24	**1.00**	0.39	0.32	**1.00**	0.48	0.39	0.86	0.54	0.22	**1.00**	0.37	0.39	0.14	0.21
FL	**1.00**	0.14	**0.25**	**0.41**	**1.00**	**0.58**	**0.44**	0.86	**0.58**	0.24	**1.00**	0.39	0.30	0.95	0.45	0.64	**1.00**	**0.78**	**0.47**	**1.00**	**0.64**	0.83	0.14	0.24
CoT	0.20	**0.19**	0.19	0.25	**1.00**	0.41	0.26	**1.00**	0.41	0.24	0.90	0.38	0.37	0.76	0.49	0.48	0.76	0.58	0.25	**1.00**	0.40	0.18	**0.38**	**0.25**
CoT+FL	**1.00**	0.14	**0.25**	0.31	0.90	0.47	0.22	0.81	0.34	**0.38**	0.90	**0.53**	**0.58**	0.95	**0.72**	**0.67**	0.76	0.71	0.41	**1.00**	0.58	**1.00**	0.14	**0.25**

recall but low precision, i.e., they classified a relationship as a law of nature too often. In both cases, low F1 scores ensue, indicating bad performance overall.

With exceptions, models with high F1 scores tend to show a large difference in performance when Few-Shot learning was employed, implying that the concept of law of nature might be less "intuitively" determinable solely from the context provided.

6 Discussion and Conclusion

This paper aims to support the classification of activity relationships according to their vulnerability and explanatory rationale. Especially the contextual origin is often not given in process models and is usually manually identified based on textual data, such as process descriptions or interviews. With their advanced text-analyzing capabilities, the emergence of LLMs presents a promising approach for addressing this automation gap by extracting contextual information directly from process documentation. Therefore, we conducted a comparative analysis and tested various LLMs in combination with different prompting techniques to investigate their performance in this task. We selected the thesis process of a German university as a use case to demonstrate the approach. The methodology for input data generation can be adapted to more specialized organizations. Official regulations referring to governmental laws are usually publicly available, and to further enhance the process description, internal documents, reports, and interview transcripts can be used. Thus, our approach can be easily applied in different domains and organizations.

Based on the results, we conclude that Few-Shot learning and Few-Shot learning in combination with Chain-of-Thought (CoT) perform best in classifying activity relationships. Due to the excellent values in precision, recall, and F1 score, these approaches show the potential of LLMs in the automated classification of relations. Providing guiding examples and encouraging the model to show its reasoning steps significantly enhances its accuracy.

Interestingly, smaller and cheaper models like GPT 4o-mini, especially when combined with Few-Shot learning, achieved similar or better performance than larger models. This means that process re-designers with limited funds and resources can leverage smaller, more efficient LLMs without compromising classification accuracy, given appropriate prompting techniques are employed. Moreover, Few-Shot learning is generally cheaper compared to CoT.

Threads to validity. Classifying activity relationships according to their contextual origin is challenging, even for humans. Particularly when differentiating between best practices and business rules, we faced problems in discerning subtle differences. One reason for this observation is the lack of contextual knowledge process participants have about the process. Sometimes, they are not aware if they are following an actual rule, especially when

they have been involved in the process for a longer time since execution orders become natural. Similarly, LLMs struggle to classify a relationship when rules are not explicitly stated, i.e., given context is ambiguous. This emphasizes the need for well-structured and detailed process descriptions to effectively guide the LLM's analysis and enhance the reliability of the results. The fact that our process description was created based on interviews that were not subject to a detailed review may be the reason for the struggles of the LLM in correctly classifying relationships. The lack of clarity among the process participants leads to ambiguities in the LLM.

Furthermore, the thesis process used for this paper primarily focuses on sequential behavior for explanatory rationales extraction of temporal constraints. More complex behavioral patterns, such as parallelization, were not covered in the experiments. However, more complex behavior introduces existential constraints that are not covered yet by our approach and, thus, require future work.

Our approach uses simple RAG, i.e., non-overlapping, fixed-size chunks with sentence-preserving. More advanced RAG techniques, such as overlapping chunks, should be considered to reduce the chance of missing context across two chunks.

Moreover, our comparative analysis includes closed-source LLMs accessed via paid APIs. The cost associated with querying these APIs, especially for resource-intensive methods like CoT, can be substantial for large processes. This reliance on proprietary technology also restricts access to the underlying model architectures and training data, making it challenging to fully understand and analyze the observed behaviors. Furthermore, the computational resources required for these LLMs contribute to a significant environmental cost, raising concerns about the sustainability of such approaches.

Future Work. Strategies for mitigating hallucinations and biases in LLM outputs, such as using more diverse input data or stricter prompting strategies, should be investigated. Further testing with, for example, the full examination regulations should be conducted in order to investigate to what extent we can reduce manual input data preprocessing. Furthermore, we recommend testing for improvements when Few-Shot examples are generated by the LLM itself, preferably with a human in the loop. We also plan to integrate the approach into business process redesign tools to further evaluate the benefit of assessing change operations based on explanatory rationales.

In summary, this paper takes a first step toward the automated extraction of contextual information about processes. The focus on extracting explanatory rationales of activity relationships has shown promising results. Particularly for preventive BPR, the approach enables a more informed assessment of change operations. For instance, it becomes easier to determine whether a relation is critical to process execution (i.e., non-violable), allowing potential and severe consequences to be anticipated in advance. Furthermore, we see great potential for compliance checking, as this approach can reveal legally mandated process relations. This enables organizations to assess the extent to which regulations are already enforced and identify areas for improvement. Additionally, providing insight into the background of activity relations enhances process participants' understanding of why certain sequences must be executed in a specific way.

References

Adamo, Greta, Stefano Borgo, Chiara Di Francescomarino, Chiara Ghidini, Nicola Guarino, and Emilio M. Sanfilippo (2018). "Business Process Activity Relationships: Is There Anything Beyond Arrows?" In: *Business Process Management Forum. BPM 2018.* Vol. 329. LNBIP. Springer.

Adamo, Greta, Chiara Di Francescomarino, Chiara Ghidini, and Fabrizio Maria Maggi (2021). "Beyond arrows in process models: A user study on activity dependences and their rationales". In: *Inf. Syst.* 100.

Andree, Kerstin and Luise Pufahl (2024). "Am I Allowed to Change an Activity Relationship? - A Metamodel for Behavioral Business Process Redesign". In: *Enterprise Design, Operations, and Computing. EDOC 2024 Workshops - iRESEARCH, MIDas4CS, Doctoral Consortium, Joint CBI-EDOC Forum and Other Joint CBI-EDOC Events*. Vol. 537. LNBIP. Springer.

Barrientos, Marisol, Karolin Winter, Juergen Mangler, and Stefanie Rinderle-Ma (2023). "Verification of Quantitative Temporal Compliance Requirements in Process Descriptions Over Event Logs". In: *Advanced Information Systems Engineering. CAiSE 2023.* Vol. 13901. LNCS. Springer.

Bellan, Patrizio, Mauro Dragoni, Chiara Ghidini, Han van der Aa, and Simone Paolo Ponzetto (2023). *Process Extraction from Text: Benchmarking the State of the Art and Paving the Way for Future Challenges*. arXiv: 2110.03754 [cs.AI]. URL: https://arxiv.org/abs/2110.03754.

Dragoni, Mauro, Serena Villata, Williams Rizzi, and Guido Governatori (2017). "Combining Natural Language Processing Approaches for Rule Extraction from Legal Documents". In: *AI Approaches to the Complexity of Legal Systems. AICOL 2015-2017.* Vol. 10791. LNCS. Springer.

Dumas, Marlon, Marcello La Rosa, Jan Mendling, and Hajo A. Reijers (2018). *Fundamentals of Business Process Management, Second Edition*. Springer. ISBN: 978-3-662-56508-7.

Governatori, Guido, Mustafa Hashmi, Ho-Pun Lam, Serena Villata, and Monica Palmirani (2016). "Semantic Business Process Regulatory Compliance Checking Using LegalRuleML". In: *Knowledge Engineering and Knowledge Management. EKAW 2016.* Vol. 10024. LNCS.

Groefsema, Heerko, N. R. T. P. van Beest, and Guido Governatori (2022). "On the Use of the Conformance and Compliance Keywords During Verification of Business Processes". In: *Business Process Management Forum. BPM 2022.* Vol. 458. LNBIP. Springer.

Grohs, Michael, Luka Abb, Nourhan Elsayed, and Jana-Rebecca Rehse (2023). "Large Language Models Can Accomplish Business Process Management Tasks". In: *Business Process Management Forum. BPM 2023.* Vol. 492. LNBIP. Springer.

Honkisz, Krzysztof, Krzysztof Kluza, and Piotr Wisniewski (2018). "A Concept for Generating Business Process Models from Natural Language Description". In: *Knowledge Science, Engineering and Management. KSEM 2018.* Vol. 11061. LNCS. Springer.

Klessascheck, Finn, Stephan A. Fahrenkrog-Petersen, Jan Mendling, and Luise Pufahl (2024). "Unlocking Sustainability Compliance: Characterizing the EU Taxonomy

for Business Process Management". In: *28th International Conference on Enterprise Design, Operations, and Computing (EDOC 2024)*.

Kunze, Matthias and Mathias Weske (2016). *Behavioural Models - From Modelling Finite Automata to Analysing Business Processes*. Springer. ISBN: 978-3-319-44958-6.

Lewis, Patrick S. H. et al. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Liu, Nelson F., Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang (2024). "Lost in the Middle: How Language Models Use Long Contexts". In: *Trans. Assoc. Comput. Linguistics* 12.

Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig (2023). "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing". In: *ACM Comput. Surv.* 55.9.

Mansar, Selma Limam and Hajo A. Reijers (2005). "Best practices in business process redesign: validation of a redesign framework". In: *Comput. Ind.* 56.5.

Naveed, Humza, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian (2023). "A Comprehensive Overview of Large Language Models". In: *CoRR* abs/2307.06435.

Revoredo, Kate (2023). "On the use of domain knowledge for process model repair". In: *Softw. Syst. Model.* 22.4.

Rosemann, Michael, Jan Recker, and Christian Flender (2008). "Contextualisation of business processes". In: *Int. J. Bus. Process. Integr. Manag.* 3.1.

Sai, Catherine, Shazia W. Sadiq, Lei Han, Gianluca Demartini, and Stefanie Rinderle-Ma (2024). "Which Legal Requirements are Relevant to a Business Process? Comparing AI-Driven Methods as Expert Aid". In: *Research Challenges in Information Science. RCIS 2024.* Vol. 513. LNBIP. Springer.

Saidani, Oumaima and Selmin Nurcan (2007). "Towards context aware business process modelling". In: *8th Workshop on Business Process Modeling, Development, and Support (BPMDS'07), CAiSE*. Vol. 7. 1.

Suriadi, Suriadi et al. (2014). "Current Research in Risk-aware Business Process Management - Overview, Comparison, and Gap Analysis". In: *Commun. Assoc. Inf. Syst.* 34.

van der Aa, Han, Claudio Di Ciccio, Henrik Leopold, and Hajo A. Reijers (2019). "Extracting Declarative Process Models from Natural Language". In: *Advanced Information Systems Engineering. CAiSE 2019.* Vol. 11483. LNCS. Springer.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17.

Vidgof, Maxim, Stefan Bachhofner, and Jan Mendling (2023). "Large Language Models for Business Process Management: Opportunities and Challenges". In: *Business Process Management Forum. BPM 2023.* Vol. 490. LNBIP. Springer.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou (2022). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: *Proceedings of the 36th International*

Conference on Neural Information Processing Systems. NIPS '22.

Weske, Mathias (2019). *Business Process Management - Concepts, Languages, Architectures, Third Edition*. Springer. ISBN: 978-3-662-59431-5.

Winter, Karolin, Han van der Aa, Stefanie Rinderle-Ma, and Matthias Weidlich (2020). "Assessing the Compliance of Business Process Models with Regulatory Documents". In: *Conceptual Modeling. ER 2020.* Vol. 12400. LNCS. Springer.

Winter, Karolin and Stefanie Rinderle-Ma (2018). "Detecting Constraints and Their Relations from Regulatory Documents Using NLP Techniques". In: *On the Move to Meaningful Internet Systems. OTM 2018.* Vol. 11229. LNCS. Springer.

Zhu, Yilun, Joel Ruben Antony Moniz, Shruti Bhargava, Jiarui Lu, Dhivya Piraviperumal, Site Li, Yuan Zhang, Hong Yu, and Bo-Hsiang Tseng (2024). "Can Large Language Models Understand Context?" In: *Findings of the Association for Computational Linguistics: EACL 2024*. Association for Computational Linguistics.