Generating RTEC Programs with LLMs

Andreas Kouvaras[1], Periklis Mantenoglou[2] and Alexander Artikis[1,3]

[1]University of Piraeus, Greece
[2]Örebro University, Sweden
[3]NCSR "Demokritos", Greece

Abstract

Composite activities are high-level patterns built from simpler, primitive events. Constructing formal definitions for recognising such activities is a challenging task: they often involve complex spatio-temporal constraints, and the labels needed to learn them automatically are rarely available. We propose a method that generates such definitions from natural-language descriptions using pre-trained Large Language Models (LLMs), together with a novel similarity metric that quantifies the human effort required to correct LLM-generated definitions. An experimental evaluation in the maritime domain demonstrates the effectiveness of our approach.

Keywords

Event Calculus, complex event recognition

1. Introduction

Composite event recognition (CER) involves the detection of composite activities by means of temporal pattern matching over streams of low-level, symbolic events. Consider, e.g., maritime situational awareness, where the CER task is to report dangerous, illegal and suspicious vessel activities with minimal latency. Detecting illegal fishing, for instance, requires combining several simpler vessel activities, i.e., entering an area and making a turn, together with spatial information, i.e., the locations of protected areas, and background knowledge, i.e., the typical velocity of a fishing vessel. Our approach generates activity definitions in the language of the "Run-Time Event Calculus" (RTEC) (https://github.com/aartikis/rtec), a logic programming implementation of the Event Calculus with windowing and caching, which has proven efficient in real applications, including maritime situational awareness and commercial fleet management [1].

Constructing composite activity definitions is challenging. Formalising a natural-language description into rules with unambiguous semantics is an arduous and error-prone task that demands specialised programming skills, while methods that learn such definitions automatically require sizeable labelled datasets that are seldom available, due to the inherent infrequency of composite activities and to privacy concerns in sensitive domains. Recent work has therefore explored pre-trained Large Language Models (LLMs) as translators from natural-language descriptions to formal expressions such as SQL queries and Answer Set Programming specifications [2], with chain-of-thought prompting further improving consistency [3]. Since LLMs are statistical models, however, the definitions they produce may include errors, in the form of, e.g., syntactic mistakes or conditions with undefined activities; a quantitative quality assessment, beyond the qualitative analyses common in the literature [2], is therefore needed.

We present a prompting method for constructing composite activity definitions in the language of RTEC using pre-trained LLMs. We propose an automated evaluation framework that quantifies the quality of LLM-generated definitions via a novel similarity metric, reflecting the human effort required to correct them. Moreover, we present an empirical analysis on maritime situational awareness using a wide range of LLMs and real maritime data, demonstrating that the generated definitions achieve high predictive accuracy and may be successfully used in practice. This paper is a summary of [4].

Joint Proceedings of the Workshops and Doctoral Consortium of the 42nd International Conference on Logic Programming (ICLP 2026), co-located with the Federated Logic Conference (FLoC 2026), July 18–19 and 24–25, 2026, Lisbon, Portugal
a.kouvaras@unipi.gr (A. Kouvaras); periklis.mantenoglou@oru.se (P. Mantenoglou); a.artikis@unipi.gr (A. Artikis)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-03

CEUR-WS.org/Vol-4311/EULP_short_01.pdf

2. Generating RTEC Programs

Our approach generates activity definitions in the language of RTEC [5, 6], a logic programming framework that extends the Event Calculus with optimisation techniques for CER. RTEC expresses an activity definition either as a *simple* fluent-value pair (FVP), defined by event-driven initiatedAt/terminatedAt rules and subject to the law of inertia, or as a *statically determined* FVP, defined by a holdsFor rule that combines the intervals of other FVPs. We refer the reader to the cited references for the formal syntax.

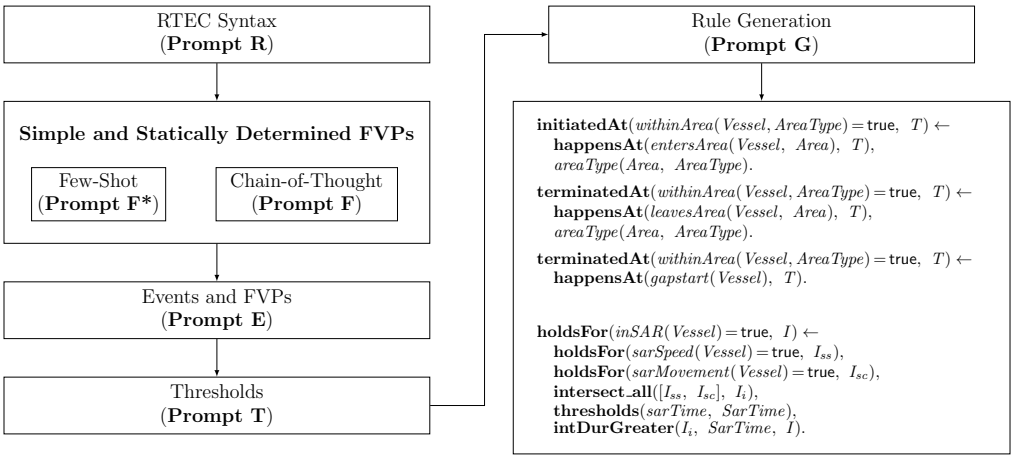


Figure 1: LLM prompting for composite activity definition generation.

Figure 1 illustrates our prompting approach. Similar to how a human would approach the task, we start by teaching the LLM the syntax of RTEC(Prompt R). We then introduce the distinction between simple and statically determined FVPs, the two ways of expressing a composite activity, via two alternative routes: *few-shot* prompting (Prompt F*), which pairs each natural-language description with the corresponding formalisation, and *chain-of-thought* prompting (Prompt F), which additionally walks the LLM through the structure of the rule, i.e., the output fluent name and arguments, the input events used, the choice between $initiatedAt/terminatedAt$ and $holdsFor$, and how Boolean combinations of intervals are formed. Zero-shot prompting produced poor results in our pilot experiments and is omitted. We then present the items of the input stream, i.e., the events and FVPs upon which RTEC reasons at run-time (Prompt E), and any domain-specific background knowledge that may appear in rule bodies—in maritime situational awareness, this includes thresholds for the expected sailing speed per vessel type and for the maximum allowed speed near the coastline (Prompt T).

```
1 Description - withinArea: We aim to identify the maximal intervals during which a
   vessel is within an area of interest.  vessel is considered to start being within
   an area of interest when it enters the area. On the other hand, the vessel is
  considered to stop being within an area of interest when (1) it leaves the area
  that it had entered, or (2) when a communication gap starts.
2
3 Description - SAR: The vessel is considered to be engaged in SAR only when it is
  sailing with a SAR speed and exhibiting movement patterns typical of SAR. Moreover,
   SAR operations must exceed a minimum duration and cannot be arbitrarily brief.
```

Prompt G

Given a natural-language description of a composite activity, we prompt the LLM to express it in the language of RTEC (Prompt G). We explicitly instruct the LLM to use the input events, FVPs and thresholds introduced earlier, and to reuse any previously formalised activity in subsequent definitions. This yields a hierarchical knowledge base in which higher-level activities depend on lower-level ones, leading to succinct event descriptions and enabling the caching of intermediate computations [5].

Note that Prompt G is domain-dependent. For instance, the aforementioned examples concerning the $withinArea$ and $inSAR$ definitions would be replaced when applying the method to a different domain.

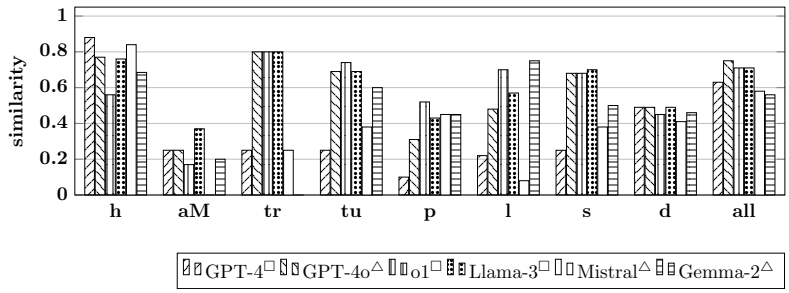
3. Empirical Analysis

We evaluated our approach on composite activity recognition for maritime situational awareness, using a publicly available dataset (https://zenodo.org/record/1167595) of 18M Automatic Identification System (AIS) position signals emitted from 5K vessels sailing in the Atlantic Ocean around the port of Brest, France, between October 2015 and March 2016 [1]. The task was to generate an RTEC rule-set of composite maritime activities, including trawling, i.e., a vessel dragging a net along the seabed, typically characterised by a specific speed range within a fishing area, and loitering, i.e., a vessel lingering in an area without a clear destination, defined as being stopped or sailing at very low speed. We applied the prompting method of Section 2 to GPT-4, GPT-4o, o1, Llama-3, Mistral and Gemma-2, querying each via the OpenAI and Groq APIs with both few-shot and chain-of-thought prompts; we write $X^{\square}$ (resp. $X^{\triangle}$) for the event description generated by LLM X via few-shot (chain-of-thought) prompting. Each generated description was compared against a hand-crafted gold standard [1], available with the source code of RTEC, using a novel similarity metric that reflects the human effort required to correct an LLM-generated definition for use with RTEC. Briefly, the metric compares an LLM-generated event description with a hand-crafted one by finding the optimal alignment between their rules and, for each aligned pair, computing the effort required to correct the generated rule's head and body conditions, where two conditions are compared based on their predicate names and arguments, with variables matched by the structural role they play in the rule rather than their name. Unmatched rules, i.e., rules that appear in the generated description but have no counterpart in the hand-crafted rule-set, are penalised by the maximum possible distance. We refer the reader to [4] for the formal definition.

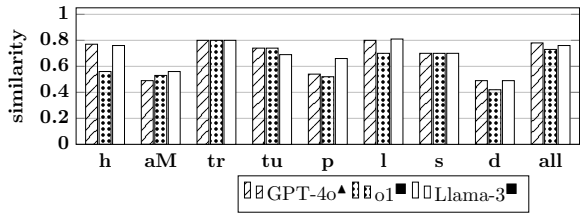
Syntactic Similarity. Figure 2a reports, for each LLM, the highest of the few-shot and chain-of-thought similarity scores across the 8 composite maritime activities and on average. Note that chain-of-thought does not always outperform few-shot: long chains-of-thought may even decrease accuracy. For trawling ('tr'), e.g., GPT-4o$^{\triangle}$, o1$^{\square}$ and Llama-3$^{\square}$ capture most of the conditions in the hand-crafted definition with at most one redundant condition, while GPT-4$^{\square}$ and Mistral$^{\triangle}$ match none of them, and Gemma-2$^{\triangle}$ expresses trawling as a simple fluent where the gold standard uses a statically determined one, scoring 0.

Qualitative Error Assessment. The errors observed in the generated event descriptions fall into four categories. First, *naming divergences*: minor differences in the names chosen for events, composite activities or background-knowledge predicates. Second, *wrong fluent type*: a composite activity is modelled with a different fluent type than the hand-crafted one, e.g., GPT-4o uses a statically determined fluent for 'movingSpeed', which the hand-crafted rules express as a simple fluent. Third, *undefined references*: a generated definition uses composite activities that are not defined in the same event description, making it unusable in practice. Fourth, *wrong interval-algebra operator*: LLMs often fail at definitions that combine activities with disjunction, conjunction or negation, e.g., GPT-4o defines 'loitering' using $intersect_all$ in place of the required $union_all$.

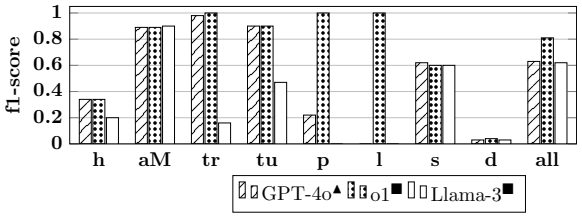
Performance on CER. We then assessed the predictive accuracy of the three highest-similarity descriptions, GPT-4o$^{\triangle}$, o1$^{\square}$ and Llama-3$^{\square}$. Each required minimal manual repair to be executable by RTEC—in o1$^{\square}$, e.g., we renamed the constant 'trawlingArea' to 'fishing' to align with the $areaType$ predicate—yielding the descriptions GPT-4o$^{\blacktriangle}$, o1$^{\blacksquare}$ and Llama-3$^{\blacksquare}$ whose average similarity remains close to that of their pre-repair counterparts (Figure 2b). We instructed RTEC to detect maritime activities over the Brest dataset using each repaired description, and computed the F1-score against the hand-crafted definitions at the granularity of seconds. Figure 2c shows that o1$^{\blacksquare}$ outperforms GPT-4o$^{\blacktriangle}$ and Llama-3$^{\blacksquare}$ on statically determined FVPs such as 'loitering' and 'pilot boarding': the gold standard for loitering requires the vessel to be 'stopped' or 'sailing at a low speed', but GPT-4o$^{\blacktriangle}$ and Llama-3$^{\blacksquare}$ encode the conjunction of these mutually exclusive conditions (confusing $union_all$ with $intersect_all$), yielding a rule that is never satisfied. o1$^{\blacksquare}$ uses the correct disjunction and achieves a perfect F1-score.



(a) Similarity values of LLM-generated definitions.



(b) Similarities after syntactic changes.



(c) Predictive accuracy.

Figure 2: 'h', 'aM', 'tr', 'tu', 'p', 'l', 's' and 'd' stand for 'high speed near coast', 'anchored or moored', 'trawling', 'tugging', 'pilot boarding', 'loitering', 'search-and-rescue' and 'drifting'. For each LLM, we report only the prompting scheme, i.e., few-shot or chain-of-thought, with the highest similarity.

4. Summary and Further Work

We proposed a method that employs pre-trained LLMs to generate RTEC specifications for composite activities based on natural-language descriptions, together with a novel similarity metric that estimates the human effort required to correct LLM-generated definitions. Unlike machine-learning approaches, our method does not require labels for composite activities. Our experimental evaluation on the maritime domain demonstrated its effectiveness. In principle, our approach may be applied to other domains, such as vehicle fleet management, by re-using Prompt R and customising prompts F, E and T with domain-specific knowledge. Testing the method in further domains, and exploiting recent advances in open foundational models such as OLMo (https://allenai.org/olmo), are directions for further work.

Acknowledgments

This work was supported partly by the University of Piraeus Research Center.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] M. Pitsikalis, A. Artikis, R. Dreo, C. Ray, E. Camossi, A. Jousselme, Composite event recognition for maritime monitoring, in: DEBS, 2019, pp. 163–174.

[2] A. Ishay, Z. Yang, J. Lee, Leveraging large language models to generate answer set programs, in: KR, 2023, pp. 374–383.

[3] E. Coppolillo, F. Calimeri, G. Manco, S. Perri, F. Ricca, Llasp: Fine-tuning large language models for answer set programming, in: KR, 2024, pp. 834–844.

[4] A. Kouvaras, P. Mantenoglou, A. Artikis, Generating activity definitions with large language models, in: EDBT, 2025, pp. 1005–1013.
[5] A. Artikis, M. J. Sergot, G. Paliouras, An event calculus for event recognition, IEEE Trans. Knowl. Data Eng. 27 (2015) 895–908.
[6] P. Mantenoglou, D. Kelesis, A. Artikis, Complex event recognition with allen relations, in: KR, 2023, pp. 502–511.